TADD: **T**opology-**A**ware **D**istributed **D**istillation in AI-Native 6G RAN

Xiaochan Xue*, Zhuosheng Zhang†, Shucheng Yu‡
*Department of Electrical and Computer Engineering, University of Hawai'i at Mānoa, HI 96822, USA
†Zoox, Inc., Foster City, CA 94404, USA
‡Department of Graduate Computer Science and Engineering, Yeshiva University, NY 10033, USA
Email: xxue@hawaii.edu, gnbenjamin1994@gmail.com, shucheng.yu@yu.edu

***Abstract*—AI-native radio access networks (AI-RAN) move learning toward wireless edge nodes and user equipments (UEs) for tasks such as beam management. Distributed distillation allows heterogeneous devices to exchange soft decisions on a shared reference set instead of raw radio data or full model weights. Existing methods, however, often assume all-to-all connectivity and do not address multi-hop sidelink dissemination under airtime budgets. We propose topology-aware multi-hop distributed distillation (TADD), which constructs a budget-induced communication graph from delay-shortest-path reachability, evaluates mixing using the second-largest eigenvalue magnitude (SLEM), and selects an effective budget T^* through a spectral-elbow rule. Experiments on four representative topologies and a DeepMIMO beam prediction task show that TADD improves airtime-to-accuracy efficiency in bottlenecked and hierarchical networks, reducing the cumulative airtime budget by 46–78% relative to full collection. In dense or well-mixed graphs, simpler policies remain competitive, indicating that topology-aware budget expansion is most useful under nontrivial dissemination constraints.**

***Index Terms*—AI-RAN, 6G Sidelink, Integrated Access and Backhaul, Distributed Distillation, Knowledge Distillation, Topology-Aware Dissemination**

I. Introduction

AI-native radio access networks (AI-RAN) aim to embed intelligence directly into the RAN, enabling adaptation at the wireless edge. In this paradigm, user equipments (UEs) learn from local radio observations such as channel state information (CSI), beam training feedback, and link quality indicators. One 6G application is mmWave or FR2 beam management, where each UE maintains a lightweight beam predictor that must adapt to mobility, blockage, and environmental changes. In such scenarios, training data are naturally decentralized, privacy-sensitive, and non-i.i.d. across devices.

Unlike centralized or cloud-RAN learning, this setting involves heterogeneous edge nodes, including UEs, O-DUs, and near-RT RICs. Sharing raw CSI or sensor data is often infeasible, while exchanging full model weights can be costly and unreliable when model architectures differ. Distributed knowledge distillation addresses these issues by exchanging *soft decisions*, i.e., logits or class probabilities evaluated on a shared unlabeled reference set [1]–[3]. Compared with model averaging, distillation only requires agreement on the output space and reference set, making it suitable for heterogeneous AI-RAN devices while reducing communication overhead.

In this work, we instantiate AI-RAN distributed distillation in a UE-to-UE sidelink setting. Each UE trains a local beam predictor from its own radio observations and exchanges soft decisions with other UEs through topology-constrained wireless links, as shown in Fig. 1. Nearby and delay-reachable UEs can observe complementary channel, blockage, and beam-state patterns, while relying only on local samples can bias each UE toward a narrow spatial region. This setting exposes the communication bottleneck that standard distillation formulations often abstract away.

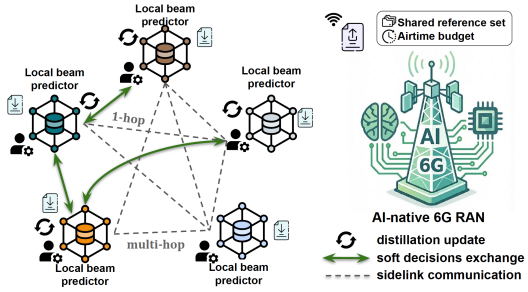

Fig. 1: UE-to-UE soft-decision distillation in AI-RAN.

Practical UE-to-UE learning faces partial and directed connectivity. Message delivery may require multi-hop forwarding, while each training round receives a limited sidelink airtime budget from the scheduler. This budget defines the communication window available for dissemination. Only peers whose shortest-path delay fits within the allocated window can contribute during that round. The *dissemination policy* therefore becomes a bottleneck because it determines how quickly knowledge propagates across devices.

The set of physical one-hop neighbors is not always the most effective exchange set. A UE can reach peers beyond its immediate neighborhood through delay-efficient multi-hop paths within the allocated airtime budget. In contrast, one-hop exchange can remain too local, while full-network collection requires a much larger budget. This creates a need for topology-aware distillation protocols that account for both network structure and communication cost. Existing decentralized distillation schemes, such as co-distillation and one-hop peer exchange, often assume all-to-all connectivity or idealized mixing conditions, including symmetric graphs and doubly stochastic matrices [2], [4]. Airtime-constrained multi-hop sidelink dissemination remains less explored.

We model each training round as airtime-constrained dissemination over a delay-sensitive sidelink graph. We introduce topology-aware multi-hop distributed distillation (TADD), which characterizes soft-decision mixing through the second-largest eigenvalue magnitude (SLEM), estimates shortest-path delays through decentralized probing, and selects an airtime budget T^* using a reproducible spectral-elbow rule. The rule identifies where further budget expansion yields diminishing gains, enabling multi-hop dissemination without reserving the full-collection window. Our main contributions are as follows:

- We formalize distributed distillation under a per-round airtime budget over directed UE-to-UE graphs and relate mixing speed to the spectrum of the induced communication matrix.
- We design a decentralized probing and budget-selection protocol that constructs the communication graph from shortest-path reachability and selects an effective dissemination budget.
- We evaluate TADD on representative topologies and a DeepMIMO beam prediction task, demonstrating improved airtime-to-accuracy efficiency in constrained graphs and competitive performance in dense graphs.

The remainder of the paper is organized as follows. Section III introduces the system model and learning objective. Section IV presents the exchange protocol and budget selection method. Section V reports the network mixing and beam prediction results. Section VI concludes the paper.

II. Related Work

Distributed distillation extends knowledge transfer and co-regularization [5], [6] to decentralized learning. A representative method is co-distillation [4], where each device acts as a student and uses the average logits from other devices as a teacher. Devices exchange soft outputs, such as logits or probability vectors, on a shared reference set, enabling knowledge transfer without synchronizing full model parameters. To avoid sharing private inputs, Jeong et al. [1] proposed a GAN-based method to synthesize aligned samples. FedMD and related methods [2] further showed that a common unlabeled reference set supports communication-efficient learning across heterogeneous models. These studies provide the foundation for soft-decision exchange, while our work focuses on delivering such messages over partial, multi-hop, and airtime-limited wireless links.

Network topology is an important factor in decentralized learning. One-hop distillation [2] shows that repeated exchanges with neighbors can approximate global soft-decision averaging under suitable mixing conditions. In wireless sidelink networks, however, a device can reach peers through delay-efficient multi-hop paths, while one-hop exchange may keep information within local neighborhoods for many rounds. We therefore study not only physical neighbors, but also devices reachable within a communication-time budget.

Recent RAN studies have used distillation for multi-tenant RAN slicing [7], digital-twin-enabled 5G network-slicing resource trading [8], O-RAN xApp conflict mitigation [9], reinforcement-learning-based RAN link adaptation [10], and AI-RAN/MEC learning under non-IID data [11]. These studies establish distillation as a useful tool for wireless and RAN intelligence, including policy sharing, model compression, xApp coordination, and edge-assisted learning. Prior wireless learning studies also optimize model update aggregation over the air [12]. In contrast, TADD studies multi-hop UE-to-UE soft-decision dissemination among heterogeneous models over a topology-constrained sidelink graph under an airtime budget.

From the communication perspective, classical gossip and distributed aggregation show that topology governs information mixing. Uniform gossip is often analyzed through Markov-chain spectra [13], while decentralized SGD studies show how communication patterns affect learning [14]. TADD applies this graph-mixing view to distributed distillation, where the exchanged objects are task-level soft decisions on a shared reference set rather than scalar states or model updates.

Overall, TADD is a communication-layer method for distributed distillation in AI-RAN, rather than a new distillation method. It focuses on topology-aware multi-hop UE-to-UE soft-decision dissemination under an airtime budget. TADD constructs a budget-induced communication graph, uses the SLEM of the receiver-normalized aggregation matrix to characterize mixing, and selects an exchange budget that balances dissemination quality and communication cost.

III. Problem Formulation and Models

This section formalizes topology-constrained distributed distillation over airtime-weighted UE sidelinks, then defines the learning objective and induced aggregation graph.

A. Distributed Distillation Formulation

We consider a K-class classification task. Each device $n \in [N]$ learns a private model with parameters $\theta^n \in \mathbb{R}^{p^n}$, where p^n may vary across devices. Given input $x \in \mathcal{X}$, the model outputs a soft decision $s_n(\theta^n, x) : \mathbb{R}^{p^n} \times \mathcal{X} \to \Delta^K$. When the device index is clear, we write $s(\theta^n, x)$. Each device n has a private labeled dataset $D_n = \{(\tilde{x}_i^n, \mathbf{y}(\tilde{x}_i^n))\}_{i=1}^{M_n}$, where $\mathbf{y}(\tilde{x}_i^n) \in \Delta^K$ is one-hot. All devices also share an *unlabeled* reference set $D_r = \{x_j\}_{j=1}^{Q}$ for distillation.

Goal of distributed distillation. The goal is to learn $\theta^1, \ldots, \theta^N$ that minimize expected loss under the unknown global distribution P:

$$\min_{\theta} \ \mathbb{E}_{(x,\mathbf{y})\sim P}\Big[\sum_{n=1}^{N} \mathcal{L}(s(\theta^n, x), \mathbf{y})\Big], \tag{1}$$

where $\mathcal{L}(\cdot,\cdot) : \Delta^K \times \Delta^K \to \mathbb{R}$ is a loss function. Since P is unknown locally, each device trains on D_n and aligns its soft decisions on D_r with other devices. The local and network-wide distillation losses are

$$\mathcal{L}_n(\theta^n) = \sum_{(\tilde{x},\mathbf{y})\in D_n} \mathcal{L}(s(\theta^n, \tilde{x}), \mathbf{y}), \tag{2}$$

$$\mathcal{L}_{\text{dist}}(\theta^n) = \sum_{x\in D_r} \Big\| \frac{1}{N-1} \sum_{m\neq n} s(\theta^m, x) - s(\theta^n, x) \Big\|^2. \tag{3}$$

With $\alpha > 0$, the standard training objective is

$$\min_{\theta} \sum_{n=1}^{N} \Big(\mathcal{L}_n(\theta^n) + \alpha \, \mathcal{L}_{\text{dist}}(\theta^n) \Big). \tag{4}$$

This objective is used in [3], [4] when each device accesses the network-wide average soft decision per round.

B. Formulation of Soft-Decision Exchange

The objective in eq. (4) assumes access to the network-wide average soft decision each round, which is unrealistic under partial connectivity, multi-hop relaying, and airtime limits. We therefore use a *network graph* to model links and per-hop costs, and a *communication graph* to represent soft decisions delivered and aggregated within each round.

Definition 1 (Network graph)**.** *Let $G = (\mathcal{V}, \mathcal{E}, W)$ be a strongly connected directed graph with $\mathcal{V} = [N]$. For each edge $(m, n) \in \mathcal{E}$, $w_{m,n} > 0$ denotes the per-hop airtime cost of delivering one compressed soft-decision message from device m to device n. If $(m, n) \notin \mathcal{E}$, the edge is unavailable and its shortest-path cost is ∞.*

The airtime weight can be instantiated as

$$w_{m,n} = \frac{B_{\text{sd}}}{R_{m,n}} + \tau_{m,n}, \qquad B_{\text{sd}} = Q_b K b, \tag{5}$$

where $R_{m,n}$ is the sidelink rate, $\tau_{m,n}$ captures per-hop scheduling or relay overhead, and B_{sd} is the payload for Q_b samples, K classes, and b quantization bits per soft value. We use normalized airtime weights in the experiments, while eq. (5) maps message size and link rate to the per-hop cost. Each dissemination round receives a sidelink resource window of length T. A source contributes to a receiver if its shortest-path cost fits within this window. Thus, T is the reserved per-round airtime budget, not the sum of individual relay transmission times. Detailed NR sidelink scheduling, spatial reuse, packet loss, and retransmission are outside the present scope.

Definition 2 (Communication graph)**.** *Let $G' = (\mathcal{V}, \mathcal{E}', W')$ be a directed graph on $\mathcal{V} = [N]$. Device m contributes its soft decision to device n iff $(m, n) \in \mathcal{E}'$. The aggregation matrix $W' = \{w'_{m,n} \geq 0\}$ is column stochastic, namely $\sum_{m=1}^{N} w'_{m,n} = 1$ for all n, and $w'_{m,n} > 0$ only if $(m, n) \in \mathcal{E}'$.*

Compared with [2], we only require W' to be column stochastic, matching UE-to-UE aggregation where each receiver normalizes over the soft decisions it receives. The resulting topology-constrained training objective is

$$\min_{\theta} \sum_{n=1}^{N} \Big(\mathcal{L}_n(\theta^n) + \alpha \sum_{x \in D_r} \|z^n(x) - s(\theta^n, x)\|^2 \Big),$$
$$z^n(x) = \sum_{m=1}^{N} s(\theta^m, x) w'_{m,n}. \tag{6}$$

IV. Topology-Aware Soft-Decision Exchange

Distributed distillation often assumes access to the network-wide average soft decision in every round, which is unrealistic over 6G UE-to-UE sidelinks with partial connectivity, multi-hop dissemination, and limited airtime. We therefore design a topology-aware scheme that selects the communication budget using a spectral mixing metric of the aggregation matrix W'.

A. Communication Graph Under a Budget

We use the physical graph $G = (\mathcal{V}, \mathcal{E}, W)$ from Section III, where $w_{m,n}$ is the per-hop airtime cost for UE m to deliver one soft-decision message to UE n. In each round, UEs disseminate soft decisions within an allocated airtime budget $\mathcal{T}$. Multi-hop forwarding uses duplicate suppression, so $\mathcal{T}$ acts as a delay-based TTL that bounds propagation.

We interpret $\mathcal{T}$ as the duration of the sidelink communication window allocated to one dissemination round. Reachability is determined by whether a message's shortest-path delay fits within this window. Thus, rT measures the cumulative airtime budget reserved over r rounds, rather than the sum of individual relay transmission times. Detailed NR sidelink scheduling, spatial reuse, packet loss, and retransmission are outside the present scope.

Physical neighbor vs. time-bounded peer. A *physical neighbor* is a one-hop neighbor in G. A *time-bounded peer* under budget $\mathcal{T}$ is any UE reachable through a multi-hop path whose shortest-path delay is at most $\mathcal{T}$. TADD targets time-bounded peers rather than only physical neighbors.

Delay shortest-path reachability. For any ordered pair (m, n), let $d_{m,n}$ denote the delay shortest-path distance from m to n. Given $\mathcal{T}$, UE m is reachable to UE n if $d_{m,n} \leq \mathcal{T}$. We construct the induced communication graph $G'(\mathcal{T}) = (\mathcal{V}, \mathcal{E}'(\mathcal{T}), W'(\mathcal{T}))$ by adding a directed edge $(m, n) \in \mathcal{E}'(\mathcal{T})$ for each reachable pair, including the lazy self-edge (n, n). Increasing $\mathcal{T}$ expands multi-hop reachability and densifies $G'(\mathcal{T})$. Given $G'(\mathcal{T})$, each UE aggregates received soft decisions using a column-stochastic matrix $W'(\mathcal{T})$ as in Definition 2. We use uniform receiver-normalized averaging:

$$W'_{m,n}(\mathcal{T}) = \begin{cases} \frac{1}{|\mathcal{N}^-(n;\mathcal{T})|}, & (m, n) \in \mathcal{E}'(\mathcal{T}), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{N}^-(n;\mathcal{T}) = \{m : (m, n) \in \mathcal{E}'(\mathcal{T})\}$ includes the receiver itself as a lazy self-edge. This indexing matches eq. (6): column n contains the weights used by receiver n. The idealized network-wide objective in eq. (3) excludes self-logits, while the implemented topology-constrained exchange uses self-inclusive aggregation for stability and primitivity.

Collect 1-Hop and *Collect All* correspond to small and saturating budgets respectively; TADD selects an intermediate T^* via the spectral-elbow rule.

B. Mixing Metric and Spectral Characterization

To isolate dissemination, we analyze the diffusion of communicated soft decisions on the reference set. Fix any $x \in D_r$ and stack node soft decisions as columns, $Z_t(x) = [z_t^1(x), \ldots, z_t^N(x)]$. The mixing-only update is

$$Z_{t+1}(x) = Z_t(x) \, W', \tag{8}$$

where W' is the column-stochastic aggregation matrix induced by a chosen exchange strategy. For directed W', we use $|\lambda_2(W')|$ to denote the second-largest eigenvalue magnitude (SLEM), which measures how quickly disagreement decays after the dominant consensus mode is removed.

Proposition 1 (Consensus and Mixing Rate)**.** *Assume W' is primitive and column stochastic. Then $W'^t \to \pi \mathbf{1}^T$, where π is the stationary right eigenvector satisfying $W'\pi = \pi$ and $\mathbf{1}^T\pi = 1$. Hence the update in eq. (8) reaches consensus, $Z_t(x) \to Z_0(x)\pi\mathbf{1}^T$, and the disagreement component decays geometrically at a rate governed by $|\lambda_2(W')|$.*

Proof sketch. Primitivity and column stochasticity imply, by Perron–Frobenius theory, a unique unit-modulus eigenvalue at 1 and all other eigenvalues strictly inside the unit circle. Therefore W'^t converges to $\pi\mathbf{1}^T$, and all disagreement modes decay geometrically with dominant rate $|\lambda_2(W')|$. Primitivity is supported by the self-inclusive lazy aggregation used in the implemented exchange.

Proposition 1 motivates using $|\lambda_2(W'(\mathcal{T}))|$ as a topology-aware mixing metric for selecting the dissemination budget. Since W' is column stochastic but not necessarily doubly stochastic, the limiting consensus is weighted by the stationary vector π rather than being the exact arithmetic average. This is acceptable for TADD because the objective is topology-aware soft-decision dissemination under airtime constraints. If exact uniform averaging is required, the aggregation layer can be replaced by push-sum or a doubly stochastic weighting rule when graph conditions permit.

Budget-selection heuristic. While the SLEM is not an instantaneous disagreement magnitude measure, for the linear mixing process induced by $W'(\mathcal{T})$, it characterizes the asymptotic rate at which disagreement modes decay after removing the consensus component. This motivates us to use SLEM as a topology-aware proxy for soft-decision mixing quality, e.g., by selecting T^* using the empirical spectral elbow of the observed SLEM curve.

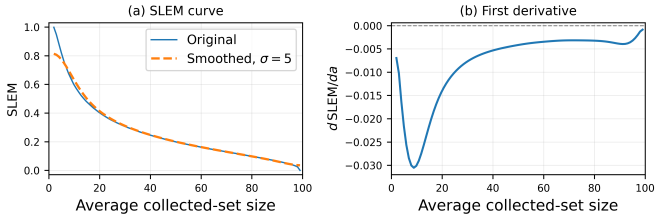

Fig. 2: Spectral-elbow budget selection. The SLEM curve and derivative show the transition to diminishing returns.

Budget rule. TADD sweeps budgets $\mathcal{T}_i \in [0, 1.05T_{\text{all}}]$, constructs $W'(\mathcal{T}_i)$, and records $\rho_i = |\lambda_2(W'(\mathcal{T}_i))|$ with average collected-set size a_i. It compresses the curve by a_i, filters out the disconnected regime with $\rho_i \approx 1$, smooths the SLEM-versus-a_i curve, and selects the first inflection point where the second derivative changes from negative to nonnegative. The selected budget T^* is the earliest budget that reaches this target collected-set size. Fig. 2 illustrates the spectral-elbow rule. The smoothed SLEM curve drops quickly at small collected-set sizes and then flattens, while the derivative highlights the transition used to select T^*.

C. Topology Probing and Budget Selection

Algorithm 1 estimates the delay matrix D, where $D_{m,n}$ approximates the delay shortest-path distance from UE m to UE n. Unlike timestamp-based probing, each probe carries an accumulated delay field, so the protocol does not require globally synchronized UE clocks. Each UE then uses D to construct $G'(\mathcal{T})$ and $W'(\mathcal{T})$ across candidate budgets and selects T^* using the spectral-elbow rule.

Algorithm 1: Distributed probing and adaptive budget selection.

Result: Adaptive dissemination budget T^*

- **Initialization at device n:**
  - Initialize local delay column $D^n_{:,n} \leftarrow \infty$ and set $D^n_{n,n} = 0$;
  - Create probe $Ping^n = \{\text{source} = n, \delta = 0\}$;
  - Send $Ping^n$ to all physical neighbors;
- **Upon receiving a probe from neighbor k:**
  - Parse $Ping^m = \{\text{source} = m, \delta\}$;
  - Set $\delta' \leftarrow \delta + w_{k,n}$;
  - **if** $\delta' < D^n_{m,n}$ **then**
    - Set $D^n_{m,n} \leftarrow \delta'$;
    - Forward $\{\text{source} = m, \delta'\}$ to physical neighbors except k;
  - **end**
- **Delay-table dissemination:**
  - Each device disseminates its local delay column $D^n_{:,n}$ using the same control channel;
  - After receiving all columns, form D by setting $D_{m,n} \leftarrow D^n_{m,n}$;
- **Budget selection:**
  - For each candidate $\mathcal{T}$, construct $G'(\mathcal{T})$ by adding (m, n) if $D_{m,n} \le \mathcal{T}$;
  - Form column-stochastic $W'(\mathcal{T})$ and compute $|\lambda_2(W'(\mathcal{T}))|$;
  - Select T^* using the spectral-elbow rule;

Assumption and overhead. During probing, link delays are quasi-static and per-hop processing delay is small relative to transmission airtime. The probing and delay-table exchange are executed at initialization or upon detected topology changes, and their cost is amortized over many subsequent distillation rounds.

Connection to the experiments. Section V first reports the average standard deviation of communicated soft decisions versus the per-round exchange budget as an observable proxy for mixing quality. It then reports distillation accuracy versus cumulative airtime under the same budget definitions. The selected T^* is expected to help most in bottlenecked or hierarchical topologies, while dense graphs may already mix effectively with one-hop exchange.

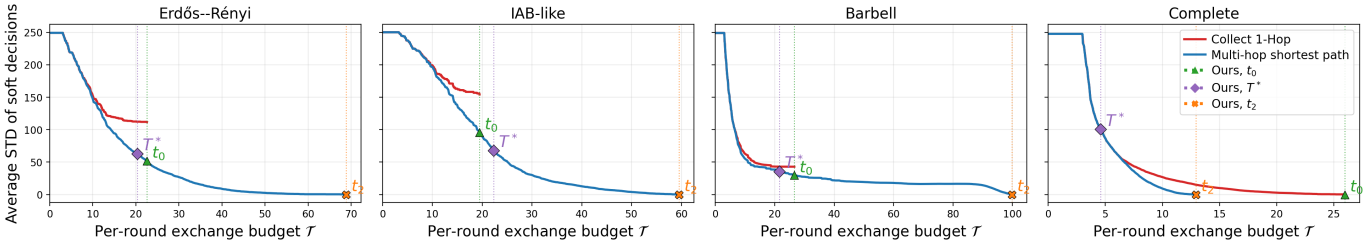

Fig. 3: Network mixing under increasing per-round exchange budget $\mathcal{T}$. One-hop exchange is compared with multi-hop delay-shortest-path exchange. Markers denote t_0, T^*, and t_2; lower average STD indicates faster soft-decision mixing.

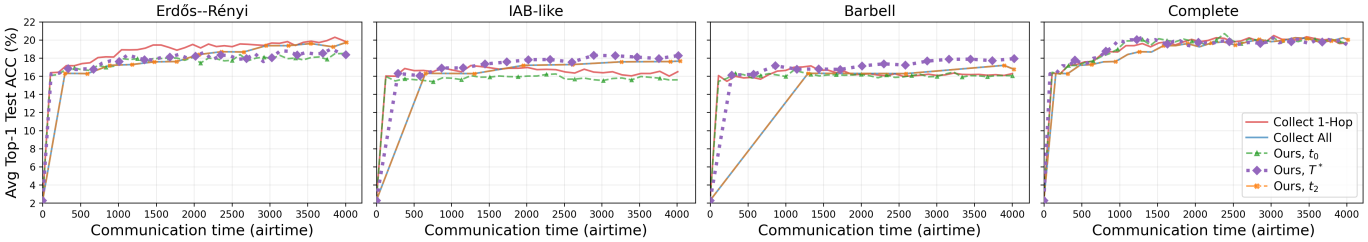

Fig. 4: Distributed distillation under increasing cumulative communication time on the DeepMIMO beam prediction task. The adaptive budget T^* is highlighted; higher Top-1 accuracy indicates better beam prediction.

V. Evaluation

A. Network Mixing Experiments

We isolate the communication effect of TADD by comparing soft-decision mixing across graph topologies and exchange budgets. We test whether time-bounded multi-hop dissemination reduces disagreement more effectively than physical-neighbor exchange, using the average standard deviation of node vectors as the disagreement metric.

Baselines and budgets. We compare three exchange policies. *Collect All* performs full-network collection under a sufficiently large budget. *Collect 1-Hop* collects soft decisions from physical neighbors and the receiver itself. *Ours* performs delay-shortest-path multi-hop dissemination under airtime budget $\mathcal{T}$. We mark the one-hop budget t_0, adaptive budget T^*, and full-mixing budget t_2. The x-axis is the per-round airtime budget $\mathcal{T}$.

Network models. We use four connected 100-node graphs in Fig. 3: Erdős–Rényi with $p = 0.05$, IAB-like hierarchical, Barbell, and Complete. They represent sparse sidelink connectivity, hierarchical access and backhaul relaying, clustered UEs connected by a relay bottleneck, and dense connectivity, respectively. Link delays follow a truncated $\mathcal{N}(10, 20)$ distribution with $w \geq 3$.

Results. We initialize each node n with a synthetic vector $z_0^n \in \mathbb{R}^{32}$ whose entries follow a Gaussian distribution with mean 500 and standard deviation 250, and apply eq. (8). This isolates graph mixing, while the DeepMIMO experiment uses task-generated soft decisions. Fig. 3 shows that *Collect 1-Hop* reduces disagreement quickly at small budgets but often plateaus as information remains within physical neighborhoods. In contrast, multi-hop shortest-path exchange continues to reduce disagreement as $\mathcal{T}$ increases, especially in sparse, hierarchical, and bottlenecked graphs. The markers t_0, T^*, and t_2 indicate that TADD selects an intermediate budget before full-network mixing is needed.

B. Distributed Distillation Experiments

We next evaluate whether the mixing advantage improves communication efficiency on a radio-native beam prediction task. The mixing experiments use 100-node graphs, while distillation uses 20 UE nodes matching the dataset partition; the same graph families are resampled at this node scale.

DeepMIMO beam dataset. We generate a DeepMIMO beam-index classification dataset. The task has 32 beam classes, with each sample represented by an 8-dimensional radio feature vector. After filtering, the dataset contains 123,884 samples across 20 UE nodes: 89,196 for training, 9,911 unlabeled reference samples for soft-decision exchange, and 24,777 for testing. The reference set covers all 32 classes. The node-level data are non-i.i.d., as different UEs observe between 3 and 30 beam classes due to spatially dependent channel and beam observations. Random Top-1 accuracy is $1/32 = 3.125\%$, so we evaluate the relative accuracy–communication tradeoff rather than standalone beam-prediction performance.

Learning and airtime accounting. Each UE trains the same beam classifier on its DeepMIMO samples. In each round, UEs exchange soft decisions on the shared unlabeled reference set and update local models using supervised and distillation losses. All exchange methods share the same model, optimizer, initialization, data split, and reference set; only the exchange policy and budget differ. For a method with per-round budget T, its cumulative reserved airtime budget after round r is rT.

Airtime efficiency metric. Since TADD targets communication-efficient dissemination rather than peak accuracy, we use *airtime-to-accuracy* (ATA) under our per-round airtime-budget accounting. For target Top-1 accuracy

τ, ATA(τ; method) is the smallest cumulative airtime budget at which a method first reaches τ. Lower ATA indicates that the method reaches the same target using a smaller reserved communication budget.

Budget structure across topologies. Across the four graphs, the one-hop budget T_{1hop} falls in the narrow range $[10.43, 12.57]$, while the full-collection budget T_{all} spans $[15.80, 129.71]$. This shows that the required dissemination budget is dominated by graph structure rather than per-link delay alone. The ratio T_{all}/T^* in Table I captures the per-round airtime-budget reduction of TADD relative to full collection. It is largest on Barbell ($4.51\times$) and IAB-like ($2.26\times$), where multi-hop relaying must traverse structural bottlenecks or hierarchical relay paths, and smallest on Complete ($1.93\times$), where the graph is already dense.

	Per-round budget		ATA at target τ				
Topology	T^*	T_{all}/T^*	τ (%)	1-Hop	Ours (T^*)	All	Saving
Barbell	28.74	$4.51\times$	17.0	–	**862**	3891	78%
IAB-like	28.91	$2.26\times$	17.5	–	**1735**	3261	47%
Erdős–Rényi	11.21	$2.63\times$	18.0	**730**	1121	2066	46%
Complete	8.17	$1.93\times$	17.5	542	**409**	790	48%

TABLE I: Top-1 ATA comparison; lower is better and "–" denotes unreached target.

Airtime efficiency results. Table I reports the cumulative airtime budget required to reach a target Top-1 accuracy τ, selected near each topology's saturation level. In bottlenecked or hierarchical topologies, namely Barbell and IAB-like graphs, *Collect 1-Hop* does not reach the target within the evaluation horizon because one-hop exchange cannot bridge structural bottlenecks or hierarchical relay paths. *Collect All* reaches the target but requires a much larger cumulative budget. In contrast, *Ours (T^*)* reaches the same targets with 78% and 47% smaller cumulative airtime budgets than *Collect All*, respectively. An intermediate topology-aware budget therefore captures useful multi-hop dissemination without reserving the full-collection window.

In well-mixed topologies, the behavior is consistent with the smaller budget gap. On Erdős–Rényi, *Collect 1-Hop* reaches $\tau = 18\%$ fastest because the graph is already close to globally mixed. *Ours (T^*)* still reduces the cumulative airtime budget by 46% relative to *Collect All*, but it does not dominate the cheaper one-hop policy. On Complete, the spectral elbow selects $T^* < T_{\mathrm{1hop}}$, and *Ours (T^*)* achieves the lowest ATA among all three policies.

Discussion. These results support TADD's intended role. The spectral-elbow rule identifies the most efficient multi-hop dissemination budget for each topology without penalizing cases where simpler policies suffice. Savings are largest under structural bottlenecks or hierarchical relay paths, reducing the cumulative airtime budget by 78% on Barbell and 47% on IAB-like relative to full collection. As a sanity check, exchange-based methods improve over the no-exchange local-only baseline ($16.41 \pm 6.08\%$ Top-1) by up to 2.59 percentage points (15.8% relative), confirming meaningful soft-decision transfer. In well-mixed graphs, T^* approaches the one-hop regime and TADD behaves like lightweight local exchange. Thus, topology-aware budget expansion is most valuable under nontrivial dissemination constraints rather than as a uniform improvement across all graphs.

VI. Conclusion

This paper revisits soft-decision exchange for distributed distillation in AI-RAN networks, where connectivity is partial, multi-hop dissemination is common, and each round has a limited airtime budget. We link dissemination efficiency to the spectral properties of the induced communication graph and propose TADD, which selects an effective budget and disseminates soft decisions along shortest-delay paths. Experiments on topologies and a DeepMIMO beam prediction task show that TADD reduces the cumulative airtime budget by 46–78% relative to full collection, with the largest gains in bottlenecked and hierarchical graphs. In dense or well-mixed graphs, simpler exchange policies remain competitive.

References

[1] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[2] I. Bistritz, A. Mann, and N. Bambos, "Distributed distillation for on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[4] R. Anil, G. Pereyra, A. T. Passos, R. Ormandi, G. Dahl, and G. Hinton, "Large scale distributed neural network training through online distillation," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[5] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proceedings of ICML workshop on learning with multiple views*, vol. 2005. Citeseer, 2005, pp. 74–79.

[6] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel, "Efficient co-regularised least squares regression," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 137–144.

[7] G. Sun, D. Ayepah-Mensah, H. Chen, G. O. Boateng, and G. Liu, "Fediststice: Federated policy distillation for collaborative intelligence in multi-tenant ran slicing," *IEEE Transactions on Services Computing*, vol. 18, no. 1, pp. 184–197, 2025.

[8] D. Ayepah-Mensah, G. Sun, G. Owusu Boateng, and G. Liu, "Federated policy distillation for digital twin-enabled intelligent resource trading in 5g network slicing," *IEEE Transactions on Network and Service Management*, vol. 22, no. 1, pp. 361–379, 2025.

[9] H. Erdol, X. Wang, R. Piechocki, G. Oikonomou, and A. Parekh, "xapp distillation: Ai-based conflict mitigation in b5g o-ran," 2024.

[10] S. Khosravi, B. Demirel, L. Zhou, J. Rasines, and P. Soldati, "Practical policy distillation for reinforcement learning in radio access networks," 2026.

[11] L. Zou, H. Oh, C. M. Thwal, A. Adhikary, S. Hong, and Z. Han, "A multi-prototype-guided federated knowledge distillation approach in ai-ran enabled multi-access edge computing system," 2026.

[12] X. Xue, M. K. Hasan, S. Yu, L. N. Kandel, and M. Song, "Over-the-air federated learning with enhanced privacy," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 4546–4551.

[13] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, 2003, pp. 482–491.

[14] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 5381–5393.