METADATA-AGNOSTIC DECENTRALIZED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized learning enables collaborative model training without sharing raw data, offering strong privacy benefits. However, many existing studies in decentralized learning research rely on an unrealistic assumption that all participants can share metadata such as class labels and the total number of categories. This assumption, which we term Metadata-Dependent Supervised Learning (MDSL), fails to reflect the diversity and autonomy of real-world participants. In contrast, we propose MAZEL: Metadata-Agnostic Zero-shot Learning, a framework that eliminates the need for shared metadata by leveraging CLIP-based zero-shot classification. MAZEL enables more realistic and flexible decentralized learning, where clients can dynamically join or leave without requiring predefined output heads. Our contributions are fourfold: (1) We formalize the distinction between MDSL and MAZEL; (2) we show that standard claims about performance degradation and slow convergence in MSDL-based decentralized learning may not hold under MAZEL; (3) we provide benchmarks using up to 8–16 diverse datasets to rigorously evaluate newly proposed decentralized learning methods under real metadata-agnostic cases; and (4) we propose two-stage and cosine gossip schedulers to optimize communication efficiency.

1 INTRODUCTION

Motivating question: *How can decentralized learning algorithms be evaluated under conditions that reflect real-world constraints, such as lack of shared metadata, heterogeneous label spaces, and dynamic participation?*

Decentralized learning offers a promising framework for peer-to-peer collaborative learning across geographically dispersed resources without sharing raw data. In such collaborative scenarios, a collective of agents, often from diverse domains, participate in joint training processes without disclosing sensitive information about their local datasets. This privacy-preserving feature is particularly useful in light of strict data protection regulations. However, we note that there is a discrepancy between how decentralized learning is experimentally evaluated in research and how it is intended to function in real-world deployments.

A notable example of this discrepancy lies in the implicit assumption of "metadata awareness." In many decentralized learning experiments, researchers commonly assume prior knowledge of dataset distributions across the participating agents. For instance, it is common practice to simulate non-IID settings by sampling from CIFAR-100 via a Dirichlet distribution with a specific parameter (e.g., $\alpha = 0.1$) (Yurochkin et al., 2019; Hsu et al., 2019). In such an experimental design, although each node's data distribution is distinct, the overarching metadata, such as the total number of classes, is still treated as shared global information. We refer to this experimental setting as Metadata-Dependent Supervised Learning (MDSL). MDSL implicitly requires that all agents know, for example, the classification categories used at every node.

In contrast, real-world decentralized systems often involve participants with unknown, disjoint, or partially overlapping label spaces. For example, one client might have 1,000 categories while another has only 5; or two clients may refer to the same class using different labels. Moreover, clients may dynamically join or leave the network, making fixed classification heads inefficient and brittle. In such settings, synchronizing metadata is not only impractical but may also violate privacy requirements.

Such constraints frequently arise in realistic applications. In healthcare, for instance, institutions often adopt different coding systems, such as ICD-10 versus SNOMED, and privacy regulations like

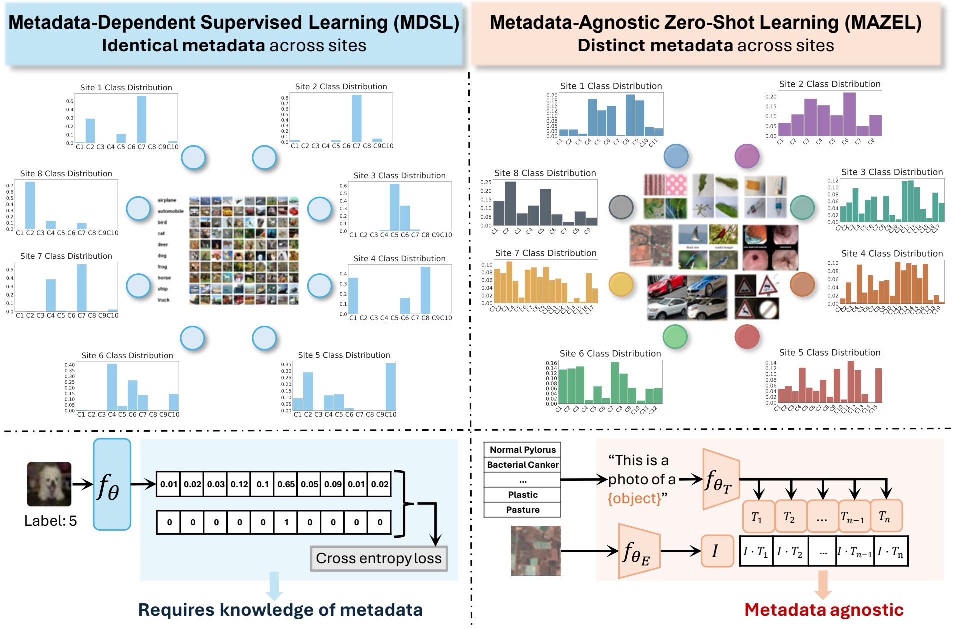
Metadata-Dependent Supervised Learning (MDSL)
Identical metadata across sites
Metadata-Agnostic Zero-Shot Learning (MAZEL)
Distinct metadata across sites
Site 1 Class Distribution
Site 2 Class Distribution
Site 3 Class Distribution
Site 4 Class Distribution
Site 5 Class Distribution
Site 6 Class Distribution
Site 7 Class Distribution
Site 8 Class Distribution
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
Label: 5
f_θ
0.01 0.02 0.03 0.12 0.1 0.65 0.05 0.09 0.01 0.02
0 0 0 0 0 1 0 0 0 0
Cross entropy loss
Requires knowledge of metadata
Normal Pylorus
Bacterial Canker
...
Plastic
Pasture
“This is a photo of a {object}”
f_{θ_T}
T_1 T_2 ... T_{n-1} T_n
f_{θ_E}
I
$I \cdot T_1$ $I \cdot T_2$ ... $I \cdot T_{n-1}$ $I \cdot T_n$
Metadata agnostic

Figure 1: Illustration of the metadata-dependent supervised learning (MDSL) and metadata-agnostic zero-shot learning (MAZEL) settings. **(Left)** Traditional decentralized learning assumes homogeneous, Dirichlet-partitioned datasets (e.g., CIFAR-10/100) with shared metadata (class labels, synchronized task alignment). Coordination relies on pre-agreed classification heads and explicit label synchronization. **(Right)** Proposed real-world scenario: nodes host truly heterogeneous datasets (8–16 distinct domains, e.g., textures, mango leaf, garbage, satellite images, medical imaging, cars, traffic signs) with no shared metadata (unknown classes, divergent label spaces). CLIP-based zero-shot alignment replaces metadata-dependent coordination by leveraging multimodal embeddings for cross-node task alignment. **(Bottom)** Comparison between traditional supervised learning using metadata-dependent cross-entropy loss and CLIP-based training that enables metadata-agnostic alignment through image-text embeddings.

HIPAA or GDPR Voigt & Von dem Bussche (2017) can prohibit sharing metadata that may indirectly reveal sensitive information about patient populations. In collaborative settings across industries, such as between e-commerce platforms and social media networks, participants may be unwilling or unable to disclose internal label structures due to business confidentiality.

To address these challenges, we introduce the **Metadata-Agnostic Zero-shot Learning (MAZEL)** framework. Instead of requiring a shared label space, we leverage CLIP-style image-text models to perform zero-shot classification at each site. This removes the need for label alignment, enabling flexible participation and scalable deployment across heterogeneous domains.

Importantly, we note some criticized claims and conclusions about decentralized learning in the current literature under the MDSL experimental setups:

> **Claim 1**: Local models in decentralized learning often generalize poorly to a global test set in highly heterogeneous scenarios (Lin et al., 2021; Vogels et al., 2021).
>
> **Claim 2**: Decentralized learning is known to have much slower convergence compared to centralized learning (Lian et al., 2017; Koloskova et al., 2020).

In examining scenarios closer to real-world requirements, we find that these claims are **not necessarily correct** for the “gradient-and-gossip” protocols but rather specific to the MDSL experimental settings. Under the MAZEL framework, there is no necessity to share predefined metadata. Preliminary findings suggest that local models can achieve strong performance on the global test set without incurring communication overhead, while effectively balancing both local and global performance.

To bridge the gap between decentralized learning experimental settings and realistic decentralized learning applications, we propose metadata-agnostic zero-shot testing baselines for both 8 sites and

16 sites. In this framework, we neither require agreement on the number of classes nor advance knowledge of each node's label sets. As a result, newly designed decentralized learning strategies or algorithms can be evaluated under conditions that more accurately reflect real-world constraints and data heterogeneity. Furthermore, for scenarios demanding strict controllability of non-IID degrees, our approach remains compatible with Dirichlet-based sampling strategies.

Our contributions in this paper are:

1. We formalize the distinction between metadata-dependent supervised learning (MDSL) and metadata-agnostic zero-shot learning (MAZEL).
2. We introduce MAZEL, a framework that leverages CLIP for zero-shot classification, enabling decentralized training without shared metadata.
3. We demonstrate that several long-held assumptions about MDSL-based decentralized learning such as poor generalization and slow convergence do not hold under MAZEL.
4. We provide benchmarks using up to 16 heterogeneous datasets, enabling evaluation under realistic metadata-agnostic settings.
5. We evaluate two-stage and cosine gossip scheduler tricks on MAZEL as an illustration.

2 Related work

Decentralized Learning. Decentralized learning enables multiple nodes to train models collaboratively without a central coordinator. A key example is Decentralized Stochastic Gradient Descent (DSGD) (Lian et al., 2017). Recent advancements address challenges like dynamic network topologies (Nedi'c & Olshevsky, 2014; Lu & Wu, 2020; Koloskova et al., 2020; Ying et al., 2021), asynchronous communication (Lian et al., 2018; Xu et al., 2021; Nadiradze et al., 2021; Bornstein et al., 2023), and heterogeneous data distributions (Tang et al., 2018; Vogels et al., 2021; Le Bars et al., 2023), improving scalability and robustness in real-world scenarios.

Zero-shot Classification. CLIP Radford et al. (2021a) introduced contrastive learning between images and texts, enabling strong zero-shot transfer. SigLIP Zhai et al. (2023) optimized loss functions for better scaling, while CLAP Wu et al. (2023) extended contrastive learning to audio. DINOv2 Oquab et al. (2023) improved self-supervised learning, producing high-quality visual features.

Model Merging. Methods including "model soups" Wortsman et al. (2022) and task vectors Ilharco et al. (2023) enable efficient model combination. DARE Yu et al. (2023) allows language models to merge capabilities without retraining. TIES-merging Yadav et al. (2023) mitigates parameter conflicts, while AdaMerging Yang et al. (2024) learns merging coefficients adaptively. MAP Li et al. (2024) efficiently finds the Pareto front, revealing trade-offs in model fusion.

For more detailed related work, please refer to Section C.

3 Preliminaries

3.1 Decentralized Learning

Decentralized learning allows collaboratively training models without the control of a central server. In a decentralized learning framework, the setting can be represented as a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ signifies the set of $|\mathcal{V}| = N$ agents involved in the learning process, and $\mathcal{E}$ indicates the communication links among these agents. We also define a mixing matrix between agents using a weighted adjacency matrix $A \in \mathbb{R}^{N\times N}$, $A_{i,j} \in [0, 1]\,\forall i, j$, where $A_{i,j}$ denotes the strength of the connection from agent j to agent i. Each agent $i \in \mathcal{V}$ is characterized by its local model $\theta_i \in \mathbb{R}^d$ and its local data distribution P_i. In decentralized learning, two primary settings are considered:

(1) Personalized Setting (Vanhaesebrouck et al., 2017; Kharrat et al., 2024): This setting focuses on optimizing models for individual agents, where each model is trained to perform well on the agent's local data distribution P_i. The corresponding objective is the **Local Population Risk**, which seeks to optimize individual models to perform well on data from their respective local distributions: $\min_{\{\theta_i\in\mathbb{R}^d\}_{i\in\mathcal{V}}} \left[F(\theta) \triangleq \frac{1}{N}\sum_{i\in\mathcal{V}} \mathbb{E}_{x_i\sim P_i} l(\theta_i; x_i)\right]$. **(2) Generic Setting** (Koloskova et al., 2020). This setting aims to train a single consensus model that performs well on the entire data distribution.

The corresponding objective is the **Global Population Risk**, which focuses on optimizing a single consensus model θ to serve the entire network: $\min_{\theta \in \mathbb{R}^d} \Big[G(\theta) \triangleq \frac{1}{N} \sum_{i \in \mathcal{V}} \mathbb{E}_{x_i \sim P_i} l(\theta; x_i) \Big]$.

Algorithm 1 Decentralized Learning

```
Require: For each node $i \in \mathcal{V}$, initialize $\theta_i^0 \in \mathbb{R}^d$, iterations $T$, mixing matrix $A$
for $t = 0$ to $T$ do
    for $i \in \mathcal{V}$ do
        (in parallel)
        Sample batch $x_{i,j}^t$ from $P_i$,
        $\theta_i^{t+\frac{1}{2}} = \text{Optimizer}(\theta_i^t; x_{i,j}^t)$        ▷ Local training
        Send $\theta_i^{t+\frac{1}{2}}$ to out-neighbors and receive $\{\theta_l^{t+\frac{1}{2}}\}_{l \in \mathcal{N}_{\text{in}}(i)}$        ▷ Communication
        $\theta_i^{t+1} = \sum_{l \in \mathcal{N}_{\text{in}}(i)} A_{i,l} \, \theta_l^{t+\frac{1}{2}}$        ▷ Gossip averaging
    end for
end for
```

In practical scenarios, the theoretical objectives of minimizing Local Population Risk and Global Population Risk are achieved through empirical risk minimization (ERM) using the available local datasets. Each agent $i \in \mathcal{V}$ possesses a local dataset $D_i = \{x_{i,1}, \ldots, xi, n_i\}$. The collective dataset across all agents is denoted as $D \triangleq \bigcup_{i=1}^{N} D_i$. Therefore, the ERM problem is formulated as: $\min_{\theta \in \mathbb{R}^d} \Big[\hat{G}_D(\theta) \triangleq \frac{1}{N} \sum_{i \in \mathcal{V}} \sum_{j=1}^{n_i} l(\theta_i; x_{i,j}) \Big]$.

Decentralized learning algorithms address the consensus model optimization problem by relying solely on local agent model updates and peer-to-peer communications within the network graph (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009). In Algorithm 1, we illustrate a typical decentralized learning process that alternates between local model updates for each agent and the integration of agents' parameters through gossip averaging with neighboring n-
odes based on the mixing matrix A.

3.2 Metrics

To evaluate model performance in decentralized learning, we adopt practical metrics derived from the theoretical objectives introduced earlier. Specifically, local test accuracy measures generalization on local data, while global test accuracy evaluates generalization across the entire data distribution. These two aspects are typically studied independently. Therefore, we aim to provide precise definitions for these concepts before delving into our work.

Definition 1 (Test Accuracy)**.** *Assuming the tasks for decentralized learning are image classification, we define the local test accuracy and global test accuracy of site i:*

$$\text{LocalTestAcc}_i(\theta_i) = \frac{\sum_{x_{i,j} \in D_{\textit{i-th site test}}} \mathbf{1}\Big(f(\theta_i, x_{i,j}) = y_{i,j} \Big)}{|D_{\textit{i-th site test}}|} \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $f(\theta_i, \cdot)$ is the model in site i, $x_{i,j} \in D_{\textit{i-th site test}} \sim P_i$ test data on site i, and $y_{i,j}$ is the ground truth label for $x_{i,j}$.

$$\text{GlobalTestAcc}(\theta_i) = \frac{\sum_{x_{i,j} \in D_{\textit{global test}}} \mathbf{1}\Big(f(\theta_i, x_{i,j}) = y_{i,j} \Big)}{|D_{\textit{global test}}|} \tag{2}$$

where $x_{i,j} \in D_{\textit{global test}} \sim P$ which is the distribution of test data over all the sites.

In addition, $\text{ALA}(\theta_1, \cdots, \theta_n)$ *denotes the* ***A****verage* ***L****ocal* *Test* ***A****ccuracy, defined as* $\frac{1}{N} \sum_{i=1}^{N} \text{LocalTestAcc}_i(\theta_i)$*;* $\text{AGA}(\theta_1, \cdots, \theta_n)$ *stands for* ***A****verage* ***G****lobal Test* ***A****ccuracy, computed as* $\frac{1}{N} \sum_{i=1}^{N} \text{GlobalTestAcc}(\theta_i)$*;* $\text{MMGA}(\theta_{\textit{merged}})$ *refers to* ***M****erged* ***M****odel* ***G****lobal Test* ***A****ccuracy, given by* $\text{GlobalTestAcc}(\theta_{\textit{merged}})$*, where* $\theta_{\textit{merged}}$ *is the model obtained by merging* $\theta_1, \cdots, \theta_n$.

In this paper, we take $\theta_{\text{merged}} = \frac{1}{N}\sum_i^N \theta_i$, which is known as model soup Wortsman et al. (2022) in the model merging community. The choice of model merging method in decentralized learning remains an open question and is not the primary focus of our work.

Definition 2 (Gossip Gain)**.** *We define Gossip Gain (GG) as* $GG = \left(\frac{MMGA(\theta_{merged})}{AGA(\theta_1,\ldots,\theta_n)} - 1\right) \times 100\%$

This metric quantifies the improvement in global test accuracy achieved by merging models from all neighboring sites at a given communication step. The motivation behind introducing this metric is that, at any communication step, maximizing information exchange across all sites provides an upper bound on the achievable global generalization performance. A low Gossip Gain indicates that expanding the communication graph has limited potential to further enhance global generalization, suggesting that additional inter-site collaboration may yield diminishing returns.

3.3 Metadata-Dependent Supervised Learning (MDSL)

In most decentralized learning studies, researchers aim to develop algorithms that enhance knowledge transferability across different sites, improve optimization efficiency, and enhance generalization capabilities, etc. To demonstrate the effectiveness of their proposed algorithms, they typically utilize benchmark datasets such as MNIST LeCun et al. (1998), CIFAR-10, CIFAR-100 Krizhevsky (2009), or TinyImageNet Le & Yang (2015). These datasets are commonly partitioned into a predetermined number of nodes by sampling from a Dirichlet distribution, allowing for the simulation of non-IID experimental settings. The degree of non-IID data distribution is controlled by the parameter α in the Dirichlet distribution, where smaller values of α lead to a higher degree of non-IID data distribution across sites.

At each site, supervised learning is conducted using gradient descent to minimize the cross-entropy loss. While this experimental setup appears reasonable, it implicitly assumes that participants across different sites have access to shared metadata. This assumption enables each site to define classification heads by aggregating class distributions across nodes and coordinating class assignments within the classification head. We define this experimental setting as Metadata-Dependent Supervised Learning (MDSL). However, this approach contradicts the fundamental privacy-preserving principles of decentralized learning and is not realistic in many real-world applications such as healthcare.

3.4 Metadata-Agnostic Zero-Shot Learning (MAZEL)

In contrast, we propose Metadata-Agnostic Zero-Shot Learning (MAZEL), which addresses the issue of requiring different sites to share metadata and coordinate classification heads before training. In this approach, each participant independently initializes a CLIP-based model. Instead of relying on shared metadata, each site constructs its own textual template, such as “This is a photo of object", for its respective classes without disclosing this information to other sites. The participant can then store the text embedding corresponding to each class label, denoted as T_i.

During training, participants evaluate their models by performing inference on local test images to obtain their representations, denoted as I. They then compute the similarity scores between these image representations and all stored text class embeddings T_i. The predicted class is determined by $c^* = \arg\max_i \text{sim}(I, T_i)$.

MAZEL eliminates the need for participants to predefine a classification head that incorporates class dimensions from other sites. Consequently, it removes the necessity of exchanging metadata, providing a more realistic setting for decentralized learning compared to MDSL.

3.5 Analysis of MAZEL and MDSL model formulation

For clarity and simplicity, we compare MAZEL with the variant of MDSL that utilizes a pretrained CLIP encoder with frozen weights and a trainable multilayer perceptron (MLP) classification head, as this configuration has shown the best empirical performance among the MDSL training variants.

MDSL Formulation Let the parameters of the trainable MLP classification head be denoted by $W \in \mathbb{R}^{d\times c}$, where d is the dimensionality of the encoder output, and c is the number of classes. For a given visual representation $v_i \in \mathbb{R}^d$, the training objective can be expressed as $\min_W y_i \log \text{softmax}(Wv_i)$, where y_i is the one-hot encoded ground-truth label corresponding to the input v_i, and the bias term is omitted for brevity. The matrix W can be written as

$W = [w_1, \dots, w_c]$, with each $w_k \in \mathbb{R}^d$ for $k = 1, \dots, c$. Consequently, the objective becomes $\min_W y_i \log \mathrm{softmax}([w_1^\top v_i, \dots, w_c^\top v_i])$.

MAZEL Formulation Let $t_k \in \mathbb{R}^d$ be the textual embedding of class k and $v_i(\theta) \in \mathbb{R}^d$ the visual representation. In both our implementation and Algorithm 2, we use *cosine similarities*, i.e., we L2-normalize both embeddings and take logits $z_k(x_i) = \left\langle \frac{t_k}{\|t_k\|}, \frac{v_i(\theta)}{\|v_i(\theta)\|} \right\rangle$. The training objective is $\min_\theta \; y_i^\top \log \mathrm{softmax}(z(x_i))$. For notational simplicity, in the rest of this section we write $t_k^\top v_i(\theta)$ to denote this cosine logit between normalized embeddings, i.e., we absorb the L2-normalization into the definition of t_k and $v_i(\theta)$. This reparameterization does not change the classifier's decision boundaries or calibration: it is equivalent to a fixed temperature scaling of the softmax. Please see 2 for the detailed algorithm.

Table 1: Comparison of decentralized learning performance under MDSL and MAZEL across different datasets and finetuning strategies. Abbreviations: **ALA**: Average local test accuracy, **AGA**: Average global test accuracy, **MMGA**: Merged model global test accuracy, **MDSL**: Metadata-dependent supervised learning setting, **MAZEL**: Metadata-agnostic zero-shot learning setting. See definitions in Definition 1. Experiments were conducted using two datasets: CIFAR-100 and Kvasir v2. We evaluate CLIP-pretrained Radford et al. (2021b) and ImageNet-pretrained Dosovitskiy et al. (2020) ViT-B/32 models under two finetuning strategies: **Full FT** (full finetuning) and **Classification head FT** (classification head-only finetuning). We include four learning topologies: random communication with uniform weights, random communication with influence-based weights Zhu et al. (2025), complete communication graph with uniform weights, and local training only. For the MDSL setting, training consists of 8,000 steps with a gossip interval of 100 steps, while for the MAZEL setting, training consists of 2,050 steps with the same gossip interval.

		CIFAR-100				Kvasir V2	
		MDSL			MAZEL	MDSL	MAZEL
		CLIP pretrained		ImageNet pretrained	CLIP pretrained	CLIP pretrained	CLIP pretrained
		Full FT	Classification head FT	Full FT	Full FT	Classification head FT	Full FT
Random (Uniform)	ALA	30.42 ± 0.03	78.95 ± 0.00	38.56 ± 0.02	90.67 ± 0.00	88.17 ± 0.01	93.54 ± 0.02
	AGA	7.29 ± 0.02	45.52 ± 0.01	23.76 ± 0.01	80.65 ± 0.00	65.10 ± 0.01	91.99 ± 0.01
	MMGA	12.54 ± 0.03	60.31 ± 0.01	29.15 ± 0.01	83.72 ± 0.02	83.33 ± 0.00	93.75 ± 0.01
	(ALA+AGA)/2	18.86	62.24	31.16	85.66	76.64	92.77
	Gossip Gain	72.02%	32.49%	22.69%	3.81%	28.00%	1.91%
	Converge Steps	>8000	5860	>8000	1100	6700	3450
Random (Influence-weighted)	ALA	33.32 ± 0.01	80.39 ± 0.00	39.31 ± 0.02	92.47 ± 0.00	84.02 ± 0.00	93.91 ± 0.01
	AGA	8.10 ± 0.01	47.40 ± 0.01	23.17 ± 0.01	81.06 ± 0.00	68.10 ± 0.01	92.47 ± 0.01
	MMGA	12.71 ± 0.02	61.88 ± 0.00	28.05 ± 0.01	83.87 ± 0.01	84.90 ± 0.02	94.21 ± 0.02
	(ALA+AGA)/2	20.71	63.90	31.24	86.77	76.06	93.19
	Gossip Gain	56.91%	30.55%	21.06%	3.47%	24.67%	1.88%
	Converge Steps	>8000	7200	>8000	1100	7100	3600
Complete (Uniform))	ALA	11.93 ± 0.01	71.22 ± 0.00	28.83 ± 0.03	87.09 ± 0.00	91.68 ± 0.00	94.56 ± 0.00
	AGA	7.81 ± 0.02	62.83 ± 0.01	29.37 ± 0.01	84.88 ± 0.01	90.63 ± 0.02	94.27 ± 0.02
	MMGA	7.70 ± 0.02	62.93 ± 0.00	29.34 ± 0.02	84.92 ± 0.01	91.15 ± 0.01	94.27 ± 0.03
	(ALA+AGA)/2	9.87	67.03	29.10	85.99	91.16	94.42
	Gossip Gain	-1.41%	0.16%	-0.10%	0.05%	0.57%	0.00%
	Converge Steps	> 8000	7200	> 8000	4350	4400	1250
Local training	ALA	33.48 ± 0.01	90.61 ± 0.00	40.43 ± 0.01	93.30 ± 0.00	94.11 ± 0.00	97.65 ± 0.00
	AGA	2.88 ± 0.03	33.70 ± 0.01	12.98 ± 0.00	52.84 ± 0.00	33.91 ± 0.03	51.50 ± 0.01
	MMGA	1.00 ± 0.04	44.25 ± 0.01	19.77 ± 0.01	79.31 ± 0.01	36.23 ± 0.02	68.75 ± 0.01
	(ALA+AGA)/2	18.18	62.16	26.71	73.07	59.01	74.58
	Gossip Gain	-65.28%	31.31%	52.31%	50.09%	51.53%	33.50%
	Converge Steps	> 8000	3300	> 8000	2100	4600	2400

4 Major Findings and Discussions

4.1 Comparing MDSL and MAZEL under the same settings

Experiment settings We run experiments on 16 decentralized sites, where each site trains a model under two different settings: (1) ViT-B-32 CLIP-pretrained on all parameters, ViT-B-32 CLIP-pretrained on the classification head, and ViT-B-32 ImageNet pretrained on all parameters; (2) MAZEL: ViT-B-32 CLIP-pretrained model with all parameters being trainable. Please find all results in Table 1. We use the Adam optimizer for experiments Kingma & Ba (2014).
All the experiments are conducted under the setting of Random-2 De Vos et al. (2024) on 16 sites, which means that each time one other site is randomly picked to gossip. The batch size is always set to 64. Random (uniform) indicates that the weights in model merging are equally weighted. In Random (softmax), the weights in model merging are not equal; we calculate the scores according to Zhu et al. (2025). In the complete graph, each model communicates with everyone else each time. Local training corresponds to the setting where there is no communication between sites at all.

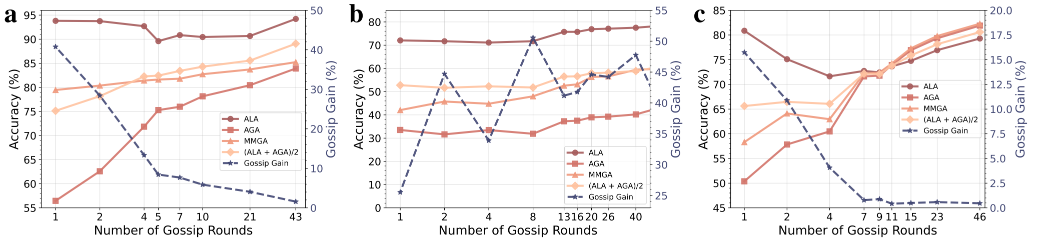

Figure 2: Impact of the number of gossip rounds on various performance metrics for CLIP-pretrained ViT-B/32 models Radford et al. (2021b) across different finetuning strategies and dataset settings. Abbreviations: **ALA**: Average local test accuracy, **AGA**: Average global test accuracy, **MMGA**: Merged model global test accuracy, **MDSL**: Metadata-dependent supervised learning, **MAZEL**: Metadata-agnostic zero-shot learning. See definitions in Definition 1. **Subfigure a**: Full finetuned under MAZEL settings using the CIFAR-100 dataset. **Subfigure b**: Classification head-only finetuned under MDSL settings using the CIFAR-100 dataset. **Subfigure c**: Fully finetuned under MAZEL settings using eight diverse datasets described in Section 5.

All experiments are conducted on one Nvidia A100 (80GB) GPU. Each set of experiments can be completed within 1-2 days. We adopt the recommended learning rate of 10^{-3} for MDSL (Zhu et al., 2025) and a learning rate of 10^{-5} for MAZEL (Ilharco et al., 2023).

Dataset To compare MDSL and MAZEL, we require a setting where MDSL can be meaningfully implemented. Since MDSL is not feasible in scenarios where different nodes possess distinct datasets, we simulate non-IID data distributions across sites by sampling from a Dirichlet distribution with $\alpha = 0.1$. We conduct our experiments using the CIFAR-100 Krizhevsky (2009) and the Kvasir V2 Pogorelov et al. (2017a). The dataset assigned to each site is biased samples drawn from these datasets. CIFAR-100 is a widely used dataset, and it is highly probable that similar images were included in the CLIP model pretraining data. To mitigate this potential overlap, we also evaluate our approach on Kvasir V2, a medical imaging dataset that is unlikely to be part of the pretraining dataset of CLIP. This hypothesis is supported by the fact: it achieves about 53% test accuracy on the CIFAR-100, whereas its performance on the Kvasir V2 test set is 0%, indicating Kvasir V2 were not part of pretraining data.

4.1.1 Results analysis

Key observation 1: *Under MAZEL, local models generalize well to a global test set under highly heterogeneous scenarios comparing to under MDSL.*

As shown by the results in Table 1, Claim 1, that local models generally fail to generalize well to a global test set in highly heterogeneous scenarios, holds true in the MDSL setting. Average Global Test Accuracy (AGA), which quantifies the global generalization performance of local models, consistently demonstrates poor performance across all MDSL configurations. However, this trend does not persist in the MAZEL setting, where local models exhibit improved generalization capabilities.

In MDSL, the best-performing configuration involves using a CLIP-pretrained model with only the classification head unfrozen during training. In the CIFAR-100 experiments, we observe that AGA improves 11.82% from local training to adopting random uniform decentralized learning. In contrast, under MAZEL, AGA increases 27.81%. This suggests that local models in MAZEL inherently possess strong global generalization capabilities.

A similar trend is observed in the Kvasir V2 experiments, where the improvement is even more pronounced. The AGA increases by 40.49%, reaching 91.99%, which is remarkably close to the ALA of 93.54%, which means the local generalization has been transferred to global generalization well.

One intriguing observation is that both MDSL and MAZEL exhibit better performance on Kvasir V2 than on CIFAR-100, despite the fact that Kvasir V2 being dissimilar to CLIP's pretraining data. We attribute this discrepancy to the difference in class cardinality: Kvasir V2 contains only 8 classes, meaning that even a random guess yields a 12.5% accuracy, whereas CIFAR-100 has 100 classes, making it a significantly more challenging classification task.

Key observation 2: *Under MAZEL, local models achieve faster convergence in terms of ALA, AGA and GG compared to under MDSL.*

Previous studies Lian et al. (2017); Koloskova et al. (2020); Kong et al. (2021) have shown that decentralized learning under MDSL exhibits slow convergence. Our findings align with these results: under MDSL with full finetuning, neither the CLIP-pretrained nor the ImageNet-pretrained model achieves convergence within 8000 training steps. Even in the case where only the classification head of the CLIP-pretrained model is finetuned, convergence still takes approximately twice as many steps as in the MAZEL setting.

Furthermore, under the MAZEL setting, the Gossip Gain (GG) score remains below 4% for both Random (Uniform) and Random (Influence-weighted) strategies. In contrast, under MDSL, the GG score ranges from 21.06% to 72.02%, indicating significant potential for further improving the global generalization performance of local models. This suggests that even when the ALA and AGA curves have reached a plateau, MDSL still has considerable room for enhancement through additional gossip.

In conclusion, the above results highlight that decentralized learning under MDSL and MAZEL can yield significantly different outcomes in key metrics, such as the global generalization of local models and convergence speed. Given that MAZEL better aligns with real-world decentralized learning applications, we strongly encourage researchers to conduct experiments within the MAZEL framework to ensure a more comprehensive evaluation of the proposed algorithms.

4.2 The impact of gossip rounds on local generalization and global generalization

Experiments settings We conduct experiments on 16 decentralized sites, where each site trains a model under two different settings: (1) MDSL: ViT-B-32 CLIP-pretrained model with classification head only finetuning; (2) MAZEL: ViT-B-32 CLIP-pretrained model with full finetuning. We follow the same data partitioning method as in Subsection 4.1, distributing CIFAR-100 among the 16 sites.

All experiments follow the Random-2 (Uniform) protocol, meaning that at each gossip round, a site randomly selects one other site for communication and models are merged using equal weights.

For MDSL, training consists of 8050 steps with gossip intervals set at 100, 200, 300, 400, 500, 600, 1000, 2000, 3000, and 6000 steps. The corresponding number of gossip rounds is 80, 40, 26, 20, 16, 13, 8, 4, 2, and 1, respectively.

Since our previous results in Table 1 indicate that convergence is faster under MAZEL, we set the total number of training steps for MAZEL at 2150. The gossip intervals are set at 50, 100, 200, 300, 400, 500, 1000, and 2000, with the corresponding number of gossip rounds being 43, 21, 10, 7, 5, 4, 2, and 1, respectively.

4.2.1 Results analysis

As illustrated in Figure 2, in MAZEL, ALA initially decreases as the number of communication rounds increases but subsequently improves. In contrast, AGA consistently increases throughout. This observation suggests that for practitioners who prioritize local generalization, there exists a critical range of communication rounds that should be avoided. Specifically, when the total communication round is near 5, local generalization is at its lowest. Either increasing or decreasing the gossip frequency improves ALA, leading to better local model performance.

However, when considering AGA, global generalization continuously improves as the number of communication rounds increases. When averaging ALA and AGA, the highest performance is observed at a gossip interval of 200 steps. Nevertheless, this should not be interpreted as a universal optimal setting, as the ideal number of communication rounds depends on the priorities and preferences of the participating site owners. Moreover, this U-shaped pattern becomes even more evident in subsequent experiments (Figure 2 **c**).

In contrast, under MDSL, both ALA and AGA exhibit a nearly monotonous increasing trend as the number of communication rounds increases.

5 MAZEL Baselines: 8-site and 16-site Benchmarks

To advance decentralized learning under MAZEL framework, we introduce two benchmark baselines:

- 8-Site Baseline: MNIST LeCun (1998), Cars Krause et al. (2013), DTD Cimpoi et al. (2014), EuroSAT Helber et al. (2019), GTSRB Stallkamp et al. (2011), RESISC45 Cheng et al. (2017), SUN397 Xiao et al. (2016), SHVN Netzer et al. (2011).
- 16-Site Baseline: MNIST, Cars, DTD, EuroSAT, GTSRB, RESISC45, SUN397, SVHN, Dogs Khosla et al. (2011), CUB-200-2011 Wah et al. (2011), Weather Xiao et al. (2021), MangoLeafBD Ahmed et al. (2023), Garbage CCHANG (2018), Beans Lab (2020), Kvasir Pogorelov et al. (2017b), and FashionMNIST Xiao et al. (2017).

Datasets for the 8-site baseline have been widely adopted in the model merging community as standard benchmarks. Model merging plays a critical role in decentralized learning, as it occurs at every gossip communication step. By aligning our dataset selection with those commonly used in the model merging community, we aim to bridge research efforts between the two fields. This alignment enables decentralized learning researchers to effectively evaluate different model merging techniques and determine which approach best enhances the performance of gossip-based updates.

However, experiments with only 8 sites may not fully capture the complexities of decentralized learning. Therefore, we expanded the benchmark to 16 sites, incorporating additional datasets inspired by those used in the EMR-merging study Huang et al. (2024). This extension ensures a more diverse and comprehensive evaluation, enabling researchers to analyze decentralized learning performance across a broader range of heterogeneous data distributions.

Below, we present experimental results evaluating various techniques on the MAZEL Baseline 8-Site in Table 2 and MAZEL Baseline 16-Site benchmarks in Table B.1. We implemented two gossip schedulers to regulate the timing of communication rounds: the Two-Stage Gossip Scheduler and the Cosine Gossip Scheduler. Check Section D for more details.

Table 2: Comparison of various communication frequencies, learning rate schedulers, and gossip schedulers under the proposed MAZEL setting. Abbreviations: **ALA**: Average local test accuracy, **AGA**: Average global test accuracy, **MMGA**: Merged model global test accuracy, **MAZEL**: Metadata-agnostic zero-shot learning, **Local**: Local training only with no gossip. See definitions in Definition 1. All experiments used CLIP-pretrained ViT-B/32 Radford et al. (2021b) on the 8 datasets described in Section 5. Experimented configurations include: **A**: Random neighbor selection with influence-based weights Zhu et al. (2025), **B**: Random neighbor selection with uniform weights, **C**: Cosine learning rate scheduler, **D**: Uniform learning rate scheduler with warmup, **E**: Two-stage gossip scheduler (9 total gossip rounds), **F**: Cosine gossip scheduler (9 total gossip rounds), **G**: Uniform gossip scheduler (9 total gossip rounds), **H**: Cosine gossip scheduler (15 total gossip rounds), **I**: Cosine gossip scheduler (46 total gossip rounds). Two-stage gossip scheduler and cosine gossip scheduler are defined in Section D.

	A+D+F	B+D+F	A+C+E	B+C+E	A+D+G	A+D+E	A+D+H	A+D+I	A+C+F	Local
ALA	72.62	72.38	76.81	74.56	74.57	80.68	76.82	80.84	73.52	88.66
AGA	73.92	70.55	76.86	71.81	76.18	75.44	79.05	82.95	75.67	38.29
MMGA	74.31	71.00	77.73	73.44	76.61	78.32	79.35	83.25	76.03	58.10
(ALA+AGA)/2	73.27	71.47	76.84	73.19	75.38	78.06	77.94	81.90	74.60	63.48
Gossip Gain	0.53%	0.64%	1.13%	2.27%	0.56%	3.82%	0.38%	0.36%	0.48%	51.74%

6 Conclusion and Limitations

We challenge conventional assumptions in decentralized learning by distinguishing Metadata-Dependent Supervised Learning (MDSL) from Metadata-Agnostic Zero-Shot Learning (MAZEL), showing that many common claims break under MAZEL. To better match real-world constraints, we release realistic benchmarks (8 and 16 datasets). These results underscore the need to evaluate decentralized methods under practical conditions. Our analysis is limited to image classification; extending to broader task families remains future work.

REFERENCES

Sarder Iftekhar Ahmed, Muhammad Ibrahim, Md Nadim, Md Mizanur Rahman, Maria Mehjabin Shejunti, Taskeed Jabid, and Md Sawkat Ali. Mangoleafbd: A comprehensive image dataset to classify diseased and healthy mango leaves. *Data in Brief*, 47:108941, 2023.

Marco Bornstein, Tahseen Rabbani, Evan Z Wang, Amrit Bedi, and Furong Huang. SWIFT: Rapid decentralized federated learning via wait-free model communication. In *The Eleventh International Conference on Learning Representations*, 2023.

CCHANG. Garbage classification. https://www.kaggle.com/ds/81794, 2018.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Martijn De Vos, Sadegh Farhadkhani, Rachid Guerraoui, Anne-Marie Kermarrec, Rafael Pires, and Rishi Sharma. Epidemic learning: Boosting decentralized learning with randomized communication. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *arXiv preprint arXiv:2405.17461*, 2024.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6t0Kwf8-jrj.

Salma Kharrat, Marco Canini, and Samuel Horvath. Decentralized personalized federated learning. *arXiv preprint arXiv:2406.06520*, 2024.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian Stich. Consensus control for decentralized deep learning. In *International Conference on Machine Learning*. PMLR, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Makerere AI Lab. Bean disease dataset, January 2020. URL https://github.com/AI-Lab-Makerere/ibean/.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. In *CS 231N*, 2015.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized SGD with heterogeneous data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lu Li, Tianyu Zhang, Zhiqi Bu, Suyuchen Wang, Huan He, Jie Fu, Yonghui Wu, Jiang Bian, Yong Chen, and Yoshua Bengio. Map: Low-compute model merging with amortized pareto fronts via quadratic approximation. *arXiv preprint arXiv: 2406.07529*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, 2018.

Tao Lin, Sai Praneeth Karimireddy, Sebastian Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2016.

Songtao Lu and Chai Wah Wu. Decentralized stochastic non-convex optimization over weakly connected time-varying digraphs. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, and Dan Alistarh. Asynchronous decentralized sgd with quantized and local updates. *Advances in Neural Information Processing Systems*, 2021.

Angelia Nedi'c and Alex Olshevsky. Distributed optimization over time-varying directed graphs. volume 60, pp. 601–615. IEEE, 2014.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Konstantin Pogorelov, Kristin R. Randel, Carsten Griwodz, Sigrun L. Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter T. Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys '17, pp. 164–169, New York, NY, USA, 2017a. ACM. ISBN 978-1-4503-5002-0. doi: 10.1145/3083187.3083212. URL http://doi.acm.org/10.1145/3083187.3083212.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 2017b.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pp. 1453–1460. IEEE, 2011.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D2: Decentralized training over decentralized data. In *International Conference on Machine Learning*. PMLR, 2018.

J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 509–517, 2017.

Thijs Vogels, Lie He, Anastasia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U Stich, and Martin Jaggi. Relaysum for decentralized deep learning on heterogeneous data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

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.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Mitchell Wortsman, Gabriel Ilharco, S. Gadre, R. Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Y. Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *International Conference on Machine Learning*, 2022. doi: 10.48550/arXiv.2203.05482.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://doi.org/10.1109/ICASSP49357.2023.10095969.

Haixia Xiao, Feng Zhang, Zhongping Shen, Kun Wu, and Jinglin Zhang. Classification of weather phenomenon from images by using deep convolutional neural network. *Earth and Space Science*, 8(5):e2020EA001604, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119: 3–22, 2016.

Jie Xu, Wei Zhang, and Fei Wang. A(dp)^{2}sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Neural Information Processing Systems*, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*, 2021.

Le Yu, Yu Bowen, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2311.03099.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.01100.

Tongtian Zhu, Wenhao Li, Can Wang, and Feng Xiang He. DICE: Data influence cascade in decentralized learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2TIYkqieKw.

A LLM Usage Declaration

We use LLMs solely for grammar refinement and LaTeX code debugging.

B Additional Experiment Results

We present experimental results evaluating various techniques on the MAZEL Baseline 16-Site in Table B.1. Please check Section D for definitions of the Two-Stage Gossip Scheduler and the Cosine Gossip Scheduler.

Table B.1: Comparison of various communication frequency, learning rate schedulers, and gossip schedulers. All experiments using random neighbor selection with influence-based weights Zhu et al. (2025) using CLIP-pretrained ViT-B/32 on 16 datasets. Experimented configurations include: **A**: cosine gossip scheduler (9 total gossips), **B**: cosine gossip scheduler (15 total gossips), **C**: cosine gossip scheduler (46 total gossips), **D**: Cosine learning rate scheduler, **E**: uniform learning rate scheduler with warmup, **F**: two-stage gossip scheduler (9 total gossips).

	D+F	A+D	A+E	B+E	C+E
ALA	63.17	81.81	84.18	83.66	83.09
AGA	58.16	57.64	55.78	59.77	70.34
MMGA	58.77	61.03	60.96	63.57	73.19
(ALA+AGA)/2	60.67	69.73	69.98	71.72	76.72
Gossip Gain	1.05%	5.88%	9.29%	6.36%	4.05%

C Detailed Related work

Decentralized Learning. Decentralized learning has emerged as a powerful paradigm for distributed optimization, enabling collaborative model training across multiple nodes without the need for a centralized coordinator. Decentralized Stochastic Gradient Descent (DSGD) (Lian et al., 2017) serves as a prominent example of decentralized learning algorithms. Building on the principles of DSGD, the field of decentralized learning has expanded rapidly, driven by the need for adaptable and efficient solutions in diverse and dynamic environments. Modern decentralized algorithms have evolved to address challenges such as time-varying network topologies (Nedi'c & Olshevsky, 2014; Lu & Wu, 2020; Koloskova et al., 2020; Ying et al., 2021), enabling robust performance even in scenarios where communication links between nodes fluctuate. Moreover, the incorporation of asynchronous communication protocols (Lian et al., 2018; Xu et al., 2021; Nadiradze et al., 2021; Bornstein et al., 2023) has empowered decentralized methods to overcome latency and synchronization barriers, further enhancing their scalability. Another critical advancement is the ability to handle heterogeneous data distributions (Tang et al., 2018; Vogels et al., 2021; Le Bars et al., 2023), which mirrors the realities of non-IID data commonly encountered in real-world decentralized systems.

Zero-shot Classification. CLIP Radford et al. (2021a) pioneers contrastive learning between images and text, demonstrating that large-scale natural language supervision enables strong zero-shot transfer across diverse vision tasks. SigLIP Zhai et al. (2023) introduces a pairwise sigmoid loss for language-image pretraining, enabling efficient scaling of batch sizes while improving zero-shot accuracy on ImageNet. CLAP Wu et al. (2023) extends contrastive learning to the audio domain, training a large-scale language-audio model using feature fusion and keyword-to-caption augmentation for superior zero-shot classification and retrieval. DINOv2 Oquab et al. (2023) advances self-supervised learning by training large ViT models on curated datasets, producing robust all-purpose visual features that surpass OpenCLIP in most benchmarks.

Model-merging "Model soups"Wortsman et al. (2022) averages the weights of multiple fine-tuned models improves accuracy and robustness without increasing inference time. Task vectors Ilharco et al. (2023) represent directions in weight space; by adding or subtracting these vectors, models can acquire or diminish specific capabilities. DARE Yu et al. (2023) demonstrated that language models could absorb new abilities by assimilating parameters from homologous models without retraining, a process facilitated by the DARE method to sparsify delta parameters. Ties-merging Yadav et al. (2023) addressed parameter interference in model merging, a method that resolves conflicts by resetting minimally changed parameters and aligning parameter signs. AdaMerging Yang et al. (2024) is an adaptive approach that autonomously learns merging coefficients without relying on original training data, enhancing performance across multiple tasks. MAP Li et al. (2024), a low-compute algorithm that efficiently identifies a Pareto set of scaling coefficients for merging models, reflecting the trade-offs involved. Collectively, these studies contribute to the evolving landscape of model merging, offering diverse strategies to combine models effectively.

D Details about the two gossip schedulers

From our experiments, we observed that allocating more communication rounds to the early training stage rather than the later stages is beneficial. Excessive communication toward the end of training may disrupt model convergence.

Two-Stage Gossip Scheduler. The Two-Stage Gossip Scheduler divides the training process into two phases: the early stage and the convergence stage, each with a distinct gossip interval hyperparameter. The early stage involves more frequent communication, while the convergence stage adopts a less frequent communication schedule to prevent unnecessary perturbations. However, even in the later stage, where the gossip interval is larger, careful hyperparameter tuning is required to ensure that no gossip occurs in the final training steps, as this could destabilize the model.

Cosine Gossip Scheduler. To further address this issue, we introduce the Cosine Gossip Scheduler, inspired by the cosine learning rate scheduler Loshchilov & Hutter (2016). This approach gradually reduces the probability of gossiping as training progresses, ensuring that the communication frequency is significantly lower toward the end of training. This minimizes potential disruptions in the final optimization steps, preventing oscillations in model performance.

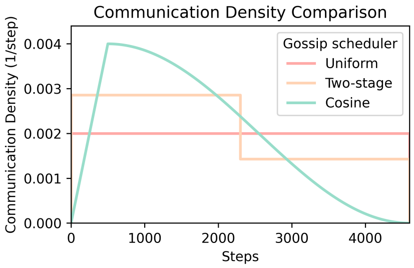

Figure D.1: Comparison of different gossip schedulers over 4,600 training steps, all with 9 total gossip rounds. The uniform gossip scheduler has fixed intervals of 500 steps. The two-stage scheduler uses a higher frequency in the first half, followed by lower frequency in the second half. The cosine gossip scheduler uses a cosine decay with an initial warmup.

In conclusion, the local training step and the gossip step in decentralized learning function like a non-zero-sum tug-of-war. The local training step promotes local models' ability to generalize locally, whereas the gossip step enhances their global generalization. Increasing the learning rate strengthens local training, thereby benefiting ALA, while more frequent gossip updates improve AGA. As shown in Table 2, the combination of influence-based weighting, a uniform learning rate scheduler with warmup, and a frequent cosine gossip scheduler leads to higher ALA and AGA while maintaining a low GG.

E Potential Q&A

Q1: Why is MAZEL advantageous over MDSL?

A1: Both MDSL and MAZEL train a classifier using cross-entropy loss, but differ in how the classifier is parameterized and what supervision they use.

MDSL uses a frozen CLIP visual encoder and learns a parametric classifier (MLP with weights $W \in \mathbb{R}^{d\times c}$):

$$\min_{W} \text{CrossEntropy}\left(\text{softmax}(Wv_i),\ y_i\right)$$

where $v_i \in \mathbb{R}^d$ is the CLIP image embedding and $y_i \in \{1, \ldots, c\}$ is the hard label.

MAZEL also uses a frozen CLIP text encoder to get fixed label embeddings $\{t_k\}_{k=1}^{c} \subset \mathbb{R}^d$, and instead of learning a classifier, it aligns the trainable image embedding $v_i(\theta)$ with the frozen text embeddings via cosine similarity:

$$\min_{\theta} \text{CrossEntropy}\left(\text{softmax}\left(\left[\frac{t_1^\top v_i(\theta)}{\|t_1\|\|v_i(\theta)\|}, \ldots, \frac{t_c^\top v_i(\theta)}{\|t_c\|\|v_i(\theta)\|}\right]\right),\ y_i\right)$$

This shifts the burden of semantic alignment to the pre-trained CLIP text encoder, removing the need to learn per-site classification heads and making the process **metadata-agnostic**.

Q2: Does the shared CLIP text embeddings imply metadata sharing across sites?

A2: No. All sites use the **publicly available CLIP text encoder** to generate the same set of label embeddings from natural language prompts (e.g., “a photo of a dog”), but no metadata such as class indices or taxonomies are exchanged between sites. This avoids privacy violations that would occur if class label mappings (e.g., “index 5 =HIV”) were shared.

Each site can independently construct the same set $\{t_k\}_{k=1}^{c}$ without coordination or sensitive disclosure.

Q3: Is the comparison with MDSL fair? Doesn’t MAZEL use more expressive supervision?

A3: The intent is not to claim superior performance of MAZEL over MDSL. We would like to point out that most decentralized learning works are evaluated under the MDSL setting, while in the MAZEL setting some of their claims/results might not be true. Our major claim is that conducting experiments within MAZEL is needed in decentralized learning scenarios that each of the participants are not willing to expose its meta-data (e.g. the class names in classification). We encourage researches in this domain to not only test in MDSL setting for their newly proposed algorithms but also test under MAZEL scenarios which is more realistic.

Q4. Could MAZEL be viewed as multi-task learning (MTL) under a different name? Why not use existing MTL methods?

A4: MAZEL and MTL differ fundamentally in both architecture and assumptions. In MTL, tasks are typically known and fixed in advance, and training is performed jointly (or in a coordinated fashion) across those tasks often on a centralized server.

MAZEL operates in a fully decentralized peer-to-peer setting, where each node has its own private task, and no global task information is available. In addition, MAZEL makes no assumption about task relationships or label space alignment, whereas MTL typically assumes either a shared backbone or joint modeling objective.

Therefore, while both deal with learning across tasks, MAZEL specifically solves the problem of decentralized learning under label-space heterogeneity and metadata constraints, which is not addressed by standard MTL frameworks. To validate this, we also compare with centralized MTL baselines, and the results are shown below:

Q5. Experiments focus only on image classification; broader tasks (e.g., segmentation, NLP, multi-modal) would better support claims of generality.

A5: We intentionally target image classification to directly compare with the dominant evaluation paradigm in decentralized learning. This work establishes the *MAZEL* setting as a foundation; extending to NLP, time-series, segmentation, and multi-modal tasks is a key direction for future work (explicitly noted in our limitations).

Table E.1: Test accuracy (%) of MTL, MAZEL variants, and MDSL baselines on CIFAR-100, KvairV2, and the 16-site medical dataset, illustrating that MAZEL provides consistent gains across diverse decentralized image classification benchmarks.

	CIFAR-100	**KvairV2**	**16-Site Baseline**
MTL	89.26	91.25	88.02
MAZEL-ALA	90.67	93.54	72.38∼80.84
MAZEL-AGA	80.65	91.99	70.55∼82.95
MDSL-ALA	78.95	88.17	-
MDSL-AGA	45.52	65.10	-

Q6. CLIP's zero-shot generalization can vary in underrepresented or out-of-domain settings; more analysis is desired.

A6: Our central claim concerns the *MAZEL framework*, not CLIP as the universally best model. We use CLIP as a proof-of-concept and deliberately include the out-of-domain Kvasir V2 medical dataset (where CLIP's zero-shot score is low). Under MAZEL, decentralized training still differs markedly from MDSL and attains strong performance even in this setting (e.g., *AGA* 91.99% vs. *ALA* 93.54%), suggesting robustness of the evaluation paradigm itself. A comprehensive comparison with alternative backbones (e.g., DINOv2, SigLIP v2) is valuable future work but beyond the scope of this initial introduction of MAZEL.

Q7. MAZEL currently relies on CLIP-style aligned text/image embeddings; compatibility beyond CLIP should be discussed.

A7: MAZEL is model-agnostic by construction: it requires a zero-shot (or promptable) evaluator to avoid shared metadata and predefined heads. In principle, other pretrained backbones with label- or prompt-conditioned scoring (e.g., vision-only or multi-modal encoders) can be substituted. We discuss this extensibility and leave systematic backbone sweeps to future work.

Q8. Be specific about mentioned Local–Global "Tug-of-War".

A8: Each agent i performs local descent on its data,

$$\theta_i^{t+\frac{1}{2}} = \theta_i^t - \eta \, \nabla \hat{L}_i(\theta_i^t),$$

increasing *Average Local Accuracy* (ALA), then gossips with neighbors via mixing matrix A,

$$\theta_i^{t+1} = \sum_{j \in N_{\text{in}}(i)} A_{ij} \, \theta_j^{t+\frac{1}{2}} = \theta_i^{t+\frac{1}{2}} + \sum_{j \in N_{\text{in}}(i)} A_{ij} \big(\theta_j^{t+\frac{1}{2}} - \theta_i^{t+\frac{1}{2}}\big),$$

promoting consensus and *Average Global Accuracy* (AGA). Early in training, parameter disparity makes the consensus term large and beneficial; late in training, excessive gossip can induce oscillations around well-tuned local optima. Both schedulers thus *front-load* communication and *anneal* it to stabilize convergence.

Q9. The paper uses a general gossip-based Decentralized Learning algorithm for MDSL. Since the experiments are done for 8 and 16 sites and the overheads are not studied, couldn't personalized decentralized learning works like FACADE that use personalized classification heads could potentially solve this problem by having one head per node?

A9: While a personalized head approach (merging only the backbones) is metadata-agnostic in the sense that nodes don't share their final layer parameters, MAZEL operates under a stricter, more realistic set of assumptions about privacy and autonomy. It's hard to create a single, unified "global model" that can perform classification across all sites' tasks (need to retrain the classification head). The practitioners can merge the backbones, but the resulting model is incomplete without a classification head. Creating a new global head would require reintroducing the metadata-sharing problem. A key advantage of MAZEL framework is its ability to produce a powerful merged model that is immediately functional on a global test set. This is possible because classification is performed via similarity to text embeddings, not a fixed output layer. This merged model represents a valuable consensus of knowledge from all participating nodes.

F Metadata-Agnostic Zero-Shot Learning

Algorithm 2 Metadata-Agnostic Zero-Shot Learning (MAZEL)

```
Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; initial parameters $\{\theta_i^0 \in \mathbb{R}^d\}_{i \in \mathcal{V}}$; iterations $T$; mixing matrix $A$;
    frozen text encoder $f_{\theta_T}$ (e.g., CLIP)

Initialization
for all $i \in \mathcal{V}$ do
    (in parallel)
    Generate textual prompts for each local class (e.g., “a photo of a {class_name}”).
    Create fixed text embeddings with the frozen encoder:
        $T_i \coloneqq \{t_k\}_{k=1}^{c_i}, \qquad t_k = f_{\theta_T}(\text{prompt}_k).$
end for

Training Loop
for $t = 0$ to $T - 1$ do
    for all $i \in \mathcal{V}$ do
        (in parallel)
        Local Training
        Sample a mini-batch $\{(x_{i,j}^t, y_{i,j}^t)\}_{j=1}^B$ from local data $P_i$.
        Obtain visual embeddings with the image encoder:
            $v_{i,j}(\theta_i^t) \leftarrow \text{ImageEnc}(x_{i,j}^t;\, \theta_i^t).$
        Compute class scores via cosine similarity:
            $s_{i,j} \coloneqq \big[\text{sim}(v_{i,j}(\theta_i^t), t_1), \ldots, \text{sim}(v_{i,j}(\theta_i^t), t_{c_i})\big].$
        Compute loss (cross-entropy on softmax scores):
            $L(\theta_i^t) \coloneqq \text{CrossEntropy}\big(\text{softmax}(s_{i,j}),\, y_{i,j}^t\big).$
        Update local model (one or more optimizer steps):
            $\theta_i^{t+\frac{1}{2}} \leftarrow \text{Optimizer}\big(\theta_i^t,\, \nabla_{\theta_i^t} L(\theta_i^t)\big).$
        Communication
        Send $\theta_i^{t+\frac{1}{2}}$ to out-neighbors; receive $\{\theta_l^{t+\frac{1}{2}}\}_{l \in \mathcal{N}_{\text{in}}(i)}$.
        Gossip Averaging
            $\theta_i^{t+1} \leftarrow \sum_{l \in \mathcal{N}_{\text{in}}(i)} A_{i,l}\, \theta_l^{t+\frac{1}{2}}.$
    end for
end for
```

G Impact Statement

This paper puts forward a new evaluation setting in decentralized learning, a topic with significant societal implications. Specifically, we aim to bridge the gap between decentralized learning's real-world applications and algorithm evaluation.

The proposed MAZEL framework has the potential for positive societal impact by enabling privacy-preserving decentralized learning in sensitive domains such as healthcare, finance, and scientific research. By eliminating the need for shared metadata, MAZEL allows institutions to collaboratively train models without disclosing sensitive information such as patient diagnoses, internal label taxonomies, or proprietary data annotations. This can promote broader collaboration across organizations that were previously unable to share data due to regulatory or legal constraints, ultimately leading to more inclusive, equitable, and effective AI systems. However, the approach also raises potential negative implications. Since MAZEL relies on a large, fixed pretrained model (CLIP), its performance and fairness may inherit biases present in the pretraining data, especially when deployed in domains (e.g., medical imaging) that differ from the original training corpus. Additionally, enabling fully decentralized, metadata-agnostic learning may lower institutional barriers for large-scale model deployment without sufficient oversight, which could amplify risks related to misuse, accountability gaps, or uneven performance across underserved populations. These risks warrant further research into model auditing and domain adaptation techniques in high-stakes applications.