EFFICIENT DECENTRALIZED MULTI-TASK DATASET VALUATION VIA MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate and efficient dataset valuation is essential for enabling fair and transparent data marketplaces, especially when multiple contributors provide data for training multi-task models. Most existing valuation methods, however, are limited to single-task settings, overlooking scenarios where a buyer aims to optimize performance across multiple downstream tasks. Moreover, traditional valuation approaches—such as Shapley-based or retraining-based methods—are computationally expensive and poorly suited for decentralized environments without a trusted central coordinator and with strict privacy constraints. We propose DMVM (Decentralized Multi-task Valuation via Model Merging), a novel framework that bypasses retraining and data sharing by leveraging task arithmetic to infer dataset contributions directly from model combinations. Instead of retraining or sharing raw data, DMVM quantifies how models trained on different datasets combine in parameter space to infer each dataset's marginal utility across multiple tasks. This formulation yields a valuation process that is scalable, computationally efficient, and explicitly aligned with multi-task generalization behavior. To support decentralized deployment, we introduce a secure aggregation protocol that enables collaborative valuation without revealing individual model parameters or private data. We also provide theoretical error bounds characterizing the approximation quality of DMVM and validate our framework through comprehensive experiments on computer vision and natural language processing tasks.

1 INTRODUCTION

Dataset valuation has emerged as a central challenge in the design of modern data marketplaces, where multiple sellers offer proprietary datasets and buyers seek to improve model performance on specific target tasks. Accurately estimating the value of a seller's data prior to purchase is vital for fair pricing, informed decision-making, and sustaining trust in the marketplace. This estimation aims to quantify the marginal utility a dataset provides to a learning algorithm for a specific task, guiding buyers toward high-impact acquisitions and incentivizing high-quality data contributions from sellers. Prior work has made substantial progress toward formalizing the problem of data valuation. Game-theoretic methods such as Data Shapley (Ghorbani & Zou, 2019) offer theoretically principled metrics by averaging marginal contributions across all data subsets. Influence-function techniques estimate the effect of training samples on model predictions (Koh & Liang, 2017; Zhang & Amiri, 2025), while gradient-based methods (e.g., reinforcement learning or utility-based proxies) seek to capture a dataset's alignment with target tasks (Yoon et al., 2020; Cai, 2024). While theoretically compelling, these methods face critical obstacles when applied to real-world data marketplaces, as discussed next.

A primary barrier to practical deployment is computational overhead: estimating marginal contributions, especially via Shapley-based methods, typically requires retraining or fine-tuning models across an exponential number of dataset combinations. This becomes infeasible at scale, particularly for large models or a large number of sellers. A second challenge is data privacy. Most valuation techniques assume centralized access to raw data, an unrealistic expectation in many real-world marketplaces involving sensitive or regulated datasets, e.g., medical imaging or personal data. In such settings, sellers are unlikely to permit raw data sharing, even for valuation purposes.

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

Beyond scalability and privacy, a third and often overlooked limitation is the narrow focus on single-task valuation. Most existing methods assess a dataset's utility with respect to a single predefined task. However, in realistic settings, buyers often pursue multi-task learning objectives (e.g., multi-task vision models performing both detection and segmentation, or multi-task language models performing sentiment analysis, topic classification, and summarization). Valuation in a multi-task learning context is inherently more complex because a dataset's utility may differ across tasks, and its value cannot be captured by a single performance metric. Despite the growing importance of multi-task models in practice, there has been little work addressing dataset valuation in this setting, leaving a significant methodological gap (Ahmadian et al., 2024; Zhu et al., 2026).

Meanwhile, model merging has recently emerged as a strategy for combining models fine-tuned on different datasets without requiring access to, or retraining on, combined data (Ilharco et al., 2023; Wortsman et al., 2022; Matena & Raffel, 2022; Wang et al., 2025c; Alipour & Amiri, 2025). Starting from a common base, models are fine-tuned independently; their parameter updates, often represented as task vectors (Ilharco et al., 2023) are then merged—via averaging, interpolation, or more advanced techniques such as task-vector trimming or balancing—to approximate the effect of joint training (Yadav et al., 2023; Yu et al., 2024). Empirically, Tao et al. (2025) showed that merging models fine-tuned on disjoint datasets can closely approximate the performance of a model trained on the combined data. They further show a strong linear correlation between the accuracy of the merged model and that of the model trained on the full data mixture, indicating that model merging can serve as an efficient and reliable proxy for mixture fine-tuning or multi-task learning. Moreover, Zhou et al. (2025) provide theoretical support for this intuition by deriving an error bound between the weights of a model trained via multi-task learning and a model merged with task-arithmetic.

In multi-task setting, the ideal way to assess a dataset's value is to measure its marginal improvement to a jointly trained multi-task model; that is, how much performance gain it contributes when all tasks are optimized jointly. However, such joint training is computationally prohibitive in decentralized marketplaces, as it requires access to all datasets and repeated retraining across combinations of sellers. To overcome this bottleneck, we draw inspiration from recent advances in model merging, which approximate the effects of joint multi-task training without the cost of training on combined datasets. Starting from a shared initialization, models are fine-tuned independently on individual datasets, and their parameter updates, often represented as task vectors, are combined through arithmetic operations in parameter space. Empirical and theoretical studies have shown that this process can closely replicate the outcomes of true multi-task learning (Li et al., 2025; Ilharco et al., 2023; Zhou et al., 2025; Na et al., 2024), suggesting that model merging can serve as an efficient proxy for evaluating cross-task contributions.

Building on this insight, we introduce DMVM (Decentralized Multi-task Valuation via Model Merging), a framework that formalizes multi-task dataset valuation through the lens of model merging. DMVM estimates each dataset's marginal utility by analyzing how its corresponding model influences the aggregated performance across a buyer's target tasks. Our formulation enables fully decentralized valuation without raw data exchange, incorporates a privacy-aware protocol to protect individual models during valuation, and provides a theoretical error bound that characterizes the approximation gap between DMVM-based valuation and the ground-truth multi-task valuation. We validate our framework through simulations of a decentralized data marketplace, showing consistent value estimates across vision and language benchmarks. Our main contributions are:

- **Efficient and decentralized valuation framework.** We introduce DMVM, a novel approach for multi-task dataset valuation through model merging. It operates without retraining or centralized data access, enabling a fully decentralized deployment, without a trusted broker or third party.
- **Privacy-aware protocol.** DMVM incorporates a secure aggregation protocol enabling collaborative valuation without revealing raw data or individual models, ensuring compatibility with privacy-sensitive applications.
- **Theoretical guarantees.** We derive error bounds that characterize the approximation quality of DMVM relative to ground-truth multi-task marginal contributions.
- **Empirical evaluation.** We demonstrate stable and reliable valuation performance of DMVM across representative vision and language benchmarks.

By bridging the domains of dataset valuation and model merging, DMVM establishes a scalable, privacy-aware, and theoretically grounded foundation for dataset valuation in multi-task market-

places. A comprehensive discussion of related work on data valuation, data marketplaces, and model merging is provided in Appendix A.2.

2 PRELIMINARIES

In a data marketplace, multiple sellers offer datasets of varying relevance and quality, and a buyer seeks to assess the contribution of each dataset to the performance of a learned model. This process, known as *dataset valuation*, provides a quantitative basis for fair pricing and informed data acquisition. A principled valuation metric should capture the *marginal utility* of each dataset; that is, how much additional benefit it brings when combined with others. We consider a data marketplace consisting of a set of n data sellers, where each seller i holds a private dataset $\mathcal{D}_i$, and the collection of all datasets is denoted by $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$. These datasets are not accessible to the buyer prior to purchase due to privacy or commercial constraints. The buyer is interested in a set of downstream tasks $\mathcal{T}$, where each task $t \in \mathcal{T}$ has a validation set $\mathcal{D}_t^{\text{val}}$. The buyer's objective is to train a model $\boldsymbol{\theta}$ that generalizes well across all tasks in $\mathcal{T}$.

If the buyer had access to the datasets of a subset of sellers $\mathcal{S} \subseteq [n] := \{1, ..., n\}$, a model could be trained via multi-task learning (MTL) as $\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}} = \arg\min_{\boldsymbol{\theta}} \sum_{i \in \mathcal{S}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_i)$, where $\mathcal{L}$ denotes the loss function. The corresponding utility of the subset $\mathcal{S}$ is defined as the average performance of the resultant model across all downstream tasks:

$$u(\mathcal{S}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Perf}_t(\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}}, \mathcal{D}_t^{\text{val}}), \tag{1}$$

where Perf_t denotes a task-specific evaluation metric (e.g., accuracy or negative loss) on task t[1]. Alternatively, utility can be expressed in terms of validation loss as $u(\mathcal{S}) = -\sum_{t \in \mathcal{T}} \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}}, \mathcal{D}_t^{\text{val}})$. The marginal contribution of seller i to a coalition $\mathcal{S}$ of other sellers is given by $u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})$, which quantifies the added utility from including dataset $\mathcal{D}_i$ when training jointly with datasets of the sellers in $\mathcal{S}$. The Shapley value (Shapley et al., 1953) provides a unique and principled method for aggregating marginal contributions across all possible coalitions. For seller i, it is defined as:

$$\phi_i = \frac{1}{n} \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}} \frac{u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}}, \tag{2}$$

where the summation averages seller i's marginal contribution across all subsets $\mathcal{S}$ that excludes i, weighted by their cardinality. This formulation satisfies key fairness axioms, including symmetry, linearity, and additivity, and is therefore widely regarded as the *ground-truth reference* for data valuation (Ghorbani & Zou, 2019). Despite its strong theoretical foundations, computing the Shapley value ϕ_i directly is computationally intractable in practice. It requires training models on all 2^{n-1} possible subsets $\mathcal{S} \subseteq [n] \setminus \{i\}$ to evaluate $u(\mathcal{S})$, which becomes exponentially expensive as n grows. Moreover, in real marketplaces, buyers do not have access to the raw datasets of sellers prior to purchase. Therefore, the utility function $u(\mathcal{S})$ cannot be computed without violating confidentiality. These challenges motivate the development of scalable and privacy-aware alternatives to the classical Shapley formulation. In the following section, we propose DMVM, a framework designed to overcome both computational and privacy limitations while maintaining theoretical faithfulness to multi-task marginal contributions.

3 METHODOLOGY

This section introduces our proposed framework for decentralized dataset valuation. Section 3.1 presents an efficient proxy for approximating the utility function $u(\cdot)$ without requiring model training. Section 3.2 then introduces a privacy-aware protocol that enables secure valuation without exposing individual models.

[1]We use uniform task weighting for simplicity; DMVM also supports buyer-specified task preferences through non-uniform task weights, as evaluated in Appendix A.8.3.

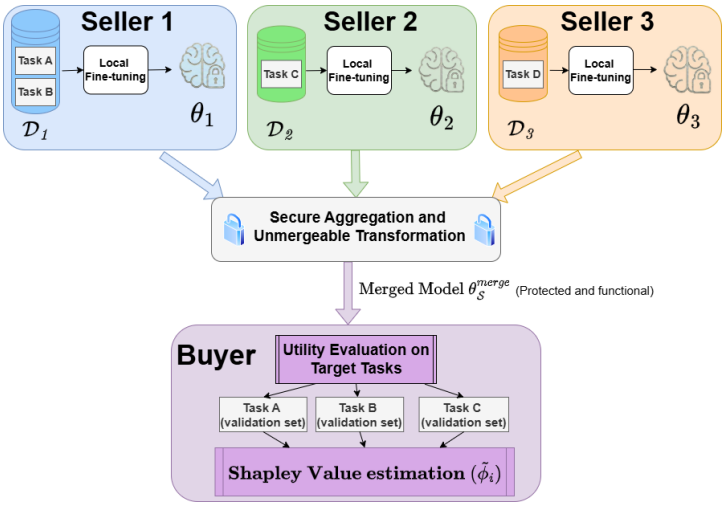

Figure 1: Overview of the DMVM framework for decentralized multi-task dataset valuation. Each seller trains a local model on its private dataset, and using a secure protocol, they produce a protected merged model for each coalition. The buyer evaluates the merged models on its target validation tasks to estimate seller contributions via Shapley values, without accessing sellers' private data or individual models.

3.1 Model Merging as a Proxy for Utility

We assume that all sellers either adopt a common pretrained model $\boldsymbol{\theta}^{\text{pre}}$ provided by the buyer or agree on a shared architecture. Each seller i locally fine-tunes this model on their private dataset $\mathcal{D}_i$, yielding a personalized model $\boldsymbol{\theta}_i$. To estimate the utility of any seller subset $\mathcal{S} \subseteq [n]$, the buyer merges the corresponding fine-tuned models $\{\boldsymbol{\theta}_i : i \in \mathcal{S}\}$ into a single model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$ using a merging strategy, specifically, *task arithmetic*. For brevity, we use $\boldsymbol{\theta}$ to denote full model parameters and their associated weight matrices. Building on recent theoretical and empirical findings (Tao et al., 2025; Zhou et al., 2025), we approximate the true utility $u(\mathcal{S})$ with the proxy $\hat{u}(\cdot)$ defined as $\hat{u}(\mathcal{S}) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{Perf}_t\big(\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}, \mathcal{D}_t^{\text{val}}\big)$, or equivalently, in terms of loss as $\hat{u}(\mathcal{S}) = -\sum_{t\in\mathcal{T}} \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}, \mathcal{D}_t^{\text{val}})$. The merged model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$ is obtained as $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}} = \boldsymbol{\theta}^{\text{pre}} + \alpha \sum_{i\in\mathcal{S}}(\boldsymbol{\theta}_i - \boldsymbol{\theta}^{\text{pre}})$, where α is a scaling hyperparameter (Ilharco et al., 2023) and $(\boldsymbol{\theta}_i - \boldsymbol{\theta}^{\text{pre}})$ denotes the task vector associated with seller i. Using this utility proxy, we approximate the Shapley value ϕ_i of each seller i as: $\hat{\phi}_i = \frac{1}{n}\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{\hat{u}(\mathcal{S}\cup\{i\}) - \hat{u}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}}$.

This formulation enables the buyer to estimate each dataset's contribution without retraining a model for all 2^{n-1} possible subsets. By approximating utility $u(\cdot)$ with $\hat{u}(\cdot)$, the computational burden is reduced from 2^n model trainings and inferences to 2^n model summations and evaluations, yielding substantial efficiency gains while preserving a close approximation to the ground-truth Shapley value, as established in the following theorem:

Theorem 3.1. *Assume that MTL is performed via full-batch gradient descent, and that the difference between the merged model $\boldsymbol{\theta}_{\mathcal{S}}^{merge}$ and the MTL-trained model $\boldsymbol{\theta}_{\mathcal{S}}^{MTL}$ is approximated up to second-order terms (i.e., higher-order components are negligible, with remainder dominated by $\mathcal{O}(\eta^3)$, where η is the learning rate for MTL). If the loss function $\mathcal{L}$ is L-Lipschitz continuous, then the approximation error between the true and proxy Shapley values satisfies:*

$$|\phi_i - \hat{\phi}_i| \leq (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC,$$

where $C = \binom{h+2}{2} H_{\max} G_{\max}$. Here, h denotes the number of optimization steps (i.e., epochs), and $G_{\max}$ and $H_{\max}$ are upper bounds on the gradient and Hessian norms, respectively.

This theorem establishes a formal link between MTL and task-arithmetic as the model merging method, in the context of Shapley-based valuation (the proof is provided in Appendix A.4). It shows that, under mild assumptions, the error between DMVM and ground-truth valuation, grows

quadratically in the number of sellers and linearly in α. The bound is worst-case and empirically, we observe substantially smaller errors than this worst-case scaling. Its purpose is to characterize how merging error propagates to valuation. Moreover, setting $\alpha = \frac{1}{n}$ makes the error grow linearly with n. However, this comes at the cost of restricting the performance of the merging algorithm, since α is typically treated as a tunable hyperparameter, e.g., selected via binary search. Furthermore, even the 2^n summations and evaluations can be replaced by an arbitrary finite number of Monte Carlo (MC) samples while retaining a controlled approximation error, as established in the following theorem:

Theorem 3.2. *Let $\bar{\phi}_i^{(m)}$ denote the MC estimator of $\hat{\phi}_i$ obtained from m independent samples of the marginal contribution terms in the Shapley expansion. Assume each sampled marginal contribution lies in an interval of width at most R. Under the assumptions of Theorem 3.1, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have: $|\phi_i - \bar{\phi}_i^{(m)}| \leq (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC + \sqrt{\frac{R^2}{2m}\log\left(\frac{2}{\delta}\right)}$.*

Theorem 3.2 shows that DMVM can admit a further sampling-based acceleration beyond the deterministic approximation in Theorem 3.1. In particular, the total error decomposes into a deterministic merging error and a statistical sampling error that decays as $\mathcal{O}(1/\sqrt{m})$. Proof for this theorem is provided in Appendix A.6. Finally, although the utility proxy $\hat{u}(\cdot)$ avoids raw data sharing, computing the merged model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$ still requires access to the individual fine-tuned models $\{\boldsymbol{\theta}_i\}_{i\in[n]}$. To prevent model leakage during this process, we introduce a privacy-aware protocol in the next section.

3.2 Privacy-Aware Merging via Secure Aggregation

While the utility proxy $\hat{u}(\cdot)$ enables efficient valuation, it still requires access to the individual fine-tuned models $\{\boldsymbol{\theta}_i\}_{i\in[n]}$ for computing the merged model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$. Directly sharing these models would violate seller privacy and undermine the decentralized design of the marketplace. To address this, we design a *two-fold privacy-aware protocol* that integrates secure MPC (Bonawitz et al., 2017) with an *unmergeable transformation* mechanism inspired by recent advances in model obfuscation (Junhao et al., 2025; Wang et al., 2025b; Chen et al., 2025).

Overview. Our protocol incorporates two components: (1) a *secure aggregation* mechanism that allows the buyer to obtain only the aggregated merged model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$ without revealing any individual model $\boldsymbol{\theta}_i$; and (2) an *unmergeable transformation* step that prevents an adversarial buyer from reconstructing any seller's model via algebraic manipulation of overlapping subsets of the merged models. We first describe the secure aggregation process, then extend it to incorporate the unmergeable mechanism.

Secure Aggregation. For brevity and simplicity, we set $\alpha = \frac{1}{|\mathcal{S}|}$; we have provided the algorithm for any arbitrary α value in Appendix A.1. For any subset $\mathcal{S} \subseteq [n]$ with $|\mathcal{S}| \geq 2$, each seller $i \in \mathcal{S}$ samples a random mask matrix $\boldsymbol{m}_i$ of the same dimension as their model parameters. Each seller shares $\boldsymbol{m}_i$ with all other sellers in $\mathcal{S}$ and computes a masked version of their contribution as $\boldsymbol{e}_i = \frac{\boldsymbol{\theta}_i}{|\mathcal{S}|} - \boldsymbol{m}_i$, which is sent to the buyer. Meanwhile, all sellers compute the aggregate mask $\boldsymbol{m} = \sum_{i\in\mathcal{S}} \boldsymbol{m}_i$ and send it to the buyer, who verifies the consistency across submissions. The buyer then reconstructs the merged model as $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}} = \boldsymbol{m} + \sum_{i\in\mathcal{S}} \boldsymbol{e}_i = \sum_{i\in\mathcal{S}} \boldsymbol{m}_i + \sum_{i\in\mathcal{S}} \left(\frac{\boldsymbol{\theta}_i}{|\mathcal{S}|} - \boldsymbol{m}_i\right) = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \boldsymbol{\theta}_i$. This ensures that the buyer receives only the aggregated model $\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}$, not individual $\boldsymbol{\theta}_i$'s. Similarly, since only random masks are shared among sellers, no seller gains access to others' models.

Unmergeable transformation. While the above protocol protects individual models, a malicious buyer could still recover a seller's model by subtracting two merged models from overlapping subsets (e.g., $\{\theta_1, \theta_2, \theta_3\}$ and $\{\theta_1, \theta_2\}$). To prevent this, we introduce a transformation that preserves functional behavior for inference but disrupts linear relationships among models. We instantiate this for two common modules: multi-layer perceptron (MLP) blocks and consecutive linear projections.

Consider two consecutive layers of an MLP: $\text{MLP}(x) = \boldsymbol{W}^{(2)}\sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$, where $\sigma(\cdot)$ is an element-wise nonlinearity. For weight matrices $\boldsymbol{W}_i^{(1)}$ and $\boldsymbol{W}_i^{(2)}$ held by seller i, the goal is for the buyer to recover $\sum_{i\in\mathcal{S}} \boldsymbol{W}_i^{(1)}$ and $\sum_{i\in\mathcal{S}} \boldsymbol{W}_i^{(2)}$ without accessing individual weights. To achieve this, all sellers in $\mathcal{S}$ jointly agree on a random permutation matrix $\boldsymbol{P}_{\mathcal{S}}$ (orthogonal with one "1" per

Table 1: Comparison of our methods and baselines (m is the number of subsets that are sampled).

Method	Operates without access to sellers' datasets	Computational time
Dataset Shapley (GT)	×	$2^n \times (T_{\text{train}} + T_{\text{inference}})$
MC-Dataset-Shapley (Ghorbani & Zou, 2019)	×	$n \times m \times (T_{\text{train}} + T_{\text{inference}})$
DU-Shapley (Garrido Lucero et al., 2024)	×	$n \times n \times (T_{\text{train}} + T_{\text{inference}})$
DMVM	✓	$2^n \times T_{\text{inference}}$
MC-DMVM	✓	$n \times m \times T_{\text{inference}}$

row/column and "0" elsewhere, satisfying $\boldsymbol{P}_{\mathcal{S}}^{-1} = \boldsymbol{P}_{\mathcal{S}}^{\top}$). For mask matrices $\boldsymbol{m}_i^{(1)}$ and $\boldsymbol{m}_i^{(2)}$, instead of sending $\frac{\boldsymbol{W}_i^{(1)}}{|\mathcal{S}|} - \boldsymbol{m}_i^{(1)}$ and $\frac{\boldsymbol{W}_i^{(2)}}{|\mathcal{S}|} - \boldsymbol{m}_i^{(2)}$, each seller i transmits $\frac{\boldsymbol{P}_{\mathcal{S}}\boldsymbol{W}_i^{(1)}}{|\mathcal{S}|} - \boldsymbol{m}_i^{(1)}$ and $\frac{\boldsymbol{W}_i^{(2)}\boldsymbol{P}_{\mathcal{S}}^{\top}}{|\mathcal{S}|} - \boldsymbol{m}_i^{(2)}$. Having received $\sum_{i\in\mathcal{S}} \boldsymbol{m}_i^{(1)}$ and $\sum_{i\in\mathcal{S}} \boldsymbol{m}_i^{(2)}$ from the sellers, the buyer can only reconstruct $\frac{\sum_{i\in\mathcal{S}} \boldsymbol{P}_{\mathcal{S}}\boldsymbol{W}_i^{(1)}}{|\mathcal{S}|}$ and $\frac{\sum_{i\in\mathcal{S}} \boldsymbol{W}_i^{(2)}\boldsymbol{P}_{\mathcal{S}}^{\top}}{|\mathcal{S}|}$ instead of $\frac{\sum_{i\in\mathcal{S}} \boldsymbol{W}_i^{(1)}}{|\mathcal{S}|}$ and $\frac{\sum_{i\in\mathcal{S}} \boldsymbol{W}_i^{(2)}}{|\mathcal{S}|}$. Since $\boldsymbol{P}_{\mathcal{S}}^{\top}\boldsymbol{P}_{\mathcal{S}} = I$ and $\sigma(\cdot)$ operates elementwise, this transformation leaves the model's inference behavior unchanged, but destroys the additive structure that would otherwise enable model reconstruction via subtraction. Thus, any negation or subtraction of merged models yields uninformative parameter combinations.

For consecutive linear layers such as $\boldsymbol{Q}\boldsymbol{K}^{\top}$ projection in transformer blocks (Vaswani et al., 2017) (for every query matrix $\boldsymbol{Q}$ and key matrix $\boldsymbol{K}$), a similar mechanism applies. Sellers agree on a random scalar $\rho \in \mathbb{R}\backslash\{0\}$, and instead of sending $\frac{\boldsymbol{Q}_i}{|\mathcal{S}|} - \boldsymbol{m}_i^{(1)}$ and $\frac{\boldsymbol{K}_i}{|\mathcal{S}|} - \boldsymbol{m}_i^{(2)}$, seller i sends $\frac{\rho\boldsymbol{Q}_i}{|\mathcal{S}|} - \boldsymbol{m}_i^{(1)}$ and $\frac{\boldsymbol{K}_i}{\rho|\mathcal{S}|} - \boldsymbol{m}_i^{(2)}$ to the buyer. This also preserves the functionality of $\boldsymbol{Q}\boldsymbol{K}^{\top}$ while making the individual components unmergeable, and therefore negations would result in useless and meaningless weights.

Secure Shapley Estimation. With these mechanisms in place, the buyer never accesses any raw data or individual model weights. We define the privacy-aware Shapley estimate of seller i as:

$$\tilde{\phi}_i = \frac{1}{n} \sum_{\mathcal{S}\subseteq[n]\backslash\{i\}, |\mathcal{S}|\geq 2} \frac{\hat{u}(\mathcal{S}\cup\{i\}) - \hat{u}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}}. \tag{3}$$

where $\hat{u}(\cdot)$ is computed via the above secure aggregation mechanism with unmergeable transformation. In the above approximation, we restrict the cardinality to $|\mathcal{S}| \geq 2$, thereby excluding the case $|\mathcal{S}| = 1$ to guarantee that no individual model is shared from a seller to the buyer, which is required for privacy. The following results bound the deviation between $\tilde{\phi}_i$ and the true Shapley value ϕ_i with the proof provided in Appendix A.5 and ablation study in Appendix A.10.

Theorem 3.3. *Let R denote the maximum marginal contribution $|u(\mathcal{S}\cup\{i\}) - u(\mathcal{S})|$ over all subsets in the Shapley expansion of ϕ_i. Then, $|\phi_i - \tilde{\phi}_i| \leq \frac{n-2}{n}(2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC + \frac{2R}{n}$, where L and C are defined as in Theorem 3.1.*

Remarks. Our privacy-aware protocol ensures that: (i) The buyer never accesses individual models or raw data from the sellers; (ii) sellers do not receive any information about other sellers' data or models; (iii) buyers cannot reconstruct individual models by differencing merged ones; and (iv) the mask consistency check additionally serves as a lightweight integrity check, flagging inconsistent aggregate masks caused by communication errors or a seller misreporting the aggregate. A high-level overview of DMVM framework is illustrated in Figure 1. The complete procedure for DMVM is presented in Algorithm 1 in Appendix A.1. In the next section, we evaluate the performance of this framework by simulating data marketplaces in a multi-task setting.

4 Numerical Experiments

This section investigates whether DMVM can serve as a reliable and efficient proxy for dataset valuation in practical multi-task settings. Since this is, to our knowledge, the first study on multi-task dataset valuation, there exists no established benchmark or baseline for comparison. Consequently, we design two types of realistic and controlled simulation environments that allow us to test whether our method can correctly recover the ground-truth ranking and values of sellers with respect to their ground-truth Shapley value. Here, we discuss the realistic scenario and the results for the controlled

setting are provided in Appendix A.11, where we construct a spectrum of sellers with varying degrees of relevance to the buyer's target tasks. We simulate a decentralized marketplace with n sellers, where each seller owns a dataset $\mathcal{D}_i$. The buyer is interested in a set of target tasks $\mathcal{T}$. For natural language processing (**NLP**) experiments, we consider three task families, each represented by three datasets: **Review classification** (SST-2 (Wang et al., 2018), IMDb (Maas et al., 2011), Yelp (Zhang et al., 2015)), **Paraphrase identification** (MRPC (Wang et al., 2018), QQP (Wang et al., 2018), PAWS (Zhang et al., 2019)), and **Question answering** (BoolQ (Clark et al., 2019), PubMedQA (Jin et al., 2019), StrategyQA (Geva et al., 2021)). Each dataset is treated as a seller. Thus, the NLP marketplace contains nine sellers in total. The buyer's target tasks can be any subset of these tasks, and the validation utility is computed by averaging performance across the corresponding task-level validation sets. This design captures a realistic setting in which multiple sellers provide data for the same task, but from different domains or distributions. For instance, SST-2, IMDb, and Yelp all correspond to review or sentiment classification, yet differ in text length, source domain, and linguistic style. We use `google/flan-t5-small` (Chung et al., 2022) as the shared pretrained model.

For **vision** experiments, we construct a marketplace using four task families: **Digit classification** (MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), USPS (Hull, 2002), EMNIST (Cohen et al., 2017)), **Animal classification** (CIFAR-10 cat/dog (Krizhevsky et al., 2009), Oxford-IIIT Pets cat/dog (Parkhi et al., 2012)), **Scene classification** (SUN397 (Xiao et al., 2010), MIT Indoor67 (Quattoni & Torralba, 2009)), and **Land-use classification** (EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017)). Again, each dataset corresponds to a seller. This yields ten vision sellers in total. The datasets within each task family share the same high-level prediction objective, but differ in visual domain, image statistics, and label distribution. For the shared pretrained model, we use `openai/clip-vit-base-patch32` (Radford et al., 2021). More implementation details are provided in Appendix A.14. The baselines are as follows:

Dataset Shapley as ground truth (GT). Dataset Shapley serves as the oracle reference. For each seller i, it computes the exact Shapley value using equation 2, where each coalition utility is obtained by training on the corresponding subset of sellers' datasets. Since this requires access to all sellers' raw datasets, we use it only for evaluation.

MC-Dataset-Shapley (Ghorbani & Zou, 2019). MC-Dataset-Shapley provides MC approximation of Dataset Shapley by sampling m coalitions or permutations instead of enumerating all possible coalitions.

DU-Shapley (Garrido Lucero et al., 2024). DU-Shapley is a dataset-level Shapley approximation based on discrete-uniform sampling. For seller i, it estimates the value as $\psi_i = \frac{1}{n}\sum_{k=0}^{n-1}\left[u\big(\mathcal{D}^{(k)} \cup \mathcal{D}_i\big) - u\big(\mathcal{D}^{(k)}\big)\right]$, where $\mathcal{D}^{(k)}$ is a dataset that its data samples are sampled uniformly without replacement from $\mathcal{D}_{-i} = \bigcup_{j\in[n]\setminus\{i\}} \mathcal{D}_j$, with size $k\mu_{-i}$, and $\mu_{-i} = \frac{1}{n-1}|\mathcal{D}_{-i}|$. ($|\mathcal{D}_{-i}|$ is the size of data samples in $\mathcal{D}_{-i}$). In our multi-task adaptation, $u(\cdot)$ is the average validation performance defined in equation 1.

Our methods. We report both DMVM and MC-DMVM as our proposed methods. In MC-DMVM, the MC sampling variant of DMVM introduced in Theorem 3.2, seller values are estimated from m sampled marginal contributions instead of all coalitions. Table 1 illustrates the computational time and privacy aspect of each method.

We evaluate each method from two complementary perspectives. First, we use rank-correlation metrics used in the literature (Wang et al., 2025a; Yang et al., 2024; Tian et al., 2022; Lu et al., 2024) to measure whether a method correctly identifies the relative ordering of sellers. Second, we use mean squared error as an end-to-end value-estimation metric to measure whether the estimated numerical values are close to the ground-truth Dataset Shapley values. **Kendall's** τ measures the fraction of *concordant* versus *discordant* seller pairs. For two sellers i and j with predicted values (v_i, v_j) and ground-truth values (g_i, g_j), the pair is concordant if $(v_i - v_j)(g_i - g_j) > 0$. Let n_c and n_d denote the number of concordant and discordant pairs among all $\binom{n}{2}$ pairs. Then, τ is defined as $\tau = \frac{n_c - n_d}{\binom{n}{2}}$. The score ranges from -1 (perfect disagreement) to 1 (perfect agreement). **Spearman's** ρ computes the Pearson correlation between the predicted and ground-truth rank vectors. Let r_i and s_i denote the predicted and ground-truth ranks of seller i, respectively. Let $\bar{r}$ and $\bar{s}$ denote the corresponding average of ranks across all sellers. Then, ρ is defined as $\rho = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$. When no

Table 2: Performance of valuation methods on vision and NLP tasks. Buyer's target task size $|\mathcal{T}| = 3$ for NLP and $|\mathcal{T}| = 4$ for vision.

n	Method	Vision Tasks			NLP Tasks		
		τ	ρ	MSE	τ	ρ	MSE
7	DU-Shapley	0.8305 ± 0.1247	0.9174 ± 0.0752	0.000770 ± 0.000490	0.7235 ± 0.2117	0.8207 ± 0.1953	0.003324 ± 0.002471
	DMVM	0.8079 ± 0.1356	0.8988 ± 0.0895	0.000756 ± 0.000368	0.7784 ± 0.1912	0.8675 ± 0.1589	0.001967 ± 0.001442
8	DU-Shapley	0.8384 ± 0.1079	0.9250 ± 0.0667	0.000612 ± 0.000391	0.7132 ± 0.1869	0.8213 ± 0.1629	0.002661 ± 0.001996
	DMVM	0.7955 ± 0.1193	0.8963 ± 0.0772	0.000761 ± 0.000344	0.7578 ± 0.1739	0.8613 ± 0.1351	0.001815 ± 0.001294
9	DU-Shapley	0.8410 ± 0.0993	0.9288 ± 0.0607	0.000476 ± 0.000315	0.7030 ± 0.1726	0.8179 ± 0.1475	0.002122 ± 0.001628
	DMVM	0.7795 ± 0.1098	0.8893 ± 0.0727	0.000747 ± 0.000322	0.7487 ± 0.1539	0.8617 ± 0.1200	0.001668 ± 0.001150

Table 3: Comparison of valuation methods for different numbers of buyer's target tasks. Number of sellers $n = 9$ for NLP and $n = 10$ for vision.

$\lvert\mathcal{T}\rvert$	Method	Vision Tasks			NLP Tasks		
		τ	ρ	MSE	τ	ρ	MSE
2	DU-Shapley	0.8094 ± 0.1028	0.9100 ± 0.0722	0.000651 ± 0.000571	0.7469 ± 0.1464	0.8657 ± 0.1137	0.002924 ± 0.002667
	DMVM	0.7738 ± 0.1827	0.8677 ± 0.1729	0.001442 ± 0.000962	0.8441 ± 0.1141	0.9315 ± 0.0616	0.002240 ± 0.001838
3	DU-Shapley	0.8367 ± 0.0876	0.9270 ± 0.0529	0.000472 ± 0.000363	0.7030 ± 0.1726	0.8179 ± 0.1475	0.002122 ± 0.001628
	DMVM	0.7837 ± 0.1404	0.8788 ± 0.1193	0.000969 ± 0.000487	0.7487 ± 0.1539	0.8617 ± 0.1200	0.001668 ± 0.001150
4	DU-Shapley	0.8399 ± 0.0938	0.9287 ± 0.0567	0.000383 ± 0.000259	0.7385 ± 0.1504	0.8530 ± 0.1196	0.001721 ± 0.001088
	DMVM	0.7697 ± 0.1086	0.8842 ± 0.0750	0.000732 ± 0.000296	0.6856 ± 0.2082	0.8004 ± 0.1910	0.001383 ± 0.000815
5	DU-Shapley	0.8397 ± 0.0929	0.9312 ± 0.0554	0.000329 ± 0.000195	0.7650 ± 0.1211	0.8792 ± 0.0796	0.001480 ± 0.000755
	DMVM	0.7556 ± 0.0940	0.8822 ± 0.0579	0.000590 ± 0.000198	0.6980 ± 0.1656	0.8267 ± 0.1496	0.001211 ± 0.000607

rank ties exist, this simplifies to $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, where $d_i = r_i - s_i$. $\rho = 1$ indicates perfect alignment, $\rho = -1$ complete reversal, and $\rho = 0$ no correlation. While Kendall's τ emphasizes pairwise ordering consistency, Spearman's ρ captures overall rank alignment; using both provides a comprehensive evaluation of ranking quality.

End-to-end value-estimation error. While rank correlations evaluate whether a method orders sellers correctly, they do not measure whether the estimated seller values are numerically accurate. Therefore, we also report the mean squared error (MSE) between the estimated seller values and the ground-truth Dataset Shapley values as $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\phi_i - \hat{\phi}_i)^2$, where ϕ_i is the ground-truth Dataset Shapley value of seller i, and $\hat{\phi}_i$ is the value estimated by the method being evaluated. We refer to this as an end-to-end metric because it evaluates the final output of the valuation pipeline: the numerical value assigned to each seller.

4.1 Results and Discussion

We evaluate whether DMVM can approximate ground-truth multi-task Dataset Shapley values while avoiding the two main limitations of existing valuation baselines: repeated retraining and access to sellers' private datasets. Unless otherwise specified, DMVM denotes the deterministic estimator that evaluates all coalitions using the merging-based utility proxy. We emphasize that all non-DMVM baselines are privileged baselines: Dataset Shapley, MC-Dataset-Shapley, and DU-Shapley require access to sellers' datasets in order to train or evaluate models on seller coalitions. Therefore, they should be interpreted as practical upper-bound references rather than deployable baselines in decentralized data marketplaces. In contrast, DMVM estimates seller values only through merged models and does not require access to raw seller datasets.

Table 2 evaluates DMVM as the number of sellers varies. Across both modalities, DMVM achieves strong agreement with the ground-truth Dataset Shapley values. In vision, DMVM obtains Kendall's τ values around 0.78–0.81 and Spearman's ρ values around 0.89, with MSE on the order of 10^{-4}. Although DU-Shapley gives slightly higher rank correlations in vision, DMVM remains close while operating under a substantially more realistic constraint. In NLP, DMVM consistently outperforms DU-Shapley in both ranking and MSE across all seller counts. For example, when $n = 9$, DMVM improves Kendall's τ from 0.7030 to 0.7487, Spearman's ρ from 0.8179 to 0.8617, and reduces MSE from 0.002122 to 0.001668. Table 3 studies the effect of the number of buyer target tasks. DMVM remains stable across different values of $|\mathcal{T}|$ for both Vision and NLP experiments, suggesting that model merging provides a reliable proxy for multi-task valuation. In NLP, DMVM achieves lower MSE than DU-Shapley for all target-task sizes, indicating that it provides accurate

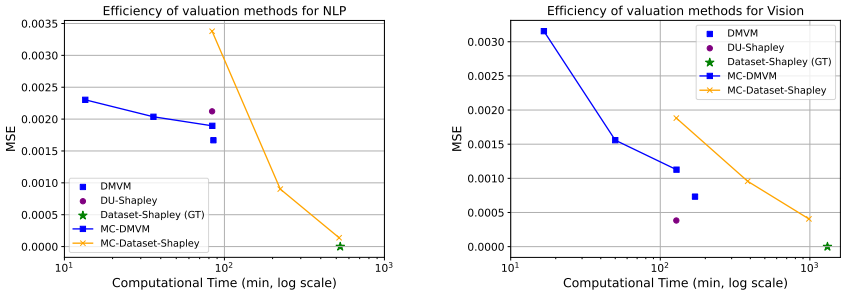

Figure 2: Efficiency-accuracy trade-off of valuation methods on NLP and vision tasks. The x-axis reports wall-clock computational time in minutes, and the y-axis reports MSE relative to ground-truth.

end-to-end numerical value estimates even when the ranking metrics are sometimes slightly lower. All reported results are averaged over all possible choices of n sellers and target task set sizes.

Figure 2 shows the efficiency-accuracy trade-off. Dataset Shapley gives the exact reference values but is computationally expensive because it requires training over all coalitions. MC-Dataset-Shapley reduces the number of coalitions but still requires retraining on sellers' datasets. In contrast, DMVM replaces coalition training with model merging and inference, making valuation substantially cheaper and compatible with decentralized settings. For the MC-based methods, the three plotted points correspond to $m \in \{9, 24, 56\}$ for NLP, and $m \in \{10, 30, 77\}$ for vision. Notably, the largest-sample MC-DMVM point matches the wall-clock time of DU-Shapley, while achieving lower MSE in NLP and competitive MSE in vision. This demonstrates that MC-DMVM provides a practical accuracy-efficiency trade-off: increasing m reduces valuation error while keeping the method computationally lighter than retraining-based Shapley estimators. More ablation studies on the convergence of MC-based methods as m increases are provided in Appendix A.3. Overall, these results show that DMVM provides accurate seller rankings and low end-to-end valuation error while avoiding the key limitations of existing baselines. Unlike Dataset Shapley, MC-Dataset-Shapley, and DU-Shapley, DMVM does not require access to sellers' datasets, making it both computationally efficient and privacy-aware for realistic multi-task data marketplaces.

Additional experiments with more model architectures and datasets for a systematic and controlled simulation, where the sellers might have datasets including more than one task in both overlapping and non-overlapping settings are provided in Appendix A.11. Scalability analysis is provided in Appendix A.13. Moreover, we evaluate whether DMVM can identify harmful sellers with increasing label noise in Appendix A.12. Additional analyses include comparisons with Leave-One-Out (LOO), Standalone, and GTG-Shapley (Appendix A.7), direct validation of the merging-based utility proxy (Appendix A.9), and robustness studies on the buyer's validation-set size, seller dataset-size imbalance, and buyer task preferences (Appendix A.8).

5 Conclusion

This work reframes multi-task dataset valuation as a model-composition problem rather than a repeated retraining problem. In decentralized data marketplaces, the challenge is not only estimating which sellers are useful, but doing so before purchase, without assuming access to their private datasets. DMVM addresses this setting by using model merging as a proxy for coalition utility, enabling Shapley-style valuation through merged models instead of training on every seller subset. This preserves the economic interpretation of marginal contribution while making valuation more practical in multi-task and privacy-sensitive environments. Our theoretical analysis characterizes the approximation gap between merging-based and ground-truth multi-task valuation, while our secure aggregation protocol prevents direct exposure of raw data or individual seller models. Empirically, DMVM and its MC variant achieve strong agreement with privileged retraining-based baselines across vision and NLP tasks, often attaining low value-estimation error with only inference-time computation. These results suggest that model merging can serve not only as a tool for multi-task adaptation, but also as an efficient mechanism for decentralized market valuation.

AI USE STATEMENT

In this work, we used generative AI tools for improving and polishing the writing of the paper. We take responsibility for the final content of this work.

REFERENCES

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pp. 701–726, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929. doi: 10.1145/3328526.3329589. URL https://doi.org/10.1145/3328526.3329589.

Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, Sara Hooker, et al. Mix data or merge models? optimizing for diverse multi-task learning. *arXiv preprint arXiv:2410.10801*, 2024.

Mohammadsajad Alipour and Mohammad Mohammadi Amiri. Towards reversible model merging for low-rank weights. *arXiv preprint arXiv:2510.14163*, 2025.

Mohammad Mohammadi Amiri, Frédéric Berdoz, and Ramesh Raskar. Fundamentals of task-agnostic data valuation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26106. URL https://doi.org/10.1609/aaai.v37i8.26106.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Huaiguang Cai. Chg shapley: Efficient data valuation and selection towards trustworthy machine learning. *arXiv preprint arXiv:2406.11730*, 2024.

Lingjiao Chen, Paraschos Koutris, and Arun Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, pp. 1535–1552, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450356435. doi: 10.1145/3299869.3300078. URL https://doi.org/10.1145/3299869.3300078.

Shilian Chen, Jie Zhou, Qin Chen, Wen Wu, Xin Li, Qi Feng, and Liang He. Can heterogeneous language models be fused? *arXiv preprint arXiv:2604.01674*, 2026.

Wei-Jia Chen, Min-Yen Tsai, Cheng-Yi Lee, and Chia-Mu Yu. Defending unauthorized model merging via dual-stage weight protection. *arXiv preprint arXiv:2511.11851*, 2025.

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.

Saurab Chhachhi and Fei Teng. A joint energy and differentially-private smart meter data market. *arXiv preprint arXiv:2412.07688*, 2024.

Hongliang Chi, Wei Jin, Charu C. Aggarwal, and Yao Ma. Precedence-constrained winter value for effective graph data valuation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tVRVE0OAyb.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Ali Falahati and Mohammad Mohammadi Amiri. Disentangled structural and featural representation for task-agnostic graph valuation. *arXiv preprint arXiv:2408.12659*, 2024.

Felipe Garrido Lucero, Benjamin Heymann, Maxime Vono, Patrick Loiseau, and Vianney Perchet. Du-shapley: A shapley value proxy for efficient dataset valuation. *Advances in Neural Information Processing Systems*, 37:1973–2000, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 2002.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Tayyebeh Jahani-Nezhad, Parsa Moradi, Mohammad Ali Maddah-Ali, and Giuseppe Caire. Private, augmentation-robust and task-agnostic data valuation approach for data marketplace. *arXiv preprint arXiv:2411.00745*, 2024.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, 12(11):1610–1623, July 2019a. ISSN 2150-8097. doi: 10.14778/3342263.3342637. URL https://doi.org/10.14778/3342263.3342637.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1167–1176. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/jia19a.html.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Wei Junhao, Yu Zhe, and Jun Sakuma. Disrupting model merging: A parameter-level defense without sacrificing accuracy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17698–17707, October 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

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

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.

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vRvVVb0NAz.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. Dealer: an end-to-end model marketplace with differential privacy. *Proc. VLDB Endow.*, 14(6):957–969, February 2021. ISSN 2150-8097. doi: 10.14778/3447689.3447700. URL https://doi.org/10.14778/3447689.3447700.

Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Trans. Intell. Syst. Technol.*, 13(4), May 2022. ISSN 2157-6904. doi: 10.1145/3501811. URL https://doi.org/10.1145/3501811.

Charles Lu, Mohammad Mohammadi Amiri, and Ramesh Raskar. Data measurements for decentralized data markets. *arXiv preprint arXiv:2406.04257*, 2024.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Clara Na, Ian Magnusson, Ananya Harsh Jha, Tom Sherborne, Emma Strubell, Jesse Dodge, and Pradeep Dasigi. Scalable data ablation approximations for language models through modular training and merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21125–21141, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Xu Ouyang, Felix Xiaozhu Lin, and Yangfeng Ji. Selectformer: Private and practical data selection for transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2cF3f9t31y.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 413–420. IEEE, 2009.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Lloyd S Shapley et al. A value for n-person games. 1953.

Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data valuation in machine learning: "ingredients", strategies, and open challenges. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5607–5614. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/782. URL https://doi.org/10.24963/ijcai.2022/782. Survey Track.

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, 2011. doi: 10.1109/IJCNN.2011.6033395.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Fusionbench: A comprehensive benchmark of deep model fusion. *Journal of Machine Learning Research*, 2025a.

Xin Tang, Yanzhe Lei, and Chao Li. Heterogeneous graph data valuation: A shapley value-based approach. In *2025 40th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 2871–2876, 2025b. doi: 10.1109/YAC66630.2025.11150137.

Zhixu Silvia Tao, Kasper Vinken, Hao-Wei Yeh, Avi Cooper, and Xavier Boix. Merge to mix: Mixing datasets via model merging. *arXiv preprint arXiv:2505.16066*, 2025.

Zhihua Tian, Jian Liu, Jingyu Li, Xinle Cao, Ruoxi Jia, Jun Kong, Mengdi Liu, and Kui Ren. Private data valuation and fair payment in data marketplaces. *arXiv preprint arXiv:2210.08723*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=HD6bWcj87Y.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In *Federated Learning: Privacy and Incentive*, pp. 153–167. Springer, 2020.

Zihao Wang, Enneng Yang, Lu Yin, Shiwei Liu, and Li Shen. Model unmerging: Making your models unmergeable for secure model sharing. *arXiv preprint arXiv:2509.01548*, 2025b.

Zijing Wang, Yongkang Liu, Yingfeng Luo, Ming Wang, Zhen Song, Shi Feng, Xiaocui Yang, Dingyang Lin, Daling Wang, Yifei Zhang, et al. Scaling intelligence through model merging: A comprehensive survey. *Authorea Preprints*, 2025c.

Shuyue Wei, Yongxin Tong, Zimu Zhou, Tianran He, and Yi Xu. Efficient data valuation approximation in federated learning: A sampling-based approach. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 2922–2934, 2025. doi: 10.1109/ICDE65448.2025.00219.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

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, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Zhengqi Xu, Han Zheng, Jie Song, Li Sun, and Mingli Song. Training-free heterogeneous model merging. *arXiv preprint arXiv:2501.00061*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Ziao Yang, Han Yue, Jian Chen, and Hongfu Liu. On the inflation of knn-shapley value. *arXiv preprint arXiv:2405.17489*, 2024.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308, 2019.

Yuchen Zhang and Mohammad Mohammadi Amiri. Toward efficient influence function: Dropout as a compression tool. *arXiv preprint arXiv:2509.15651*, 2025.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. *arXiv preprint arXiv:2508.16082*, 2025.

Mingzhi Zhu, Boris Sobolev, Rahul Krishna, Raju Pavuluri, Stacy Patterson, and Michele Merler. Multi-task code llms: Data mix or model merge? *arXiv preprint arXiv:2601.21115*, 2026.

A APPENDIX

A.1 ALGORITHM

Algorithm 1 DMVM: Decentralized Multi-task Valuation via Secure Model Merging

Input: Buyer queries a potential group of sellers $[n]$, base model weights $\boldsymbol{W}_{\text{pre}}$, scaling hyper-parameter α
for each subset $\mathcal{S} \subseteq [n]$ with $|\mathcal{S}| \geq 2$ **do**
  for each layer pair $(\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)})$ **do**
    if $(\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)})$ correspond to a 2-layer MLP **then**
      Sellers agree on a random permutation matrix $\boldsymbol{P}$.
      Define the layer transformation: $(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) = (\boldsymbol{P}, \boldsymbol{P}^\top)$.
    else if $(\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)})$ are two consecutive linear layers **then**
      Sellers agree on a random scalar ρ.
      Define the layer transformation: $(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) = (\rho \boldsymbol{I}, \rho^{-1} \boldsymbol{I})$.
    end if
    for each seller $i \in \mathcal{S}$ **do**
      if $\alpha = \frac{1}{|\mathcal{S}|}$ **then**
        Generate masked shares: $\boldsymbol{e}_1^{(i)} = \frac{\boldsymbol{\Psi}_1 \boldsymbol{W}_i^{(1)}}{|\mathcal{S}|} - \boldsymbol{m}_i^{(1)},\ \boldsymbol{e}_2^{(i)} = \frac{\boldsymbol{W}_i^{(2)} \boldsymbol{\Psi}_2}{|\mathcal{S}|} - \boldsymbol{m}_i^{(2)}$
      else
        Compute local secrets:

$$\boldsymbol{s}_i^{(1)} = \frac{\boldsymbol{W}_{\text{pre}}^{(1)}(1 - |\mathcal{S}|\alpha)}{|\mathcal{S}|} + \alpha \boldsymbol{W}_i^{(1)},\ \boldsymbol{s}_i^{(2)} = \frac{\boldsymbol{W}_{\text{pre}}^{(2)}(1 - |\mathcal{S}|\alpha)}{|\mathcal{S}|} + \alpha \boldsymbol{W}_i^{(2)}$$

        Generate masked shares: $\boldsymbol{e}_1^{(i)} = \boldsymbol{\Psi}_1 \boldsymbol{s}_i^{(1)} - \boldsymbol{m}_i^{(1)},\ \boldsymbol{e}_2^{(i)} = \boldsymbol{s}_i^{(2)} \boldsymbol{\Psi}_2 - \boldsymbol{m}_i^{(2)}$
      end if
      Send $(\boldsymbol{e}_1^{(i)}, \boldsymbol{e}_2^{(i)})$ to the buyer and $(\boldsymbol{m}_i^{(1)}, \boldsymbol{m}_i^{(2)})$ to all other sellers.
    end for
    Sellers compute aggregate masks: $\boldsymbol{m}^{(1)} = \sum_{i \in \mathcal{S}} \boldsymbol{m}_i^{(1)}, \quad \boldsymbol{m}^{(2)} = \sum_{i \in \mathcal{S}} \boldsymbol{m}_i^{(2)}$
    Each seller sends $(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)})$ to the buyer for verification.
    Buyer performs sanity check on consistency of $(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)})$.
    Buyer reconstructs the merged model weights:

$$\boldsymbol{W}_{\text{merge}}^{(1)} = \boldsymbol{m}^{(1)} + \sum_{i \in \mathcal{S}} \boldsymbol{e}_1^{(i)},\ \boldsymbol{W}_{\text{merge}}^{(2)} = \boldsymbol{m}^{(2)} + \sum_{i \in \mathcal{S}} \boldsymbol{e}_2^{(i)}$$

  end for
end for
Output: Privacy-aware merged models for all the subsets ready to use for valuation using equation 3 for each of the sellers.

A.2 RELATED WORK

The proliferation of data-driven machine learning has created a need to quantify the economic value of data, both to fairly compensate contributors and to efficiently acquire new, high-quality data. This research area can be broadly categorized into methods for valuing individual data points, techniques for valuing entire datasets, and the design and operation of data marketplaces.

Data Valuation: Valuing Individual Points. A dominant approach for data valuation is the Shapley value (Shapley et al., 1953), a concept from cooperative game theory that fairly assigns value to each data point based on its average marginal contribution to a machine learning model's performance across all possible data subsets (Ghorbani & Zou, 2019). This method is axiomatically fair, satisfying properties such as symmetry and additivity. The primary drawback of the standard Data Shapley is its exponential computational complexity, $O(2^N)$, which is prohibitive for modern-scale datasets of size N. To overcome this, several approximations have been developed (Jia et al., 2019b). Sampling-based methods estimate the Shapley value by sampling random permutations of

the data, but they may still require numerous model retrainings (Ghorbani & Zou, 2019; Maleki et al., 2013). Jia et al. (2019a) proposed an exact and efficient algorithm for computing Shapley values specifically for K-nearest neighbour models, reducing the complexity to $O(N \log N)$. Recent works further extend Shapley-based valuation to graph-structured data (Chi et al., 2025; Tang et al., 2025b).

Dataset Valuation. While point-level valuation supports curation and debugging, dataset valuation addresses how much an entire contributor's dataset adds to utility within a cooperative game among data owners. This line of work focuses on valuing entire datasets as a single unit, which is distinct from aggregating individual data point values. The Shapley value can also be applied at the dataset level, measuring the marginal contribution of one seller's entire dataset to a coalition of other sellers. To address the computational cost, DU-Shapley (Garrido Lucero et al., 2024) was proposed as an efficient proxy specifically for dataset valuation, using a discrete uniform approximation that is more efficient than generic Monte Carlo sampling. Another line of work is task-agnostic dataset valuation methods that assess data without relying on a specific model or validation set, avoiding issues such as unavailable validation data. Instead, they estimate the value based on the dataset's intrinsic distributional properties. The PriArTa framework evaluates sellers by computing the Wasserstein distance between augmentation-invariant latent distributions of the buyer's and seller's datasets (Jahani-Nezhad et al., 2024). Another related study introduces a fundamental valuation framework that characterizes datasets through two complementary metrics: relevance, capturing statistical similarity, and diversity, capturing statistical difference (Amiri et al., 2023). These are estimated by comparing the dataset's second-moment properties, which can be done privately by sharing only principal components. This methodology was later extended to graph data (Falahati & Amiri, 2024).

Recent work has also studied participant valuation in federated learning (FL), where each client's contribution is estimated using information from the federated training process (Wang et al., 2020; Liu et al., 2022; Wei et al., 2025). For example, GTG-Shapley (Liu et al., 2022) efficiently approximates client Shapley values using model updates collected across federated training rounds. These methods typically assume a coordinating server and value participants with respect to a shared global training objective. In contrast, DMVM considers a data-marketplace setting in which sellers retain their private datasets, fine-tune a shared pretrained model locally, and are valued according to a buyer-defined multi-task objective through model merging, without relying on a federated training trajectory.

Data Marketplace. A growing line of research explores the design and economic principles of data marketplaces. Building such markets is challenging because data possess unique characteristics: they can be replicated at near-zero cost, their value arises from complex interactions among datasets, and their usefulness for a given task is difficult to evaluate beforehand (Liu et al., 2021; Tian et al., 2022; Chhachhi & Teng, 2024; Chen et al., 2019; Agarwal et al., 2019; Lu et al., 2024). To address this, some frameworks propose a model-based pricing approach, where the marketplace sells machine learning model instances rather than the raw data itself (Chen et al., 2019). The Dealer marketplace framework (Liu et al., 2021) formalizes an end-to-end system where data owners' compensation is a function of both their Shapley value and their privacy sensitivity. These mechanisms, however, often rely on a trusted broker to access all datasets and perform the valuation. This introduces a fundamental dilemma: data owners must share their data to have it valued, but a buyer no longer needs to purchase data they have already observed. Privatizing these computations via multi-party computation (MPC)-style protocols closes the loop for deployable and trustable data markets (Tian et al., 2022; Ouyang et al., 2025).

Model Merging. Model merging builds on the observation that neural network parameters often exhibit approximately linear structure across tasks. Ilharco et al. (2023) first introduced task arithmetic, showing that subtracting a pre-trained model from its fine-tuned counterpart yields a task vector whose addition or subtraction can edit model behavior, enabling multi-task composition or unlearning without retraining. Li et al. (2025) provided the first theoretical justification for task arithmetic, proving that the effectiveness of adding or subtracting task vectors depends on how related the underlying tasks are and showing that properly scaled combinations can generalize to unseen tasks. In an empirical study, Tao et al. (2025) proposed Merge to Mix, which leverages merging as a surrogate for fine-tuning on dataset mixtures. By correlating the performance of merged models with that of mixture-trained ones, they enabled efficient and data-free selection of optimal dataset combinations. Zhou et al. (2025) reframed task-arithmetic as an efficient form of approximate multi-task learning. They showed that merging task vectors is equivalent to performing one step of gradi-

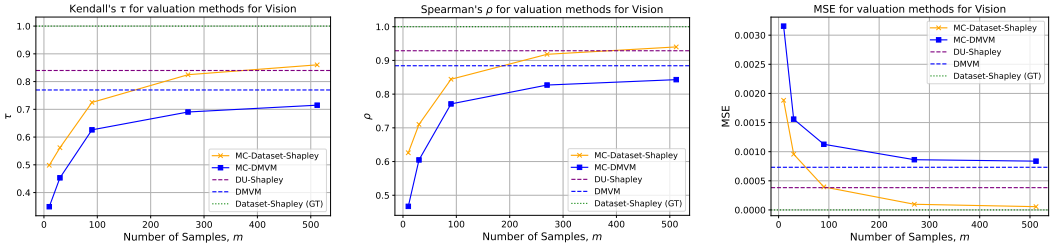

Figure 3: Convergence behavior of MC-based valuation methods on vision tasks with $n = 10$ sellers. The plots report Kendall's τ, Spearman's ρ, and MSE as the number of MC samples m increases. MC-Dataset-Shapley converges toward Dataset Shapley, while MC-DMVM converges toward deterministic DMVM.

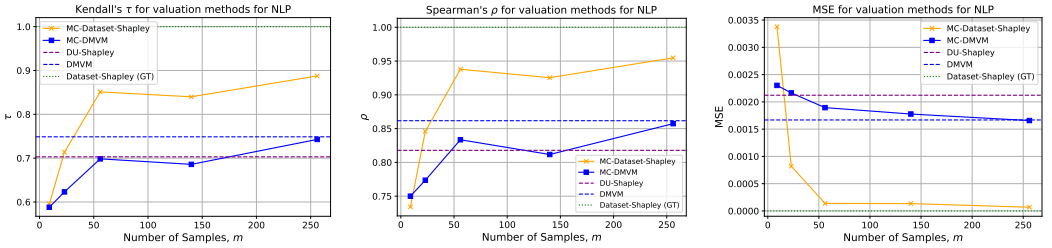

Figure 4: Convergence behavior of MC-based valuation methods on NLP tasks with $n = 9$ sellers. The plots report Kendall's τ, Spearman's ρ, and MSE as the number of MC samples m increases. MC-Dataset-Shapley approaches Dataset Shapley with larger m, while MC-DMVM approaches deterministic DMVM using only merging-based utility estimates.

ent descent on a combined multi-task objective, and derived a second-order bound quantifying the difference between the merged model and the model obtained from full multi-task training. This analysis formally explains why task arithmetic can closely approximate joint optimization. Despite these advances, existing model merging studies have not been explored in the context of dataset valuation, nor leveraged to quantify marginal data contributions in decentralized multi-task settings.

A.3 Convergence Analysis of MC-based methods

We further study the convergence behavior of the MC variants as the number of sampled coalitions m increases. We focus on the setting with $n = 10$ sellers for vision and $n = 9$ sellers for NLP. For vision, we evaluate $m \in \{10, 30, 90, 270, 512\}$, and for NLP, we evaluate $m \in \{9, 23, 56, 140, 256\}$. The goal is to examine whether the MC estimators approach their corresponding deterministic counterparts as m increases, while using only a fraction of the full coalition space.

Figures 3 and 4 show that both MC-Dataset-Shapley and MC-DMVM improve consistently as m increases. MC-Dataset-Shapley converges toward Dataset Shapley: its MSE decreases substantially, and both Kendall's τ and Spearman's ρ improve with more samples. This is expected, since increasing m gives a better MC approximation to the exact Shapley value. However, MC-Dataset-Shapley still requires access to sellers' datasets and repeated coalition training.

MC-DMVM converges toward deterministic DMVM, since it samples marginal contributions using the merging-based utility proxy rather than the retraining-based utility. In both vision and NLP, increasing m reduces the gap between MC-DMVM and DMVM, especially in terms of MSE. The ranking metrics show small non-monotonic fluctuations, which are expected for sampling-based estimators, but they generally stabilize near the deterministic DMVM values as m becomes larger.

These results support the error decomposition in Theorem 3.2: increasing m reduces the statistical sampling error, while the remaining gap is governed by the deterministic merging approximation. Importantly, MC-DMVM provides two levels of computational savings. First, it replaces coalition training with model merging and inference. Second, it reduces the number of required coalition evaluations from exponential in the number of sellers, i.e., 2^{n-1}, to only m sampled evaluations.

Therefore, MC-DMVM provides a practical accuracy-efficiency trade-off for decentralized valuation, achieving stable seller rankings and low value-estimation error without raw-data access or repeated retraining.

A.4 Error bound of our method

Lemma A.1 (Convex Combination Upper Bound)**.** *Let* $a_1, a_2, \ldots, a_n \in \mathbb{R}$ *and let* $w_1, w_2, \ldots, w_n \geq 0$ *satisfy* $\sum_{i=1}^{n} w_i = b$*. Then the weighted sum is bounded by the largest element, i.e.,*

$$\sum_{i=1}^{n} w_i a_i \;\leq\; b \max_{1\leq i\leq n} a_i \;\leq\; b \sup_{1\leq i\leq n} a_i.$$

Zhou et al. (2025) proposed the following error bound between multi-task learning and task-arithmetic merging:

$$\|\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}} - \boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}}\|_2 \leq s(\alpha s + 1)C,$$

where s is the cardinality of $\mathcal{S}$, i.e., $s = |\mathcal{S}|$, and $C = \binom{h+2}{2} H_{\max} G_{\max}$ is a constant, where h denotes the number of optimization steps (i.e., epochs), and $G_{\max}$ and $H_{\max}$ are upper bounds on the gradient and Hessian norms, respectively and α is the hyperparameter for task-arithmetic. Assuming the loss function is L-Lipschitz continuous, we have:

$$|\mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}) - \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}})| \leq L\|\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}} - \boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}}\|_2 \leq Ls(\alpha s + 1)C. \tag{4}$$

Assuming that utility $u(\cdot)$ has a negative relation with loss function, we set $u = -\mathcal{L}$, then the error between our method and ground-truth Shapley will be:

$$\begin{aligned}
|\phi_i - \hat{\phi}_i| &= \frac{1}{n}\left|\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{u(\mathcal{S}\cup\{i\}) - u(\mathcal{S}) - \hat{u}(\mathcal{S}\cup\{i\}) + \hat{u}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}}\right| \\
&\leq \frac{1}{n}\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{|u(\mathcal{S}\cup\{i\}) - u(\mathcal{S}) - \hat{u}(\mathcal{S}\cup\{i\}) + \hat{u}(\mathcal{S})|}{\binom{n-1}{|\mathcal{S}|}} \\
&\leq \frac{1}{n}\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{\left|-\mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}\cup\{i\}}^{\text{MTL}}) + \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}}) + \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}\cup\{i\}}^{\text{merge}}) - \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}})\right|}{\binom{n-1}{|\mathcal{S}|}} \\
&\leq \frac{1}{n}\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{\left|\mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}\cup\{i\}}^{\text{merge}}) - \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}\cup\{i\}}^{\text{MTL}})\right| + \left|\mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{merge}}) - \mathcal{L}(\boldsymbol{\theta}_{\mathcal{S}}^{\text{MTL}})\right|}{\binom{n-1}{|\mathcal{S}|}} \\
&\overset{(a)}{\leq} Ln(\alpha n + 1)C + L(n-1)(\alpha(n-1)+1)C = (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)LC.
\end{aligned}$$

Inequality (a) follows from Lemma A.1. Since $\mathcal{S} \subseteq [n]\setminus\{i\}$, we have $|\mathcal{S}\cup\{i\}| \leq n$ and $|\mathcal{S}| \leq n-1$. Hence, the two loss-difference terms are uniformly bounded by $Ln(\alpha n+1)C$ and $L(n-1)(\alpha(n-1)+1)C$, respectively, using inequality equation 4. Moreover, the Shapley coefficients form a convex combination, i.e., $\frac{1}{n}\sum_{\mathcal{S}\subseteq[n]\setminus\{i\}} \frac{1}{\binom{n-1}{|\mathcal{S}|}} = 1$. Therefore, applying Lemma A.1 with $b = 1$ yields inequality (a).

A.5 Error bound of our privacy-aware method

Assuming that we want to exclude subsets of size less than k. Let R be the maximum marginal utility, i.e., $\forall \mathcal{S} \subseteq [n] \setminus \{i\}, \quad |u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})| \leq R$. Then, the error will be:

$$
\begin{aligned}
|\phi_i - \tilde{\phi}_i| &= \frac{1}{n} \Bigg| \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| \geq k} \frac{u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) - \hat{u}(\mathcal{S} \cup \{i\}) + \hat{u}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}} \\
&\quad + \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| < k} \frac{u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}} \Bigg| \\
&\leq \frac{1}{n} \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| \geq k} \left| \frac{u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) - \hat{u}(\mathcal{S} \cup \{i\}) + \hat{u}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}} \right| \\
&\quad + \frac{1}{n} \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| < k} \left| \frac{u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}} \right| \\
&\leq \frac{1}{n} \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| \geq k} \frac{\left| \mathcal{L}(\boldsymbol{\theta}^{\text{merge}}_{\mathcal{S} \cup \{i\}}) - \mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S} \cup \{i\}}) \right| + \left| \mathcal{L}(\boldsymbol{\theta}^{\text{merge}}_{\mathcal{S}}) - \mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S}}) \right|}{\binom{n-1}{|\mathcal{S}|}} \\
&\quad + \frac{1}{n} \sum_{\mathcal{S} \subseteq [n] \setminus \{i\}, |S| < k} \frac{|\mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S}}) - \mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S} \cup \{i\}})|}{\binom{n-1}{|\mathcal{S}|}} \\
&\overset{(a)}{\leq} \frac{(n-k)}{n} (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1) LC + \frac{k}{n} R.
\end{aligned}
$$

Inequality (a) follows by applying Lemma A.1 separately to the two sums. For the first sum, the same argument as in Appendix A.4 gives the uniform bound:

$$
\left| \mathcal{L}(\boldsymbol{\theta}^{\text{merge}}_{\mathcal{S} \cup \{i\}}) - \mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S} \cup \{i\}}) \right| + \left| \mathcal{L}(\boldsymbol{\theta}^{\text{merge}}_{\mathcal{S}}) - \mathcal{L}(\boldsymbol{\theta}^{\text{MTL}}_{\mathcal{S}}) \right| \leq \left(2\alpha n^2 - 2\alpha n + 2n + \alpha - 1 \right) LC.
$$

The total weight of coalitions with $|\mathcal{S}| \geq k$ is $\frac{1}{n} \sum_{\substack{\mathcal{S} \subseteq [n] \setminus \{i\} \\ |\mathcal{S}| \geq k}} \frac{1}{\binom{n-1}{|\mathcal{S}|}} = \frac{1}{n} \sum_{s=k}^{n-1} 1 = \frac{n-k}{n}$. For the excluded coalitions with $|\mathcal{S}| < k$, the marginal utility is bounded by R, and their total weight is: $\frac{1}{n} \sum_{\substack{\mathcal{S} \subseteq [n] \setminus \{i\} \\ |\mathcal{S}| < k}} \frac{1}{\binom{n-1}{|\mathcal{S}|}} = \frac{1}{n} \sum_{s=0}^{k-1} 1 = \frac{k}{n}$. Applying Lemma A.1 with these two values of b gives inequality (a).

A.6 MC Sampling approximation proof

Let ϕ_i denote the true Shapley value, $\hat{\phi}_i$ the deterministic approximation with error

$$
|\hat{\phi}_i - \phi_i| \leq (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC,
$$

and $\bar{\phi}_i^{(m)}$ the MC estimator of $\hat{\phi}_i$ based on m samples. Assume each sampled marginal contribution lies in an interval of width R, i.e., $\forall \mathcal{S} \subseteq [n] \setminus \{i\}, \quad |\hat{u}(\mathcal{S} \cup \{i\}) - \hat{u}(\mathcal{S})| \leq R$.

By Hoeffding's inequality, for any $\varepsilon > 0$,

$$
\Pr\left(|\bar{\phi}_i^{(m)} - \hat{\phi}_i| \geq \varepsilon \right) \leq 2 \exp\left(-\frac{2m\varepsilon^2}{R^2} \right).
$$

Using the triangle inequality,

$$
|\bar{\phi}_i^{(m)} - \phi_i| \leq |\bar{\phi}_i^{(m)} - \hat{\phi}_i| + |\hat{\phi}_i - \phi_i| \leq |\bar{\phi}_i^{(m)} - \hat{\phi}_i| + (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC.
$$

Thus, with probability at least $1 - \delta$, we have:

$$
|\bar{\phi}_i^{(m)} - \phi_i| \leq (2\alpha n^2 - 2\alpha n + 2n + \alpha - 1)\, LC + \sqrt{\frac{R^2}{2m} \log\left(\frac{2}{\delta} \right)}.
$$

Table 4: Comparison with additional valuation baselines on the NLP setting with $|\mathcal{T}| = 3$. We report Kendall's τ and Spearman's ρ with respect to ground-truth Dataset Shapley values.

# Sellers (n)	Method	Kendall's $\tau \uparrow$	Spearman's $\rho \uparrow$
7	DU-Shapley	0.724	0.821
	DMVM	**0.778**	**0.867**
	GTG-Shapley	0.619	0.714
	LOO	0.546	0.682
	Standalone	0.601	0.679
8	DU-Shapley	0.713	0.821
	DMVM	**0.758**	**0.861**
	GTG-Shapley	0.357	0.619
	LOO	0.528	0.663
	Standalone	0.591	0.680
9	DU-Shapley	0.703	0.818
	DMVM	**0.749**	**0.862**
	GTG-Shapley	0.389	0.533
	LOO	0.560	0.712
	Standalone	0.574	0.673

A.7 Additional Baselines

We additionally compare DMVM with several alternative valuation approaches.

GTG-Shapley (Liu et al., 2022) is a Shapley-based participant valuation method originally developed for federated learning. We adapt it to our multi-task setting by using the buyer's multi-task utility when evaluating participant contributions.

Leave-One-Out (LOO) (Sim et al., 2022; Ghorbani & Zou, 2019) estimates the value of seller i by measuring the decrease in the utility of the full coalition when that seller is removed:

$$\phi_i^{\mathrm{LOO}} = u(\mathcal{N}) - u(\mathcal{N} \setminus \{i\}).$$

Unlike Shapley value, LOO considers only the seller's marginal contribution to the full coalition and therefore ignores its contributions to smaller coalitions.

Standalone values each seller independently according to the utility obtained from that seller alone:

$$\phi_i^{\mathrm{Standalone}} = u(\{i\}).$$

This provides a simple baseline that does not account for interactions or complementarities among sellers.

Table 4 reports the comparison on the NLP setting with $|\mathcal{T}| = 3$. DMVM consistently achieves stronger agreement with ground-truth Dataset Shapley values than GTG-Shapley, LOO, and Standalone valuation across all evaluated marketplace sizes.

A.8 Robustness Analyses

We further evaluate the robustness of DMVM under several practical variations of the marketplace and buyer setup. Appendix A.8.1 studies the effect of the buyer's validation-set size, Appendix A.8.2 examines robustness to unequal seller dataset sizes, and Appendix A.8.3 evaluates DMVM when the buyer assigns different preferences to its target tasks.

A.8.1 Robustness to Buyer's Validation-Set Size

Since DMVM evaluates coalition utilities on the buyer's validation sets, we examine whether the resulting valuations are sensitive to the amount of validation data available. We vary the validation-set size from 800 to $2{,}000$ examples per target task and evaluate robustness both across different numbers of target tasks and different marketplace sizes.

DMVM remains highly stable across the evaluated validation-set sizes. Reducing the number of validation examples from $2{,}000$ to 800 changes Kendall's τ by at most 0.01, with similarly small changes in Spearman's ρ, while MSE changes by less than 10^{-4}. This behavior is consistent across both target-task sizes and seller counts, indicating that DMVM does not require a large buyer validation set to recover stable seller rankings.

Table 5: Robustness of DMVM to the buyer's validation-set size. We report Kendall's τ, Spearman's ρ, and MSE with respect to exact Dataset Shapley.

Setting	Val. Size	Kendall's $\tau \uparrow$	Spearman's $\rho \uparrow$	MSE $\downarrow$
Varying target-task size ($n = 9$)				
$\lvert\mathcal{T}\rvert = 2$	800	0.838	0.928	0.00232
	1400	0.840	0.929	0.00227
	2000	0.844	0.931	0.00224
$\lvert\mathcal{T}\rvert = 3$	800	0.743	0.856	0.00172
	1400	0.745	0.858	0.00170
	2000	0.749	0.862	0.00167
$\lvert\mathcal{T}\rvert = 4$	800	0.676	0.793	0.00141
	1400	0.679	0.796	0.00141
	2000	0.686	0.800	0.00138
$\lvert\mathcal{T}\rvert = 5$	800	0.694	0.824	0.00123
	1400	0.692	0.822	0.00123
	2000	0.698	0.827	0.00121
Varying number of sellers ($\lvert\mathcal{T}\rvert = 3$)				
$n = 7$	800	0.770	0.862	0.00201
	1400	0.777	0.866	0.00200
	2000	0.778	0.867	0.00197
$n = 8$	800	0.749	0.853	0.00185
	1400	0.752	0.857	0.00184
	2000	0.758	0.861	0.00181
$n = 9$	800	0.743	0.856	0.00172
	1400	0.745	0.858	0.00170
	2000	0.749	0.862	0.00167

Table 6: Robustness to imbalanced seller dataset sizes on vision tasks with $n = 6$ sellers. Results are reported as mean $\pm$ standard deviation across buyer target-task sets of size $|\mathcal{T}|$.

$\lvert\mathcal{T}\rvert$	Regime	Kendall's $\tau \uparrow$	Spearman's $\rho \uparrow$	MSE $\downarrow$
2	Balanced	0.9022 ± 0.1065	0.9505 ± 0.0606	0.00279 ± 0.00161
	Moderate	0.8756 ± 0.1065	0.9352 ± 0.0643	0.00285 ± 0.00154
	Strong	0.8044 ± 0.1321	0.9048 ± 0.0769	0.00314 ± 0.00138
3	Balanced	0.8067 ± 0.1183	0.8943 ± 0.0725	0.00208 ± 0.00071
	Moderate	0.7667 ± 0.1359	0.8629 ± 0.1005	0.00213 ± 0.00067
	Strong	0.7467 ± 0.1051	0.8743 ± 0.0631	0.00232 ± 0.00067

A.8.2 Robustness to Seller Dataset-Size Imbalance

Our main experiments use approximately balanced seller dataset sizes. To examine whether DMVM remains reliable when sellers contribute different amounts of data, we construct an imbalanced vision marketplace with $n = 6$ sellers: MNIST, Oxford-IIIT Pets, SVHN, RESISC45, USPS, and EuroSAT. We consider three regimes. In the *balanced* setting, every seller has $5{,}000$ training examples. The *moderate* setting uses $(5000, 5000, 3750, 2500, 5000, 5000)$ examples, while the *strong* setting uses $(5000, 3750, 2500, 1250, 1250, 1000)$ examples, following the dataset order above. For each regime, we train both the individual seller models used by DMVM and all coalition models required to compute exact Dataset Shapley. We keep the DMVM merging rule unchanged, i.e., $\alpha = 1/|\mathcal{S}|$, so this experiment directly tests the robustness of the original method to unequal seller dataset sizes.

Table 6 shows that increasing seller-size imbalance causes a gradual degradation in valuation accuracy, but DMVM remains robust even under the strongest imbalance considered. For $|\mathcal{T}| = 2$, Kendall's τ decreases from 0.902 to 0.804 and Spearman's ρ from 0.950 to 0.905 between the balanced and strongly imbalanced settings. For $|\mathcal{T}| = 3$, the corresponding changes are smaller, from 0.807 to 0.747 in τ and from 0.894 to 0.874 in ρ. MSE also increases only modestly in both cases. These results indicate that, although unequal seller dataset sizes make the valuation problem more challenging, DMVM continues to preserve useful agreement with exact Dataset Shapley without modifying its merging rule.

Table 7: Robustness of DMVM to non-uniform buyer task preferences for $n = 10$ and $|\mathcal{T}| = 3$. We report agreement with the corresponding weighted Dataset Shapley values.

Task Weighting	Kendall's $\tau \uparrow$	Spearman's $\rho \uparrow$	MSE $\downarrow$
Uniform	0.784	0.879	0.00097
$2\times$ Preference	0.782	0.882	0.00109
$4\times$ Preference	0.785	0.886	0.00144
Random Simplex	0.779	0.878	0.00143

A.8.3 Robustness to Buyer Task Preferences

Our main experiments use a uniform average across the buyer's target tasks. More generally, a buyer may prioritize some tasks over others. We therefore consider a weighted multi-task utility of the form

$$u(\mathcal{S}) = \sum_{t \in \mathcal{T}} w_t \, \mathrm{Perf}_t\big(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{MTL}}, \mathcal{D}_t^{\mathrm{val}}\big), \qquad w_t \geq 0, \quad \sum_{t \in \mathcal{T}} w_t = 1,$$

with the corresponding DMVM proxy

$$\hat{u}(\mathcal{S}) = \sum_{t \in \mathcal{T}} w_t \, \mathrm{Perf}_t\big(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{merge}}, \mathcal{D}_t^{\mathrm{val}}\big).$$

Thus, task weighting only changes the buyer's utility definition; the model merging and Shapley-value computation remain unchanged.

We evaluate robustness to buyer preferences in the vision setting with $n = 10$ sellers and $|\mathcal{T}| = 3$ target tasks. In addition to uniform weighting, we consider settings in which one target task is weighted $2\times$ or $4\times$ more than each of the remaining tasks, cycling the preferred task so that every target task is emphasized in turn. We also sample 20 random preference vectors from a symmetric Dirichlet distribution, $\boldsymbol{w} \sim \mathrm{Dirichlet}(1, 1, 1)$.

DMVM remains highly stable as the buyer's task preferences become increasingly non-uniform. Increasing the relative preference for one task from uniform weighting to $4\times$ leaves Kendall's τ essentially unchanged (0.784 to 0.785), while Spearman's ρ slightly improves (0.879 to 0.886). The randomly sampled preference vectors show similarly strong agreement, with $\tau = 0.779$ and $\rho = 0.878$. Although the MSE increases modestly for more skewed preferences, the seller rankings remain stable, indicating that DMVM can accommodate heterogeneous buyer objectives without modifying the underlying valuation framework.

A.9 Quality of the Merging-Based Utility Proxy

DMVM replaces the expensive joint-training utility $u(\mathcal{S})$ with the merged-model utility $\hat{u}(\mathcal{S})$. Importantly, dataset valuation does not require the merged model to exactly reproduce the utility of every coalition; rather, the proxy should preserve the relative marginal contributions that determine seller values. We therefore directly examine both the magnitude and the coalition-dependent variation of the utility approximation error. Unless otherwise stated, we consider deterministic DMVM without singleton exclusion, so the discrepancy measured here isolates the effect of the merging-based utility proxy from Monte Carlo sampling and the privacy-aware estimator.

We first measure the absolute utility error $|u(\mathcal{S}) - \hat{u}(\mathcal{S})|$ over all evaluated coalitions. As shown in Table 8, the average gap is 0.0492 for vision and 0.0925 for NLP. The gap increases as the number of target tasks grows, particularly for NLP, reflecting the greater difficulty of approximating joint training as the multi-task objective becomes more heterogeneous.

Absolute error alone, however, can overstate its effect on valuation. A systematic component shared by coalitions of the same size does not determine their relative ordering; what is more consequential is the residual variation among comparable coalitions. To characterize this rank-relevant component, we define the signed proxy error $\Delta(\mathcal{S}) = u(\mathcal{S}) - \hat{u}(\mathcal{S})$, and report its mean and within-cell standard deviation at fixed target-task and coalition sizes. Table 9 presents this analysis for NLP, which exhibits the larger overall approximation gap and therefore provides the more challenging setting.

The results reveal an encouraging structure in the approximation error. Although the merged utility is not an exact substitute for joint training, its coalition-dependent variation decreases sharply as the

Table 8: Average absolute utility error $|u(\mathcal{S}) - \hat{u}(\mathcal{S})|$ across different numbers of buyer target tasks.

$\lvert\mathcal{T}\rvert$	Vision ↓	NLP ↓
2	0.0311	0.0505
3	0.0404	0.0718
4	0.0495	0.0927
5	0.0582	0.1134
6	0.0668	0.1341
Average	0.0492	0.0925

Table 9: Mean ± standard deviation of the signed proxy error $\Delta(\mathcal{S})$ for NLP, grouped by target-task and coalition sizes. The standard deviation captures variation among coalitions of equal size, which is particularly relevant to preserving relative marginal contributions.

$\lvert\mathcal{T}\rvert\backslash\lvert\mathcal{S}\rvert$	2	3	4	5	6	7	8
2	0.0132 ± 0.0473	0.0360 ± 0.0280	0.0469 ± 0.0168	0.0482 ± 0.0118	0.0465 ± 0.0086	0.0454 ± 0.0072	0.0443 ± 0.0037
3	0.0197 ± 0.0722	0.0540 ± 0.0494	0.0704 ± 0.0342	0.0724 ± 0.0239	0.0698 ± 0.0155	0.0681 ± 0.0114	0.0665 ± 0.0060
4	0.0263 ± 0.0918	0.0720 ± 0.0611	0.0938 ± 0.0431	0.0965 ± 0.0287	0.0930 ± 0.0184	0.0908 ± 0.0125	0.0886 ± 0.0076
5	0.0329 ± 0.0992	0.0900 ± 0.0629	0.1173 ± 0.0468	0.1206 ± 0.0307	0.1163 ± 0.0204	0.1136 ± 0.0147	0.1108 ± 0.0077
6	0.0395 ± 0.1447	0.1080 ± 0.0962	0.1408 ± 0.0736	0.1447 ± 0.0426	0.1395 ± 0.0255	0.1363 ± 0.0155	0.1330 ± 0.0053

coalition grows. For example, at $|\mathcal{T}| = 3$, the within-size standard deviation drops from 0.0722 at $|\mathcal{S}| = 2$ to only 0.0060 at $|\mathcal{S}| = 8$. Similar behavior is observed across target-task sizes. Thus, much of the approximation error for larger coalitions behaves as a shared bias rather than highly irregular coalition-specific noise.

This distinction is important for DMVM: the goal of the proxy is not pointwise reconstruction of joint-training utilities, but faithful recovery of seller marginal contributions. Despite the nonzero utility gap, DMVM retains strong agreement with exact Dataset Shapley in the main experiments. Taken together, the direct utility analysis and the resulting valuation accuracy show that model merging provides a useful surrogate for the otherwise prohibitively expensive coalition-training process, while reducing each coalition evaluation to merging and inference.

A.10 Effect of Singleton Exclusion

The privacy-aware version of DMVM excludes singleton coalitions so that the buyer never directly observes an individual seller model. This restriction modifies the Shapley estimator and can therefore introduce an accuracy–privacy trade-off. To quantify its effect, we compare DMVM-FULL, which uses all coalitions, with DMVM-EXCLUDE, which removes singleton coalitions, against exact Dataset Shapley. Table 10 illustrates the comparison. As expected, excluding singleton coalitions introduces a measurable loss in valuation accuracy, with the largest effect in very small marketplaces. The impact decreases substantially as the number of sellers grows: the MSE of DMVM-EXCLUDE drops from 0.0483 at $n = 3$ to 0.0066 at $n = 9$, while rank correlations remain relatively stable. This trend is natural since singleton coalitions account for a progressively smaller portion of the coalitions considered as the marketplace grows. Importantly, this trade-off is specific to the privacy-aware estimator: when individual merged models need not be hidden through singleton exclusion, standard DMVM can retain the full coalition information.

A.11 Controlled Simulations

The experiments in Section 4 evaluate DMVM in practical marketplace scenarios where sellers correspond to different datasets and domains. However, in such settings, the true relevance of each seller to the buyer's target tasks is not directly controllable, since datasets may differ along several factors simultaneously, including domain, distribution, label space, and sample complexity. Therefore, we further conduct controlled simulations in which the relationship between each seller's data and the buyer's target tasks is explicitly designed. These experiments allow us to test whether DMVM can recover known relevance patterns, including highly useful, partially useful, irrelevant, and harmful sellers.

We consider two complementary experimental regimes. First, we study *single-domain multi-task* vision benchmarks, which provide an intuitive and interpretable setting for analyzing seller relevance.

Table 10: Effect of excluding singleton coalitions. We report Kendall's τ, Spearman's ρ, and MSE with respect to exact Dataset Shapley. DMVM-FULL uses all coalitions, while DMVM-EXCLUDE corresponds to the privacy-aware estimator.

n	DMVM-FULL			DMVM-EXCLUDE		
	$\tau \uparrow$	$\rho \uparrow$	MSE $\downarrow$	$\tau \uparrow$	$\rho \uparrow$	MSE $\downarrow$
3	0.899	0.921	0.0022	0.794	0.829	0.0483
4	0.848	0.898	0.0028	0.807	0.872	0.0272
5	0.853	0.912	0.0030	0.775	0.855	0.0185
6	0.853	0.919	0.0029	0.785	0.875	0.0134
7	0.856	0.928	0.0026	0.797	0.893	0.0102
8	0.849	0.928	0.0024	0.801	0.899	0.0082
9	0.844	0.931	0.0022	0.790	0.899	0.0066

Second, we extend our evaluation to *multi-domain multi-task* experiments in both vision and NLP, where tasks may correspond to distinct domains. Detailed task distribution schemes for different buyer task set sizes and seller constructions are provided in Appendix A.15. Scalability analysis is provided in Appendix A.13.

A.11.1 Single-Domain Multi-Task Vision Experiments

To establish intuition, we begin with vision experiments where all tasks belong to a single domain. We consider MNIST and CIFAR-10, and define the buyer's objective as learning a subset of classes, treated as multiple target tasks. Multiple sellers provide datasets with varying relevance to the buyer's task set. For these experiments, we compute exact multi-task Shapley values by explicitly training models on all subsets of sellers. The resulting Shapley values serve as the reference ordering against which DMVM-induced rankings are compared.

Non-Overlapping Sellers. We first consider a non-overlapping seller setting, where each seller provides data for a disjoint subset of classes. As an illustrative MNIST example, we define the buyer's target tasks as learning to recognize digits 0-6, and construct four sellers as follows:

- Seller 1: classes (0, 1, 2, 3)
- Seller 2: classes (4, 5, 6)
- Seller 3: classes (7, 8, 9)
- Seller 4: mislabeled or noisy samples

Sellers 1 and 2 contribute directly to the buyer's tasks, with Seller 1 expected to have higher Shapley value due to broader task coverage. Seller 3 provides irrelevant data, while Seller 4 is harmful. The reference ranking is therefore induced by the corresponding Shapley values.

Overlapping Sellers. We next consider an overlapping seller setting, where sellers provide partially intersecting subsets of the buyer's tasks. For MNIST, the buyer's tasks correspond to recognizing digits 0-3, and we simulate five sellers with progressively decreasing overlap:

- Seller 1: classes (0, 1, 2, 3)
- Seller 2: classes (0, 1, 2, 4)
- Seller 3: classes (0, 1, 4, 5)
- Seller 4: classes (0, 4, 5, 6)
- Seller 5: classes (4, 5, 6, 7)

This setting induces a smooth spectrum of Shapley values, allowing us to evaluate whether DMVM preserves relative ordering as task relevance decreases.

A.11.2 Multi-Domain Multi-Task Language Modeling

We next evaluate DMVM in a multi-domain multi-task language modeling setting, where each task corresponds to a distinct reasoning domain and models are generative. We consider the following benchmarks: `CommonsenseQA` (Talmor et al., 2019) (commonsense multiple-choice reasoning), `OpenBookQA` (Mihaylov et al., 2018) (open-book science QA with multi-step reasoning), `MedMCQA` (Pal et al., 2022) (medical multiple-choice question answering), `LogiQA` (Liu et al.,

Table 11: Rank correlation between DMVM-induced rankings and multi-task Shapley-value-based rankings for single-domain multi-task vision experiments with non-overlapping sellers.

$\lvert\mathcal{T}\rvert$	MNIST		CIFAR-10	
	τ	ρ	τ	ρ
5	1.0	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0

Table 12: Rank correlation between DMVM-induced rankings and multi-task Shapley-value-based rankings for single-domain multi-task vision experiments with overlapping sellers.

$\lvert\mathcal{T}\rvert$	MNIST		CIFAR-10	
	τ	ρ	τ	ρ
3	0.92 ± 0.16	0.95 ± 0.1	0.67 ± 0.0	0.8 ± 0.0
4	0.8 ± 0.0	0.9 ± 0.0	0.85 ± 0.1	0.92 ± 0.05
5	0.77 ± 0.13	0.87 ± 0.08	0.83 ± 0.2	0.87 ± 0.16
6	0.95 ± 0.1	0.97 ± 0.05	0.8 ± 0.16	0.88 ± 0.13

2020) (logical reasoning over short passages), `GSM8K` (Cobbe et al., 2021) (grade-school mathematical problem solving), and `ARC-Easy` (Clark et al., 2018) (elementary-level science question answering). Each task is treated as a separate domain-specific learning objective. We again adopt non-overlapping and overlapping seller designs. For non-overlapping, given a buyer with $|\mathcal{T}|$ target tasks, sellers partition the task set disjointly. For example, when the buyer has six target tasks, Seller 1 provides data for three tasks, Seller 2 for two tasks, and Seller 3 for the remaining task. An additional seller provides data from an irrelevant task outside the buyer's target set. This design induces a clear hierarchy of Shapley values based on task coverage, enabling quantitative comparison with DMVM-induced rankings.

Results and Discussion. For single-domain multi-task vision experiments, we employed ViT-B/16 (Wu et al., 2020) as the shared pretrained model initialized from the HuggingFace checkpoint `google/vit-base-patch16-224-in21k`. Each seller fine-tunes this model for two epochs on its own dataset, after which we compute the valuation scores using $\hat{\phi}$. Fine-tuning is performed with AdamW using a learning rate of 1×10^{-4} and weight decay 0.01, providing stable adaptation across all seller datasets. Experiments for multi-domain multi-task vision are provided in Appendix A.11.3. For multi-domain multi-task language experiments, we used two pretrained large language models with different scales and architectures: **LLaMA-3.2-1B** (Grattafiori et al., 2024) and **Qwen2.5-3B** (Qwen et al., 2025). Tables 11, 12, and 13 collectively demonstrate that DMVM induces seller rankings closely aligned with ground-truth Shapley-value-based orderings across both vision and language modalities. Performance is strongest in non-overlapping settings, while remaining robust under partial overlap and across diverse language domains.

In vision experiments, DMVM exactly recovers the Shapley ordering in non-overlapping settings (Table 11), and remains highly correlated under partial overlap (Table 12). In language modeling, DMVM achieves strong rank correlation for both LLaMA-3.2-1B and Qwen2.5-3B as the shared model, demonstrating that merging-based valuation extends naturally to generative, multi-domain scenarios (Table 13). Table 15 compares the wall-clock time required to compute dataset Shapley values using DMVM and exact ground-truth Shapley in the language domain, highlighting the computational efficiency of DMVM. Taken together, these results demonstrate that DMVM remains effective across a wide range of shared model architectures, model scales, data modalities, and marketplace seller topologies, including both vision and language models.

The consistent recovery of seller rankings indicates that DMVM is not tied to a specific architecture or parameter regime, but instead leverages transferable representation shifts induced by seller-specific fine-tuning. Importantly, DMVM achieves this while requiring only a single fine-tuning per seller, avoiding the combinatorial retraining costs associated with Shapley-value estimation and enabling scalable deployment in multi-seller settings. Moreover, valuation is performed through model merging rather than direct access to seller data, preserving data locality and supporting privacy-aware evaluation.

Table 13: Rank correlation between DMVM-induced rankings and Shapley-value-based rankings for multi-domain multi-task language modeling.

$\lvert\mathcal{T}\rvert$	LLaMa-3.2-1B		Qwen2.5-3B	
	τ	ρ	τ	ρ
3 (overlapping)	0.78	0.87	0.84	0.9
5	0.89	0.93	1.0	1.0
6	1.0	1.0	1.0	1.0

A.11.3 Multi-Domain Multi-Task Vision

Table 14: DMVM ranking performance in the multi-domain multi-task vision setting for different numbers of target tasks under non-overlapping and overlapping task distributions.

$\lvert\mathcal{T}\rvert$	Non-overlapping		Overlapping	
	τ	ρ	τ	ρ
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0

We evaluate DMVM in a multi-domain multi-task vision setting using **CLIP-ViT-B/32** (Radford et al., 2021) as the shared pretrained model. To construct a set of heterogeneous vision tasks, we follow the dataset selection used in the Fusion benchmark (Tang et al., 2025a), which was designed for evaluating model merging across diverse visual domains.

Specifically, we consider: MNIST (LeCun et al., 1998) (handwritten digit classification), CIFAR-10 (Krizhevsky et al., 2009) (natural object classification), GTSRB (Stallkamp et al., 2011) (traffic sign recognition), SVHN (Netzer et al., 2011) (street-view house number digit classification), Stanford Cars (Krause et al., 2013) (fine-grained car model classification), EuroSAT (Helber et al., 2019) (satellite land-use and land-cover classification), Oxford-IIIT (Parkhi et al., 2012) (pet category), DTD (Cimpoi et al., 2014) (describable texture classification), SUN397 (Xiao et al., 2010) (scene recognition across diverse environments), RESISC45 (Cheng et al., 2017) (Remote sensing image scene classification), Fashion-MNIST (Xiao et al., 2017) (clothing category), and Food101 (Bossard et al., 2014) (food categories). These datasets span multiple visual domains and data distributions, forming a heterogeneous multi-task setting suitable for evaluating multi-domain multi-task dataset valuation.

Results And Discussion. As summarized in Table 14, DMVM consistently recovers the ground-truth seller ranking with perfect agreement across all multi-domain settings we consider. Both correlations remain equal to 1.0, under both non-overlapping and overlapping task-distribution schemes. This indicates that, even in the presence of substantial domain heterogeneity—ranging from digits and traffic signs to satellite imagery, textures, scenes, and fine-grained object categories—DMVM preserves the relative importance ordering of sellers exactly. The results demonstrate that our valuation proxy remains stable as the number of target tasks grows and as inter-seller task overlap increases, highlighting its robustness in realistic multi-domain multi-task vision scenarios.

Table 15: Total wall-clock time required to compute dataset Shapley values using MTL (ground-truth) versus DMVM on language tasks. Measured averages: training = 1 hour, evaluation = 0.1 hour (6 minutes), merging (negligible ≈ 0).

# Sellers (n)	Dataset-Shapley (GT) Time (hours)	DMVM Time (hours)
4	$16 \times 1.1 = 17.6$	$16 \times 0.1 = 1.6$
5	$32 \times 1.1 = 35.2$	$32 \times 0.1 = 3.2$

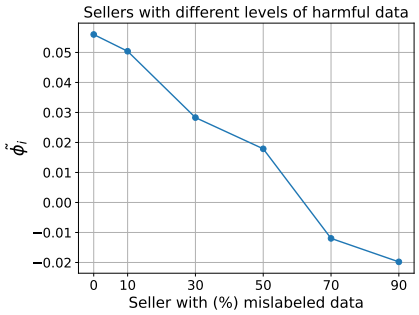

Figure 5: DMVM-induced Shapley values of sellers with different percentages of mislabeled data. As label noise increases, the seller's contribution decreases and eventually becomes negative, showing that DMVM successfully identifies and penalizes harmful data providers.

A.12 Harmful Dataset Detection

To examine whether DMVM can distinguish between helpful and harmful data providers, we design a controlled synthetic experiment in which a seller's dataset is progressively degraded by randomly flipping an increasing fraction of its labels. This setup allows us to systematically vary data quality while keeping all other factors fixed. Figure 5 illustrates the resulting Shapley value of the seller as a function of the mislabeling rate. We observe a clear and consistent monotonic decline in valuation as the noise level increases, indicating that DMVM is sensitive to degradation in data quality. This behavior highlights an important practical property of DMVM: beyond ranking relevant contributors highly, it can also reliably detect and penalize harmful or low-quality data sources in a multi-task learning setting.

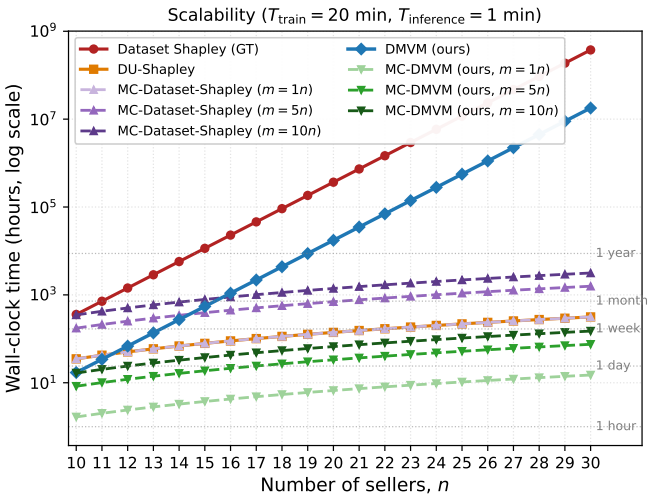

Figure 6: Projected wall-clock cost of each valuation method as the marketplace size n grows, computed from the asymptotic formulas in Table 1 with $T_{\text{train}} = 20$ min and $T_{\text{inference}} = 1$ min.

A.13 Scalability Analysis

To assess how DMVM scales beyond the regimes we measured directly, we project the wall-clock cost of each method in Table 1 using $T_{\text{train}} = 20$ minutes and $T_{\text{inference}} = 1$ minute, a reasonable assumption for regular downstream tasks and 0.5–1B parameter models. Figure 6 reports the projected runtime as the marketplace size grows from $n = 10$ to $n = 30$ sellers. For the MC variants, we sweep $m \in \{n, 5n, 10n\}$ to span different possible sample budgets. The figure shows two qualitatively different regimes. Exact Dataset Shapley and deterministic DMVM both scale as

$\mathcal{O}(2^n)$ and become infeasible past $n \approx 11$: at $n = 30$, exact Shapley requires more than 10^8 hours of computation, while deterministic DMVM, despite avoiding all training, still exceeds 10^7 hours. The polynomial methods are tractable in this regime, but DU-Shapley and MC-Dataset-Shapley both require retraining on sellers' raw data, which is incompatible with the decentralized setting we target. MC-DMVM is the only method that is simultaneously polynomial in n, dependent only on $T_{\text{inference}}$, and privacy-aware: at $n = 30$ it completes in approximately 15, 75, and 150 hours for $m = n$, $5n$, and $10n$, respectively. Real-world data marketplaces could involve tens of sellers per valuation round, and Figure 6 shows that this is precisely the regime where MC-DMVM remains practical while requiring a single local fine-tuning per seller and no raw-data exchange.

Moreover, to better study the effect of scaling number of sellers, n, and number of buyer's target tasks, $|\mathcal{T}|$, on the valuation performance, we have provided an extended version of Table 2 and Table 3. Figures 7 and 8 report DMVM's empirical performance across the full $(n, |\mathcal{T}|)$ grid, for vision and NLP, respectively. Across all configurations, Kendall's τ remains above 0.69, Spearman's ρ above 0.80, and MSE stays on the order of 10^{-3} in both modalities. MSE decreases as $|\mathcal{T}|$ grows, since averaging utility over more validation tasks reduces variance in the estimated coalition values, while rank-correlation metrics decline mildly with both n and $|\mathcal{T}|$: an expected effect, since larger marketplaces introduce more pairwise orderings to recover. The decline is gradual rather than abrupt, indicating that DMVM remains stable across the parameter range relevant to multi-task data marketplaces.

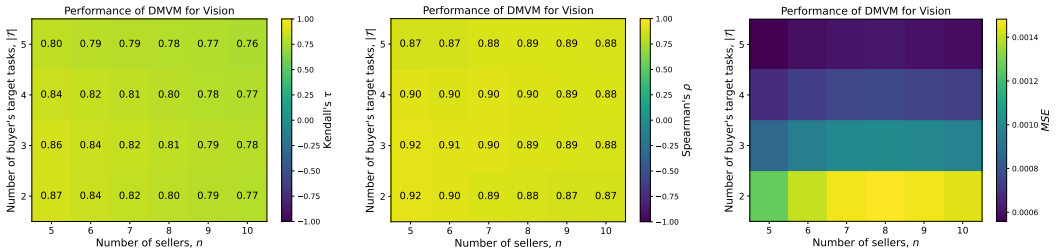

Figure 7: DMVM performance on vision tasks across the full $(n, |\mathcal{T}|)$ grid. Left to right: Kendall's τ, Spearman's ρ, and MSE relative to ground-truth Dataset Shapley.

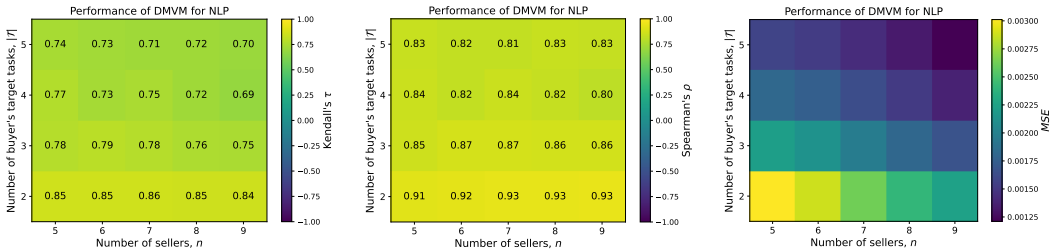

Figure 8: DMVM performance on NLP tasks across the full $(n, |\mathcal{T}|)$ grid. Left to right: Kendall's τ, Spearman's ρ, and MSE relative to ground-truth Dataset Shapley.

A.14 Implementation Details

All experiments are implemented in PyTorch. For both vision and NLP experiments, seller models are fine-tuned from a shared pretrained initialization using the Adam optimizer. Since computing the ground-truth Dataset Shapley values requires training models for an exponential number of seller coalitions, we use a lightweight fine-tuning setup and train most coalition models for only 2–3 epochs. This makes the ground-truth computation feasible while still providing a consistent reference for comparing valuation methods.

For the NLP experiments, we use `google/flan-t5-small` as the shared pretrained model. The main NLP results are obtained using a batch size of 32 and learning rate 3×10^{-4}. For the vision experiments, we use `openai/clip-vit-base-patch32` as the shared pretrained model. The main vision results are obtained using a batch size of 64, learning rate 10^{-5}, and weight decay 10^{-4}.

Table 16: Task distribution for $|\mathcal{T}| = 3$.

	$\mathcal{T} = \{t_1, t_2, t_3\}$	
Seller	Non-overlapping	Overlapping
1	$\{t_1, t_2\}$	$\{t_1, t_2, t_3\}$
2	$\{t_3\}$	$\{t_1, t_2, t_4\}$
3	$\{t_4\}$	$\{t_1, t_4, t_5\}$
4	noisy/mislabeled/no data	$\{t_4, t_5, t_6\}$

Table 17: Task distribution for $|\mathcal{T}| = 4$.

	$\mathcal{T} = \{t_1, t_2, t_3, t_4\}$		
Seller	Non-overlapping		Overlapping
	Scheme A	Scheme B	
1	$\{t_1, t_2, t_3\}$	$\{t_1, t_2\}$	$\{t_1, t_2, t_3, t_4\}$
2	$\{t_4\}$	$\{t_3\}$	$\{t_1, t_2, t_3, t_5\}$
3	$\{t_5\}$	$\{t_5\}$	$\{t_1, t_2, t_5, t_6\}$
4	noisy/mislabeled/no data	noisy/mislabeled/no data	$\{t_1, t_5, t_6, t_7\}$
5	—	—	$\{t_5, t_6, t_7, t_8\}$

For merging the seller models in each subset $\mathcal{S}$, we use task arithmetic with scaling coefficient $\alpha = \frac{1}{|\mathcal{S}|}$, corresponding to averaging the task vectors within the subset. For each buyer's target task, we evaluate utility on at most 2000 validation examples, using the full test set when it contains fewer than 2000 examples. The buyer's utility is then computed by averaging validation performance across the target tasks. All main experiments can be run on a single NVIDIA H100 GPU. The main computational bottleneck is the retraining-based ground-truth and baseline computation, whereas DMVM only requires model merging over sellers and inference over the buyer's validation tasks.

A.15 Task Distribution Schemes

In the controlled simulation experiments, for buyers with target task sets of varying cardinalities, we design task allocation schemes for both non-overlapping and overlapping seller constructions. In the non-overlapping schemes, each fully relevant seller is assigned a distinct number of tasks from $\mathcal{T}$. This avoids ties or near-ties in Shapley rankings and ensures a clear separation in seller relevance. Sellers with irrelevant data may contain data from multiple irrelevant tasks or domains and can represent different combinations of such irrelevant sources. In addition, a set of irrelevant data may come from only a subset of the irrelevant domains. Tables 16 to 20 illustrate the high-level structure of these task distribution schemes.

Table 18: Task distribution for $|\mathcal{T}| = 5$.

	$\mathcal{T} = \{t_1, t_2, t_3, t_4, t_5\}$		
Seller	Non-overlapping		Overlapping
	Scheme A	Scheme B	
1	$\{t_1, t_2, t_3\}$	$\{t_1, t_2, t_3\}$	$\{t_1, t_2, t_3, t_4, t_5\}$
2	$\{t_4, t_5\}$	$\{t_4\}$	$\{t_1, t_2, t_3, t_4, t_6\}$
3	$\{t_6\}$	$\{t_6\}$	$\{t_1, t_2, t_3, t_6, t_7\}$
4	noisy/mislabeled/no data	noisy/mislabeled/no data	$\{t_1, t_2, t_6, t_7, t_8\}$
5	—	—	$\{t_1, t_6, t_7, t_8, t_9\}$
6	—	—	$\{t_6, t_7, t_8, t_9, t_{10}\}$

Table 19: Task distribution for $|\mathcal{T}| = 6$.

	$\mathcal{T} = \{t_1, t_2, t_3, t_4, t_5, t_6\}$	
Seller	Non-overlapping	Overlapping
1	$\{t_1, t_2, t_3\}$	$\{t_1, t_2, t_3, t_4, t_5, t_6\}$
2	$\{t_4, t_5\}$	$\{t_1, t_2, t_3, t_4, t_5, t_7\}$
3	$\{t_6\}$	$\{t_1, t_2, t_3, t_4, t_7, t_8\}$
4	$\{t_7\}$	$\{t_1, t_2, t_3, t_7, t_8, t_9\}$
5	noisy/mislabeled/no data	$\{t_1, t_2, t_7, t_8, t_9, t_{10}\}$
6	—	$\{t_1, t_7, t_8, t_9, t_{10}, t_{11}\}$
7	—	$\{t_7, t_8, t_9, t_{10}, t_{11}, t_{12}\}$

Table 20: Task distribution for $|\mathcal{T}| = 7$.

	$\mathcal{T} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$			
Seller	Non-overlapping			Overlapping
	Scheme A	Scheme B	Scheme C	
1	$\{t_1, t_2, t_3\}$	$\{t_1, t_2, t_3, t_4\}$	$\{t_1, t_2, t_3, t_4\}$	$\{t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$
2	$\{t_4, t_5\}$	$\{t_5, t_6\}$	$\{t_5, t_6, t_7\}$	$\{t_1, t_2, t_3, t_4, t_5, t_6, t_8\}$
3	$\{t_6\}$	$\{t_7\}$	$\{t_8\}$	$\{t_1, t_2, t_3, t_4, t_5, t_8, t_9\}$
4	$\{t_7\}$	$\{t_8\}$	noisy/mislabeled/no data	$\{t_1, t_2, t_3, t_4, t_8, t_9, t_{10}\}$
5	noisy/mislabeled/no data	noisy/mislabeled/no data	—	$\{t_1, t_2, t_3, t_8, t_9, t_{10}, t_{11}\}$
6	—	—	—	$\{t_1, t_2, t_8, t_9, t_{10}, t_{11}, t_{12}\}$
7	—	—	—	$\{t_1, t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}\}$
8	—	—	—	$\{t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}, t_{14}\}$

A.16 LIMITATIONS

The current formulation of DMVM makes several assumptions that define its scope. First, task-vector merging assumes that seller models are fine-tuned from a shared pretrained checkpoint. Thus, each seller performs a one-time local fine-tuning that can be reused across coalition evaluations and buyers sharing the same base model, but a buyer adopting a different checkpoint or architecture may require a new seller-side fine-tuning. Extending DMVM to recent heterogeneous model-merging techniques (Xu et al., 2024; Chen et al., 2026) is a promising direction for relaxing this requirement.

Second, deterministic DMVM replaces expensive coalition retraining with model merging and inference, but still enumerates an exponential number of coalitions. MC-DMVM is therefore the variant intended for larger marketplaces, trading a controllable sampling error for substantially improved computational and communication scalability. The privacy-aware protocol also introduces communication overhead, since each coalition evaluation requires sellers to transmit masked model updates and coordinate secure-aggregation masks. Similarly, the quality of DMVM ultimately depends on how faithfully the underlying merging method approximates joint training. Our experiments directly characterize this approximation and show strong valuation fidelity, but larger task interference can increase the proxy error. We use task arithmetic because of its simplicity and theoretical tractability; incorporating stronger merging methods provides a natural extension of the framework. Moreover, our analysis assumes full-batch gradient descent for tractability and to isolate the deterministic approximation error. Practical training may instead use stochastic optimizers, as in our experiments; extending the bound to explicitly account for stochastic-gradient noise is left to future work. Finally, our privacy protocol targets a semi-honest setting in which participants follow the prescribed protocol but may attempt to infer private models from their observations; protection against fully malicious participants or arbitrary collusion is outside the current scope.