FASTDV: SCALABLE MODEL-AGNOSTIC DATA VALUATION VIA DENSITY RATIO LEAVE-ONE-OUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Leave-One-Out (LOO) data valuation (DV) quantifies the contribution of an individual sample through the finite utility change induced by its removal, providing useful guidance for data pruning, procurement, and selection. While model-agnostic DV avoids repeated model retraining by evaluating distributional utilities, existing methods typically rely on sensitivity, influence, or hierarchical approximations rather than explicit finite deletion. Moreover, most DV methods are developed for centralized settings, leaving independently computed values across multiple data holders generally incomparable. In this work, we propose **FastDV**, a scalable model-agnostic valuation framework. Specifically, we first derive an additive-statistic formulation that decomposes target set dependence into sample-wise additive statistics, making individual deletions analytically decomposable and distributed data partitions naturally composable. Guided by this principle, FastDV defines sample value through a Pearson-divergence-based density-ratio utility and achieves efficient LOO valuation through exact rank-one updates. Leveraging the partition composability of the above formulation, FastDV further extends to multi-party global valuation and produces globally comparable values without pooling raw client data. Extensive experiments on typical valuation tasks and data marketplaces demonstrate the effectiveness and scalability of FastDV. Notably, FastDV achieves up to **15.7**$\times$ speedup over the fastest baseline on million-scale datasets.

1 INTRODUCTION

With the paradigm shift from model-centric to data-centric AI, data quality has become a decisive factor governing model performance (Liang et al., 2023), generalizability (Yan et al., 2025), and fairness (Pessach & Shmueli, 2022). However, real-world data inevitably exhibit heterogeneity, redundancy, and noise, caused by diverse collection sources and susceptibility to inconsistent processing or labeling errors (Li et al., 2025b; Jiang et al., 2023a;b). Therefore, data valuation (DV) (Ghorbani & Zou, 2019; Koh & Liang, 2017; Sim et al., 2022) is proposed to quantify data quality and evaluate the intrinsic properties of data samples. By identifying and leveraging high-value data, DV methods facilitate the development of high-performance and trustworthy models. This quantification paradigm has promoted broad adoption in practical scenarios, such as noise detection (Wang & Jia, 2023), dataset pruning (Yoon et al., 2020), and data pricing (Cong et al., 2022).

Among existing DV methods, Leave-One-Out (LOO) (Cook, 1977) remains a cornerstone method, assessing sample value through the finite change in a predefined data utility after removing that sample. Driven by distinct optimization objectives, later work mainly falls into two lines: 1) Model-dependent valuation. Shapley-based methods (Ghorbani & Zou, 2019) measure marginal contributions over training subsets, while influence-based methods (Koh & Liang, 2017) approximate the effect of sample removal through parameter perturbations. Subsequent works improve their scalability through sampling or alternative utility functions (Jia et al., 2019; Watson et al., 2022; Pruthi et al., 2020; Kokhlikyan et al., 2025; Askari et al., 2025), but the resulting values remain coupled with a particular model or training procedure. 2) Model-agnostic valuation. Recent methods instead evaluate samples through distributional discrepancies between the target and support sets, such as optimal transport or kernel-based distances (Just et al., 2023; Zhu et al., 2025; Liu et al., 2026). Although they decouple valuation from model optimization, they typically rely on perturbation- or influence-based approximations to estimate sample contributions, rather than explicitly evaluating the finite utility change, and therefore do not recover the exact LOO value under their respective distributional utilities.

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.

Despite this progress, two challenges remain. **(i) Tension between finite-deletion fidelity and scalability.** For model-agnostic DV, directly evaluating the LOO value requires measuring the distributional utility both before and after removing each candidate sample. Repeating this procedure over all samples can be prohibitively expensive at scale. Existing methods therefore commonly replace explicit finite deletion with local sensitivity, influence, or hierarchical approximations (Just et al., 2023; Kessler et al., 2025; Zhu et al., 2025). While effective computationally, these approximations no longer directly evaluate the finite utility change that defines LOO. This raises a fundamental question: *can a model-agnostic valuation principle preserve explicit finite deletion while remaining computationally scalable?*

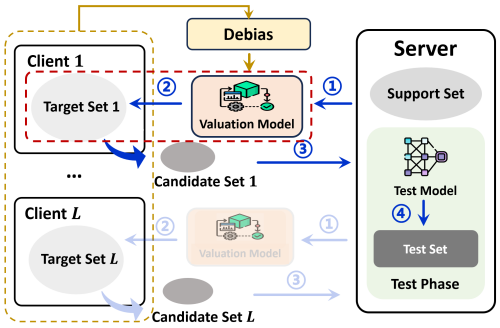


Figure 1: Illustration of the overall pipeline for data valuation. ① Given the support set provided by the data purchaser (server), ② the valuation model quantifies the target sets from data sellers (clients). Subsequently, ③ the purchaser procures high-value candidate data from each client and ④ then uses the server side test set to verify the quality of the purchased data via a model trained on them. (**Challenge**: Existing DV methods suffer from low valuation efficiency in stage ②, and lack scalability to multi-party data markets, as highlighted by the red and yellow boxes, respectively.)

(ii) Lack of global comparability in multi-party data valuation. Most sample-level DV methods assume that candidate data are centrally accessible. In realistic multi-party data marketplaces, however, datasets held by different institutions often remain isolated due to commercial and data-governance constraints (Li et al., 2025a). As a result, valuation is typically performed separately on each client using its own local target set. Because each client is evaluated with respect to different local target statistics, the resulting sample values may become incomparable when client datasets are heterogeneous or redundant. This prevents reliable global value ranking and raises another question: *can sample values be made globally comparable across distributed data holders without pooling their raw data?*

To address these challenges, we propose **FastDV**, a scalable model-agnostic framework that takes finite deletion as the starting point of valuation design. Rather than approximating the effect of sample removal after specifying a distributional utility, we propose an additive-statistic formulation that summarizes target-set dependence through sample-wise additive statistics. This formulation makes individual deletions decomposable and distributed data partitions composable. Based on this, FastDV instantiates this formulation with a Pearson-divergence-based density-ratio utility, whose risk-based interpretation connects distributional alignment to downstream prediction. Under a fixed kernel representation, its least-squares realization expresses the target-set dependence through an empirical second-moment matrix, to which each sample contributes a rank-one outer product. Therefore, deleting a sample reduces to a rank-one correction of a shared matrix, allowing exact empirical FastDV LOO updates while reducing repeated computation and scaling efficiently to large datasets. Meanwhile, partition composability of the above formulation allows client-side statistics to be aggregated under shared anchors. Their aggregation recovers the centralized FastDV statistic in the common representation space, yielding globally comparable sample rankings without pooling raw client data. Our main contributions are summarized as follows:

- We introduce an additive-statistic formulation to characterize model-agnostic finite-deletion valuation through sample-wise additive target statistics, revealing deletion decomposability and partition composability under a unified view.
- Guided by this formulation, FastDV instantiates a Pearson-divergence-based density-ratio utility for explicit LOO valuation and achieves efficient valuation through exact rank-one updates.
- The partition composability property further extends FastDV to multi-party global valuation. Shared-anchor aggregation exactly recovers the centralized FastDV statistic in a common representation space, making sample values globally comparable without pooling raw client data.
- Extensive experiments across sample removal, label-free data selection, large-scale valuation, data procurement, and federated learning demonstrate the effectiveness and scalability of FastDV in both centralized and distributed settings.

Related Work. Our work is primarily related to model-agnostic data valuation (Just et al., 2023; Kessler et al., 2025; Zhu et al., 2025), which evaluates data through distributional discrepancies rather than a specific downstream model (Ghorbani & Zou, 2019; Koh & Liang, 2017; Wang & Jia, 2023). FastDV differs by combining explicit finite-deletion valuation with an additive empirical realization that enables exact low-rank updates and partition-additive global valuation. It also differs from multi-center dataset valuation (Tay et al., 2022; Li et al., 2025a), which mainly values entire datasets or clients rather than individual samples. Further discussion is provided in Appendix A.1.

2 PRELIMINARIES

In this paper, $\mathcal{D}$ and $\mathcal{T}$ denote finite datasets, while $P_{\mathcal{D}}$ and $P_{\mathcal{T}}$ denote their associated data distributions, respectively. Let $\mathcal{D} = \{z_i\}_{i=1}^{N}$ and $\mathcal{T} = \{\tilde{z}_j\}_{j=1}^{M}$ denote the target set to be valued and the support set serving as a reference standard, respectively. We use $z = (x, y)$ for task-aware valuation and $z = x$ for label-free valuation.

Definition 2.1 (Data Utility)**.** A data utility $\mathrm{U}(\mathcal{D}, \mathcal{T}) \in \mathbb{R}$ measures the quality of the finite target set $\mathcal{D}$ with respect to the support set $\mathcal{T}$, where a larger value indicates higher utility.

Definition 2.2 (LOO Score)**.** For a sample $z_i \in \mathcal{D}$, its Leave-One-Out (LOO) value is defined as

$$\mathcal{V}(z_i) := \mathrm{U}(\mathcal{D}, \mathcal{T}) - \mathrm{U}(\mathcal{D}_{-i}, \mathcal{T}), \tag{1}$$

where $\mathcal{D}_{-i} := \mathcal{D} \setminus \{z_i\}$ denotes the target set after removing z_i.

Different choices of U induce different LOO valuation principles. Model-dependent methods define utility through downstream predictive performance and therefore depend on a particular model, whereas model-agnostic methods construct utility from distributional discrepancies between the target and support sets (Jiang et al., 2023a; Just et al., 2023; Zhu et al., 2025).

Pearson Divergence. Assuming $P_{\mathcal{T}} \ll P_{\mathcal{D}}$, let $r^*(z) := \frac{\mathrm{d}P_{\mathcal{T}}}{\mathrm{d}P_{\mathcal{D}}}(z)$ denote the density ratio between the support and target distributions. When both distributions admit densities, this reduces to $r^*(z) = p_{\mathcal{T}}(z)/p_{\mathcal{D}}(z)$. The Pearson divergence of $P_{\mathcal{T}}$ relative to $P_{\mathcal{D}}$ is defined as

$$D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{D}}) = \frac{1}{2}\mathbb{E}_{z\sim P_{\mathcal{D}}}\left[(r^*(z) - 1)^2\right]. \tag{2}$$

It is non-negative and equals zero if $P_{\mathcal{T}} = P_{\mathcal{D}}$ almost everywhere. We use this population discrepancy to construct the density-ratio valuation principle in Section 3.1.

3 METHOD

3.1 DENSITY-RATIO LEAVE-ONE-OUT VALUATION

By decoupling valuation from a particular downstream model, recent model-agnostic data valuation methods improve the flexibility of sample valuation through distributional discrepancies between the target and support sets (Just et al., 2023; Kessler et al., 2025; Zhu et al., 2025). However, to make sample-level valuation computationally tractable, existing approaches typically rely on sensitivity, influence, or hierarchical approximations rather than explicitly evaluating the finite utility change induced by sample deletion. This motivates us to retain the model-agnostic distributional paradigm while taking finite deletion itself as the starting point of valuation design.

To instantiate this principle, the utility should capture a meaningful notion of alignment between the target and support distributions while retaining a connection to downstream prediction. Therefore, we adopt the Pearson divergence $D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{D}})$, whose risk-based interpretation is formalized below:

Proposition 3.1 (Risk Interpretation of Pearson Divergence)**.** *Let $P_{\mathcal{Q}}$ denote an arbitrary candidate distribution satisfying $P_{\mathcal{T}} \ll P_{\mathcal{Q}}$ and $D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}}) < \infty$. Assume that the loss is bounded as $0 \le \ell(f; z) \le L$. Let $f_{\mathcal{Q}} \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{Q}}(f)$ and $f_{\mathcal{T}} \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f)$, where $\mathcal{R}_{\mathcal{Q}}(f) := \mathbb{E}_{z\sim P_{\mathcal{Q}}}[\ell(f; z)]$ and $\mathcal{R}_{\mathcal{T}}(f) := \mathbb{E}_{z\sim P_{\mathcal{T}}}[\ell(f; z)]$. Then,*

$$\mathcal{R}_{\mathcal{T}}(f_{\mathcal{Q}}) - \mathcal{R}_{\mathcal{T}}(f_{\mathcal{T}}) \le 2L\sqrt{2D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}})}. \tag{3}$$

In particular, the result applies to $P_{\mathcal{Q}} = P_{\mathcal{D}}$.

The proof is provided in Appendix A.2.1. Proposition 3.1 shows that Pearson divergence controls the support set excess risk, providing a model-independent measure of alignment with the support distribution. Motivated by this interpretation, we can define the population density-ratio utility as

$$\mathrm{U}_{\mathrm{DR}}^{\mathrm{pop}}(P_{\mathcal{Q}}, P_{\mathcal{T}}) := -D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}}). \tag{4}$$

For finite datasets, let $\widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}, \mathcal{T})$ denote an empirical surrogate of $\mathrm{U}_{\mathrm{DR}}^{\mathrm{pop}}(P_{\mathcal{D}}, P_{\mathcal{T}})$. The valuation principle itself is independent of the particular density-ratio estimator used for this realization. Accordingly, the density-ratio LOO value is defined through the finite change of this empirical utility.

Definition 3.1 (Density-Ratio LOO Value)**.** For $z_i \in \mathcal{D}$, its density-ratio LOO value is defined as

$$\mathcal{V}_{\mathrm{DR}}(z_i) := \widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}, \mathcal{T}) - \widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}_{-i}, \mathcal{T}). \tag{5}$$

A positive value indicates that removing z_i decreases the empirical density-ratio utility.

3.2 An Additive-Statistic Formulation for Finite Deletion Valuation

Definition 3.1 specifies the finite-deletion objective but not how to evaluate it efficiently. Ideally, the empirical utility should separate shared target set dependence from the sample-specific deletion effect, allowing computation to be reused across all LOO candidates. Therefore, we realize this property through a fixed-dimensional additive statistic, so that each deletion differs only by removing one sample contribution. This motivates the following additive-statistic formulation.

Definition 3.2 (Additive-Statistic Valuation)**.** Let ψ denote a representation fixed throughout valuation, and let $\mathbf{S}_\psi(z)$ be a fixed-dimensional sample-wise statistic induced by this representation. An empirical data utility is additive-statistic realizable if it can be written as

$$\widehat{\mathrm{U}}(\mathcal{D}, \mathcal{T}) = F_N(\mathbf{H}_{\mathcal{D}}, \mathbf{h}_{\mathcal{T}}), \qquad \mathbf{H}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{S}_\psi(z_i), \tag{6}$$

where $\mathbf{h}_{\mathcal{T}}$ summarizes the support set and remains fixed across candidate deletions, and F_N denotes a possibly sample-size-dependent utility functional.

Remark 3.1. Definition 3.2 exposes two structural properties of the additive target statistic: *deletion decomposability* and *partition composability*. (i) For a finite deletion $\mathcal{D}_{-i} = \mathcal{D} \setminus \{z_i\}$, we obtain $\mathbf{H}_{\mathcal{D}_{-i}} = \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} - \frac{1}{N-1}\mathbf{S}_\psi(z_i)$. Thus, all candidate deletions share the same rescaled target statistic and differ only through a single sample-specific correction. (ii) Suppose that $\mathcal{D}_g = \bigcup_{l=1}^{L} \mathcal{D}^{(l)}$ is partitioned across L data holders, with $N_g = \sum_l N_l$. If all parties use the same fixed representation ψ, then $\mathbf{H}_{\mathcal{D}_g} = \frac{1}{N_g}\sum_{l=1}^{L} N_l \mathbf{H}_{\mathcal{D}^{(l)}}$. Hence, the statistic obtained from the union of all samples can be reconstructed exactly from partition-level statistics without pooling the raw data.

3.3 Additive-Statistic Realization of Density-Ratio Utility

Having defined the density-ratio finite-deletion objective and the additive-statistic principle, we construct an empirical realization. Following uLSIF (Kanamori et al., 2009), for the density ratio $r_{\mathcal{D}}(z) = \mathrm{d}P_{\mathcal{T}}/\mathrm{d}P_{\mathcal{D}}$, Pearson divergence is associated with the population least-squares objective

$$\mathcal{J}_{\mathcal{D}}(r) = \frac{1}{2}\mathbb{E}_{z\sim P_{\mathcal{D}}}\left[r(z)^2\right] - \mathbb{E}_{z\sim P_{\mathcal{T}}}\left[r(z)\right], \tag{7}$$

which satisfies $D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{D}}) = -\mathcal{J}_{\mathcal{D}}(r_{\mathcal{D}}) - \frac{1}{2}$ at the true density ratio.

We use a fixed kernel dictionary $\mathcal{B} = \{b_j\}_{j=1}^{B}$ and parameterize $r_{\boldsymbol{\mu}}(z) = \boldsymbol{\mu}^\top \boldsymbol{\phi}(z)$, where $\boldsymbol{\phi}(z) = [\kappa_z(z, b_1), \ldots, \kappa_z(z, b_B)]^\top$. With ridge regularization $\lambda > 0$, the empirical least-squares objective is

$$\widehat{\mathcal{J}}_{\mathcal{D}}(\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}^\top(\mathbf{H}_{\mathcal{D}} + \lambda\mathbf{I})\boldsymbol{\mu} - \mathbf{h}^\top\boldsymbol{\mu}, \tag{8}$$

where $\mathbf{H}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}(z_i)\boldsymbol{\phi}(z_i)^\top \in \mathbb{R}^{B\times B}$ and $\mathbf{h} = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{\phi}(\tilde{z}_j) \in \mathbb{R}^B$.

Comparing this objective with Definition 3.2, the objective is additive-statistic realizable with $\mathbf{S}_\psi(z) = \boldsymbol{\phi}(z)\boldsymbol{\phi}(z)^\top$, which is a rank-one outer-product statistic. Therefore, all target set dependence is summarized by $\mathbf{H}_{\mathcal{D}}$, while the support set enters through the shared statistic $\mathbf{h}$.

Let $\mathbf{G}_{\mathcal{D}} = \mathbf{H}_{\mathcal{D}} + \lambda \mathbf{I}$. Since $\lambda > 0$ and $\mathbf{H}_{\mathcal{D}} \succeq 0$, $\mathbf{G}_{\mathcal{D}}$ is positive definite. Therefore, the empirical objective admits the closed-form optimum $\boldsymbol{\mu}^*_{\mathcal{D}} = \mathbf{G}_{\mathcal{D}}^{-1}\mathbf{h}$ and $\widehat{\mathcal{J}}^*_{\mathcal{D}} = -\frac{1}{2}\mathbf{h}^\top \mathbf{G}_{\mathcal{D}}^{-1}\mathbf{h}$. Rather than identifying the regularized empirical optimum directly with the Pearson divergence, we define the corresponding regularized empirical density-ratio utility as

$$\widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}, \mathcal{T}) := \widehat{\mathcal{J}}^*_{\mathcal{D}} + \frac{1}{2} = -\frac{1}{2}\mathbf{h}^\top \mathbf{G}_{\mathcal{D}}^{-1}\mathbf{h} + \frac{1}{2}. \tag{9}$$

Substituting this empirical utility into Definition 3.1 gives the FastDV score

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}, \mathcal{T}) - \widehat{\mathrm{U}}_{\mathrm{DR}}(\mathcal{D}_{-i}, \mathcal{T}) = \frac{1}{2}\mathbf{h}^\top \left(\mathbf{G}_{\mathcal{D}_{-i}}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1}\right)\mathbf{h}. \tag{10}$$

3.4 Exact LOO via Low-Rank Statistic Updates

Although Eq. (10) gives the empirical FastDV finite-deletion score, it requires inverting $\mathbf{G}_{\mathcal{D}_{-i}}$ for every candidate deletion. Recall that the additive-statistic formulation isolates each sample deletion effect from the shared target statistic. In FastDV, it can avoid repeated computation by exploiting this formulation. With $\mathbf{S}_\psi(z_i) = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$ and $\boldsymbol{\phi}_i = \boldsymbol{\phi}(z_i)$, the deletion identity in Remark 3.1 gives

$$\mathbf{G}_{\mathcal{D}_{-i}} = \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} + \lambda \mathbf{I} - \frac{1}{N-1}\boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top. \tag{11}$$

Define the shared base matrix $\mathbf{A} = \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} + \lambda \mathbf{I}$ and $\bar{\boldsymbol{\phi}}_i = \frac{\boldsymbol{\phi}_i}{\sqrt{N-1}}$. Then, $\mathbf{G}_{\mathcal{D}_{-i}} = \mathbf{A} - \bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top$. This decomposition shows that all candidate deletions share the same base matrix $\mathbf{A}$, while their sample-specific dependence is isolated in the rank-one correction $\bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top$. Therefore, applying the Sherman–Morrison identity (Niño et al., 2012) yields an exact inverse update without independently solving the density-ratio estimation problem for each deletion, leading to the following result.

Theorem 3.1 (Exact Low-Rank Update of the Empirical FastDV Score)**.** *Assume that the kernel dictionary $\mathcal{B}$, kernel hyperparameters, and regularization coefficient $\lambda > 0$ remain fixed throughout valuation. Then the empirical FastDV score in Eq. (10) can be computed exactly as*

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \frac{1}{2}\mathbf{h}^\top \left(\mathbf{A}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1}\right)\mathbf{h} + \frac{1}{2}\frac{\left(\bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\mathbf{h}\right)^2}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i}. \tag{12}$$

The derivation and validity of the Sherman–Morrison update are provided in Appendix A.2.2.

Corollary 3.1 (Ranking-Equivalent FastDV Score)**.** *The first term in Eq. (12) is independent of z_i. Therefore, ranking samples by $\mathcal{V}_{\mathrm{FastDV}}(z_i)$ is equivalent to ranking them by*

$$\widetilde{\mathcal{V}}_{\mathrm{FastDV}}(z_i) = \frac{\left(\bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\mathbf{h}\right)^2}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i}. \tag{13}$$

With a fixed dictionary size B, constructing $\mathbf{H}_{\mathcal{D}}$ costs $\mathcal{O}(NB^2)$, computing the shared inverse $\mathbf{A}^{-1}$ costs $\mathcal{O}(B^3)$, and evaluating each sample-specific score costs $\mathcal{O}(B^2)$. The overall complexity is therefore $\mathcal{O}(B^3 + NB^2)$, which is linear in N for a fixed dictionary size B. Importantly, this efficiency is achieved without approximating the finite-deletion effect, since all candidate deletions reuse the same shared inverse and differ only through rank-one sample corrections.

3.5 Global Valuation via Partition Composability

The above additive structure also provides a natural way to establish globally comparable valuation across distributed data. Consider L clients, where client l holds $\mathcal{D}^{(l)} = \{z_i^{(l)}\}_{i=1}^{N_l}$ and raw samples remain local. Independent local valuation produces scores conditioned on different target statistics and hence not directly comparable across clients. FastDV addresses this issue by placing all client statistics in a shared kernel representation defined by the same dictionary $\mathcal{B}$ and kernel hyperparameters.

Let $N_g = \sum_{l=1}^{L} N_l$ and $\mathcal{D}_g = \bigcup_{l=1}^{L} \mathcal{D}^{(l)}$. Client l computes its local FastDV statistic as $\mathbf{H}^{(l)} = \frac{1}{N_l}\sum_{i=1}^{N_l} \boldsymbol{\phi}(z_i^{(l)})\boldsymbol{\phi}(z_i^{(l)})^\top$. Applying the partition composability identity gives the global statistic

$$\mathbf{H}_g = \frac{1}{N_g}\sum_{l=1}^{L} N_l\, \mathbf{H}^{(l)}. \tag{14}$$

Corollary 3.2 (Centralized Equivalence from Partition Additivity)**.** *Under the shared kernel representation, the aggregated statistic in Eq. (14) satisfies*

$$\mathbf{H}_g = \frac{1}{N_g} \sum_{z_i \in \mathcal{D}_g} \boldsymbol{\phi}(z_i)\boldsymbol{\phi}(z_i)^\top = \mathbf{H}_{\mathcal{D}_g}. \tag{15}$$

Consequently, given the same support statistic $\mathbf{h}$*, regularization coefficient* λ*, and total sample size* N_g*, the empirical FastDV score constructed from* $\mathbf{H}_g$ *is identical to that obtained by applying centralized FastDV to* $\mathcal{D}_g$ *under the same shared representation.*

The proof follows directly from the partition-additive statistic and is provided in Appendix A.2.3. Thus, multi-party FastDV reuses the centralized ranking rule in Corollary 3.1 with $\mathbf{A}_g = \frac{N_g}{N_g - 1}\mathbf{H}_g + \lambda \mathbf{I}$ and $\bar{\boldsymbol{\phi}}_i^{(l)} = \boldsymbol{\phi}(z_i^{(l)})/\sqrt{N_g - 1}$. Since all clients share $\mathbf{H}_g$ and the server-side support statistic $\mathbf{h}$, the resulting sample scores are directly comparable across clients without pooling raw data.

Communication Protocol. The server broadcasts the shared dictionary and $\mathbf{h}$, clients upload $(\mathbf{H}^{(l)}, N_l)$, and the server returns $(\mathbf{A}_g^{-1}, N_g)$ after aggregation. Each client then evaluates its samples locally using Corollary 3.1. Therefore, FastDV requires only two communication rounds to obtain globally comparable sample values. Detailed pseudocode is provided in Appendix A.3.

3.6 Practical Instantiations

Anchor Construction. The FastDV instantiation of the additive-statistic formulation requires the kernel representation to remain fixed throughout valuation. Accordingly, FastDV constructs a dictionary $\mathcal{B} = \{b_j\}_{j=1}^{B}$ before valuation from the support set and keeps it fixed across all candidate deletions. The anchors can be constructed by representative sampling, clustering (Yuan et al., 2026), or dataset condensation (Yan et al., 2025). In the multi-party setting, the server constructs a single shared dictionary that defines a common feature map for all clients. In our experiments, the server constructs synthetic anchors by clustering, which provides a common representation without distributing raw support samples. Once the dictionary is fixed, the low-rank update remains exact with respect to the resulting empirical FastDV objective.

Task-Aware and Label-Free Valuation. FastDV supports both task-aware and label-free valuation through the choice of sample representation and kernel, with κ_x instantiated as a Gaussian RBF kernel. For task-aware valuation, $z = (x, y)$ and $\kappa_z(z_i, z_j) = \kappa_x(x_i, x_j)\kappa_y(y_i, y_j)$, where $\kappa_y(y_i, y_j) = \mathbb{I}[y_i = y_j]$ for classification. For label-free valuation, $z = x$ and $\kappa_z = \kappa_x$. Both settings preserve the same additive-statistic structure and FastDV computations.

4 Experiment Results and Analysis

4.1 Experimental Settings

Datasets. To comprehensively evaluate the valuation performance of FastDV, four representative datasets (Cifar10 (Krizhevsky et al., 2009), Products (Hu et al., 2020), Arxiv (Zhang et al., 2025), and Wiki (Mernyei & Cangea, 2020)) from diverse domains are adopted in experiments. Among them, the Products dataset, with a scale of 2 million, is used to validate valuation efficiency, while the Arxiv and Wiki datasets are simulated as data marketplaces and divided into three clients and one server.

Baselines. In this work, FastDV is compared with four representative DV methods, i.e., model-based influence function (InFun) (Koh & Liang, 2017), DVRL (Yoon et al., 2020), BetaShapley (BetaS.) (Kwon & Zou, 2022), and DataBanzhaf (DataB.) (Wang & Jia, 2023) and three model-agnostic valuation algorithms (LAVA (Just et al., 2023), SAVA (Kessler et al., 2025), Kairos (Zhu et al., 2025)). The hyperparameters of baselines are set according to the original paper if available. Otherwise, we adopt the default setting from OpenDataVal (Jiang et al., 2023a). Furthermore, to simulate federated training, we additionally employ four mainstream federated learning (FedAVG (McMahan et al., 2017), FedProx (Li et al., 2020), Scaffold (Karimireddy et al., 2020), and FedDC (Gao et al., 2022)).

Implementations. We evaluate all methods in two key scenarios. (1) Traditional valuation, encompassing sample removal, valuation efficiency, and data selection. (2) Valuation in data marketplaces, including data procurement and federated learning. Data procurement aims to centralize training

by procuring high-value candidate sets from clients, while the latter examines global valuation performance with multi-client federated training. All experiments report the average accuracy (%) and standard deviation across 10 independent trials. Darker shades indicate the best results, whereas lighter shades denote the second-best. Detailed data statistics, baseline descriptions, implementation details, and task settings are provided in Appendix A.4.

4.2 Experimental Results

Sample Removal. An effective DV method should demonstrate a performance degradation for the trained model upon the removal of high-value samples. Conversely, the removal of low-value samples should lead to an improvement in model performance. Table 1 reports high-value sample removal results. FastDV consistently causes the largest performance degradation on both datasets as the removal ratio increases. In particular, removing 40% of FastDV-selected samples reduces accuracy to 42.21% on Cifar10 and 14.63% on Products, outperforming the strongest competing methods. These results indicate that FastDV effectively identifies samples that are important for downstream learning.

Table 1: The high-value sample removal results. Lower performance indicates better valuation.

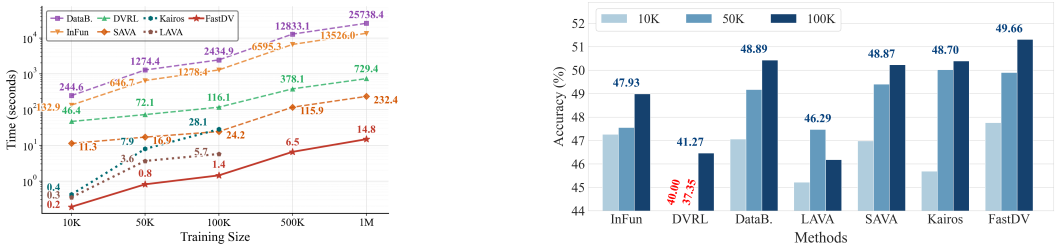
	Cifar10				Products			
	10%	20%	30%	40%	10%	20%	30%	40%
InFun	63.61±0.47	61.09±0.24	57.41±0.74	48.25±0.44	35.55±2.04	32.10±0.78	23.77±0.38	22.76±0.27
DVRL	64.96±0.50	59.21±0.45	56.16±0.49	50.96±0.31	33.82±1.59	31.04±1.61	31.97±0.62	22.12±1.25
BetaS.	65.22±0.58	64.69±0.47	62.88±0.41	60.51±0.37	33.74±0.94	31.88±1.29	30.07±1.58	26.33±1.03
DataB.	65.33±0.84	64.96±0.40	62.37±0.65	61.72±0.73	36.40±1.59	33.01±2.33	33.05±0.74	32.68±1.90
LAVA	65.60±0.37	63.82±0.37	60.73±0.38	58.13±0.62	34.73±1.70	34.02±1.22	31.96±2.69	32.02±2.14
SAVA	66.71±0.28	65.89±0.60	62.92±0.23	62.46±0.48	36.17±1.47	36.06±1.44	35.35±1.63	31.04±1.25
Kairos	63.75±0.56	59.71±0.32	52.38±0.79	44.12±1.05	35.00±1.10	31.56±1.68	23.10±2.19	16.87±1.13
FastDV	62.38±0.99	55.79±0.90	48.51±0.71	42.21±0.20	30.82±1.64	24.16±0.43	21.92±0.37	14.63±0.22

Table 2 further reports the results of low-value sample removal, where higher performance indicates better valuation quality. Compared with training on the whole dataset, removing low-value samples selected by FastDV consistently improves model performance on both datasets. Specifically, FastDV improves the accuracy from 66.87% to 67.99% and 69.17% on Cifar10 under 10% and 20% removal ratios, respectively. On Products, FastDV also increases the performance from 36.10% to 37.87% and 37.33%. Although FastDV does not achieve the best result in every setting, it remains highly competitive. These results suggest that FastDV can effectively identify low-value samples, such as redundant or potentially harmful data, whose removal contributes to better downstream generalization.

Table 2: The low-value sample removal experiment. Higher performance indicates better valuation.

	Cifar10		Products	
	10%	20%	10%	20%
Whole	*66.87±0.30*	*66.87±0.30*	*36.10±2.03*	*36.10±2.03*
InFun	65.75±0.68	62.29±0.53	38.09±1.48	36.93±1.48
DVRL	66.50±0.52	62.81±1.60	35.71±1.36	35.32±1.32
BetaS.	66.28±0.18	65.45±0.58	37.75±1.41	36.46±1.53
DataB.	66.30±0.31	66.20±0.26	37.03±1.60	36.10±2.31
LAVA	66.85±0.27	66.92±0.26	37.29±2.00	36.55±0.49
SAVA	66.68±0.29	66.17±0.43	36.29±1.54	36.85±0.66
Kairos	67.66±0.21	69.42±0.17	35.98±1.20	36.92±1.07
FastDV	67.99±1.11	69.17±0.28	37.87±1.67	37.33±1.26

Runtime and Scalability. To intuitively demonstrate the efficiency of DV methods, Fig. 2a compares the runtime of different DV methods on the Products dataset as the target-set size increases from 10K to 1M samples. BetaShapley is omitted because it does not finish within 24 hours, while LAVA and Kairos encounter out-of-memory errors beyond 100K samples. FastDV uses a fixed budget of 1,000 anchors, with anchor construction time included in the reported runtime. As the dataset size increases, FastDV exhibits substantially slower runtime growth than the competing methods. Compared to the fastest completed baseline SAVA, FastDV achieves a 15.7× speedup on the 1M-scale dataset. These results empirically support the scalability of the rank-one update and fixed-anchor representation. Fig. 2b further examines the efficiency–effectiveness trade-off introduced by the anchor representation. Using only 1,000 anchors, FastDV achieves the best performance on the 10K and 100K settings and remains competitive on 50K samples in the 40% low-value removal task. This indicates that a moderate, fixed anchor budget can retain strong valuation quality while substantially reducing computational cost.



(a) Runtime comparison on Products.

(b) Low-value sample removal under different scales.

Figure 2: Efficiency and scalability analysis of FastDV. (a) Runtime comparison across different data scales. (b) Valuation performance with 1,000 representative anchors under different data scales.

Label-Free Data Selection. Since most data valuation methods are model- and task-driven, their heavy reliance on label information limits their applicability in label-scarce scenarios. FastDV naturally supports label-free valuation by setting $z = x$ and estimating the feature-level density ratio. We compare FastDV with the model-agnostic baselines LAVA, SAVA, and Kairos, with all methods performing valuation without access to target-set labels. After ranking the unlabeled samples, the selected high-value samples are labeled and used for downstream training. As shown in Fig. 3, FastDV achieves the best performance across all labeling ratios on Cifar10, with a larger advantage at higher labeling budgets. These results indicate that FastDV can identify informative samples from feature distributions alone, demonstrating its applicability to label-efficient data selection.

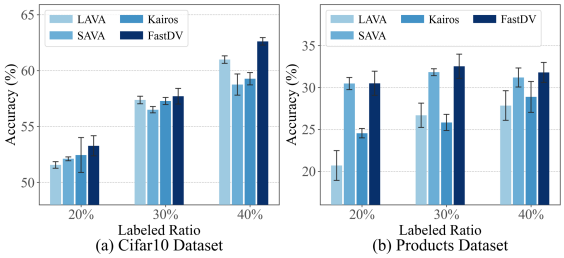


Figure 3: High-value data selection.

Data Procurement. The aim of data procurement is for the server to purchase high-value data from clients for centralized training. For existing valuation methods, independently computed client-side scores are not comparable. Therefore, we consider two post-hoc procurement strategies: *(S1) Uniform Procurement*, where each client contributes its local top-K% samples, and *(S2) Centralized Procurement*, where local scores are normalized, merged, and globally re-ranked before selecting the top-K% samples. In contrast, FastDV directly produces globally comparable scores through the shared-anchor protocol and therefore performs a single global procurement strategy by ranking all candidate samples in the centralized-equivalent valuation space. Table 3 illustrates the performance for centralized training using 60% of the high-value data. The two post-hoc strategies S1 and S2 are applied to the baseline methods, whereas FastDV directly performs global procurement using its globally comparable scores. These results further support the benefit of performing valuation in a common global space, particularly when samples from different data holders must be compared and selected jointly. Moreover, FastDV does not require pooling client data or distributing the complete support set to clients.

Table 3: Performance in data procurement.

	Arxiv		Wiki	
	S1	*S2*	*S1*	*S2*
InFun	60.77±2.23	64.95±1.66	70.30±0.94	72.36±0.56
DVRL	61.97±1.72	67.38±0.67	66.91±0.36	64.91±0.79
BetaS.	66.66±0.55	65.81±0.44	71.22±0.57	66.17±1.02
DataB.	66.70±0.49	66.83±0.38	71.51±0.20	70.66±0.31
LAVA	64.86±0.43	64.65±0.52	71.23±0.55	71.64±0.30
SAVA	66.78±0.53	66.54±0.36	71.42±0.38	71.22±0.52
Kairos	65.03±0.62	65.23±0.68	71.16±0.32	72.48±0.13
FastDV	68.09±0.39		72.61±0.19	

Federated Learning. Under this setup, the baseline methods select the top 60% of high-value data from each client for training after local valuation, whereas FastDV performs global valuation, jointly ranks samples across all clients, and selects the global top 60% samples. Model test and evaluation are performed on the server side. As shown in Table 4, where *Avg.* denotes the average across the four federated optimizers, FastDV achieves the best average performance on both datasets. Although Kairos performs better under Scaffold on Wiki, FastDV achieves the best performance under the other three federated optimizers. These results indicate that evaluating samples with globally aggregated statistics can provide more reliable data rankings for distributed training than purely local valuation.

Table 4: Performance of 40% low-value sample removal across federated learning methods.

		InFun	DVRL	BetaS.	DataB.	LAVA	SAVA	Kairos	FastDV
Arxiv	FedAVG	41.94±1.35	63.05±1.56	64.23±0.72	64.83±0.45	63.80±0.66	64.36±0.71	64.21±0.66	66.71±0.52
	FedProx	39.15±1.98	58.44±1.34	63.60±0.83	64.09±0.56	63.13±0.71	63.80±0.76	63.85±0.74	65.84±0.48
	Scaffold	41.45±2.03	60.17±1.54	59.82±3.23	59.29±2.23	59.76±2.39	60.92±1.93	60.95±1.66	61.20±1.71
	FedDC	42.41±1.03	61.85±1.62	63.35±1.17	62.25±0.80	62.30±1.11	63.56±1.17	62.59±1.02	63.93±1.05
	Avg.	*41.24*	*60.88*	*62.75*	*62.62*	*62.25*	*63.16*	*62.90*	*64.42*
Wiki	FedAVG	69.69±0.54	63.81±0.38	70.59±0.63	70.06±0.50	68.22±0.60	70.27±0.27	71.05±0.12	71.84±0.37
	FedProx	69.82±0.33	63.82±0.32	70.39±0.61	69.98±0.45	69.00±0.91	70.28±0.27	71.09±0.11	71.88±0.33
	Scaffold	56.46±0.42	54.16±0.77	60.81±0.31	61.93±0.32	59.76±0.63	60.20±0.53	62.00±0.38	60.73±0.34
	FedDC	40.47±0.48	46.94±0.35	41.64±1.14	48.25±1.24	48.59±2.41	47.73±2.22	48.40±1.46	48.76±1.10
	Avg.	*59.11*	*57.18*	*60.86*	*62.56*	*61.39*	*62.12*	*63.14*	*63.30*

Correlation with Model-based LOO. To empirically examine the relationship between FastDV and conventional model-based LOO, we conduct a small-scale analysis on 200 Cifar10 samples, where model-based LOO can be explicitly computed through repeated model retraining. We compare the scores produced by LAVA, SAVA, Kairos, and FastDV with model-based LOO using Spearman's rank correlation ρ. As shown in Fig. 4, FastDV achieves the strongest agreement with model-based LOO, with $\rho = 0.80$. The competing model-agnostic methods exhibit lower correlations, suggesting that their approximate valuation scores may deviate more from the finite-deletion ranking induced by model retraining. Although FastDV is defined through a distributional utility rather than model performance, this result provides empirical evidence that its density-ratio LOO scores can capture sample importance consistent with model-based LOO.

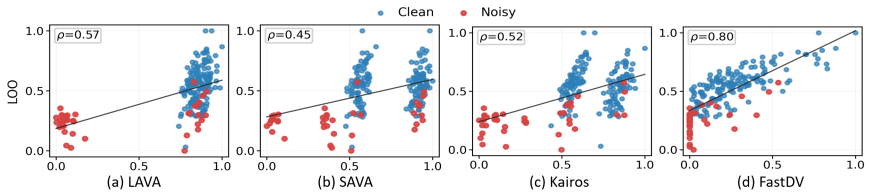


Figure 4: Correlation between model-agnostic valuation scores and model-based LOO scores.

Sensitivity Analysis. We further investigate the sensitivity of FastDV to the number of anchors on the 2-million-scale Products dataset. Fig. 5 reports both valuation time and performance under 40% low-value sample removal. As the number of anchors increases from 200 to 5,000, the accuracy improves from 51.42% to 52.43%, indicating that a larger anchor set provides a richer kernel representation and can improve valuation quality. Meanwhile, valuation time increases with the anchor size, revealing a clear trade-off between efficiency and performance. Conveniently, the calculation of $\mathbf{H}_{\mathcal{D}}$ is a one-time preprocessing cost, rendering the subsequent valuation stage linear in the number of samples. These results show that FastDV achieves strong valuation performance with a moderate number of anchors.

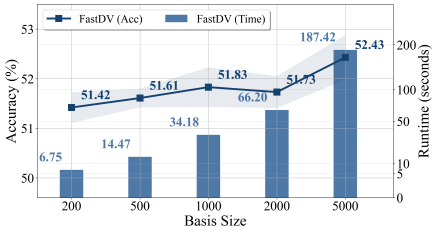


Figure 5: Performance of FastDV with different numbers of anchors.

5 Conclusion

In this work, we revisit model-agnostic data valuation from the perspective of finite deletion and show that scalability need not require approximating the deletion effect. FastDV achieves this by organizing target-set dependence into an additive-statistic principle, under which sample removal can be handled through exact low-rank updates. More importantly, its partition composability extends beyond centralized valuation, allowing distributed data holders to be evaluated in a common global space through shared aggregation. The resulting framework reconciles finite-deletion fidelity, computational scalability, and cross-client comparability within a unified valuation paradigm. Extensive experiments further validate its effectiveness in large-scale and multi-party settings.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and assistance with code implementation and debugging. All AI-assisted text and code were reviewed and revised by the authors, and the resulting implementations were tested and verified for correctness. We also independently verified the mathematical derivations, experimental results, and references. We take responsibility for the final content of this work, including text, code, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Hadi Askari, Shivanshu Gupta, Fei Wang, Anshuman Chhabra, and Muhao Chen. Layerif: Estimating layer quality for large language models using influence functions. *arXiv preprint arXiv:2505.23811*, 2025.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger B Grosse. Training data attribution via approximate unrolling. *Neural Information Processing Systems (NeurIPS)*, 37:66647–66686, 2024.

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion schrödinger bridge for density ratio estimation. *arXiv preprint arXiv:2505.05034*, 2025.

Wei Chen, Jiacheng Li, Shigui Li, Zhiqi Lin, Junmei Yang, John Paisley, and Delu Zeng. A minimum variance path principle for accurate and stable score-based density ratio estimation. In *International Conference on Learning Representations (ICLR)*, 2026.

Kristy Choi, Madeline Liao, and Stefano Ermon. Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*, pp. 172–182, 2021.

Zicun Cong, Xuan Luo, Jian Pei, Feida Zhu, and Yong Zhang. Data pricing in machine learning pipelines. *Knowl. Inf. Syst.*, 64(6):1417–1455, 2022.

R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2022.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10112–10121, 2022.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, pp. 2242–2251, 2019.

Lukas Gruber, Markus Holzleitner, Johannes Lehner, Sepp Hochreiter, and Werner Zellinger. Overcoming saturation in density ratio estimation by iterated regularization. In *International Conference on Machine Learning (ICML)*, pp. 16502–16529, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Neural Information Processing Systems (NeurIPS)*, 33:22118–22133, 2020.

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 *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 1167–1176, 2019.

Kevin Jiang, Weixin Liang, James Y Zou, and Yongchan Kwon. Opendataval: a unified benchmark for data valuation. *Neural Information Processing Systems (NeurIPS)*, 36:28624–28647, 2023a.

Yifeng Jiang, Weiwen Zhang, and Yanxi Chen. Data quality detection mechanism against label flipping attacks in federated learning. *IEEE Transactions on Information Forensics and Security*, 18:1625–1637, 2023b.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. In *International Conference on Learning Representations (ICLR)*, 2023.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, pp. 5132–5143, 2020.

Samuel Kessler, Tam Le, and Vu Nguyen. Sava: Scalable learning-agnostic data valuation. In *International Conference on Learning Representations (ICLR)*, 2025.

SungYub Kim, Kyungsu Kim, and Eunho Yang. Gex: A flexible method for approximating influence via geometric ensemble. *Neural Information Processing Systems (NeurIPS)*, 36:5888–5911, 2023.

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

Narine Kokhlikyan, Kamalika Chaudhuri, and Saeed Mahloujifar. Z0-inf: Zeroth order approximation for data influence. *arXiv preprint arXiv:2510.11832*, 2025.

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

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 8780–8802, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Machine Learning and Systems (MLSys)*, 2: 429–450, 2020.

Wenqian Li, Youjia Yang, Ruoxi Jia, and Yan Pang. Data valuation and selection in a federated model marketplace. *arXiv preprint arXiv:2509.18104*, 2025a.

Xunkai Li, Yinlin Zhu, Boyang Pang, Guochen Yan, Yeyu Yan, Zening Li, Zhengyu Wu, Wentao Zhang, Ronghua Li, and Guoren Wang. Openfgl: A comprehensive benchmark for federated graph learning. *Proc. VLDB Endow.*, 18(5):1305–1320, 2025b.

Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Zou. Accuracy on the curve: On the nonlinear correlation of ml performance between data subpopulations. In *International Conference on Machine Learning (ICML)*, pp. 20706–20724, 2023.

Song Liu, Akiko Takeda, Taiji Suzuki, and Kenji Fukumizu. Trimmed density ratio estimation. *Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Wenqin Liu, Weizhi Quan, Aoqi Zuo, Erdun Gao, Vu Nguyen, Dino Sejdinovic, Howard Bondell, and Mingming Gong. TimeLAVA: Learning-agnostic valuation for time series data. In *International Conference on Machine Learning (ICML)*, 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *ArXiv preprint arXiv:2007.02901*, 2020.

Elias D. Niño, Adrian Sandu, and Jeffrey L. Anderson. An effcient implementation of the ensemble kalman filter based on iterative sherman morrison formula. In *International Conference on Computational Science (ICCS)*, volume 9, pp. 1064–1072, 2012.

Yanzhou Pan, Huawei Lin, Yide Ran, Jiamin Chen, Xiaodong Yu, Weijie Zhao, Denghui Zhang, and Zhaozhuo Xu. Alinfik: Learning to approximate linearized future influence kernel for scalable third-parity llm data valuation. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 11756–11771, 2025.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Neural Information Processing Systems (NeurIPS)*, 33: 19920–19930, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3980–3990, 2019.

Andrea Schioppa, Katja Filippova, Ivan Titov, and Polina Zablotskaia. Theoretical and practical perspectives on what influence functions do. *Neural Information Processing Systems (NeurIPS)*, 36:27560–27581, 2023.

Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data valuation in machine learning: "ingredients", strategies, and open challenges. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5607–5614, 2022.

Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Incentivizing collaboration in machine learning via synthetic data rewards. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 9448–9456, 2022.

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.

Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 6388–6421, 2023.

Lauren Watson, Rayna Andreeva, Hao-Tsung Yang, and Rik Sarkar. Differentially private shapley values for data evaluation. *arXiv preprint arXiv:2206.00511*, 2022.

Xinyi Xu, Shuaiqi Wang, Chuan-Sheng Foo, Bryan K Low, and Giulia Fanti. Data distribution valuation. *Neural Information Processing Systems (NeurIPS)*, 37:2407–2448, 2024.

Yeyu Yan, Shuai Zheng, Wenjun Hui, Xiangkai Zhu, Dong Chen, Zhenfeng Zhu, Yao Zhao, and Kunlun He. Towards pre-trained graph condensation via optimal transport. In *Neural Information Processing Systems (NeurIPS)*, 2025.

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

Yuan Yuan, Lin Zhao, Shenfei Pei, and Feiping Nie. Efficient greedy optimization method for k-means. *Pattern Recognit.*, 171:112140, 2026.

Zhongjian Zhang, Xiao Wang, Huichi Zhou, Yue Yu, Mengmei Zhang, Cheng Yang, and Chuan Shi. Can large language models improve the adversarial robustness of graph neural networks? In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2008–2019, 2025.

Jie Zhao, Xinghua Zhu, Jianzong Wang, and Jing Xiao. Efficient client contribution evaluation for horizontal federated learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3060–3064, 2021.

Jiongli Zhu, Parjanya Prashant, Alex Cloninger, and Babak Salimi. KAIROS: scalable model-agnostic data valuation. In *Neural Information Processing Systems (NeurIPS)*, 2025.

A APPENDIX

This is the appendix of our work: **"FastDV: Scalable Model-Agnostic Data Valuation via Density Ratio Leave-One-Out"**. The appendix is organized as follows:

A.1 Related Work.

A.2 Detailed Proofs.

A.3 Distributed Valuation Protocol.

A.4 Implementation Details.

A.5 Additional Experiments.

A.6 Limitations and Future Work.

A.1 RELATED WORK

A.1.1 DATA VALUATION

Mainstream data valuation aims to quantify data quality by measuring the impact of individual samples on model performance. Among them, the Leave-One-Out (LOO) score is a representative method, defined as the change in performance when a single sample is removed. However, LOO is computationally prohibitive (Jiang et al., 2023a). To address this, Influence Functions (InFun) (Koh & Liang, 2017) utilize second-order derivatives to approximate the value of individual samples. While InFun significantly improves efficiency, the cost of computing second-order derivatives remains immense for deep neural networks (DNNs) (Zhu et al., 2025). Consequently, works such as TracIn (Pruthi et al., 2020) and Z0 (Kokhlikyan et al., 2025) further simplify InFun by bypassing the Hessian matrix calculation to accelerate the process. Nevertheless, Schioppa et al. (Schioppa et al., 2023) pointed out that InFun approximations in DNNs can deviate significantly from the exact LOO solutions. To mitigate this, Gex (Kim et al., 2023) employs a non-linear loss to calibrate these biases, while SOURCE (Bae et al., 2024) and LayerIF (Askari et al., 2025) adopt multi-stage or multi-layer IF computations to reduce errors arising from a single valuation. In contrast to these Influence-oriented methods, another line of research, such as DVRL (Yoon et al., 2020) and ALinFiK (Pan et al., 2025), seeks to simulate the overall data value distribution through proxy models or data subsets. However, the generalization capability of proxy models and the selection strategies for data subsets remain persistent challenges.

A.1.2 MODEL-AGNOSTIC DATA VALUATION

The heavy dependence on specific models motivates model-agnostic data valuation, which evaluates samples through distributional discrepancies rather than model-specific utilities (Just et al., 2023; Kessler et al., 2025; Zhu et al., 2025). LAVA (Just et al., 2023) uses optimal transport (OT) and estimates sample contributions through sensitivity analysis of the Wasserstein discrepancy, while SAVA (Kessler et al., 2025) improves scalability through hierarchical, batch-wise OT. Kairos (Zhu et al., 2025) instead adopts Maximum Mean Discrepancy (MMD) and derives an influence score that approximates the exact MMD-based LOO ranking with $O(1/N^2)$ error. Although these methods are motivated by finite-deletion distributional utilities, their practical scores rely on sensitivity, influence, or hierarchical approximations to avoid repeatedly recomputing the utility after each sample removal. These methods typically specify a distributional utility first and then approximate the effect of sample removal for computational tractability. FastDV instead takes finite deletion as the starting point and introduces an additive-statistic realization whose sample-wise structure makes deletion analytically decomposable. Moreover, because such distributional utilities generally differ from retraining-based model LOO, their relationship to downstream retrained performance has not been explicitly characterized. FastDV instead derives an explicit finite-deletion update of its empirical density-ratio utility and provides an indirect theoretical connection to the corresponding retrained predictor through Pearson divergence.

A.1.3 DATA VALUATION IN DATA MARKETPLACES

Conventional data valuation paradigms focus on a single-client setting where the support set is transparently accessible. However, such assumptions often fail in decentralized data marketplaces

(Tian et al., 2022). In these environments, data across different clients may suffer from redundancy and misaligned information (Xu et al., 2024), while the prohibition of data sharing further complicates the valuation process (Li et al., 2025a). To address this, F-RCCE (Zhao et al., 2021) leverages reinforcement learning to evaluate client value based on selection probabilities during multi-client training. CGM (Tay et al., 2022) and CombineWad (Li et al., 2025a) adopt model-agnostic approaches, utilizing distributional distances to measure inter-client disparities. Despite these extensive efforts, these methods are primarily restricted to dataset valuation of the entire clients rather than valuing individual samples.

A.1.4 Density Ratio Estimation

Density ratio estimation (DRE) is a fundamental statistical tool for handling distribution shifts. The field was significantly advanced by the unconstrained Least-Squares Importance Fitting (uLSIF) (Kanamori et al., 2009), which enables direct ratio estimation without separate density modeling. Recent advancements in DRE have addressed a variety of distributional challenges. For example, Trimmed DRE (Liu et al., 2017) was proposed to identify and mitigate the impact of outliers in unbounded ratio estimation. Addressing data heterogeneity, Duan et al. (Duan et al., 2022) implemented a communication-efficient distributed inference framework that employs density ratio tilting to approximate global scores. To adapt to high-dimensional and complex distributions, Choi et al. (Choi et al., 2021) introduced featurized DRE, leveraging invertible generative models to map distributions into a latent space for robust matching. Furthermore, recent approaches (Gruber et al., 2024; Chen et al., 2025; 2026) have utilized iterated regularization and dequantified diffusion-Schrödinger bridges to enhance estimation stability. While these estimators based on deep learning (Choi et al., 2021; Chen et al., 2025) and iterated regularization (Gruber et al., 2024) offer superior expressiveness and accuracy, they generally necessitate computationally expensive training. Consequently, we adopt uLSIF not merely for computational efficiency, but because its least-squares formulation yields an analytical solution with sample-wise additive statistics, which is essential for the exact low-rank updates in FastDV.

Notably, uLSIF also provides an analytical leave-one-out cross-validation (LOOCV) procedure (Kanamori et al., 2009). Despite the shared leave-one-out terminology, uLSIF LOOCV and FastDV compute fundamentally different quantities. In uLSIF, leave-one-out is an estimator-selection protocol to evaluate the generalization error of the density-ratio estimator, and the resulting held-out losses are averaged into a single cross-validation criterion for selecting hyperparameters. In FastDV, by contrast, the support set $\mathcal{T}$, kernel dictionary, and hyperparameters are fixed throughout valuation, and only a target sample z_i is removed. The deletion itself defines the sample value, yielding one valuation score for each target sample rather than a single cross-validation criterion. Thus, Pearson divergence defines the distributional valuation principle of FastDV, while uLSIF provides its regularized finite-sample surrogate for efficient computation. This valuation formulation further enables the sample-dependent deletion update to be expressed through a shared second-moment statistic, which in turn supports centralized-equivalent valuation across multiple clients through shared anchors.

A.2 Detailed Proofs

A.2.1 Proof of Proposition 3.1

Proposition 3.1 (Risk Interpretation of Pearson Divergence)**.** *Let $P_{\mathcal{Q}}$ denote an arbitrary candidate distribution satisfying $P_{\mathcal{T}} \ll P_{\mathcal{Q}}$ and $D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}}) < \infty$. Assume that the loss is bounded as $0 \le \ell(f; z) \le L$. Let $f_{\mathcal{Q}} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{Q}}(f)$ and $f_{\mathcal{T}} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{T}}(f)$, where $\mathcal{R}_{\mathcal{Q}}(f) := \mathbb{E}_{z \sim P_{\mathcal{Q}}}[\ell(f; z)]$ and $\mathcal{R}_{\mathcal{T}}(f) := \mathbb{E}_{z \sim P_{\mathcal{T}}}[\ell(f; z)]$. Then,*

$$\mathcal{R}_{\mathcal{T}}(f_{\mathcal{Q}}) - \mathcal{R}_{\mathcal{T}}(f_{\mathcal{T}}) \le 2L\sqrt{2 D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}})}. \tag{16}$$

In particular, the result applies to $P_{\mathcal{Q}} = P_{\mathcal{D}}$.

Proof. For any $f \in \mathcal{F}$, since $P_{\mathcal{T}} \ll P_{\mathcal{Q}}$, define

$$r_{\mathcal{Q}}(z) := \frac{\mathrm{d}P_{\mathcal{T}}}{\mathrm{d}P_{\mathcal{Q}}}(z).$$

Then,

$$\mathcal{R}_{\mathcal{T}}(f) - \mathcal{R}_{\mathcal{Q}}(f) = \mathbb{E}_{z \sim P_{\mathcal{Q}}}\left[(r_{\mathcal{Q}}(z) - 1)\, \ell(f; z)\right]. \tag{17}$$

By the Cauchy–Schwarz inequality and $0 \le \ell(f; z) \le L$,

$$|\mathcal{R}_{\mathcal{T}}(f) - \mathcal{R}_{\mathcal{Q}}(f)| \le L\sqrt{2D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}})}. \tag{18}$$

Using the optimality of $f_{\mathcal{Q}}$ under $P_{\mathcal{Q}}$,

$$\begin{aligned} &\mathcal{R}_{\mathcal{T}}(f_{\mathcal{Q}}) - \mathcal{R}_{\mathcal{T}}(f_{\mathcal{T}}) \\ &= [\mathcal{R}_{\mathcal{T}}(f_{\mathcal{Q}}) - \mathcal{R}_{\mathcal{Q}}(f_{\mathcal{Q}})] + [\mathcal{R}_{\mathcal{Q}}(f_{\mathcal{Q}}) - \mathcal{R}_{\mathcal{Q}}(f_{\mathcal{T}})] \\ &\quad + [\mathcal{R}_{\mathcal{Q}}(f_{\mathcal{T}}) - \mathcal{R}_{\mathcal{T}}(f_{\mathcal{T}})] \\ &\le 2L\sqrt{2D_{\mathrm{PE}}(P_{\mathcal{T}} \| P_{\mathcal{Q}})}, \end{aligned} \tag{19}$$

where $\mathcal{R}_{\mathcal{Q}}(f_{\mathcal{Q}}) \le \mathcal{R}_{\mathcal{Q}}(f_{\mathcal{T}})$. □

A.2.2 Proof of Theorem 3.1

Theorem 3.1 (Exact Low-Rank Update of the Empirical FastDV Score)**.** *Assume that the kernel dictionary $\mathcal{B}$, kernel hyperparameters, and regularization coefficient $\lambda > 0$ remain fixed throughout valuation. Then the empirical FastDV score in Eq. (10) can be computed exactly as*

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \frac{1}{2}\mathbf{h}^\top \left(\mathbf{A}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1}\right)\mathbf{h} + \frac{1}{2}\frac{\left(\bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\mathbf{h}\right)^2}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i}. \tag{20}$$

Proof. Since the kernel dictionary and kernel hyperparameters are fixed throughout valuation, the feature map $\boldsymbol{\phi}(\cdot)$ remains unchanged after deleting a target sample. Let $\boldsymbol{\phi}_i := \boldsymbol{\phi}(z_i)$. After removing z_i, the target statistic becomes

$$\mathbf{H}_{\mathcal{D}_{-i}} = \frac{1}{N-1}\sum_{t \ne i} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top = \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} - \frac{1}{N-1}\boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top. \tag{21}$$

Hence,

$$\mathbf{G}_{\mathcal{D}_{-i}} = \mathbf{H}_{\mathcal{D}_{-i}} + \lambda \mathbf{I} = \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} + \lambda \mathbf{I} - \frac{1}{N-1}\boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top. \tag{22}$$

Define

$$\mathbf{A} := \frac{N}{N-1}\mathbf{H}_{\mathcal{D}} + \lambda \mathbf{I}, \qquad \bar{\boldsymbol{\phi}}_i := \frac{\boldsymbol{\phi}_i}{\sqrt{N-1}}. \tag{23}$$

Then,

$$\mathbf{G}_{\mathcal{D}_{-i}} = \mathbf{A} - \bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top. \tag{24}$$

Since $\lambda > 0$ and $\mathbf{H}_{\mathcal{D}} \succeq 0$, $\mathbf{A}$ is positive definite. Likewise, $\mathbf{G}_{\mathcal{D}_{-i}} = \mathbf{H}_{\mathcal{D}_{-i}} + \lambda \mathbf{I}$ is positive definite. By the matrix determinant lemma,

$$\det(\mathbf{G}_{\mathcal{D}_{-i}}) = \det(\mathbf{A})\left(1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i\right). \tag{25}$$

Therefore,

$$1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i > 0. \tag{26}$$

Applying the Sherman–Morrison identity gives

$$\mathbf{G}_{\mathcal{D}_{-i}}^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i}. \tag{27}$$

Since the support set remains unchanged under target-sample deletion, the support statistic $\mathbf{h}$ is shared by $\mathcal{D}$ and $\mathcal{D}_{-i}$. From Eq. (10),

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \frac{1}{2}\mathbf{h}^\top \left(\mathbf{G}_{\mathcal{D}_{-i}}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1}\right)\mathbf{h}. \tag{28}$$

Substituting Eq. (27) yields

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \frac{1}{2}\mathbf{h}^\top \left(\mathbf{A}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1}\right)\mathbf{h} + \frac{1}{2}\frac{\mathbf{h}^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\mathbf{h}}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1}\bar{\boldsymbol{\phi}}_i}. \tag{29}$$

Since

$$\mathbf{h}^\top \mathbf{A}^{-1} \bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1} \mathbf{h} = \left(\bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1} \mathbf{h} \right)^2, \tag{30}$$

we obtain

$$\mathcal{V}_{\mathrm{FastDV}}(z_i) = \frac{1}{2} \mathbf{h}^\top \left(\mathbf{A}^{-1} - \mathbf{G}_{\mathcal{D}}^{-1} \right) \mathbf{h} + \frac{1}{2} \frac{\left(\bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1} \mathbf{h} \right)^2}{1 - \bar{\boldsymbol{\phi}}_i^\top \mathbf{A}^{-1} \bar{\boldsymbol{\phi}}_i}, \tag{31}$$

which is exactly the expression in Theorem 3.1. □

A.2.3 Proof of Corollary 3.2

Corollary 3.2 (Centralized Equivalence from Partition Additivity)**.** *Under the shared kernel representation, the aggregated statistic in Eq. (14) satisfies*

$$\mathbf{H}_g = \frac{1}{N_g} \sum_{z_i \in \mathcal{D}_g} \boldsymbol{\phi}(z_i) \boldsymbol{\phi}(z_i)^\top = \mathbf{H}_{\mathcal{D}_g}. \tag{32}$$

Consequently, given the same support statistic $\mathbf{h}$*, regularization coefficient* λ*, and total sample size* N_g*, the empirical FastDV score constructed from* $\mathbf{H}_g$ *is identical to that obtained by applying centralized FastDV to* $\mathcal{D}_g$ *under the same shared representation.*

Proof. By the definition of the local FastDV statistic,

$$N_l \mathbf{H}^{(l)} = \sum_{i=1}^{N_l} \boldsymbol{\phi}(z_i^{(l)}) \boldsymbol{\phi}(z_i^{(l)})^\top. \tag{33}$$

Substituting this into $\mathbf{H}_g = \frac{1}{N_g} \sum_{l=1}^{L} N_l \mathbf{H}^{(l)}$ gives

$$\begin{aligned} \mathbf{H}_g &= \frac{1}{N_g} \sum_{l=1}^{L} \sum_{i=1}^{N_l} \boldsymbol{\phi}(z_i^{(l)}) \boldsymbol{\phi}(z_i^{(l)})^\top \\ &= \frac{1}{N_g} \sum_{z_i \in \mathcal{D}_g} \boldsymbol{\phi}(z_i) \boldsymbol{\phi}(z_i)^\top = \mathbf{H}_{\mathcal{D}_g}. \end{aligned} \tag{34}$$

Since both distributed and centralized FastDV use the same shared representation, each sample has the same feature vector $\boldsymbol{\phi}(z_i)$ in both settings. Together with the same support statistic $\mathbf{h}$, regularization coefficient λ, and total sample size N_g, they therefore construct the same $\mathbf{G}_{\mathcal{D}_g}$ and shared base matrix $\mathbf{A}_g$. By Theorem 3.1, the resulting empirical FastDV score for every sample is identical in the two settings. □

A.3 Distributed Valuation Protocol

Alg. 1 summarizes the two-round shared-anchor protocol, where raw client data remain local and client-side additive statistics are aggregated for global valuation.

A.4 Implementation Details

A.4.1 Dataset Description

We evaluate FastDV on four datasets from diverse domains: Cifar10 (Krizhevsky et al., 2009), Products (Hu et al., 2020), Arxiv (Zhang et al., 2025), and Wiki (Mernyei & Cangea, 2020). Throughout this work, we refer to the dataset whose samples are valued as the target set (sourced from the data sellers, i.e., clients), and to the held-out dataset used as a reference during valuation as the support set (provided by the data purchaser, i.e., server). Therefore, the support set plays a role analogous to the validation set in conventional data valuation methods such as OpenDataVal (Jiang et al., 2023a). It participates in valuation but is never used as the final test set. Downstream performance is evaluated on a separate test set that remains unseen during both valuation and hyperparameter selection.

Cifar10 (Krizhevsky et al., 2009) is a 10-class image classification dataset. Following the experimental protocol of OpenDataVal (Jiang et al., 2023a), we randomly select 1,000 samples as the target set

Algorithm 1 Two-Round Shared-Anchor Valuation Protocol

Require: Client datasets $\{\mathcal{D}^{(l)}\}_{l=1}^{L}$ with $N_l = |\mathcal{D}^{(l)}|$; server support set $\mathcal{T} = \{\tilde{z}_j\}_{j=1}^{M}$; anchor size B; shared kernel hyperparameters; regularization coefficient $\lambda > 0$.
Ensure: Globally comparable ranking scores $\{\widetilde{\mathcal{V}}^{(l)}\}_{l=1}^{L}$.

/* *On the Server Side (Round 1)* */
Construct the shared anchor dictionary $\mathcal{B} = \{b_j\}_{j=1}^{B}$ from $\mathcal{T}$.
Compute the support statistic $\mathbf{h} \leftarrow \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{\phi}(\tilde{z}_j)$.
Broadcast $(\mathcal{B}, \mathbf{h})$ and shared kernel hyperparameters to all clients.
/* *On the Client Side (Round 1)* */
for $l = 1, \ldots, L$ **do**
    Compute $\mathbf{H}^{(l)} \leftarrow \frac{1}{N_l}\sum_{i=1}^{N_l} \boldsymbol{\phi}(z_i^{(l)})\boldsymbol{\phi}(z_i^{(l)})^\top$.
    Upload $(\mathbf{H}^{(l)}, N_l)$ to the server.
end for
/* *On the Server Side (Round 2)* */
$N_g \leftarrow \sum_{l=1}^{L} N_l$.
Aggregate $\mathbf{H}_g \leftarrow \frac{1}{N_g}\sum_{l=1}^{L} N_l \mathbf{H}^{(l)}$.
Compute $\mathbf{A}_g \leftarrow \frac{N_g}{N_g - 1}\mathbf{H}_g + \lambda \mathbf{I}$ and $\mathbf{A}_g^{-1}$.
Broadcast $(\mathbf{A}_g^{-1}, N_g)$ to all clients.
/* *On the Client Side (Round 2)* */
for $l = 1, \ldots, L$ **do**
    for $i = 1, \ldots, N_l$ **do**
        $\bar{\boldsymbol{\phi}}_i^{(l)} \leftarrow \frac{\boldsymbol{\phi}(z_i^{(l)})}{\sqrt{N_g - 1}}$.
        $\widetilde{\mathcal{V}}_i^{(l)} \leftarrow \frac{\left((\bar{\boldsymbol{\phi}}_i^{(l)})^\top \mathbf{A}_g^{-1}\mathbf{h}\right)^2}{1 - (\bar{\boldsymbol{\phi}}_i^{(l)})^\top \mathbf{A}_g^{-1} \bar{\boldsymbol{\phi}}_i^{(l)}}$.
    end for
    Upload $\widetilde{\mathcal{V}}^{(l)}$ to the Server.
end for
return $\{\widetilde{\mathcal{V}}^{(l)}\}_{l=1}^{L}$.

and 500 samples as the support set, while the remaining samples are used as the test set. This dataset is mainly used to evaluate sample removal and label-free data selection.

Products (Hu et al., 2020) contains Amazon product listings categorized into 47 classes, and the downstream task is product-category classification. For large-scale valuation, we use up to 2M samples as the target set, with 10,000 support samples and 190,000 test samples. For the sample-removal experiments, we use 5,000 target samples and 200 support samples, together with a separate test subset for downstream evaluation. The target-set size is varied from 10,000 to 2M in the runtime experiments.

Arxiv is constructed from OGB-ArXiv (Hu et al., 2020), where each sample corresponds to an Arxiv paper and citation edges are ignored. The downstream task is to predict the primary subject category of each paper. To simulate a temporally heterogeneous data marketplace, papers are partitioned into three client target sets according to publication year: papers published up to 2016, from 2017 to 2018, and in 2019. Papers published in 2020 are held by the server and further divided into disjoint support and test sets. This setting evaluates whether data from historical client repositories remain useful with respect to a more recent server-side distribution.

Wiki (Mernyei & Cangea, 2020) contains computer-science-related Wikipedia articles from 10 categories corresponding to different subfields. The dataset is randomly partitioned into four subsets, with three serving as client target sets and one held by the server. Each client contains 2,000 target samples, while 200 server-side samples are used as the support set and the remaining server samples form the test set.

Unless otherwise specified, downstream performance is measured using a fixed feature extractor followed by a lightweight MLP task head trained on the selected target samples. For the text datasets Arxiv and Wiki, we extract 768-dimensional representations from titles/abstracts or article text using Sentence-BERT (Reimers & Gurevych, 2019). For the image dataset Cifar10, features are

extracted from the penultimate layer of a ResNet50 (He et al., 2016). Products uses its precomputed dimensionality-reduced bag-of-words features.

A.4.2 Compared Baselines

In this work, seven baselines have been selected for comparison.

Influence Function (Koh & Liang, 2017) estimates the impact of a data point on the model's predictive performance by computing the sensitivity of the model parameters to infinitesimal perturbations in the weight of that individual sample. In this work, we adopt the model-based InFun score.

DVRL (Yoon et al., 2020): This is a method that utilizes reinforcement learning to compute data value. It trains an agent model to predict the importance weight of samples.

BetaShapley (Kwon & Zou, 2022): This is an improved version of DataShapley (Ghorbani & Zou, 2019). It enhances performance in noise detection by redefining subset weights through the introduction of a Beta distribution function.

DataBanzhaf (Wang & Jia, 2023): It evaluates the value of a data sample by computing its average marginal contribution across all possible subsets. In addition, it accelerates the evaluation by enhancing sampling efficiency through Maximum Sample Reuse (MSR).

LAVA (Just et al., 2023): LAVA is a model-agnostic valuation method based on optimal transport. It estimates sample importance from the sensitivity of the transport-based discrepancy between the target and support distributions.

SAVA (Kessler et al., 2025): SAVA extends transport-based model-agnostic valuation to larger datasets through a hierarchical strategy that performs valuation across and within data batches, thereby reducing the computational cost of direct optimal-transport-based valuation.

Kairos (Zhu et al., 2025): Kairos is a model-agnostic valuation method based on Maximum Mean Discrepancy (MMD). It estimates sample contributions through the change in the MMD-based discrepancy between the target and support distributions, providing an efficient approximation to finite-deletion valuation.

A.4.3 Federated Learning Methods

In the distributed setting, we simulate multi-data-center collaborative training without pooling raw client data. To this end, four representative federated learning methods are employed to evaluate the effectiveness of data valuation.

FedAVG (McMahan et al., 2017): It is a foundational method in FL. Initiated by the central server, the global model is distributed to each client, and clients independently perform local updates.

FedProx (Li et al., 2020): Based on FedAVG, it incorporates a model alignment term to mitigate local model drift.

Scaffold (Karimireddy et al., 2020): It introduces global drift and local drift terms to rectify the direction of local gradient updates, thereby addressing the issue of client drift.

FedDC (Gao et al., 2022): It introduces a local drift variable to constrain the discrepancy between the local and global models without incurring additional communication overhead.

Our goal is to evaluate the robustness of data valuation across different federated optimizers rather than to compare federated learning algorithms themselves. Therefore, the training configuration is kept identical across valuation methods for each federated optimizer. Specifically, each communication round uses three local epochs, and all experiments run for 100 communication rounds. The corresponding correction coefficients in FedProx and FedDC are set to 0.01.

A.4.4 Experimental Setup

Valuation Setting. Following OpenDataVal (Jiang et al., 2023a), we inject 20% label noise into the target sets for the noisy-data valuation experiments, while keeping the support and test sets clean. Hyperparameters of the baselines follow their original implementations whenever available. Otherwise, we use the default configurations provided by OpenDataVal. BetaShapley employs the

Gelman–Rubin truncated Monte Carlo sampler for efficient estimation. In the data selection task, we remove label-dependent terms and use only feature information for LAVA, SAVA, and Kairos to ensure a fair comparison. Since the other baselines require access to labels for valuation, we do not include them in data selection.

For FastDV, we use the Gaussian RBF kernel for κ_x throughout all experiments. The joint kernel κ_z is used by default, while the feature-only kernel κ_x is adopted for label-free data selection. In centralized experiments, representative anchors are selected by sampling from the support set, whereas in multi-party experiments the server constructs synthetic anchors using class-wise clustering on the support set, with each cluster centroid inheriting its class label, thereby avoiding the direct distribution of raw support samples. The Gaussian-kernel bandwidth σ and regularization coefficient λ are searched over $[0.5, 10]$ and $[10^{-9}, 10^{-3}]$, respectively. For task-aware valuation, σ and λ are selected once by downstream validation on the support set over the predefined candidate grids and remain fixed throughout valuation. For label-free data selection, labels are excluded from the FastDV scoring function, and valuation scores are computed solely from feature representations. The kernel hyperparameters are selected once through support-set validation and subsequently fixed for all candidate samples.

Test Setting. To ensure fairness, each experiment is repeated 10 times, and the average results along with their standard deviations are reported. In the data selection experiment, valuation is first performed on all data samples with their labels removed. The samples are then ranked, and those with the highest values are selected for labeling. The performance is subsequently evaluated based on these newly labeled high-value samples. For the computational cost experiments, all valuation methods are executed on the CPU to ensure comparability.

Data Marketplace Task Setting In data marketplaces, the data purchaser (server side) possesses a support set, while each data seller (client side) holds an independent dataset. Data cannot be shared among clients. Based on task requirements, we categorize the scenarios into **Data Procurement** and **Federated Learning**.

(1) Data procurement aims to evaluate the data value of each client, after which the server purchases high-value data from the clients for centralized training. Since existing DV methods do not directly produce globally comparable sample scores under independent local valuation, we designed two procurement strategies for them. *(S1) Uniform Procurement*: Each client selects and sends the top-K% highest-value data to the server. *(S2) Centralized Procurement*: After normalizing the data value scores from all clients, the server re-ranks all data samples and selects the global top-K% highest-value data.

(2) Federated learning involves each client participating in federated training using the samples selected by the valuation method, without uploading raw data to the server. For the baseline methods, each client independently selects its local top-$K\%$ samples according to locally computed values. For FastDV, all candidate samples across clients are jointly ranked using globally comparable values, and the global top-$K\%$ samples are selected in total. The selected samples remain on their original clients and are used for federated training.

Notably, in both data procurement and federated learning scenarios, existing DV methods require distributing the support set of the server to each client to enable valuation. In contrast, FastDV communicates a shared anchor dictionary and support statistic instead of distributing the raw support set.

Experiment Environment. The experiments are conducted on the machine with Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, and NVIDIA GeForce RTX 3090 with 24GB memory and CUDA 11.8. The operating system is Ubuntu 20.04.2 with 512GB memory.

A.5 Additional Experiments

A.5.1 Sample Removal

Table 5 provides additional high-value sample removal results at intermediate removal ratios. Overall, FastDV exhibits increasingly pronounced performance degradation as more of its top-ranked samples are removed. On Cifar10, FastDV achieves the lowest downstream accuracy from the 15% removal ratio onward, while on Products it consistently achieves the lowest accuracy across all reported ratios.

These results complement the main-text experiments and further indicate that FastDV effectively identifies samples that are important for downstream learning.

Table 5: The high-value sample removal results. Lower performance indicates better valuation.

	Cifar10				Products			
	5%	15%	25%	35%	5%	15%	25%	35%
InFun	65.27±0.60	61.75±0.48	56.88±1.36	51.30±2.37	35.05±1.34	33.15±1.55	30.01±0.48	22.99±0.34
DVRL	64.55±0.97	59.35±0.70	57.82±0.30	53.51±0.51	36.22±1.61	31.99±1.69	31.57±1.41	29.10±1.84
BetaS.	66.66±0.83	65.07±0.42	64.33±0.58	61.73±0.31	34.50±1.41	33.51±2.40	31.82±1.52	29.06±1.36
DataB.	66.28±0.93	64.59±0.66	63.35±0.53	61.55±0.80	34.75±2.21	33.93±2.30	33.72±1.01	32.93±1.50
LAVA	66.08±0.24	64.66±0.69	62.15±0.36	59.80±0.29	36.49±1.96	34.87±0.74	33.61±1.85	33.93±0.61
SAVA	66.82±0.40	66.00±0.25	63.99±0.57	63.08±0.33	36.29±0.63	34.32±1.44	36.41±1.36	30.99±0.70
Kairos	65.55±0.54	61.57±0.76	56.77±0.42	47.37±0.73	36.72±0.93	32.07±1.32	27.62±2.02	18.64±0.70
FastDV	65.61±0.58	58.78±0.47	52.70±1.17	44.45±0.40	33.04±1.95	25.32±1.07	23.01±0.79	17.42±0.48

A.5.2 Federated Learning

Tables 6 and 7 provide additional federated learning results under 60% low-value sample removal. Overall, FastDV achieves the highest average performance on both datasets, reaching 64.21% on Arxiv and 62.15% on Wiki. On Arxiv, FastDV performs best under FedAVG, FedProx, and FedDC, and remains competitive under Scaffold. A similar trend is observed on Wiki, where FastDV achieves the best results under FedAVG, FedProx, and FedDC, while Kairos performs better under Scaffold.

Compared with the 40% removal results reported in the main text, these results show that FastDV remains effective even when a larger fraction of locally stored data is discarded. The consistent performance across multiple federated optimizers further suggests that the globally comparable rankings produced from shared client statistics are robust to different downstream federated training strategies.

Table 6: The performance of 60% low-value sample removal across different federated learning methods on the Arxiv dataset.

	InFun	DVRL	BetaS.	DataB.	LAVA	SAVA	Kairos	FastDV
FedAVG	41.52±0.58	57.52±0.74	64.36±0.74	64.82±0.88	63.22±0.78	64.81±0.63	64.36±0.39	65.73±0.63
FedProx	40.61±0.64	57.31±1.02	63.03±0.98	64.68±0.55	61.52±1.10	64.32±0.64	64.07±0.59	65.10±0.52
Scaffold	40.83±1.29	54.79±2.40	58.50±3.37	60.25±1.66	59.13±2.36	61.14±1.14	62.66±1.30	61.99±1.74
FedDC	41.73±0.60	56.31±1.02	63.37±1.05	63.35±0.82	62.33±1.22	63.38±0.70	63.87±0.73	64.01±0.36
Avg.	*41.17*	*56.48*	*62.32*	*63.28*	*61.55*	*63.41*	*63.74*	*64.21*

Table 7: The performance of 60% low-value sample removal across different federated learning methods on the Wiki dataset.

	InFun	DVRL	BetaS.	DataB.	LAVA	SAVA	Kairos	FastDV
FedAVG	64.02±0.58	53.14±0.40	69.24±0.39	67.88±0.25	67.17±0.83	67.29±0.58	68.19±0.13	69.64±0.13
FedProx	64.26±0.43	53.11±0.39	69.33±0.34	67.92±0.27	66.80±0.99	67.46±0.43	68.18±0.10	69.60±0.09
Scaffold	51.72±0.39	40.87±0.61	61.32±0.96	61.43±0.40	57.70±0.54	60.44±0.58	62.40±0.32	59.20±2.24
FedDC	40.76±0.77	44.70±2.33	46.94±2.85	47.15±1.44	48.35±1.68	45.12±1.24	45.62±0.54	50.15±0.77
Avg.	*55.19*	*47.96*	*61.71*	*61.10*	*60.01*	*60.08*	*61.10*	*62.15*

A.5.3 Visualization

Cifar10 Visualization. We visualize Cifar10 examples ranked by FastDV and Kairos to qualitatively analyze the valuation behavior of FastDV, as shown in Fig. 6. Specifically, the top-5 high-value and low-value samples from the horse, truck, and deer classes are visualized (To better focus on the semantic quality of selected samples, examples with noisy labels are excluded from the visualization.). Overall, the high-value samples identified by FastDV are generally clearer and more class-discriminative, while the low-value samples tend to exhibit ambiguous shapes, weak object semantics, or visually confusing patterns. The red boxes mark samples that are commonly identified as high-value by both FastDV and Kairos, suggesting that some informative and representative samples

can be consistently recognized by different valuation criteria. In contrast, the non-overlapping samples reflect the key differences between the two methods and may partly explain their performance gap. Moreover, several low-value samples selected by FastDV, highlighted with blue boxes, are also difficult for humans to recognize, providing further qualitative evidence that FastDV can identify redundant or less informative samples.

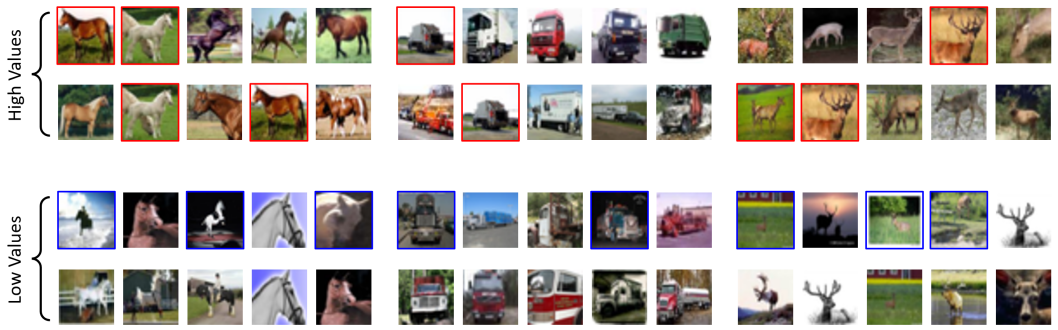


Figure 6: Visualization of Cifar10 examples ranked by FastDV and Kairos. Top-2 rows / bottom-2 rows display top-ranked / bottom-ranked examples of FastDV and Kairos, respectively.

Visualization of Value Scores. In Fig. 7, we present normalized histograms of valuation distributions under different methods. It can be clearly observed that DVRL and SAVA do not distinctly differentiate clean samples from noisy samples. In contrast, Kairos exhibits a relatively uniform distribution of noisy samples across all value ranges. However, FastDV assigns very low values to noisy samples, while the majority of clean samples are concentrated in the high-value region. It is noteworthy that some clean samples still appear in the low-value region (close to 0). These samples may be either highly dissimilar to the support set or redundant instances.

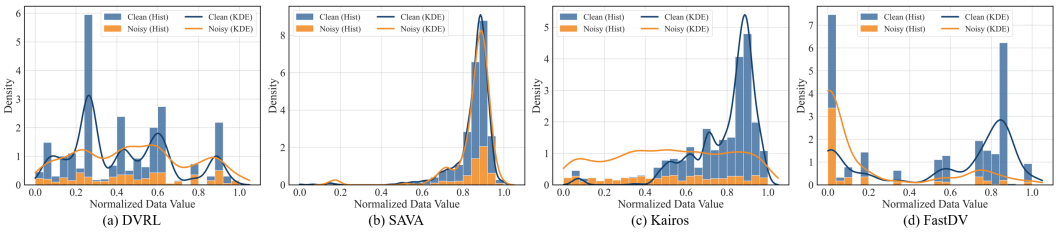


Figure 7: Histograms of value distribution on the Products dataset.

Analysis of Valuation Bias. In this section, we further examine the discrepancy between locally computed and globally computed sample values. The *merged valuation*, corresponding to strategy *S2* in the data procurement experiment, first performs valuation independently on each client and then normalizes and merges the resulting local scores. As a centralized oracle, the *global valuation* instead pools all client datasets before valuation and therefore evaluates every sample using the statistics of the complete target set.

Fig. 8 compares the resulting value distributions on Wiki. Existing methods exhibit noticeable discrepancies between merged and centralized valuations, indicating that normalization alone cannot fully reconcile values computed from different client-specific statistics. In contrast, FastDV aggregates client-side second-moment statistics in the shared anchor space and, by Corollary 3.2, recovers the centralized FastDV valuation without pooling raw client data. These results empirically support the importance of globally comparable valuation in multi-party data marketplaces.

A.6 Limitations and Future Work

Despite these strengths, it must be acknowledged that FastDV's valuation scores do not necessarily coincide with model-based LOO, which is also a common limitation of existing model-agnostic DV methods. Exploring how model-agnostic valuation relates to model-based valuation remains

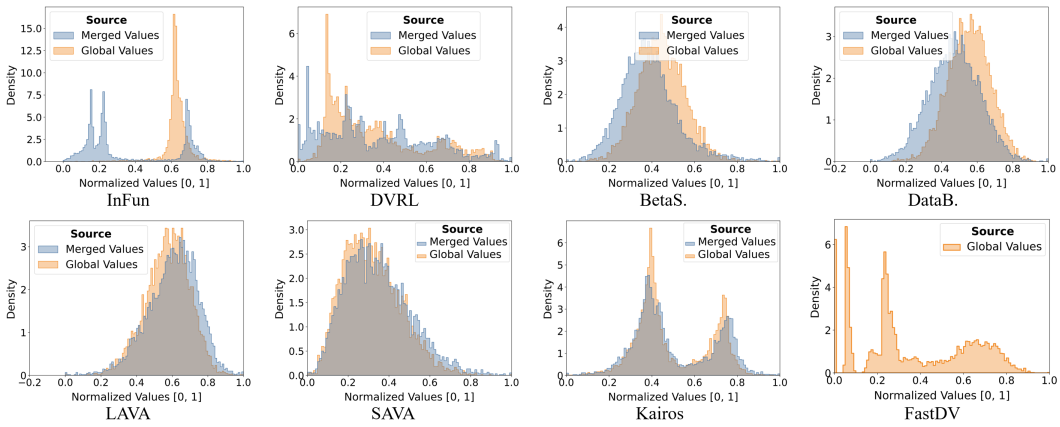


Figure 8: The distribution discrepancy between the merged valuation and the global valuation on the Wiki dataset.

an interesting direction for future work. Moreover, in multi-party settings, clients upload the data-dependent statistic $\mathbf{H}^{(l)}$, which may still reveal information about their local data and therefore does not constitute a formal privacy guarantee. Nevertheless, compared with directly pooling raw client data or distributing the complete support set, FastDV only exchanges aggregated statistics and a shared representation, substantially reducing the direct exposure of original samples. Moreover, the focus of this work is on globally comparable sample-level valuation under distributed data ownership. We leave the integration of differential privacy, secure aggregation, and other privacy-preserving mechanisms to future work.