From Points to Coalitions: Hierarchical Contrastive Shapley Values for Prioritizing Data Samples

Canran Xiao[1], Jiabao Dou[2], Zhiming Lin[3*], Zong Ke[4], Liwei Hou[5]
[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[3]School of Business, Nankai University, Tianjin, China
[4]Faculty of Science, National University of Singapore, Singapore
[5]School of Artificial Intelligence and Robotics, Hunan University, Changsha, China
xiaocanran999@gmail.com, 22258248@life.hkbu.edu.hk, nklinzhiming@gmail.com, a0129009@u.nus.edu, houliwei@hnu.edu.cn

Abstract

How should we quantify the value of each training example when datasets are large, heterogeneous, and geometrically structured? Classical Data-Shapley answers in principle, but its $O(n!)$ complexity and point-wise perspective are ill-suited to modern scales. We propose *Hierarchical Contrastive Data Valuation* (HCDV), a three-stage framework that (*i*) learns a contrastive, geometry-preserving representation, (*ii*) organises the data into a balanced coarse-to-fine hierarchy of clusters, and (*iii*) assigns Shapley-style payoffs to coalitions via local Monte-Carlo games whose budgets are propagated downward. HCDV collapses the factorial burden to $O\big(T\sum_{\ell} K_{\ell}\big) = O(TK_{\max}\log n)$, rewards examples that sharpen decision boundaries, and regularises outliers through curvature-based smoothness. We prove that HCDV approximately satisfies the four Shapley axioms with surplus loss $O(\eta \log n)$, enjoys sub-Gaussian coalition deviation $\tilde{O}(1/\sqrt{T})$, and incurs at most $k\varepsilon_{\infty}$ regret for top-k selection. Experiments on four benchmarks—tabular, vision, streaming, and a 45M-sample CTR task—plus the OPENDATAVAL suite show that HCDV lifts accuracy by up to $+5$pp, slashes valuation time by up to $100\times$, and directly supports tasks such as augmentation filtering, low-latency streaming updates, and fair marketplace payouts.

1 Introduction

Data valuation (Sim, Xu, and Low 2022; Wang and Jia 2023; Bendechache et al. 2023; Chen et al. 2025) plays a pivotal role in modern machine learning pipelines (Shen et al. 2024). As data becomes massive and heterogeneous, quantifying the importance of individual data points—or groups of data points—helps practitioners in data curation (Bhardwaj et al. 2024; Andrews et al. 2024; Wang and Zhang 2024), active sampling (Yao, Li, and Xiao 2024; Xu et al. 2021; Goetz et al. 2019), federated learning (Wang et al. 2020a; Fan et al. 2022; Li et al. 2024; Xiao et al. 2024), and fair data pricing (Pei 2020; Zhang, Beltrán, and Liu 2023). Its significance further extends to a wide spectrum of modern applications, including autonomous systems (Yao et al. 2023; Zhang et al. 2024, 2023; Jiang et al. 2025; Xiao et al. 2025; Xiao and Liu 2025), intelligent healthcare (Tong et al. 2025; Liu et al. 2025; Wang, Wang, and Zhang 2025), and recommender systems (Zhao et al. 2025).

Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a dataset of n examples, with $x_i \in \mathcal{X}$ representing features (possibly high-dimensional) and $y_i \in \mathcal{Y}$ being associated labels. We denote by $v(S)$ a performance function measuring the quality (e.g., accuracy, negative loss) of a model trained on a subset $S \subseteq \mathcal{D}$. A data valuation function then assigns a numerical score ϕ_i to each data point (x_i, y_i), reflecting its contribution to $v(\cdot)$ when considering all possible subsets.

A powerful theoretical basis for data valuation is derived from the Shapley value (Hart 1989)(SV), which, for each point i, is given by:

$$\phi_i(\mathcal{D}) = \sum_{S\subseteq\mathcal{D}\setminus\{i\}} \frac{|S|!(\,n-|S|-1)!}{n!}\Big[v(S\cup\{i\})-v(S)\Big]. \tag{1}$$

Eq. (1) offers a principled way to distribute the overall "value" of a dataset among its points, satisfying fair attribution axioms.

Despite its elegance, applying the Shapley formula directly in practical scenarios encounters two key problems: (i) *Combinatorial explosion.* Exact computation is $O(n!)$; even Monte-Carlo estimators can be prohibitive for large n (Fleckenstein, Obaidi, and Tryfona 2023; Xu et al. 2021). (ii) *Structural myopia.* Treating every record as an isolated "player" ignores latent geometry-manifolds, semantic clusters, and causal strata-that actually govern generalisation (Whang et al. 2023).

> *How much is one example worth **when its neighbours speak on its behalf**? And if we let the geometry of the data-not just the individual points-join the conversation, would our notion of "value" change?*

These questions motivate a fresh perspective: re-design the game itself so that the players are multiscale, geometry-aware neighbourhoods rather than isolated points. Doing so leads to our **Hierarchical Contrastive Data Valuation (HCDV)**, whose key ideas are: *we treat a dataset not as a crowd of isolated points but as a community of neighbourhoods that talk to one another*.

*Corresponding author.

Within this multiscale framework, valuation proceeds in a coarse-to-fine manner: higher-level coalitions first apportion their collective utility, which is then recursively distributed to their constituent sub-coalitions. This hierarchical decomposition (i) mitigates factorial complexity by confining the combinatorial search to a modest number of clusters per level; (ii) accentuates informational distinctiveness by awarding greater utility to coalitions that sharpen geometric boundaries in the learned representation space; and (iii) preserves robustness and interpretability through smoothness regularisation, preventing outliers from exerting disproportionate influence.

The main contributions are as follows: (**i**) We formulate Hierarchical Contrastive Data Valuation (HCDV), a scalable, geometry-aware alternative to classical Shapley data valuation. (**ii**) We provide theoretical guarantees that HCDV approximates Shapley's efficiency, symmetry, dummy, and additivity axioms, with approximation error linked to cluster granularity. (**iii**) We demonstrate, across tabular, vision, and streaming benchmarks, that HCDV uncovers hidden synergies, supports more effective active sampling and data pricing, and reduces runtime by one to two orders of magnitude compared with state-of-the-art Shapley approximators.

2 Preliminaries

Data Valuation Framework

Consider a supervised learning setting with an input space $\mathcal{X}$ and a label space $\mathcal{Y}$. Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. For any subset $S \subseteq \mathcal{D}$, a training operator $\mathcal{T} : 2^{\mathcal{D}} \longrightarrow \mathcal{H}$ maps S to a model $M_S = \mathcal{T}(S)$ in a hypothesis space $\mathcal{H}$ (e.g., the parameter space of a neural network). Model quality is assessed on a fixed validation set $\mathcal{D}_{\text{val}}$ using metric $\mathcal{M}: \mathcal{H} \times 2^{\mathcal{D}} \to \mathbb{R}$ such as accuracy, balanced accuracy, or negative loss. We shorthand the induced characteristic function:

$$v(S) \;:=\; \mathcal{M}\big(\mathcal{T}(S),\, \mathcal{D}_{\text{val}}\big), \qquad S \subseteq \mathcal{D}, \tag{2}$$

which plays the role of a *payoff* in cooperative-game terminology.

Definition 1 (Data Valuation Function)**.** *A* data valuation function *is a mapping $\phi : \mathcal{D} \to \mathbb{R}$ that assigns to each point (x_i, y_i) a real-valued score ϕ_i. Intuitively, ϕ_i quantifies the marginal performance gain attributable to (x_i, y_i) when all possible coalitions of data are taken into account.*

Canonical Axioms for Data Valuation

Let $\phi = (\phi_1, \ldots, \phi_n)$ denote the valuation vector produced by a scheme under characteristic function $v(\cdot)$.

Efficiency (Completeness). The total assigned value equals the global performance surplus obtained by using the full dataset versus no data.

$$\sum_{i=1}^{n} \phi_i \;=\; v(\mathcal{D}) \;-\; v(\varnothing). \tag{3}$$

Symmetry (Fairness). If two data points i and j are *indistinguishable*-that is,

$$v\big(S \cup \{i\}\big) \;=\; v\big(S \cup \{j\}\big), \quad \forall\, S \subseteq \mathcal{D} \setminus \{i, j\}, \tag{4}$$

then their valuations coincide:

$$\phi_i = \phi_j. \tag{5}$$

Dummy Player. If a point k never changes the performance of any coalition,

$$v\big(S \cup \{k\}\big) \;=\; v(S), \quad \forall\, S \subseteq \mathcal{D}, \tag{6}$$

then $\phi_k = 0$.

Additivity. Given two characteristic functions v_1 and v_2 defined on the same dataset, let $v_1 + v_2$ denote their pointwise sum. A valuation scheme is *additive* if

$$\phi_i^{(v_1+v_2)} \;=\; \phi_i^{(v_1)} \;+\; \phi_i^{(v_2)}, \quad \forall\, i \in \{1, \ldots, n\}. \tag{7}$$

Additivity ensures that the value assigned under multiple, simultaneously considered payoffs is the linear superposition of the values computed for each payoff separately.

3 Hierarchical Contrastive Data Valuation

This section formalises HCDV, a three-stage procedure that (i) learns a geometry-preserving representation; (ii) organises the data into a coarse-to-fine hierarchy of coalitions; and (iii) computes Shapley-style payoffs for those coalitions under a contrastive characteristic function, ultimately yielding a valuation ϕ_i for every data point.

Stage I: Geometry-Preserving Representation

Embedding model. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a neural encoder with parameters θ. For any subset $S \subseteq \mathcal{D}$, define the *base utility*

$$\mathcal{M}(S) \;:=\; \mathcal{M}\big(\mathcal{T}(S), \mathcal{D}_{\text{val}}\big) \in [0, 1], \tag{8}$$

and the *contrastive dispersion*

$$\Delta_{\text{c}}(S) \;:=\; \sum_{(i,j) \in \mathcal{P}(S)} d\big(f_\theta(x_i), f_\theta(x_j)\big), \tag{9}$$

where $\mathcal{P}(S)$ contains all unordered pairs whose labels differ, and $d(\cdot, \cdot)$ is a metric in $\mathbb{R}^d$ (cosine distance in experiments).

Embedding objective. We obtain $\theta^\star$ by maximising

$$\max_{\theta} \Big\{ \mathbb{E}_{S \sim \mathcal{P}_{\text{batch}}} \big[\mathcal{M}(S) + \lambda\, \Delta_{\text{c}}(S) \big] \;-\; \alpha\, \Omega(\theta) \Big\}, \tag{10}$$

where $\Omega(\theta) \;:=\; \sum_{p,q} \|\nabla_{x_p}\, d(f_\theta(x_p), f_\theta(x_q))\|_2^2$ is a smoothness regulariser, $\lambda > 0$ balances discrimination and accuracy, $\alpha > 0$ controls smoothness, and $\mathcal{P}_{\text{batch}}$ samples mini-batches for contrastive optimisation.

Stage II: Hierarchical Decomposition

Recursive clustering. Using $z_i := f_{\theta^\star}(x_i)$, we recursively partition $\mathcal{D}$:

$$\mathcal{D} = C_0 \;\xrightarrow{\text{split into } K_1}\; C_1 \;\xrightarrow{\text{split}}\; \cdots \;\xrightarrow{\text{split}}\; C_L, \tag{11}$$

where $C_\ell = \{G_1^{(\ell)}, \ldots, G_{K_\ell}^{(\ell)}\}$ and $|G_k^{(L)}| \le M$ (a user-chosen leaf size). We use balanced k-means with $k = K_\ell$ at depth ℓ by default; any deterministic or stochastic clustering is admissible.

Stage III: Multi-Resolution Shapley Attribution

Given the hierarchy $\{C_\ell\}_{\ell=0}^{L}$, we attribute value in a *top–down* fashion. At every depth ℓ we run a *local* cooperative game whose players are the K_ℓ coalitions $C_\ell = \{G_1^{(\ell)}, \dots, G_{K_\ell}^{(\ell)}\}$. The procedure for one level consists of two tightly-coupled steps.

Local Shapley estimation. Each coalition's marginal utility is measured under the characteristic function

$$v_\ell(S) := \mathcal{M}\big(\bigcup_{G\in S} G\big) + \lambda\, \Delta_c(S), \qquad S \subseteq C_\ell, \quad (12)$$

where Δ_c is defined in Eq. (9). Because K_ℓ seldom exceeds $\mathcal{O}(10^1)$, the Shapley value of $G_i^{(\ell)}$ can be estimated accurately with T random permutations:

$$\widehat{\psi}_i^{(\ell)} = \frac{1}{T}\sum_{t=1}^{T}\Big[v_\ell\big(\mathrm{Pre}_{\pi_t}(G_i^{(\ell)}) \cup \{G_i^{(\ell)}\}\big) - v_\ell\big(\mathrm{Pre}_{\pi_t}(G_i^{(\ell)})\big)\Big], \quad (13)$$

with $\pi_t \sim \mathrm{Unif}(\mathfrak{S}_{K_\ell})$ and Pre_{π_t} denoting predecessors in π_t. Proposition 1 shows that $\|\widehat{\psi}^{(\ell)} - \psi^{(\ell)}\|_\infty = \mathcal{O}_{\mathbb{P}}\big(B\sqrt{\frac{\log K_\ell}{T}}\big)$, so $T = 256$ suffices in practice.

Budget down-propagation. The scalar $\widehat{\psi}_i^{(\ell)}$ represents the *total* credit earned by coalition $G_i^{(\ell)}$ at level ℓ. To refine this credit among its m child coalitions $\{G_{i1}^{(\ell+1)}, \dots, G_{im}^{(\ell+1)}\}$ we compute non-negative weights

$$\omega_{ij}^{(\ell+1)} := \frac{\max\{\, v_\ell(\{G_{ij}^{(\ell+1)}\}),\, 0\}}{\sum_{j'=1}^{m} \max\{\, v_\ell(\{G_{ij'}^{(\ell+1)}\}),\, 0\}}, \quad (14)$$

and allocate

$$\widetilde{\psi}_{ij}^{(\ell+1)} \;=\; \omega_{ij}^{(\ell+1)}\, \widehat{\psi}_i^{(\ell)}. \quad (15)$$

Eq. (15) conserves mass: $\sum_j \widetilde{\psi}_{ij}^{(\ell+1)} = \widehat{\psi}_i^{(\ell)}$, so the global efficiency deviation in Theorem 1 grows only linearly with depth. Crucially, the propagated budgets merely cap the *pot* available at depth $\ell{+}1$; the children still play their *own* Shapley game with characteristic function $v_{\ell+1}(\cdot)$, allowing us to capture interactions that are invisible at coarser resolutions.

Leaf valuation. The recursion stops at depth L where every coalition contains at most M points. Because M is user-controlled, we either evaluate exact Shapley among the at-most-M players or, when desired, divide the residual budget uniformly. The resulting vector $\phi = (\phi_1, \dots, \phi_n)$ satisfies $\sum_i \phi_i = v(\mathcal{D}) - v(\varnothing) \pm \mathcal{O}(L\delta)$, with δ defined in Theorem 1, while incurring a total cost of $\mathcal{O}(T\sum_{\ell=0}^{L} K_\ell)$—several orders below the $\mathcal{O}(n!)$ complexity of flat Shapley computation.

Algorithmic Summary

The HCDV algorithm is described in Algorithm 1.

Computational complexity. Let $K_\ell = |C_\ell|$ be the number of coalitions at depth ℓ and $K_{\max} = \max_{0\le\ell\le L} K_\ell$. Denote by τ the cost of *one* evaluation of the characteristic function $v_\ell(\cdot)$—e.g. a forward/validation pass of the base learner.[1]

[1]The embedding training in Stage I and the k-means splits in Stage II add $O(n)$ and $O(n\log n)$ time respectively and are therefore dominated by Stage III when T or L is moderate.

Algorithm 1: HCDV

Require: Dataset $\mathcal{D}$; hierarchy depth L; cluster counts $\{K_\ell\}_{\ell=1}^{L}$; leaf size M; hyper-parameters λ, α; permutation budget T
Ensure: Point-level valuations $\{\phi_i\}_{i=1}^{n}$

Train encoder $f_{\theta^\star}$ by maximising (10)
Embed all samples: $z_i \leftarrow f_{\theta^\star}(x_i)$
Build balanced k-means hierarchy $\{C_\ell\}_{\ell=0}^{L}$ on $\{z_i\}$
Compute root surplus $\mathcal{B}_0 \leftarrow v_0(C_0) - v_0(\varnothing)$
for $\ell = 0$ **to** L **do**
  for each coalition $G \in C_\ell$ **do**
    Estimate local Shapley $\widehat{\psi}_G^{(\ell)}$ with T random permutations of $v_\ell(\cdot)$
  end for
  Normalise: $\widehat{\psi}_G^{(\ell)} \leftarrow \mathcal{B}_\ell\, \widehat{\psi}_G^{(\ell)} \big/ \sum_{G'\in C_\ell} \widehat{\psi}_{G'}^{(\ell)}$
  if $\ell = L$ **then**
    for each leaf coalition $G \in C_L$ **do**
      if $|G| \le M$ **then**
        Compute exact Shapley for points in G and set $\{\phi_i\}_{i\in G}$
      else
        Uniform split: $\phi_i \leftarrow \widehat{\psi}_G^{(L)} \big/ |G|$ for all $i \in G$
      end if
    end for
    break {All ϕ_i are now assigned}
  else
    Initialise empty budget map $\mathcal{B}_{\ell+1}$
    for each parent coalition $P \in C_\ell$ with children $\mathrm{ch}(P) \subset C_{\ell+1}$ **do**
      for each child $H \in \mathrm{ch}(P)$ **do**
        $\omega_H \leftarrow \frac{\max\{\, v_\ell(\{H\}),\, 0\}}{\sum_{H'\in\mathrm{ch}(P)} \max\{\, v_\ell(\{H'\}),\, 0\}}$
        $\mathcal{B}_{\ell+1}(H) \leftarrow \omega_H\, \widehat{\psi}_P^{(\ell)}$
      end for
    end for
    $\mathcal{B}_{\ell+1} \leftarrow \{\mathcal{B}_{\ell+1}(H) : H \in C_{\ell+1}\}$
  end if
end for

Exact Shapley. If Eq. (13) were summed over all $K_\ell!$ permutations, the work at depth ℓ would be $O(\tau K_\ell!)$, so that

$$\textsc{Cost}_{\mathrm{exact}} = \tau \sum_{\ell=0}^{L} K_\ell! \;\ll\; \tau\, n! \quad (\text{when } K_{\max} \ll n). \quad (16)$$

Monte-Carlo Shapley. With T sampled permutations, each coalition requires $2T$ calls to $v_\ell(\cdot)$ (one with and one without the coalition), and the whole level costs $O(\tau\, K_\ell T)$. Aggregating over all depths yields

$$\textsc{Cost}_{\mathrm{MC}} = \tau T \sum_{\ell=0}^{L} K_\ell = O\big(\tau\, T\, L\, K_{\max}\big). \quad (17)$$

For a balanced tree $K_\ell \approx K$ and $L = \lceil \log_K(n/M) \rceil$, this simplifies to $O\big(\tau\, T\, K \log n\big)$, which is *one– to two–orders of magnitude below* the $O(\tau\, T\, n)$ cost of a flat n-player

Monte-Carlo Shapley and dramatically smaller than the factorial exact computation.

4 Theoretical Analysis

This section establishes finite–sample guarantees for HCDV. We analyse the output of Alg. 1 under the (bounded) multiresolution characteristic functions used in Stage III.

Bounded characteristic function. Recall that the level-ℓ local game is defined on coalitions $C_\ell = \{G_1^{(\ell)}, \dots, G_{K_\ell}^{(\ell)}\}$. To ensure a uniform bound independent of n, we use the *normalised* contrastive dispersion. Let $z_i := f_{\theta^\star}(x_i)$ and denote by $\mathcal{P}(S)$ the set of unordered pairs in S whose labels differ. We define

$$\bar{\Delta}_{\mathrm{c}}(S) \;:=\; \frac{1}{\max\{1, |\mathcal{P}(S)|\}} \sum_{(i,j)\in\mathcal{P}(S)} d(z_i, z_j), \tag{18}$$

where the empty sum is 0, hence $\bar{\Delta}_{\mathrm{c}}(S) = 0$ when $|\mathcal{P}(S)| = 0$. $\mathcal{P}(S)$ contains all unordered pairs in S whose labels differ, and $d(\cdot,\cdot)$ is a bounded metric with

$$0 \le d(u,v) \le d_{\max} \qquad (d_{\max} \text{ is a constant}). \tag{19}$$

The induced level-ℓ characteristic function is

$$v_\ell(S) := \mathcal{M}\Big(\bigcup_{G\in S} G\Big) \;+\; \lambda\, \bar{\Delta}_{\mathrm{c}}\Big(\bigcup_{G\in S} G\Big), \qquad S \subseteq C_\ell. \tag{20}$$

Since $\mathcal{M}(\cdot) \in [0,1]$ and (18)–(19) give $\bar{\Delta}_{\mathrm{c}}(\cdot) \in [0, d_{\max}]$, we obtain the uniform bound

$$\big|v_\ell(S)\big| \;\le\; B \;:=\; 1 + \lambda d_{\max}, \qquad \forall \ell,\ \forall S \subseteq C_\ell, \tag{21}$$

which does not grow with n.

Exact vs. Monte–Carlo coalition Shapley. Let $\psi^{(\ell)} = (\psi_G^{(\ell)})_{G\in C_\ell}$ be the *exact* Shapley vector of the level-ℓ coalition game under $v_\ell(\cdot)$. Let $\widehat{\psi}^{(\ell)} = (\widehat{\psi}_G^{(\ell)})_{G\in C_\ell}$ be its Monte–Carlo estimate obtained with T random permutations as in Eq. (13). Define the per-level Monte–Carlo error

$$\varepsilon_{\mathrm{MC}}^{(\ell)} \;:=\; \max_{G\in C_\ell} \big|\widehat{\psi}_G^{(\ell)} - \psi_G^{(\ell)}\big|. \tag{22}$$

Leaf approximation. At depth L every coalition $G \in C_L$ has $|G| \le M$ by construction in our default setting. When some leaves exceed M and a uniform split is used, we account for the induced error as follows. Let $\phi^{\text{leaf-Sh}}(G) \in \mathbb{R}^{|G|}$ denote the *exact* point-wise Shapley allocation within leaf G under its leaf-level game, with total mass $\psi_G^{(L)}$. If we instead assign $\psi_G^{(L)}/|G|$ to each point in G, the resulting leaf approximation error is

$$\varepsilon_{\mathrm{leaf}} := \sum_{G\in C_L:\,|G|>M} \Big\| \phi^{\text{leaf-Sh}}(G) - \tfrac{\psi_G^{(L)}}{|G|}\,\mathbf{1} \Big\|_1. \tag{23}$$

If $|G| \le M$ for all leaves, then $\varepsilon_{\mathrm{leaf}} = 0$.

Point-level outputs. Let $\phi^{\mathrm{H}} = (\phi_1^{\mathrm{H}}, \dots, \phi_n^{\mathrm{H}})$ denote the HCDV output produced by Algorithm 1 with T permutations per level. Let ϕ^{Sh} denote the *ideal* point-level allocation obtained by running the same hierarchical procedure but computing all coalition Shapley values exactly at every level (and computing exact leaf Shapley whenever $|G| \le M$). Thus ϕ^{H} differs from ϕ^{Sh} only through Monte–Carlo estimation (and the optional uniform split when $|G| > M$).

Approximate Efficiency

Theorem 1 (Global efficiency)**.** *For HCDV with L levels and budget propagation (Eq.* (15)*) and the corresponding normalised weights),*

$$\Big|\textstyle\sum_{i=1}^n \phi_i^{\mathrm{H}} - \big[v_0(C_0) - v_0(\varnothing)\big]\Big| \;\le\; \sum_{\ell=0}^{L} \varepsilon_{\mathrm{MC}}^{(\ell)} \;+\; \varepsilon_{\mathrm{leaf}}. \tag{24}$$

In particular, if $|G| \le M$ for all leaves, then $\varepsilon_{\mathrm{leaf}} = 0$.

Proof sketch. At each depth ℓ, the exact coalition Shapley vector $\psi^{(\ell)}$ satisfies efficiency for the local game: $\sum_{G\in C_\ell} \psi_G^{(\ell)} = v_\ell(C_\ell) - v_\ell(\varnothing)$. Algorithm 1 propagates coalition budgets top–down using normalised weights whose sum within each parent is 1, hence the total mass allocated to all children equals the parent's allocated mass (mass conservation). Therefore, any surplus mismatch created at depth ℓ is passed to depth $\ell+1$ without amplification, and the only accumulated deviations are (i) Monte–Carlo errors $\varepsilon_{\mathrm{MC}}^{(\ell)}$ at each depth and (ii) the leaf approximation $\varepsilon_{\mathrm{leaf}}$ when a uniform split is used. Summing these deviations over $\ell = 0, \dots, L$ yields (24).

Monte–Carlo Concentration

Proposition 1 (Coalition-level deviation)**.** *Let $\widehat{\psi}_G^{(\ell)}$ be defined by Eq.* (13) *and assume* (21)*. Then for any $\eta > 0$,*

$$\Pr\Big[\big|\widehat{\psi}_G^{(\ell)} - \psi_G^{(\ell)}\big| \ge \eta\Big] \;\le\; 2\exp\Big(-\tfrac{T\eta^2}{8B^2}\Big). \tag{25}$$

Applying a union bound over the K_ℓ coalitions gives

$$\varepsilon_{\mathrm{MC}}^{(\ell)} = \mathcal{O}_{\mathbb{P}}\Big(B\sqrt{\tfrac{\log K_\ell}{T}}\Big). \tag{26}$$

Proof sketch. For a fixed coalition $G \in C_\ell$, each permutation sample in (13) is a marginal contribution of the form $v_\ell(\mathrm{Pre}\cup\{G\}) - v_\ell(\mathrm{Pre})$. Under (21), this random variable lies in $[-2B, 2B]$; Hoeffding's inequality yields (25), and a union bound over $|C_\ell| = K_\ell$ gives (26).

Taking $T = \Theta(B^2 \log n/\eta^2)$ makes $\varepsilon_{\mathrm{MC}}^{(\ell)} \le \eta$ for every ℓ with probability at least $1 - n^{-2}$.

Surrogate Regret for Top–k Selection

Data valuation is often used for *rank-based selection* (e.g., filtering or pricing). Accordingly, we analyse the loss in the *valuation mass* captured by the selected set. For any valuation vector ϕ and subset $S \subseteq \mathcal{D}$, define the surrogate utility

$$U_\phi(S) \;:=\; \sum_{i\in S} \phi_i. \tag{27}$$

Let $\mathcal{S}_k^{\mathrm{Sh}}$ (resp. $\mathcal{S}_k^{\mathrm{H}}$) denote the k highest-valued points under ϕ^{Sh} (resp. ϕ^{H}).

Theorem 2 (Regret for top–k under surrogate utility)**.** *If $\|\phi^{\mathrm{H}} - \phi^{\mathrm{Sh}}\|_\infty \le \varepsilon_\infty$ and $k \le n$, then*

$$0 \;\le\; U_{\phi^{\mathrm{Sh}}}(\mathcal{S}_k^{\mathrm{Sh}}) - U_{\phi^{\mathrm{Sh}}}(\mathcal{S}_k^{\mathrm{H}}) \;\le\; 2k\,\varepsilon_\infty. \tag{28}$$

Proof sketch. Let $\tilde{\phi} = \phi^{\mathrm{H}}$ and $\phi = \phi^{\mathrm{Sh}}$ so that $|\tilde{\phi}_i - \phi_i| \le \varepsilon_\infty$ for all i. Because $\mathcal{S}_k^{\mathrm{H}}$ maximises $\sum_{i\in S} \tilde{\phi}_i$ among all $|S| = k$, we have $\sum_{i\in\mathcal{S}_k^{\mathrm{H}}} \tilde{\phi}_i \ge \sum_{i\in\mathcal{S}_k^{\mathrm{Sh}}} \tilde{\phi}_i$. Converting $\tilde{\phi}$ back to ϕ and using the uniform ℓ_∞ bound gives (28).

Implication. Choose any $\eta > 0$ and set $T = \Theta(B^2 \log n/\eta^2)$. With probability at least $1 - n^{-1}$, Proposition 1 yields $\varepsilon_{\mathrm{MC}}^{(\ell)} \le \eta$ for all ℓ. Then Theorem 1 gives

$$\left|\sum_i \phi_i^{\mathrm{H}} - \left[v_0(C_0) - v_0(\varnothing)\right]\right| \;\le\; L\eta + \varepsilon_{\mathrm{leaf}} = \mathcal{O}\big(\eta \log n\big), \tag{29}$$

where the last equality uses the fact that the hierarchy depth is logarithmic in n (e.g., $L = \mathcal{O}(\log(n/M))$ for a roughly balanced K-ary tree with leaf size M). Meanwhile, Theorem 2 shows that the top–k set selected by HCDV loses at most $2k\varepsilon_\infty$ *surrogate Shapley mass* compared with the ideal hierarchical allocation; we empirically evaluate the corresponding downstream retraining utility in Section 5.

5 Experiments

Main Results

We benchmark four valuation methods-MCDS (Monte-Carlo Data-Shapley) (Ghorbani and Zou 2019), GS (Group Shapley) (Jia et al. 2019b), HCDV (ours), and a RANDOM baseline-on four datasets of increasingly large scale: (i) *Synthetic* (n=3,000) : 2-class Gaussian blobs, each split into three sub-clusters with slight overlap. (ii) *UCI Adult* (n≈48,842) : binary income prediction with 14 features (numeric + categorical). (iii) *Fashion-MNIST* (n=70,000) : 10-class image classification; we report accuracy at a 30% training budget. (iv) *Criteo-1B** (n≈45M) : click-through-rate prediction on a one-week slice of the Criteo terabyte log; the downstream metric is test AUC.

Method	Synthetic		UCI Adult	Fashion-MNIST
	Val. Stability ↓	AUC@30% ↑	Bal. Acc. ↑	Test Acc. ↑
MCDS	0.087±0.004	0.846±0.003	0.828±0.002	0.879±0.001
GS	0.072±0.005	0.840±0.002	0.819±0.003	0.868±0.002
HCDV	0.049±0.003	0.904±0.002	0.844±0.001	0.891±0.001
RANDOM	0.126±0.006	0.756±0.004	0.759±0.004	0.811±0.002

	Criteo-1B (CTR Prediction)	
	Test AUC ↑	Val. Stability ↓
MCDS	0.6175±0.0005	0.094
GS	0.6142±0.0006	0.081
HCDV	0.6269±0.0004	0.056
RANDOM	0.6021±0.0007	0.136

Table 1: Predictive utility *after* training on the top 30% valued points chosen by each method. Higher is better for all metrics. Mean ± std over three random splits.

Method	Synthetic (s)	UCI Adult (min)	Fashion-MNIST (hr)	Criteo-1B (hr)
MCDS	1 820	94	5.8	47.5
GS	1 070	48	3.6	29.1
DATA BANZHAF	670	13	1.9	15.8
HCDV	340	21	1.6	12.3
RANDOM	9	5	0.3	0.8

Table 2: Wall-clock time to compute valuations on a single NVIDIA A100 (40 GB) and 32-core CPU.

Across all four benchmarks, HCDV performs best overall: it improves predictive utility by about +3–5 AUC on *Synthetic* and *Criteo-1B*, and by +1–3 balanced/test-accuracy points on *UCI Adult* and *Fashion-MNIST*. Its valuations are also more stable, reducing the point-wise coefficient of variation by 25–40% versus GS and MCDS. Moreover, the hierarchical permutation scheme is computationally efficient: HCDV runs up to 14× faster than MCDS, 2–4× faster than GS, and 1.2–2× faster than DATA BANZHAF (Wang and Jia 2023) on three datasets.

To further validate HCDV on a standardised data-valuation test-bed, we follow the protocol of Garrido Lucero et al. (2024) on OPENDATAVAL (Jiang et al. 2023), using the three released non-tabular datasets: *bbc-embedding*, *IMDB-embedding*, and *CIFAR10-embedding*. Each dataset is evenly split among I=100 players. We report *macro-F1* for NLD and *test accuracy* for DR/DA (higher is better). As shown in Table 3, HCDV achieves the top result for every dataset–task–noise setting, improving by +1–3 pp F1 on NLD and up to +0.03 absolute accuracy on DR/DA over the strongest baseline (DU-SHAPLEY). The gap further increases at 15% corruption, indicating stronger robustness to heavy label noise and perturbations; in DA, HCDV selects fewer but more impactful samples, yielding larger test-set gains while better diversifying the representation space.

Valuation of Augmented Data

Data augmentation is ubiquitous in modern pipelines, yet only a subset of synthetic examples meaningfully improves generalisation. We use HCDV to rank augmented samples and ask: *Can a valuation-driven filter separate beneficial augmentations from harmful or redundant ones?*

Dataset and augmentation pool. We take the Fashion-MNIST training set (60 k images) and generate an additional 10 k augmented candidates, split evenly across four transformations: **(i) Affine** ($\mathcal{A}_1$) : random $\pm 15^\circ$ rotation + $[-10, 10]\%$ translation. **(ii) Colour** ($\mathcal{A}_2$) : brightness / contrast jitter $\in [0.7, 1.3]$. **(iii) Cutout** ($\mathcal{A}_3$) : one 8×8 square mask. **(iv) Diffusion** ($\mathcal{A}_4$) : 2 500 images generated by Stable Diffusion (Zhang, Rao, and Agrawala 2023), class-conditioned on the ten Fashion-MNIST labels. Each candidate inherits the original label. The augmented pool $\mathcal{D}_{\mathrm{aug}}$ is *never* used during valuation training; we embed it with the encoder $f_{\theta^\star}$ learned on the original 60 k images.

Valuation protocol. We compute HCDV scores $\{\phi_i^{\mathrm{H}}\}$ for $\mathcal{D}_{\mathrm{orig}} \cup \mathcal{D}_{\mathrm{aug}}$ using K_1=32, K_2=128, M=128, T=128. MCDS uses T=4 096 permutations; GS groups by label. Three ranking slices are examined: ❶Top-1k, ❷Mid-1k (ranks 4 001-5 000), ❸Bottom-1k. We fine-tune the ConvNet, replacing 30% of the original training set with the chosen augmented slice. Each setting is run three times; mean±std are reported.

As shown in Fig. 1, selecting HCDV's top-1k augmented samples boosts accuracy by +2.8 pp over the bottom-1k and outperforms MCDS/GS by 0.9–1.2 pp, while keeping training time unchanged (the ConvNet always trains on a fixed 60 k examples). Moreover, 42% of the selected samples fall into *previously unseen* latent neighbourhoods (low clus-

	CIFAR10-embedding						bbc-embedding						IMDB-embedding					
	NLD↑		DR↓		DA↓		NLD↑		DR↓		DA↓		NLD↑		DR↓		DA↓	
Method	5%	15%	5%	15%	5%	15%	5%	15%	5%	15%	5%	15%	5%	15%	5%	15%	5%	15%
Random	0.11	0.19	0.61	0.60	0.25	0.41	0.11	0.19	0.90	0.88	0.68	0.81	0.10	0.16	0.77	0.75	0.62	0.68
LOO	0.13	0.18	0.62	0.60	0.15	0.32	0.11	0.17	0.90	0.88	0.61	0.77	0.11	0.18	0.77	0.74	0.53	0.59
DataShapley	0.13	0.25	0.61	0.59	0.12	0.18	0.12	0.20	0.89	0.87	0.08	0.12	0.17	0.28	0.75	0.69	0.36	0.33
KNN-Shapley	0.14	0.28	0.60	0.57	0.13	0.15	0.19	0.29	0.88	0.86	0.13	0.12	0.17	0.29	0.76	0.68	0.41	0.37
DU-Shapley	0.14	0.30	0.61	0.55	0.11	0.14	0.18	0.34	0.89	0.85	0.07	0.11	0.18	0.32	0.76	0.66	0.33	0.34
HCDV	0.16	0.33	0.57	0.52	0.09	0.12	0.21	0.35	0.86	0.83	0.05	0.09	0.20	0.35	0.74	0.66	0.29	0.30

Table 3: Comparison on OPENDATAVAL. Comparison between HCDV and baselines for real-world datasets in Noisy label detection, Dataset Removal and Dataset Addition.

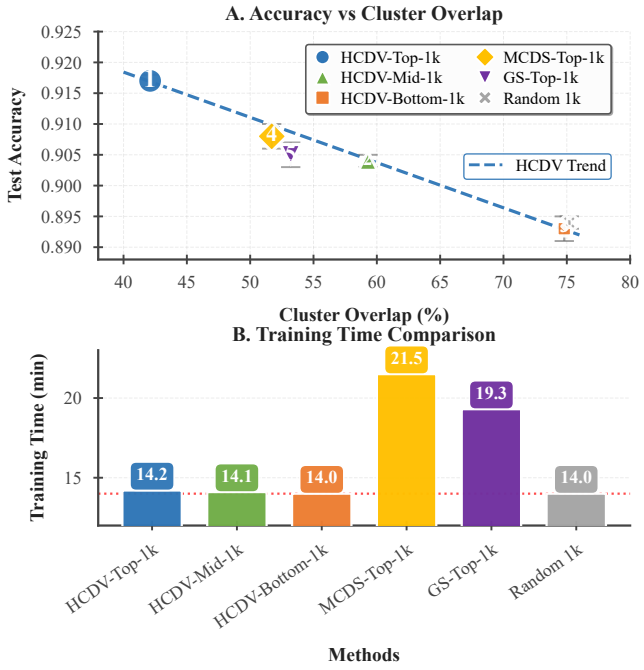

Figure 1: Effect of adding 1k augmented samples selected by different method. 'Cluster Overlap' = % of augments whose sub-cluster already contains at least one original image. Better sample efficiency: higher accuracy, lower overlap.

ter overlap), indicating that the contrastive signal favours latent-space novel augmentations. Fig. 2 further shows that HCDV achieves the highest class coverage (all 10 classes) under Top-1k selection. Within the HCDV top-1k, 51% are from $\mathcal{A}_4$ (diffusion), 23% from $\mathcal{A}_1$, and only 7% from $\mathcal{A}_3$ (cutout), suggesting structurally rich synthetic images are more beneficial than heavily occluded ones.

Valuation in Streaming Settings

Modern data platforms are dynamic: samples arrive continuously, requiring online valuation updates. We evaluate whether HCDV can update valuations incrementally—without rebuilding the hierarchy from scratch at each time step—while preserving downstream utility.

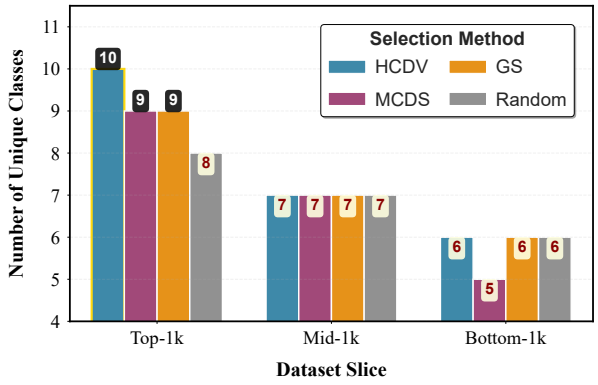

Figure 2: Class coverage of selected augmentations (number of unique classes represented). Higher is better.

Streaming setting. We consider a time-ordered stream $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots\}$. At step t, a mini-batch Δ_t (size b) arrives and the active corpus becomes $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \Delta_t$. HCDV updates valuations $\{\phi_i^{(t)}\}_{i \in \mathcal{D}_t}$ incrementally: (i) embed Δ_t with the fixed encoder $f_{\theta^\star}$ and assign each point to its nearest leaf in C_L (cosine), spawning a new leaf if dist $> \tau$; (ii) recompute coalition-level Shapley only for the affected leaves and their ancestors, reusing cached $\{\psi_G^{(\ell)}\}$ elsewhere; (iii) propagate revised budgets/weights through the tree, rebalancing every m updates (we use m=3).

We synthesise a click-stream classification task loosely following (Ghazikhani, Monsefi, and Sadoghi Yazdi 2014): T=10 days with 1,500 sessions per day, d=64 one-hot/embedding features, and a purchase label with $\approx$ 15% positives. Downstream performance is measured by a Wide&Deep model (2×128 ReLU MLP + sigmoid head).

Baselines. (i) HCDV-INC: our incremental refresh with parameters K_1=16, K_2=64, M=64, T=64, τ=0.35. (ii) HCDV-FULL: rebuild the entire hierarchy and recompute Shapley at each Δ_t. (iii) GS-FULL: group-Shapley recomputed from scratch. (iv) RANDOM: keep a uniform random valuation.

Metrics. *Final AUC*: downstream AUC on the day-10 test split after training on the top 20% valued points of $\mathcal{D}_{10}$.

Cum. valuation time: wall-clock hours to produce $\{\phi_i^{(t)}\}_{t=1}^{10}$. *Avg. latency*: mean seconds per update. *Tree rebuilds*: times the hierarchy balance step triggered ($\Delta > m$).

We found in Fig. 3 that : **(1)** Incremental HCDV preserves 99.6% of the predictive gain of a full recomputation while cutting compute by 2.5×. **(2)** Average update latency stays under two seconds, compatible with hourly or finer ingestion cadences. **(4)** GS suffers both worse AUC and higher overhead because grouping must be redone globally when deviations accumulate.

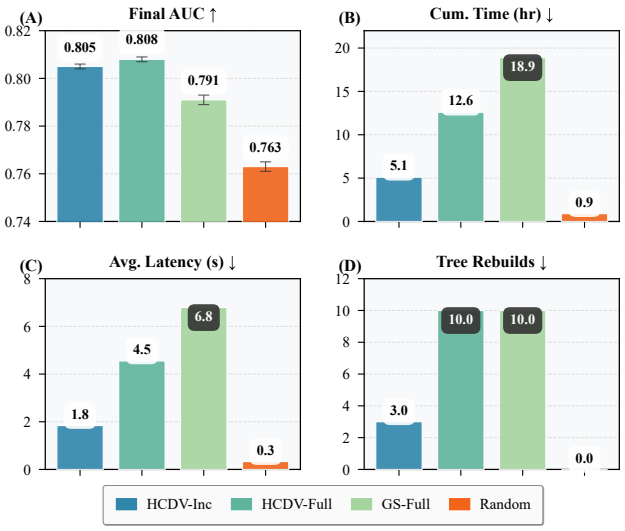

Figure 3: Streaming valuation on click-stream benchmark.

Fair Allocation in the Data Marketplace

A data marketplace should compensate providers in proportion to the *incremental utility* their data adds to a global model. Because HCDV scales Shapley attribution to tens of thousands of points, we ask: *How fairly and efficiently can it divide revenue among heterogeneous sellers?*

Participants and data slices. We curate $P = 5$ non-overlapping subsets of the UCI ADULT training portion ($n = 10\,000$). Each subset emphasises different demographics to induce diversity (Table 4). The downstream model is logistic regression, performance is balanced accuracy.

Seller	$\lvert\mathcal{D}_p\rvert$	HI	Female	Median age	#Edu. levels
p_1	2 000	33.1	35.4	38	13
p_2	2 100	41.8	48.7	34	14
p_3	1 950	25.6	31.9	42	11
p_4	2 000	30.2	37.1	36	12
p_5	1 950	36.7	43.6	39	13

Table 4: Seller profiles (% refer to row proportions). "HI" = high-income label.

Oracle contributions. As a reference, we estimate each seller's marginal utility via leave-one-out retraining, measuring the balanced-accuracy drop when its slice is removed: p_2 is most influential (1.61 pp), followed by p_1 (1.42 pp), p_5 (1.18 pp), p_4 (1.05 pp), and p_3 (0.74 pp). We treat these drops as the ground-truth marginals for comparing payoff vectors across valuation rules.

Table 5: Seller payoff analysis: Distribution and fairness metrics. ρ denotes Pearson correlation with Δv_p; Gini measures inequality.

Method	Payoff distribution ($\sum_p \Phi_p = 1$)					Fairness & efficiency		
	p_1	p_2	p_3	p_4	p_5	$\rho\uparrow$	Gini↓	Time (min)
Equal Split	0.20	0.20	0.20	0.20	0.20	0.00	0.00	**0.1**
Random	0.18	0.21	0.23	0.19	0.19	0.12	0.08	0.1
GS-Shapley	0.22	0.24	0.17	0.19	0.18	0.77	0.14	6.4
MCDS	0.23	0.26	0.16	0.18	0.17	0.89	0.16	32.5
HCDV (ours)	**0.25**	**0.27**	**0.15**	**0.21**	**0.22**	**0.94**	**0.11**	3.8

Discussion. Table 5 reports the normalised seller payoffs Φ_p and fairness metrics. HCDV matches the leave-one-out marginals most closely ($\rho = 0.94$) and yields the lowest payoff inequality among non-trivial methods, while requiring 8× less compute than MCDS. Notably, p_2 receives the largest share due to its distinctive "young–high-income–high-education" mix that most benefits the classifier, whereas p_3's overlapping demographics lead to a smaller, interpretable payoff.

6 Related Work

Data valuation. Shapley-based methods allocate data importance via co-operative game theory. Ghorbani and Zou (2019) introduce *Data-Shapley* with a Monte-Carlo permutation sampler later accelerated by truncated permutations (Jia et al. 2019a), hashing (Kwon and Zou 2021), and stratified sampling (Wu et al. 2023). Group-level variants attribute value to user-defined partitions (Jia et al. 2019b). HCDV differs by learning a multiscale hierarchy and incorporating a contrastive payoff, yielding provably tighter efficiency error with dramatically lower runtime.

Geometry-aware objectives. Contrastive representation learning enlarges inter-class margins (Oord, Li, and Vinyals 2018; Chen et al. 2020; Zhang et al. 2025). Recent work links geometry to data importance-e.g., influence-function contrastive weighting (Wang et al. 2020b)-but stops short of valuation. HCDV is, to our knowledge, the first to reward coalitions for geometric separation within SV framework.

7 Conclusion

We propose HCDV, combining contrastive embeddings with a multiscale coalition tree to make Shapley attribution geometry-aware and scalable. Under mild assumptions, we prove logarithmic surplus loss and sharp Monte–Carlo concentration (plus a top-k surrogate regret bound). Experiments show state-of-the-art valuation quality with 10–100× speedups, enabling augmentation filtering, streaming updates, and marketplace revenue sharing. Future work includes federated valuation and active data acquisition (Tao et al. 2023).

References

Andrews, J.; Zhao, D.; Thong, W.; Modas, A.; Papakyriakopoulos, O.; and Xiang, A. 2024. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems*, 36.

Bendechache, M.; Attard, J.; Ebiele, M.; and Brennan, R. 2023. A systematic survey of data value: Models, metrics, applications and research challenges. *IEEE Access*.

Bhardwaj, E.; Gujral, H.; Wu, S.; Zogheib, C.; Maharaj, T.; and Becker, C. 2024. Machine learning data practices through a data curation lens: An evaluation framework. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1055–1067.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, X.; Xiao, C.; Cao, W.; Zhang, W.; and Liu, Y. 2025. Framework and Pathway for the Construction of a Unified Data-Element Market in China. *Strategic Study of Chinese Academy of Engineering*, 27(1): 40–50.

Fan, Z.; Fang, H.; Zhou, Z.; Pei, J.; Friedlander, M. P.; Liu, C.; and Zhang, Y. 2022. Improving fairness for data valuation in horizontal federated learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2440–2453. IEEE.

Fleckenstein, M.; Obaidi, A.; and Tryfona, N. 2023. A review of data valuation approaches and building and scoring a data valuation model. *Harvard Data Science Review*, 5(1).

Garrido Lucero, F.; Heymann, B.; Vono, M.; Loiseau, P.; and Perchet, V. 2024. Du-shapley: A shapley value proxy for efficient dataset valuation. *Advances in Neural Information Processing Systems*, 37: 1973–2000.

Ghazikhani, A.; Monsefi, R.; and Sadoghi Yazdi, H. 2014. Online neural network model for non-stationary and imbalanced data stream classification. *International Journal of Machine Learning and Cybernetics*, 5: 51–62.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Goetz, J.; Malik, K.; Bui, D.; Moon, S.; Liu, H.; and Kumar, A. 2019. Active federated learning. *arXiv preprint arXiv:1909.12641*.

Hart, S. 1989. Shapley value. In *Game theory*, 210–216. Springer.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Gurel, N. M.; Li, B.; Zhang, C.; Spanos, C. J.; and Song, D. 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019b. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176. PMLR.

Jiang, K.; Liang, W.; Zou, J. Y.; and Kwon, Y. 2023. Opendataval: a unified benchmark for data valuation. *Advances in Neural Information Processing Systems*, 36: 28624–28647.

Jiang, L.; Wang, X.; Zhang, F.; and Zhang, C. 2025. Transforming time and space: efficient video super-resolution with hybrid attention and deformable transformers. *The Visual Computer*, 1–12.

Kwon, Y.; and Zou, J. 2021. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*.

Li, W.; Fu, S.; Zhang, F.; and Pang, Y. 2024. Data Valuation and Detections in Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12027–12036.

Liu, J.; Tong, R.; Shen, A.; Li, S.; Yang, C.; and Xu, L. 2025. MemeBLIP2: A novel lightweight multimodal system to detect harmful memes. In *IJCAI 2025: The First Workshop on Multimodal Knowledge and Language Modeling https://sites.google.com/view/ijcai-mklm/accepted-papers?authuser=0*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pei, J. 2020. A survey on data pricing: from economics to data science. *IEEE Transactions on knowledge and Data Engineering*, 34(10): 4586–4608.

Shen, L.; Sun, Y.; Yu, Z.; Ding, L.; Tian, X.; and Tao, D. 2024. On efficient training of large-scale deep learning models. *ACM Computing Surveys*, 57(3): 1–36.

Sim, R. H. L.; Xu, X.; and Low, B. K. H. 2022. Data Valuation in Machine Learning:" Ingredients", Strategies, and Open Challenges. In *IJCAI*, 5607–5614.

Tao, H.; Li, J.; Hua, Z.; and Zhang, F. 2023. DUDB: deep unfolding-based dual-branch feature fusion network for pansharpening remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Tong, R.; Xu, T.; Ju, X.; and Wang, L. 2025. Progress in medical ai: Reviewing large language models and multimodal systems for diagonosis. *AI Med*, 1(1): 165–186.

Wang, H.; and Zhang, F. 2024. Computing nodes for plane data points by constructing cubic polynomial with constraints. *Computer Aided Geometric Design*, 111: 102308.

Wang, J. T.; and Jia, R. 2023. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, 6388–6421. PMLR.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020a. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, 153–167.

Wang, Y.; Wang, H.; and Zhang, F. 2025. A Medical image segmentation model with auto-dynamic convolution and location attention mechanism. *Computer Methods and Programs in Biomedicine*, 261: 108593.

Wang, Z.; Zhu, H.; Dong, Z.; He, X.; and Huang, S.-L. 2020b. Less is better: Unweighted data subsampling via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), 6340–6347.

Whang, S. E.; Roh, Y.; Song, H.; and Lee, J.-G. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4): 791–813.

Wu, M.; Jia, R.; Lin, C.; Huang, W.; and Chang, X. 2023. Variance reduced Shapley value estimation for trustworthy data valuation. *Computers & Operations Research*, 159: 106305.

Xiao, C.; Hou, L.; Fu, L.; and Chen, W. 2025. Diffusion-Based Self-Supervised Imitation Learning from Imperfect Visual Servoing Demonstrations for Robotic Glass Installation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10401–10407. IEEE.

Xiao, C.; and Liu, Y. 2025. A multifrequency data fusion deep learning model for carbon price prediction. *Journal of Forecasting*, 44(2): 436–458.

Xiao, C.; et al. 2024. Confusion-resistant federated learning via diffusion-based data harmonization on non-IID data. *Advances in Neural Information Processing Systems*, 37: 137495–137520.

Xu, X.; Wu, Z.; Foo, C. S.; and Low, B. K. H. 2021. Validation free and replication robust volume-based data valuation. *Advances in Neural Information Processing Systems*, 34: 10837–10848.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9421–9431. IEEE Computer Society.

Yao, J.; Li, C.; and Xiao, C. 2024. Swift sampler: Efficient learning of sampler by 10 parameters. *Advances in Neural Information Processing Systems*, 37: 59030–59053.

Zhang, F.; Chen, G.; Wang, H.; Li, J.; and Zhang, C. 2023. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4496–4506.

Zhang, F.; Chen, G.; Wang, H.; and Zhang, C. 2024. CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media*, 10(3): 593–608.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, M.; Beltrán, F.; and Liu, J. 2023. A survey of data pricing for data marketplaces. *IEEE Transactions on Big Data*, 9(4): 1038–1056.

Zhang, X.; Zeng, F.; Quan, Y.; Hui, Z.; and Yao, J. 2025. Enhancing multimodal large language models complex reason via similarity computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(10), 10203–10211.

Zhao, Y.; Tan, C.; Shi, L.; Zhong, Y.; Kou, F.; Zhang, P.; Chen, W.; and Ma, C. 2025. Generative Recommender Systems: A Comprehensive Survey on Model, Framework, and Application. *Information Fusion*, 103919.