BEYOND POINTWISE ATTRIBUTION: RELATIONSHIP-AWARE TRAINING DATA ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training data attribution is commonly formulated by assigning each training example a scalar influence score for a target model behavior. This formulation treats influence as an intrinsic property of an individual sample, yet model parameters are learned by jointly optimizing over the entire training distribution. As a result, the contribution of one example is inherently conditional on the presence of others: perturbing one sample can change not only its own effect, but also how other samples contribute to the learned model. Existing pointwise attribution methods collapse these coupled effects into marginal scores and cannot explicitly represent substitution, reinforcement, or dependency among training examples. Motivated by this, we reformulate training data attribution from estimating isolated sample importance to characterizing how training examples influence one another's contribution to a target behavior. Instead of asking only which samples are important, we ask how the importance of one part of the training data changes when another part is perturbed. This leads to a novel *Relationship-Aware Training Data Attribution* task that captures compensatory, reinforcing, and suppressive dependencies among training samples. Specifically, we organize training samples into coherent groups according to their attribution-relevant representations and construct a query-conditioned asymmetric attribution graph over these groups. Each node represents a coherent region of the training data, while each directed edge measures how perturbing one group changes the attribution assigned to another. Since inter-sample effects are inherently asymmetric, we explicitly preserve their directionality rather than treating relationships as reciprocal. By modeling training influence at the level of structured and asymmetric group interactions rather than isolated examples, our approach provides a more faithful characterization of how different parts of the training data collectively shape model behavior.

1 INTRODUCTION

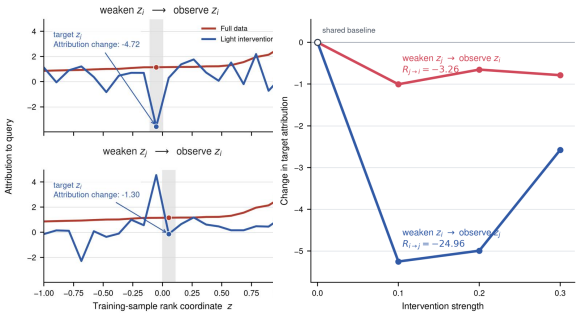
Training data attribution (TDA) has become an important tool for understanding how a model's learned behavior is shaped by its training data. Given a trained model and a target behavior, TDA seeks to identify which training examples are most responsible for that behavior. This objective underlies a broad family of approaches, including retraining-based methods (Ghorbani & Zou, 2019; Ilyas et al., 2022) and gradient-based estimators (Koh & Liang, 2017; Pruthi et al., 2020; Bae et al., 2024). By assigning influence to individual training examples, these methods provide useful tools for model interpretation, debugging, data valuation, and dataset analysis.

Despite their methodological differences, most existing TDA methods summarize each training example with a scalar attribution score for a given target behavior (Park et al., 2026). While useful for identifying important examples under a fixed training dataset, such scores do not reveal whether their importance is self-contained or supported by relationships with other training data. This distinction matters because learning is inherently joint, different samples or subsets may provide redundant evidence, compensate for one another, specialize to different input distribution (Guo et al., 2024).

Consequently, two examples with the same pointwise attribution can play very different roles, while an example with only moderate attribution may still be important if perturbing it substantially changes the contributions assigned to others. To expose this hidden structure, we intervene on one training example and measure how the attribution of another responds under the resulting

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.

counterfactual training context. Fig. 1 illustrates this phenomenon using two samples with identical baseline attribution scores: weakening z_i induces a much larger change in the attribution of z_j than the reverse intervention, and this discrepancy persists across increasing intervention strengths. These observations show that scalar attribution alone cannot fully characterize the role of a training example; understanding its dependencies on, and effects on, other training data is also necessary.



(a) (b)

Figure 1: Identical attribution scores can hide distinct directional relationships. (a) Weakening each sample produces different changes in the attribution of the other. (b) The discrepancy persists across various intervention strengths.

This observation suggests that training-data influence is not fully characterized by a collection of isolated attribution scores. Conventional TDA answers *which training data important*. We instead ask *how the contribution of each part of the training data is conditioned by, and related to, the contributions of others*. Such relationships characterize whether an example's apparent importance is largely self-contained or depends on the presence and role of other training data, and they are directional because different data can play different roles in shaping one another's contributions. We therefore formulate ***Relationship-Aware Training Data Attribution***, which extends pointwise attribution from isolated importance estimation to the characterization of directional dependencies among training data.

Importantly, Relationship-Aware TDA is distinct from conventional interaction analysis, which typically asks whether multiple subsets exhibit a joint effect beyond their individual contributions (Hammoudeh & Lowd, 2024). Our interest is instead in the directional dependency between their attributed roles: whether and how the contribution of one group is conditioned by another. We represent this structure as a query-conditioned directed attribution graph, where nodes retain conventional group-level attribution and directed edges encode inter-group dependencies. The resulting graph therefore complements the pointwise view of *which data are important* with a relational view of *how their contributions are connected*.

To recover these otherwise hidden relationships, we instantiate the formulation at the group level, where coherent groups of training examples serve as stable and interpretable units of analysis. While pairwise sample relationships are conceptually valid, exhaustively estimating them is both computationally prohibitive and statistically noisy; grouping examples with similar attribution-relevant representations provides a tractable mesoscopic view of how different regions of the training data depend on one another. We then use controlled source-group interventions to probe how target-group attribution evolves across counterfactual training contexts. Progressive source weakening yields a response trajectory for each ordered group pair, from which we estimate the corresponding directional relation, while a source-conditioned fixed reference population prevents changes in the normalization set from being mistaken for genuine attribution responses. Our analysis shows that attribution is context-dependent and that these relationships need not be reciprocal, while experiments demonstrate that groups with similar pointwise attribution can exhibit markedly different and reproducible relational patterns. Our main contributions can be summarized as:

- We reveal a fundamental limitation of conventional training data attribution: similar scalar scores can conceal substantially different dependencies on the surrounding training data. This motivates ***Relationship-Aware Training Data Attribution***, which extends attribution from isolated importance to the relational roles of training data.
- We develop a query-conditioned group-level framework for characterizing these relationships. It represents conventional group importance as node attributes and directional inter-group dependencies as edges, and estimates the latter through graded source-group interventions with source-conditioned fixed-reference normalization.
- We theoretically establish the context dependence and non-reciprocity of training-data attribution, and empirically show that the resulting relationships are reproducible, cannot be

recovered from pointwise attribution alone, and provide complementary information for anticipating how attribution reorganizes under dataset interventions.

2 Related Works

TDA aims to quantify how training examples contribute to a learned model or a particular model behavior. Existing approaches span several methodological families. Retraining-based methods directly characterize the effect of modifying the training set. For example, Data Shapley (Ghorbani & Zou, 2019) estimates the value of individual training samples through their marginal contributions across training subsets, while Datamodels (Ilyas et al., 2022) learn predictive models that approximate how different training subsets affect model behavior. Although conceptually direct, these approaches require training many models and are therefore computationally expensive.

A second family estimates training influence from model gradients or optimization information. Influence Functions (Koh & Liang, 2017) approximate the effect of infinitesimally upweighting a training example using local curvature of the training objective, while TracIn (Pruthi et al., 2020) measures influence by accumulating gradient alignment over training checkpoints. Moreover, TRAK (Park et al., 2023) combines gradient-based representations with random projections and kernel-style approximations to efficiently attribute model predictions. Motivated by this work, related extenstions have adapted such attribution mechanisms to generation models, and their differnces are the designed estimators (Georgiev et al., 2023; Zheng et al., 2024).

Beyond improving the attribution estimator itself, several works expand the attribution unit or training setting. Butler et al. (2026) proposed to find the extend marginal attribution by identifying non-additive joint effects among training examples. Basu et al. (2020) develop a second-order group influence approximation to better estimate the effect of removing a subset of training data, while Generalized Group Data Attribution (GGDA) (Ley et al., 2024) treats groups as attribution units to improve the efficiency of existing TDA estimators. Distributional Training Data Attribution (d-TDA) (Mlodozeniec et al., 2026) instead accounts for stochasticity in training by characterizing how distributions of model outputs vary with the training data. Despite these extensions, existing TDA methods still focus primarily on data-to-output influence and implicitly treat each attribution score as independently characterizable. Consequently, they provide limited insight into how relationships among training examples shape and redistribute attribution.

3 Preliminaries and Theoretical Motivation

Pointwise Training Data Attribution. Let $\mathcal{D} = \{z_i\}_{i=1}^N$ denote a training dataset, where $z_i \in \mathcal{Z}$ is the i-th training example. We denote by $\theta^*(\mathcal{D}) \in \mathbb{R}^{d_{\text{param}}}$ the model parameters obtained by training on $\mathcal{D}$. Given a target query q, conventional TDA methods assign each training example z_i a scalar attribution score $a_i(q; \theta^*) \in \mathbb{R}$, which measures the contribution of z_i to the model behavior associated with q. Different TDA methods differ in how a_i is estimated, but they typically represent the role of each training example through an individual scalar score.

Why Attribution Is Context-Dependent. The scalar formulation above is convenient for ranking individual training examples, but model parameters are learned through a joint optimization over the entire training set. This raises a fundamental question:

Can the attribution of the training data remain independent of the other examples that jointly determine the learned model?

To formalize this question, consider the weighted empirical-risk objective can be formulated as:

$$\theta^*(\mathbf{w}) = \arg\min_{\theta} \frac{1}{N} \sum_{k=1}^{N} w_k \, \ell(z_k; \theta), \tag{1}$$

where $\ell(z_k; \theta)$ is the training loss of z_k, $w_k \geq 0$ is its sample weight, and $\mathbf{w} = (w_1, \ldots, w_N)$. The original training problem corresponds to $\mathbf{w} = \mathbf{1}$.

Theorem 1 (Context Dependence of Training Attribution) *Assume that the weighted training objective in Eq. (1) is twice continuously differentiable around a local optimum $\theta^*(\mathbf{w})$, with a non-*

singular Hessian $H(\mathbf{w})$. Then, we can obtain:

$$\frac{\partial \theta^*}{\partial w_i} = -\frac{1}{N} H(\mathbf{w})^{-1} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big). \tag{2}$$

Consequently, whenever $\nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big) \neq 0$, changing the weight of z_i necessarily changes the jointly optimized model parameters. For any differentiable attribution function $a_j(q; \theta^*)$, its response to the weight of z_i is therefore

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_i; \theta^*). \tag{3}$$

Hence, the attribution of z_j generally depends on the contribution of other training examples.

Theorem 1 shows that perturbing z_i changes the shared model parameters through the optimization geometry induced by the entire training set. Consequently, even if z_j itself is unchanged, its attribution may vary because it is evaluated under a different jointly optimized model. The contribution of a training example is therefore not an intrinsic property of that example alone, but is conditioned on the surrounding training data. This theorem further immediately raises a second question: if two training examples influence each other's attribution through the jointly optimized model, must these dependencies be reciprocal? In other words, does perturbing one example affect another in the same way as the reverse perturbation?

Theorem 2 (Asymmetry of Cross-Example Attribution Dependence) *Under the conditions of Theorem 1, the attribution response of z_j to a perturbation of z_i is*

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_i; \theta^*), \tag{4}$$

whereas the reverse response is

$$\frac{\partial a_i}{\partial w_j} = -\frac{1}{N} \nabla_\theta a_i(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_j; \theta^*). \tag{5}$$

Because the two responses involve different training gradients and different attribution sensitivities, they are not generally equal, which can be formulated as $\frac{\partial a_j}{\partial w_i} \neq \frac{\partial a_i}{\partial w_j}$. Hence, cross-example attribution dependence is generally directional rather than symmetric.

Theorem 2 shows that the relationship between training examples is not a simple pointwise quantity, but a directional dependency between pairs of samples. Directly modeling such dependencies at the sample level is therefore fundamentally difficult to scale. For a dataset with N training examples, a complete characterization requires up to $N(N-1)$ directed relationships. More importantly, these relationships are not directly observed: obtaining the supervision for a single relation requires perturbing one part of the training data and reevaluating how the attribution of another part changes. Exhaustively constructing such supervision would therefore require an impractical number of data perturbations and attribution evaluations.

Taken together, these theorems show that TDA is context-dependent and directional, and cannot be fully captured by isolated sample-level scores. Since exhaustively modeling sample-wise relationships is prohibitively expensive, we instead pursue a tractable relationship-aware formulation over structured groups of training examples. The detailed proof can be found in ***Appendix***.

4 Empirical Motivation

The previous section establishes the context dependence and asymmetry of training data attribution. We now empirically examine whether these properties emerge in practice and whether relational effects can be explained by conventional pointwise attribution scores.

Attribution Dependence. We first examine whether perturbing one part of the training data produces structured changes in the attribution assigned to other parts. Specifically, we evaluate the effect in CIFAR-2 (Krizhevsky & Hinton, 2009) based on TRAK. For each fixed query images, we identify the top-ranked training samples under the attribution scores. We then restrict the sample-level interaction analysis to the query-specific high-attribution set and compute a directed response matrix.

For a fixed query, we compute the directed response $R_{i\to j}$, where the row index i denotes the removed source sample and the column index j denotes the affected target sample. Collecting these responses yields a query-conditioned response matrix, and the results can be found in Fig. 2. We first observe that removing a single training sample can substantially change the attribution assigned to other samples. Importantly, these changes are not simply determined by the original attribution ranking: samples with similar initial importance can respond very differently, while samples with substantially different initial importance may exhibit comparable changes. This suggests that the relational response cannot be explained by a simple linear relationship with pointwise attribution. Moreover, the response matrix is clearly asymmetric, with $R_{i\to j}$ generally differing from $R_{j\to i}$, indicating that the dependency between training samples is directional rather than reciprocal. Together, these empirical observations support the context dependence and asymmetry established by our theoretical analysis.

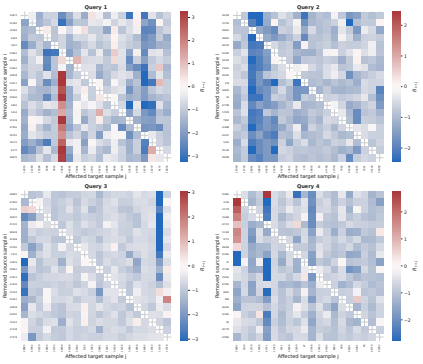
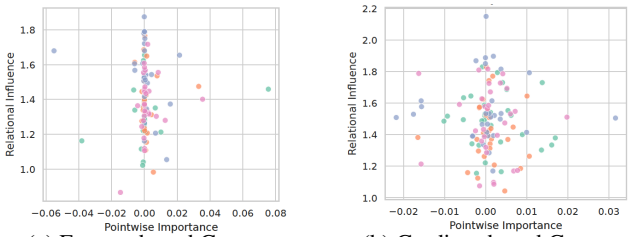


Figure 2: Directed attribution response matrices.

Pointwise Importance vs. Relational Influence. These results raise a further question: can the observed relational effects be inferred from conventional pointwise attribution scores? Unlike traditional TDA, which assigns each training sample a single scalar score, relationship-aware attribution characterizes pairwise dependencies among samples and therefore induces an $N(N-1)$ directed relation matrix. Directly comparing such high-dimensional relational information with pointwise attribution at the sample level is both computationally expensive and difficult to interpret. We therefore perform this comparison at the group level.

We define the pointwise importance P of G as its mean TRAK score as $P(G;q) = \frac{1}{|G|}\sum_{i\in G} a_i(q)$, where $a_i(q)$ denotes the TRAK attribution score of training sample z_i to q. Its relational influence O as the total magnitude of its outgoing responses, $O(G;q) = \sum_{G'\neq G} |R_{G\to G'}(q)|$. The former measures how important G itself is to the query, while the latter measures how strongly perturbing G changes the attribution of other groups. We partition the training samples into 30 groups using either feature- or gradient-based representations, with the results shown in Fig. 3. Groups with high pointwise attribution are not necessarily the groups that most strongly reorganize attribution over other groups. This indicates that relational influence captures a structure distinct from scalar pointwise attribution scores.



(a) Feature-based Group (b) Gradient-based Group

Figure 3: Pointwise importance vs. relational influence under different groupings. Colors denote different groupings.

5 Methodology

5.1 Overview

Conventional TDA explains a model prediction by assigning an individual importance score to each training example. However, as established above, the contribution of a training example is conditioned on the rest of the training data: changing one part of the training set can alter how another part contributes to the same query prediction. Therefore, explaining a prediction solely through isolated importance scores may overlook how training examples jointly contribute to the model behavior. We aim to extend TDA from pointwise importance to relationship-aware attribution, and our overview is illustrated in Fig. 4. Specifically, for a given query q, we seek to attribute not only *which* parts of the training data are important, but also *how the contribution of one part is affected by another*.

To operationalize this formulation, we first partition the training data into K groups based on their projected-gradient representations. For each source group, we then construct controlled counterfac-

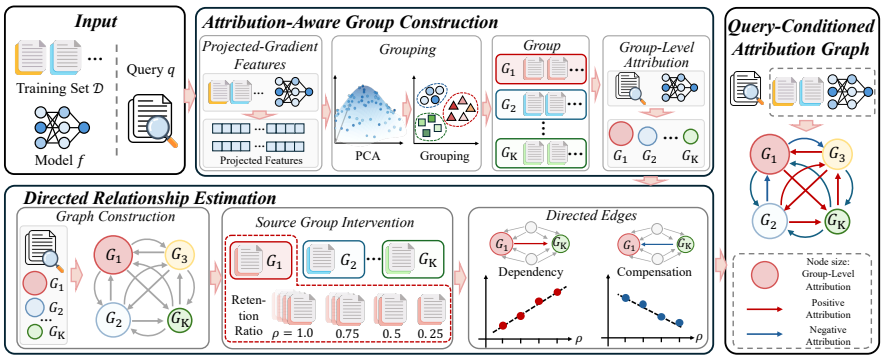


Figure 4: The overview of our proposed method

tual training contexts by progressively varying its presence while keeping the remaining data fixed, and evaluate the attribution of every target group to the same query. The sensitivity of a target group's attribution to this intervention defines a directed edge from the source group to the target group. Repeating this procedure for all source groups yields a query-conditioned directed attribution graph. In this way, the resulting graph extends conventional attribution from isolated importance scores to a relational explanation: node-level attribution identifies which regions of the training data contribute to the query, while directed edges reveal how these attributed contributions are modulated by the surrounding training data. Together, they provide a relationship-aware attribution of how different regions of the training set collectively support the query prediction.

5.2 Attribution-Aware Group Construction

The preceding analysis shows that training-data attribution is inherently context-dependent. We therefore use a query-conditioned directed graph, where nodes represent groups of training examples and directed edges capture how perturbing one group changes the attribution of another. Constructing such a graph at the sample level would require $O(N^2)$ directed relationships, so we group training examples according to their attribution roles. For each training example z_i, we use the projected-gradient representation (Park et al., 2023) as:

$$\phi_i = \frac{1}{\sqrt{d_{\text{param}}}} P^\top \nabla_\theta m(z_i; \theta^*), \tag{6}$$

where P is a random projection matrix and $m(z_i; \theta^*)$ is the classification margin under the baseline model. The representation ϕ_i captures the parameter directions through which z_i affects the model output, so examples with similar projected gradients are expected to play similar attribution roles.

To reduce the influence of gradient magnitude and high-dimensional variation, we normalize ϕ_i and apply PCA, which can be formulated as $\bar{\phi}_i = \text{PCA}\left(\frac{\phi_i}{\|\phi_i\|_2}\right)$. We then apply K-means clustering, with the assignment of z_i given by $c_i = \arg\min_{k \in \{1,\dots,K\}} \|\bar{\phi}_i - \mu_k\|_2^2$, where μ_k is the centroid of the k-th cluster. This yields the partition $\mathcal{G} = \{G_1, \dots, G_K\}$, where $G_k = \{z_i : c_i = k\}$. The partition is constructed once from the baseline model and shared across queries, while group attributions and inter-group relationships are query-specific.

After constructing the groups, we quantify their contributions to a query q. Let $\mathcal{D}'$ denote a training context before or after intervention, and let $\theta^*(\mathcal{D}')$ be the corresponding learned model. The pointwise attribution of $z_i \in \mathcal{D}'$ is

$$a_i(q; \mathcal{D}') = \mathcal{A}\big(z_i, q; \theta^*(\mathcal{D}'), \mathcal{D}'\big), \tag{7}$$

where $\mathcal{A}$ denotes a base pointwise TDA estimator; we use TRAK in our implementation.

Because attribution scores from different learned models may differ in scale, we use normalized within-context ranks rather than absolute values:

$$\widetilde{a}_i(q; \mathcal{D}') = \text{PerRank}\big(a_i(q; \mathcal{D}')\big) = \frac{1}{N'} \sum_{z_j \in \mathcal{D}'} \mathbb{I}\left[a_j(q; \mathcal{D}') \le a_i(q; \mathcal{D}')\right], \tag{8}$$

where $N' = |\mathcal{D}'|$. A larger $\widetilde{a}_i(q; \mathcal{D}')$ indicates a higher relative contribution to the query. The attribution of target group G_t is then defined as the average attribution of its members as:

$$A_t(q; \mathcal{D}') = \frac{1}{|G_t|} \sum_{z_i \in G_t} \widetilde{a}_i(q; \mathcal{D}'). \tag{9}$$

Under the original training context, $A_t(q; \mathcal{D})$ serves as the node weight of G_t, indicating its contribution to the query.

After defining the query-specific node weights, we further characterize the relationships between groups. The node weight only measures how strongly a group G_t contributes to the query in isolation, but does not reveal whether this contribution depends on the presence of other groups. To capture such inter-group dependencies, we construct directed edges that describe how the attribution of a target group changes when a source group is perturbed. These relationships are generally asymmetric, since perturbing G_s and observing the response of G_t is not equivalent to perturbing G_t and observing the response of G_s.

Specifically, for a source group G_s, we progressively weaken its presence while keeping all other groups unchanged. Let $\rho \in (0, 1]$ denote the retention ratio of G_s, and let $\widetilde{G}^{(r)}_{s,\rho} \subseteq G_s$ be a randomly retained subset satisfying $|\widetilde{G}^{(r)}_{s,\rho}| \approx \rho |G_s|$, where r indexes different random realizations. The resulting training context can be formulated as:

$$\mathcal{D}^{(r)}_{s,\rho} = (\mathcal{D} \setminus G_s) \cup \widetilde{G}^{(r)}_{s,\rho}. \tag{10}$$

Here, $\rho = 1$ corresponds to the original training set, while smaller values of ρ represent stronger interventions on G_s.

Each intervention defines a counterfactual training context $\mathcal{D}^{(r)}_{s,\rho}$, which induces a corresponding learning outcome under the same training procedure. For the same query q, we evaluate the attribution of each target group under this interventional context as:

$$A^{(s,\rho,r)}_t(q) = A_t\big(q; \mathcal{D}^{(r)}_{s,\rho}\big). \tag{11}$$

The change from the baseline attribution $A_t(q; \mathcal{D})$ captures how the contribution of G_t responds to perturbations of G_s, and therefore provides the basis for the directed relationship $G_s \to G_t$.

5.3 Directed Relationship Estimation

The relationship-aware attribution defined above describes how a target group G_t responds when a source group G_s is weakened. We use this response to construct the directed edge from G_s to G_t. Specifically, we define $\Delta_{s\to t}(q; \rho, r) = A^{(s,\rho,r)}_t(q) - A_t(q; \mathcal{D})$, which measures the change in the contribution of G_t relative to its baseline node weight.

Rather than relying on a single intervention, we evaluate the response across multiple retention ratios and fit its variation with respect to the intervention strength $1 - \rho$:

$$\Delta_{s\to t}(q; \rho, r) = \alpha_{s\to t}(q) + R_{s\to t}(q)(1 - \rho) + \epsilon_{s\to t}(q; \rho, r), \tag{12}$$

where $\alpha_{s\to t}(q)$ is the intercept, $R_{s\to t}(q)$ measures the local first-order sensitivity of the target-group attribution to the source-group intervention, and $\epsilon_{s\to t}(q; \rho, r)$ denotes the residual variation not explained by the linear response. We use $R_{s\to t}(q)$ as the directed edge weight. A positive value indicates that G_t becomes more influential as G_s is weakened, suggesting a compensatory effect, whereas a negative value indicates that the contribution of G_t decreases, suggesting a reinforcing or dependency-like effect. Because the reverse direction is obtained through a different intervention, the relationship is generally asymmetric, i.e., $R_{s\to t}(q) \neq R_{t\to s}(q)$. Repeating this procedure over all ordered group pairs yields the query-conditioned relationship matrix as:

$$\mathbf{R}(q) = [R_{s\to t}(q)]^K_{s,t=1}. \tag{13}$$

Together with the group partition $\mathcal{G}$ and the baseline node attributions $\mathbf{A}(q) = [A_1(q; \mathcal{D}), \ldots, A_K(q; \mathcal{D})]$, this defines the relationship-aware attribution graph $\mathcal{H}(q) = \big(\mathcal{G}, \mathbf{A}(q), \mathbf{R}(q)\big)$.

In this way, the constructed graph jointly captures both node-level importance and edge-level relational influence, providing a unified view of how different training-data groups contribute to and interact in shaping the query prediction. The detailed procedure is provided in the ***Appendix***.

6 EXPERIMENT

6.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate on six image-classification benchmarks: CIFAR-2, CIFAR-10, ArtBench-2, ArtBench-5, ArtBench-10, and CelebA-Male. CIFAR-2 is constructed from the *automobile* and *horse* classes of CIFAR-10 (Krizhevsky & Hinton, 2009); ArtBench subsets are derived from ArtBench-10 (Liao et al., 2022); and CelebA-Male is derived from CelebA (Liu et al., 2015). CIFAR-2 uses a ResNet-9-style model (He et al., 2016), while the remaining datasets use a compact CNN. All baseline and intervention models use identical training configurations and random seeds. Full implementation details are provided in ***Appendix***.

Evaluation metrics. We report three groups of diagnostics. For group-level ranking behavior, G-rank ρ is the Spearman correlation between target-group rankings before and after source-group perturbation, and G-Rep@k is the fraction of originally top-k target groups replaced after the perturbation. For directional structure, F/R ρ compares forward and reverse responses, while Opp. reports the fraction of directed group pairs with opposite signs. For reproducibility and pointwise reducibility, Repeat ρ measures the rank correlation of relational edge strengths across repeated perturbation estimates, and Pt.-rel. ρ compares scalar group attribution with outgoing relational influence. Formal definitions are provided in ***Appendix***.

6.2 EXPERIMENTAL RESULTS

We first ask whether perturbing a source group produces a measurable and reproducible change in the attribution ranking of target groups. The results can be found in Tab. 1, it can be noticed that our method can effectively extract the attribution relationships among the training data. The consistent changes in target-group attribution under source-group interventions, together with their reproducibility across repeated perturbations, further indicate that such relationships are not artifacts of a particular intervention draw, but reflect genuine dependencies in how different regions of the training set jointly shape model behavior.

Table 1: Group-level ranking and repetition correlations across different datasets.

Dataset	G-rank $\rho\uparrow$	G-Rep@10 $\downarrow$	G-Rep@20 $\downarrow$	Repeat $\rho\uparrow$
CIFAR-2	0.434	0.452	0.246	0.476
CIFAR-10	0.800	0.256	0.140	0.867
ArtBench-2	0.869	0.174	0.109	0.738
ArtBench-5	0.765	0.290	0.156	0.783
ArtBench-10	0.798	0.263	0.122	0.885
CelebA-Male	0.878	0.159	0.099	0.809

We test whether different source groups induce distinct attribution-response patterns, with results summarized in Tab. 2. For each source group, its response profile over the remaining groups is used to retrieve the source. Top-1 retrieval measures whether the correct source ranks first, while MRR reflects its overall ranking quality. If responses were dominated by a generic data-removal effect, different sources would yield similar profiles and poor retrieval. Instead, the strong Top-1 and MRR scores indicate distinct and reproducible source-specific response patterns, supporting that the directed graph captures meaningful relational structure.

Table 2: Source-specific response retrieval.

Dataset	Top-1 retrieval $\uparrow$	MRR $\uparrow$
CIFAR-10	0.925	0.950
ArtBench-2	0.597	0.708
ArtBench-5	0.583	0.716
ArtBench-10	0.914	0.945
CelebA-Male	0.828	0.891

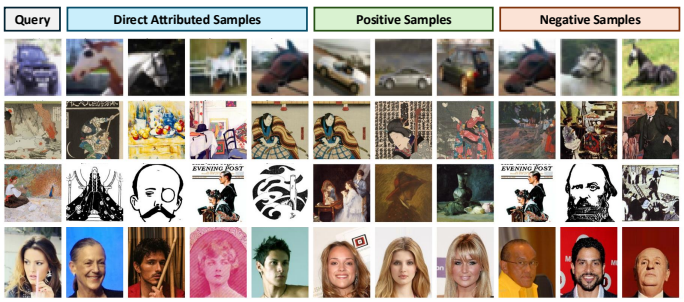


Figure 5: The qualitative results.

Fig. 5 qualitatively compares pointwise with our method. Direct attribution highlights individual training examples with high scores for the query, but these examples can still be visually or semantically mismatched with the query. We use the group containing the query as the target group, and then identify source groups whose perturbation induces strong positive or negative responses in that target group. Strong positive and strong

negative responses correspond to compensation-like and dependency-like relationships, respectively. The representative samples from these source groups are often more consistent with the query-level structure and reveal relational patterns missed by pointwise attribution.

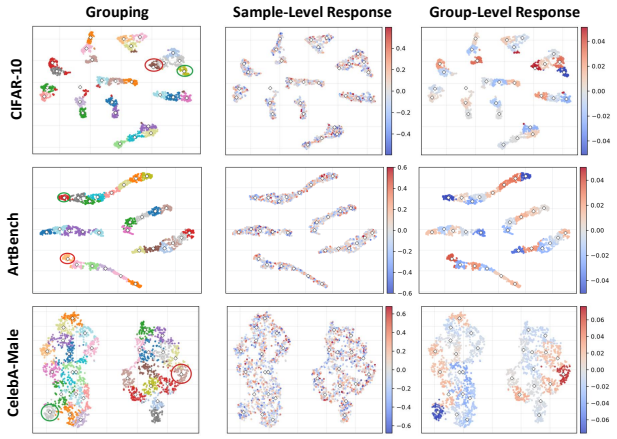


Figure 6: Visualization of attribution responses. In the grouping column, colors denote different groups; red and green circles mark the groups involved in the strongest positive and negative relations, and "◇" denotes the group centroid. In the response columns, red and blue indicate positive and negative attribution changes, respectively.

Fig. 6 visualizes the attribution response structure in the shared t-SNE space. At the sample level, attribution changes are scattered across individual examples, making stable relational patterns difficult to identify. After group-level aggregation, however, the response structure becomes substantially clearer. Some groups exhibit increased attribution after the source group is perturbed, meaning that they become relatively more important for the query when part of the source-group evidence is removed; we interpret this as a compensatory or substitutive response. In contrast, groups with decreased attribution become less important after the source perturbation, suggesting that their contribution is coupled with, or partly supported by, the presence of the source group. These results support our group-level formulation by showing that perturbing one training region can induce coherent changes in the relative attribution of other regions.

We further evaluate the robustness of our relationship-aware attribution method under graded source-group interventions. For each selected relation, we progressively vary the source-group retention ratio ρ, retrain the model, and measure the mean attribution rank of the target group. As shown in Fig. A5, strong positive and negative relations exhibit clear and systematic response trends, while low-magnitude relations show weaker or less structured changes. We use linear, quadratic, and isotonic fits to characterize the shape of these response trajectories. Although the exact response form varies across relations, the overall direction remains stable across intervention strengths and random subsets, indicating that the identified relationships are reproducible. The other experiments can be found in ***Appendix***.

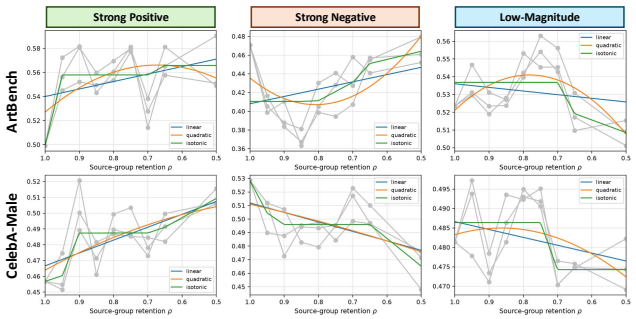


Figure 7: The robustness experiments of our method under different ρ.

7 Conclusions

In this work, we rethink training data attribution from a relational perspective. Conventional TDA asks which training examples are important, but largely treats their contributions as isolated scalar quantities. We argue that this view is incomplete because the contribution of one part of the training data can change when another part is perturbed, making attribution inherently contextual and relational. Based on this perspective, we introduce Relationship-Aware Training Data Attribution to characterize such inter-data dependencies. We realize this formulation through a query-conditioned directed graph over training-data groups, where edges capture asymmetric attribution responses. Theoretical analysis and experiments across multiple datasets show that these relationships are reproducible, directional, and distinct from conventional pointwise attribution.

AI USE STATEMENT

In this work, we used generative AI tools only for limited writing assistance, including language polishing, grammar correction, and stylistic refinement. We did not use generative AI tools for technical content, method design, experiments, or result analysis. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

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

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2020.

Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms. *Advances in Neural Information Processing Systems*, 38:72306–72340, 2026.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

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

Shangmin Guo, Yi Ren, Stefano Albrecht, and Kenny Smith. lpntk: Better generalisation with less data via sample interaction during learning. In *International Conference on Learning Representations*, volume 2024, pp. 51541–51574, 2024.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

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

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Dan Ley, Suraj Srinivas, Shichang Zhang, Gili Rusak, and Himabindu Lakkaraju. Generalized group data attribution. *arXiv preprint arXiv:2410.09940*, 2024.

Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Bruno Mlodozeniec, Isaac Reid, Sam Power, David Krueger, Murat Erdogdu, Richard Turner, and Roger Grosse. Distributional training data attribution: What do influence functions sample? *Advances in Neural Information Processing Systems*, 38:10057–10091, 2026.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Yong-Hyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Naoki Murata, WeiHsiang Liao, Woosung Choi, Kin Wai Cheuk, Junghyun Tony Koo, and Yuki Mitsufuji. Concept-trak: Understanding how diffusion models learn concepts through concept attribution. In *International Conference on Learning Representations*, volume 2026, pp. 93014–93041, 2026.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in neural information processing systems*, volume 33, pp. 19920–19930, 2020.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 18417–18452, 2024.

APPENDIX

A PROOFS OF MAIN THEOREMS

This section provides the theoretical analysis for modeling TDA as a context-dependent and directional quantity. We first show that the attribution assigned to a training example is generally not an intrinsic property of that example alone. Perturbing the contribution of another training example changes the jointly optimized model and can consequently alter its attribution. We then establish that such cross-example dependence is not, in general, symmetric, since the response in each direction is governed by a different combination of training gradients and attribution sensitivities.

Theorem 1 (Context Dependence of Training Attribution) *Assume that the weighted training objective*

$$\theta^*(\mathbf{w}) = \arg\min_{\theta} \frac{1}{N} \sum_{k=1}^{N} w_k \, \ell(z_k; \theta) \tag{A1}$$

is twice continuously differentiable around a local optimum $\theta^*(\mathbf{w})$, with a nonsingular Hessian $H(\mathbf{w})$. Then,

$$\frac{\partial \theta^*}{\partial w_i} = -\frac{1}{N} H(\mathbf{w})^{-1} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big). \tag{A2}$$

Consequently, whenever $\nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big) \neq 0$, changing the weight of z_i changes the jointly optimized model parameters to first order. For any differentiable attribution function $a_j(q; \theta^*)$, its response to the weight of z_i is

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_i; \theta^*). \tag{A3}$$

Hence, the attribution of z_j generally depends on the presence and contribution of other training examples.

Proof. Define the weighted empirical-risk objective as:

$$F(\theta, \mathbf{w}) = \frac{1}{N} \sum_{k=1}^{N} w_k \, \ell(z_k; \theta). \tag{A4}$$

By definition, $\theta^*(\mathbf{w})$ is a local minimizer of $F(\theta, \mathbf{w})$. Therefore, it satisfies the first-order stationarity condition as follows:

$$\nabla_\theta F\big(\theta^*(\mathbf{w}), \mathbf{w}\big) = \frac{1}{N} \sum_{k=1}^{N} w_k \nabla_\theta \ell\big(z_k; \theta^*(\mathbf{w})\big) = 0. \tag{A5}$$

The Hessian of the weighted objective at the local optimum can be formulated as:

$$H(\mathbf{w}) = \nabla_\theta^2 F\big(\theta^*(\mathbf{w}), \mathbf{w}\big) = \frac{1}{N}\sum_{k=1}^{N} w_k \nabla_\theta^2 \ell\big(z_k; \theta^*(\mathbf{w})\big). \tag{A6}$$

Since $H(\mathbf{w})$ is nonsingular by assumption, the implicit function theorem guarantees that $\theta^*(\mathbf{w})$ is locally differentiable with respect to $\mathbf{w}$.

Differentiating Eq. (A5) with respect to w_i and applying the chain rule can be given as:

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_i} \nabla_\theta F\big(\theta^*(\mathbf{w}), \mathbf{w}\big) \\ &= \nabla_\theta^2 F\big(\theta^*(\mathbf{w}), \mathbf{w}\big) \frac{\partial \theta^*(\mathbf{w})}{\partial w_i} + \frac{\partial}{\partial w_i} \nabla_\theta F(\theta, \mathbf{w})\Big|_{\theta=\theta^*(\mathbf{w})}. \end{aligned} \tag{A7}$$

The first term in Eq. (A7) can be formulated as $H(\mathbf{w})\,\partial\theta^*/\partial w_i$. For the second term, holding θ fixed yields $\frac{1}{N}\nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big)$. Therefore, we can reformulate Eq. (A7) as follows:

$$H(\mathbf{w}) \frac{\partial \theta^*(\mathbf{w})}{\partial w_i} + \frac{1}{N} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big) = 0. \tag{A8}$$

Then, we can get:

$$\frac{\partial \theta^*}{\partial w_i} = -\frac{1}{N} H(\mathbf{w})^{-1} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big). \tag{A9}$$

When $\nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big) \neq 0$, the nonsingularity of $H(\mathbf{w})$ implies $H(\mathbf{w})^{-1} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big) \neq 0$, and hence $\partial\theta^*/\partial w_i \neq 0$. Thus, perturbing the weight of z_i changes the learned parameters to first order.

Now consider a differentiable attribution function $a_j(q; \theta^*(\mathbf{w}))$. Its dependence on w_i is mediated through the learned parameters, so the chain rule gives as:

$$\frac{\partial a_j}{\partial w_i} = \nabla_\theta a_j\big(q; \theta^*(\mathbf{w})\big)^\top \frac{\partial \theta^*(\mathbf{w})}{\partial w_i}. \tag{A10}$$

Substituting the above equations, we can obtain:

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j\big(q; \theta^*(\mathbf{w})\big)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell\big(z_i; \theta^*(\mathbf{w})\big). \tag{A11}$$

Therefore, even when z_j itself is unchanged, its attribution can vary as the contribution of another training example z_i is perturbed. This establishes the context dependence of training attribution. □

Theorem 2 (Asymmetry of Cross-Example Attribution Dependence) *Under the conditions of the preceding theorem, the attribution response of z_j to a perturbation of z_i is*

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_i; \theta^*), \tag{A12}$$

whereas the reverse response is

$$\frac{\partial a_i}{\partial w_j} = -\frac{1}{N} \nabla_\theta a_i(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_j; \theta^*). \tag{A13}$$

Because the two responses depend on different training gradients and different attribution sensitivities, they are not generally equal. Hence, cross-example attribution dependence is generally directional rather than symmetric.

Proof. For notational simplicity, let $\theta^* = \theta^*(\mathbf{w})$. From the preceding theorem, the response of the attribution of z_j to a perturbation of w_i can be formulated as:

$$\frac{\partial a_j}{\partial w_i} = -\frac{1}{N} \nabla_\theta a_j(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_i; \theta^*). \tag{A14}$$

Reversing the roles of i and j gives:

$$\frac{\partial a_i}{\partial w_j} = -\frac{1}{N} \nabla_\theta a_i(q; \theta^*)^\top H(\mathbf{w})^{-1} \nabla_\theta \ell(z_j; \theta^*). \tag{A15}$$

For the two directional responses to be equal, the following condition must hold:

$$\nabla_\theta a_j(q;\theta^*)^\top H(\mathbf{w})^{-1}\nabla_\theta \ell(z_i;\theta^*) = \nabla_\theta a_i(q;\theta^*)^\top H(\mathbf{w})^{-1}\nabla_\theta \ell(z_j;\theta^*). \tag{A16}$$

However, the assumptions of the preceding theorem impose no such constraint. The forward response couples the attribution sensitivity of z_j, $\nabla_\theta a_j(q;\theta^*)$, with the training gradient of z_i, $\nabla_\theta \ell(z_i;\theta^*)$, whereas the reverse response couples $\nabla_\theta a_i(q;\theta^*)$ with $\nabla_\theta \ell(z_j;\theta^*)$. These two pairs of vectors are generally different, and therefore Eq. (A16) cannot not guaranteed to hold.

To make the lack of symmetry explicit, we consider a simplified one-dimensional example with $N = 2$, and we design the loss function as $\ell(z_1;\theta) = \frac{1}{2}\theta^2 + \theta$ and $\ell(z_2;\theta) = \frac{1}{2}\theta^2 - \theta$ for the two samples, respectively. Then, for arbitrary weights w_1 and w_2, the weighted objective is

$$F(\theta, \mathbf{w}) = \frac{1}{2}\left[w_1\ell(z_1;\theta) + w_2\ell(z_2;\theta)\right] = \frac{w_1 + w_2}{4}\theta^2 + \frac{w_1 - w_2}{2}\theta. \tag{A17}$$

The first-order optimality condition gives

$$\frac{\partial F}{\partial \theta} = \frac{w_1 + w_2}{2}\theta + \frac{w_1 - w_2}{2} = 0, \tag{A18}$$

and therefore

$$\theta^*(w_1, w_2) = \frac{w_2 - w_1}{w_1 + w_2}. \tag{A19}$$

At the reference point $w_1 = w_2 = 1$, we have $\theta^*(1,1) = 0$. Then, we define the differentiable attribution functions $a_1(q;\theta) = \theta$, and $a_2(q;\theta) = 2\theta$.

Differentiating Eq. (A19) with respect to the two weights and evaluating at $w_1 = w_2 = 1$ yields $\frac{\partial\theta^*}{\partial w_1}\Big|_{(1,1)} = -\frac{1}{2}$, and $\frac{\partial\theta^*}{\partial w_2}\Big|_{(1,1)} = \frac{1}{2}$. It follows:

$$\frac{\partial a_2}{\partial w_1}\bigg|_{(1,1)} = 2\,\frac{\partial\theta^*}{\partial w_1}\bigg|_{(1,1)} = -1, \tag{A20}$$

whereas

$$\frac{\partial a_1}{\partial w_2}\bigg|_{(1,1)} = \frac{\partial\theta^*}{\partial w_2}\bigg|_{(1,1)} = \frac{1}{2}. \tag{A21}$$

Thus,

$$\frac{\partial a_2}{\partial w_1}\bigg|_{(1,1)} \neq \frac{\partial a_1}{\partial w_2}\bigg|_{(1,1)}. \tag{A22}$$

This explicit construction considers the simplest possible setting, with only two training examples and a scalar model parameter, yet it already yields unequal forward and reverse attribution responses while satisfying the differentiability and nonsingular-Hessian assumptions. In realistic learning systems, the parameter space and training interactions are substantially more complex, making such cross-example dependencies potentially even more heterogeneous. Therefore, symmetry is not a structural property of cross-example attribution dependence, and the dependence should generally be regarded as directional. □

B Relationship-Aware Training Data Attribution Procedure

Our method converts a conventional pointwise training-data attribution method into a query-conditioned relational explanation. We first train a baseline model and use its attribution-relevant projected-gradient representations to partition the training set into coherent groups. For a fixed query, we compute the baseline attribution of every group. We then intervene on one source group at a time: only a controlled fraction of that group's samples is retained, while every other group and every training hyperparameter remain fixed. After retraining on the resulting counterfactual data set, we recompute the base attribution method and measure the response of every target group. Repeating the intervention over several retention ratios and random draws yields a response trajectory for each ordered source–target pair. The local slope of this trajectory is the directed edge weight. The output is therefore a query-conditioned graph whose node weights describe group-level attribution and whose directed edges describe how one group's presence modulates another group's attribution. To help readers follow our method, we illustrate the main steps in Algorithm 1.

Algorithm 1 Relationship-Aware Training Data Attribution

Require: Training set $\mathcal{D}$, query q, retention ratios $\mathcal{P}$, and draws $\mathcal{R}$
Ensure: Group partition $\mathcal{G}$, node attributions $\mathbf{A}(q)$, and relation matrix $\mathbf{R}(q)$

```
Train the baseline model θ*(D)
Construct G = {G_1, ..., G_K} from the projected-gradient features {φ_i}_{i=1}^N
Compute A(q) = [A_1(q; D), ..., A_K(q; D)]
for G_s ∈ G do
    for ρ ∈ P and r ∈ R do
        Sample G̃_{s,ρ}^{(r)} ⊆ G_s with |G̃_{s,ρ}^{(r)}| ≈ ρ|G_s|
        Form D_{s,ρ}^{(r)} = (D \ G_s) ∪ G̃_{s,ρ}^{(r)}
        Train θ*(D_{s,ρ}^{(r)}) and compute A_t^{(s,ρ,r)}(q) for all G_t ∈ G
    end for
    for G_t ∈ G with t ≠ s do
        Fit Δ_{s→t}(q; ρ, r) against 1 − ρ
        Store the fitted slope as R_{s→t}(q)
    end for
end for
return G, A(q), and R(q)
```

C Experimental Details and Metric Definitions

This sections provides additional implementation details and formal definitions of the evaluation metrics used in the main paper. We follow the relationship-aware attribution formulation introduced in the main text and focus here on the experimental configurations, intervention protocol, and evaluation criteria.

C.1 Experimental Settings

Datasets and training. We evaluate CIFAR-2, CIFAR-10, ArtBench-2, ArtBench-5, ArtBench-10, and CelebA-Male. CIFAR-2 consists of the *automobile* and *horse* classes from CIFAR-10 (Krizhevsky & Hinton, 2009); ArtBench subsets are derived from ArtBench-10 (Liao et al., 2022); and CelebA-Male is constructed from the binary `Male` attribute of CelebA (Liu et al., 2015). We use 10,000 training examples for all datasets except ArtBench-2, which uses 5,000, and select four fixed correctly classified queries per dataset.

CIFAR-2 mechanism experiments use a ResNet-9-style CNN with a 512-dimensional TRAK projection. Cross-dataset group experiments use a compact four-layer CNN with a 256-dimensional TRAK projection. All models are trained with cross-entropy and SGD using learning rate 0.03, momentum 0.9, and weight decay 5×10^{-4}. Baseline and counterfactual models within each experiment use identical architectures, optimization settings, and random seeds.

Grouping and interventions. We use TRAK (Park et al., 2023) as the base attribution estimator and construct $K = 30$ groups from normalized projected-gradient representations using PCA followed by MiniBatch k-means. The partition is fixed within each dataset and shared across queries. For complete group graphs, each source group is evaluated at $\rho \in \{0.9, 0.8, 0.7\}$ using three independently sampled retained subsets at each retention level.

C.2 Evaluation Metrics

Group-attribution reorganization. We first measure how much a source-group intervention reorganizes the attribution ranking over target groups. G-rank ρ is the Spearman correlation between the rankings induced by $\mathbf{A}(q)$ and $\mathbf{A}^{(s,\rho,r)}(q)$. Lower G-rank ρ indicates that the relative ordering of target-group attribution changes more substantially. Let $\mathcal{T}_k(q)$ denote the set of the k groups with the largest baseline attribution values in $\mathbf{A}(q)$, and let $\mathcal{T}_k^{(s,\rho,r)}(q)$ denote the corresponding top-k

set after the intervention. We define group replacement as:

$$\text{G-Rep@}k = 1 - \frac{\left|\mathcal{T}_k(q) \cap \mathcal{T}_k^{(s,\rho,r)}(q)\right|}{k}. \tag{A23}$$

Thus, G-Rep@$k = 0$ means that all originally top-k groups remain in the top k, whereas G-Rep@$k = 1$ means that none remain. Larger values indicate stronger changes in group-level attribution structure. We report $k \in \{10, 20\}$.

Directionality and reproducibility. For each query, the directed relationship matrix is $\mathbf{R}(q) = [R_{s\to t}(q)]_{s,t=1}^{K}$. F/R ρ is the Spearman correlation between all off-diagonal edge weights $\{R_{s\to t}(q) : s \neq t\}$ and their reverse-direction counterparts $\{R_{t\to s}(q) : s \neq t\}$. A value close to one would indicate nearly symmetric relations, whereas a value close to zero indicates that the forward and reverse effects are largely unrelated.

Opp. measures the fraction of directed pairs for which the two directions have opposite signs:

$$\text{Opp.} = \frac{1}{K(K-1)} \sum_{s\neq t} \mathbb{I}\left[\text{sign}(R_{s\to t}(q)) \neq \text{sign}(R_{t\to s}(q))\right]. \tag{A24}$$

A larger Opp. value provides direct evidence that the two directions imply qualitatively different attribution responses.

Pointwise reducibility. For source group G_s, its scalar pointwise importance is its baseline group attribution $P_s(q) = A_s(q; \mathcal{D})$. Its outgoing relational influence is the total magnitude of the changes it induces in other groups, $O_s(q) = \sum_{t\neq s} |R_{s\to t}(q)|$.

Pt.-rel. ρ is the Spearman correlation between $\{P_s(q)\}_{s=1}^{K}$ and $\{O_s(q)\}_{s=1}^{K}$. A small correlation indicates that conventional scalar attribution does not directly recover which groups most strongly redistribute the attribution of other groups.

Source-specific response retrieval. For each source group G_s, we denote its directed response profile by $\mathbf{r}_s(q) = [R_{s\to t}(q)]_{t\neq s}$, which summarizes how perturbing G_s changes the attribution of all other target groups.We estimate response profiles from two independent retention draws and retrieve each source by profile correlation. Top-1 retrieval is the fraction of cases where the correct source ranks first. To further account for the full retrieval ranking, we report Mean Reciprocal Rank (MRR), defined as:

$$\text{MRR} = \frac{1}{K} \sum_{s=1}^{K} \frac{1}{\text{rank}(G_s)}, \tag{A25}$$

where $\text{rank}(G_s)$ denotes the rank of the correct source group among all candidate sources. A larger MRR indicates that the correct source is consistently ranked closer to the top. Together, Top-1 retrieval and MRR measure whether different source groups induce distinct and reproducible attribution-response profiles.

Response-trajectory diagnostics. For selected source–target pairs, we evaluate the response $\Delta_{s\to t}(q; \rho, r)$ over a denser grid of retention ratios. We compare three summaries: a linear fit, a quadratic fit that additionally includes $(1-\rho)^2$, and an isotonic fit that permits a flexible monotone response. We report ordinary and adjusted R^2 for the linear fit.

These metrics provide a comprehensive evaluation of the effectiveness, robustness, and relational effects captured by our method.

D Pointwise approximation of relational attribution.

We first investigate whether relational information can be approximated directly from conventional pointwise attribution scores, without explicit counterfactual retraining. Specifically, we examine whether a training-free relational TRAK approximation can recover the attribution redistribution caused by removing an individual training sample. For one representative CIFAR-2 query, we select

the top-20 TRAK training samples as candidate sources. For each source sample i, we remove only i, retrain the model using the same seed and training configuration, and define the ground-truth response of each remaining target sample j as $R^{\mathrm{GT}}_{i\to j}(q) = s_j(\mathcal{D}\setminus\{i\}, q) - s_j(\mathcal{D}, q)$. Then, we compare this response with the fixed-geometry estimator $\widehat{R}_{i\to j}(q) = \frac{K_{q\to i}K_{i\to j}}{1-K_{i\to i}}$, which is computed solely from the baseline projected-gradient kernel and requires no counterfactual retraining.

Each point in Fig. A1 corresponds to one ordered source–target pair, with the ground-truth response on the x-axis and the no-retraining prediction on the y-axis. For visualization, both are separately z-score normalized. If the static estimator captured the true redistribution, the points would align along the diagonal. Instead, the points show no clear relationship, with Pearson correlation 0.014 and Spearman correlation 0.040. It indicates that attribution redistribution after retraining cannot be inferred from the baseline kernel alone. Removing one sample changes the learned parameters and, consequently, the attribution geometry. In contrast, pairwise similarities computed from the original model remain fixed and cannot reflect this change in training context. This motivates our controlled-intervention formulation, which measures relational influence under changed training-data contexts rather than approximating it with static pairwise similarity.

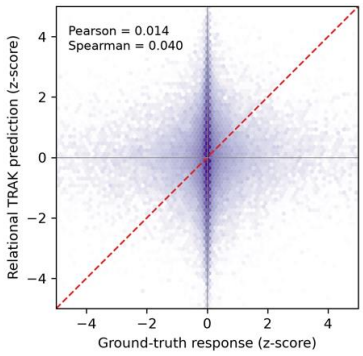


Figure A1: The relational response prediction through pointwise attribution scores.

E Directedness and Pointwise Non-Reducibility

We evaluate whether the recovered responses behave like genuine directed relations rather than a symmetric similarity score or a reparameterization of pointwise TRAK attribution. Tab. A1 summarizes three diagnostics. F/R ρ tests whether the response $G_s \to G_t$ agrees with the reverse response $G_t \to G_s$. Opp. measures how often the two directions have opposite signs. Pt.-rel. ρ compares each group's scalar attribution with its outgoing relational influence. The directional diagnostics are consistent across datasets. Forward/reverse correlations remain close to zero, and approximately half of the directed group pairs have opposite signs. Thus, the effect of perturbing G_s and measuring G_t is not interchangeable with perturbing G_t and measuring G_s. The pointwise-reducibility diagnostic gives the same conclusion from a different angle: Pt.-rel. correlations are small on all datasets. Therefore, groups with high scalar attribution are not necessarily the groups whose perturbation most strongly reorganizes other groups' attributions.

Table A1: Directed Relationship Analysis across datasets.

Datasets	F/R ρ	Opp.	Pt.-rel. ρ
CIFAR-2	-0.029	0.511	-0.030
CIFAR-10	0.008	0.499	0.007
ArtBench-2	0.023	0.495	-0.178
ArtBench-5	0.011	0.490	-0.019
ArtBench-10	-0.004	0.505	-0.041
CelebA-Male	-0.002	0.507	-0.010

F Response Consistency

We further examine how the identified relations evolve under different source-group intervention strengths. As shown in Tab. A2, some response trajectories exhibit nonlinear variation, with isotonic fits explaining more variation than linear fits. Nevertheless, the response direction remains largely consistent across retention levels and independent perturbation draws. These results suggest that the fitted slope provides a useful first-order summary of the dominant attribution response, without requiring the full response trajectory to be strictly linear.

F.1 Comparison with Pairwise Prediction Interaction

We further examine whether the relationships captured by our method can be explained by pairwise interaction effects on the query prediction. For each pair of training-data groups, we compare the prediction change caused by removing the two groups together with the combined changes obtained by removing each group individually. If the joint effect is already explained by the sum of the two

Table A2: Response consistency across source-group intervention strengths.

Dataset	Relation	Local slope	Linear fit R^2	Monotonic fit R^2	Sign agreement
ArtBench-2	Strong negative	0.0086	0.060	0.343	0.889
	Strong positive	0.0041	0.072	0.380	0.778
CelebA-Male	Strong negative	-0.0279	0.154	0.511	1.000
	Strong positive	0.0466	0.194	0.535	1.000

individual effects, the pair exhibits little additional interaction at the prediction level. Otherwise, the remaining difference reflects a non-additive interaction between the two groups.

We perform this comparison on CIFAR-10 using the same group partition and fixed queries as in the main experiments. For each group pair (G_s, G_t), we compare the prediction-level interaction with four quantities derived from our relational responses: the response from G_s to G_t, the reverse response from G_t to G_s, the average of the two responses, and their absolute difference. These comparisons test whether prediction-level interaction is associated with either directional response, their overall tendency, or the difference between the two directions.

Table A3: Correlation between prediction interaction and relational attribution.

Comparison	Pearson	Spearman
Interaction vs. $G_s \to G_t$ response	0.011	-0.013
Interaction vs. $G_t \to G_s$ response	-0.028	-0.055
Interaction vs. average response	-0.005	0.005
Interaction vs. absolute response difference	-0.003	0.016

As shown in Tab. A3, all correlations are close to zero. This indicates that pairwise interaction on the query prediction provides little information about the relational attribution response. In particular, a pair of groups may exhibit a noticeable non-additive effect on the prediction without inducing a strong attribution redistribution, while another pair with weak prediction interaction may still show a substantial relational response. These results suggest that prediction-level interaction and relational attribution capture different aspects of training-data dependence: the former measures whether two groups jointly affect the model output beyond their individual effects, whereas the latter measures how perturbing one group changes the attributed role of another.

G Relational attribution in diffusion models

Our previous experiments focus on classification models. We further investigate whether similar relational attribution effects also emerge in generative models by extending the controlled intervention protocol to an unconditional DDPM (Ho et al., 2020) trained on CIFAR-2. For each generated query, we first compute standard D-TRAK (Zheng et al., 2024) scores and cluster the 5,000 training examples into six groups using baseline projected-gradient features. We then independently downsample one source group while keeping all other training samples and the training protocol fixed. After retraining each counterfactual model, we recompute D-TRAK and summarize each target group by its mean normalized attribution rank.

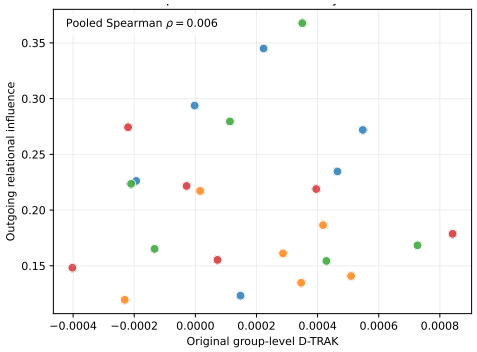


Figure A2: The relationship among the output relational influence and the original group-level attribution scores.

Fig. A2 examines whether pointwise D-TRAK importance captures the relational role of a training group. The horizontal axis shows each group's average D-TRAK attribution in the original model, while the vertical axis measures its total outgoing relational influence, i.e., how strongly perturbing that group redistributes attribution across the remaining groups after retraining. If pointwise importance reflected relational influence, groups with higher attribution would also induce stronger outgoing responses. Instead, the pooled Spearman correlation across all query–group pairs is close to zero, indicating little agreement between the two rankings. Groups with similar pointwise importance can therefore exhibit substantially different relational effects,

while moderately attributed groups may still strongly reorganize the attribution of others. This distinction arises because pointwise D-TRAK measures contribution under a fixed training context, whereas relational influence measures how the attribution landscape changes under a counterfactual intervention. Thus, the directed relational graph captures information that is not recoverable from conventional one-score-per-sample attribution alone.

Moreover, we further examine whether the recovered relationships are asymmetric in diffusion models. As shown in Fig. A5, the directed response matrix exhibits clearly different patterns across opposite directions. This indicates that perturbing one training-data group can change the attribution of another group differently from the reverse intervention. These observations provide empirical support for our theoretical analysis that relation-aware attribution dependencies are generally asymmetric, and further show that this property extends beyond classification models to generative settings.

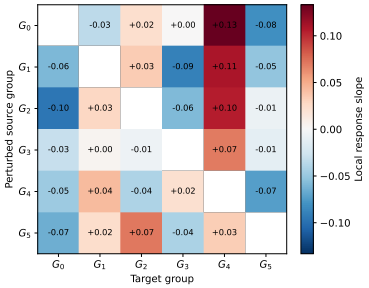


Figure A3: Asymmetric attribution responses in diffusion models.

H Sensitivity to the Percentile-Rank Reference Set

Our normalization computes percentile ranks within each intervention context. Consequently, downsampling a source group slightly changes the reference population used for ranking, which may introduce a mechanical rank shift independent of model retraining. We therefore quantify this potential confound to assess whether it materially affects the recovered relationships. We expect its magnitude to be limited, since it depends primarily on the fraction of source samples removed and on how different the attribution distribution of G_s is from that of the remaining training data. Because G_s constitutes only a subset of the reference population, partially removing it induces only a correspondingly limited change in the empirical percentile distribution.

Table A4: Sensitivity to the percentile-rank reference set.

Dataset	Rank corr.	Sign (%)	Ref.-set effect (%)
CIFAR-2	0.999	98.6	1.8
CIFAR-10	0.999	99.1	2.1
ArtBench-2	0.999	98.8	2.5
ArtBench-5	0.999	98.7	2.3
ArtBench-10	0.999	98.9	2.5
CelebA-Male	0.999	98.7	2.5
DDPM CIFAR-2	0.960	95.8	6.7

To isolate this effect, we recompute the relational responses using a source-specific fixed reference set that remains unchanged across retention ratios. This analysis is used only as a diagnostic and does not alter the definition of our primary method. For each source group G_s, we define the fixed reference set as $C_s = \mathcal{D} \setminus G_s$. When evaluating the effect of G_s on another group G_t, percentile ranks are always computed with respect to the same set C_s, regardless of the retention ratio of G_s. This removes rank changes caused purely by variation in the reference population.

As shown in Tab. A4, the estimated relationships remain highly consistent under the fixed-reference normalization. Rank correlation measures the consistency of edge ordering, while sign agreement evaluates whether the relationship direction is preserved. The reference-set effect further quantifies the variation caused solely by changing the percentile reference population relative to the response after counterfactual retraining. Overall, the edge rankings and directions are largely preserved, and the reference-set effect remains substantially smaller than the retraining response, indicating that it has only a limited influence on the estimated relationships.

I Relational Sensitivity Across Group-Attribution Ranks

To examine whether groups with high direct attribution are also the most structurally influential, we rank the $K = 30$ groups separately for each query q according to their baseline node attribution $A_s(q; D)$. For each group, we then summarize the query-conditioned relationship matrix $R(q)$ using its mean absolute outgoing and incoming sensitivities: $S_{\text{out}}(s; q) = \frac{1}{K-1} \sum_{t \neq s} |R_{s \to t}(q)|$, $S_{\text{in}}(t; q) = \frac{1}{K-1} \sum_{s \neq t} |R_{s \to t}(q)|$. Here, $S_{\text{out}}(s; q)$ measures how strongly perturbing source group G_s changes the attribution of the remaining groups, whereas

$S_{\text{in}}(t; q)$ measures how strongly the attribution of target group G_t responds to perturbations of the other groups.

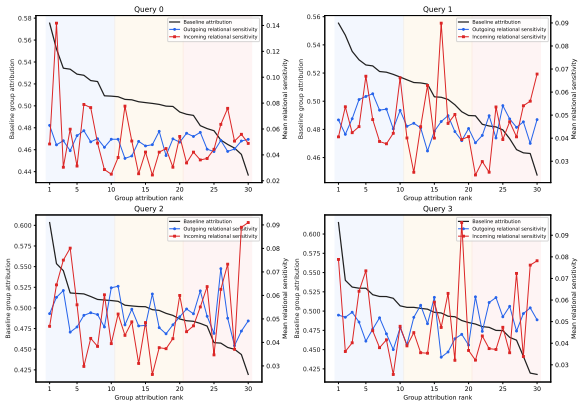


Figure A4: Relational sensitivity across group-attribution ranks for the four evaluation queries.

Fig. A5 illustrates these quantities against the baseline attribution rank for each query. Although $A_s(q; D)$ decreases by construction with rank, neither outgoing nor incoming relational sensitivity exhibits a corresponding monotonic decline. Averaged across the four queries, the mean outgoing sensitivities of the top-, middle-, and bottom-ranked thirds are 0.0510, 0.0468, and 0.0502, respectively. Moreover, the Spearman correlation between baseline attribution rank and relational sensitivity is not statistically significant for any query, while individual low-ranked groups can exhibit some of the largest outgoing or incoming sensitivities.

These results further demonstrate that node-level attribution and edge-level relational influence capture distinct properties of the training data. A group with modest baseline attribution $A_s(q; D)$ can nevertheless have a large $S_{\text{out}}(s; q)$, indicating that perturbing it substantially redistributes attribution across other groups. Conversely, a group with a large $S_{\text{in}}(t; q)$ can be highly sensitive to changes elsewhere in the training set even when its own baseline attribution is modest. Thus, scalar group attribution alone does not characterize the structural role of a group within the query-conditioned relationship graph.

J Pointwise Attribution Misses Relational Roles

Fig. A5 provides a controlled group-level example of information hidden by pointwise attribution. For the same CIFAR-2 query, the two source groups, shown on the left and right respectively, receive nearly identical baseline attribution scores of 0.5030 and 0.5037. Conventional attribution therefore regards them as almost equally important. We progressively weaken each source using retention levels $w \in \{1.0, 0.9, 0.8, 0.7\}$ and track how the attribution of every other target group changes. The directional response $R_{s \to t}$ summarizes the change in target-group attribution per unit increase in source-intervention strength $1 - w$.

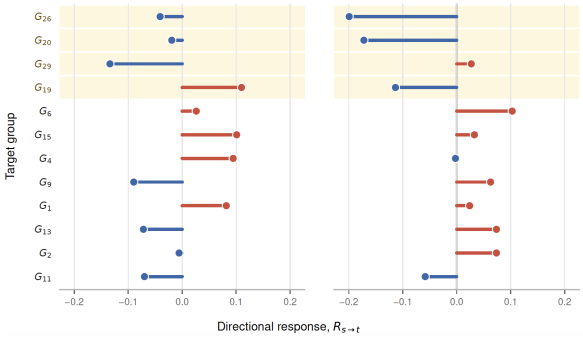


Figure A5: Relational sensitivity across group-attribution ranks for the four evaluation queries.

Despite their almost identical scalar importance, the two groups exhibit markedly different outgoing response results. For example, weakening G_8 increases the attribution of G_{19}, whereas weakening G_{16} decreases it. Conversely, G_{16} produces substantially stronger negative responses in G_{20} and G_{26}. This discrepancy is not restricted to a single target: the two complete outgoing response profiles have correlation approximately zero. Thus, pointwise attribution identifies how important a group is to the query, while relationship-aware attribution additionally reveals how that group supports, suppresses, or redistributes importance across the rest of the training data.

K Pointwise Similarity Does Not Imply Relational Similarity

We further examine this phenomenon at the dataset level by asking whether groups with similar baseline attribution also exhibit similar relational roles across all group pairs. We first select attribution-

matched group pairs using increasingly strict thresholds on their baseline attribution difference, $|A_s - A_t|$.

For each selected pair, we measure its directional asymmetry as $|R_{s\to t} - R_{t\to s}|$, where a larger value indicates a larger discrepancy between the forward and reverse relationships. Fig. A6 plots the empirical cumulative distribution of this quantity: for a given horizontal value x, the vertical axis gives the fraction of matched group pairs whose directional asymmetry is no larger than x. Thus, a curve concentrated near zero would indicate that similar pointwise attribution also implies similar directional relationships. Instead, substantial asymmetry remains even among closely attribution-matched pairs. In particular, among the 164 pairs satisfying $|A_s - A_t| < 0.005$, only 50.6% have directional asymmetry below 0.05, meaning that 49.4% exceed 0.05; similarly, 22.0% exceed 0.10. Comparable behavior persists under stricter matching thresholds. These dataset-level statistics show that groups that appear nearly indistinguishable under scalar attribution can still exhibit substantially different directional relationships, demonstrating that pointwise attribution similarity does not imply relational similarity.

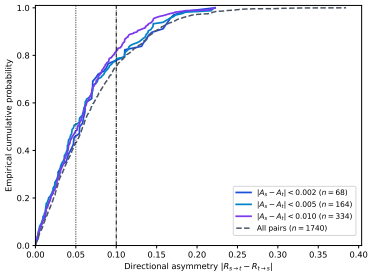


Figure A6: Dataset-level evidence that pointwise similarity does not imply relational similarity.

L Predictive Utility under Unseen Interventions

We compare five predictors with different available information. "Attribution Only" only predicts the held-out change using only the baseline source and target attribution scores, without relational information. "Attribution + Relation" augments the same baseline attribution features with the estimated directional response $R_{s\to t}$. "Relation Only" directly extrapolates the learned directional response to the stronger intervention without additional calibration. "Symmetric projection" removes directionality by averaging the forward and reverse relational predictions, while "No Change / Random" assumes zero attribution change for RMSE and randomly selects target groups for Top-3 evaluation.

As shown in Tab. A5, incorporating relational information substantially improves the prediction of attribution changes under the unseen intervention. Compared with the A-only baseline, $A+R$ reduces RMSE from 0.0207 to 0.0162, corresponding to an absolute reduction of 0.0045 and a relative improvement of 21.7%. The R-only predictor achieves the highest Top-3 overlap (53.6%), indicating that directional responses are particularly informative for identifying the most affected target groups. In contrast, symmetrizing the relations degrades both RMSE and Top-3 overlap, further supporting the value of preserving directionality.

Table A5: Prediction of target-attribution changes under held-out 30% source-group removal on CIFAR-2.

Predictor	RMSE ↓	Top-3 overlap ↑
No Change / Random	0.0239	10.3%
Symmetric Relation	0.0235	45.8%
Relation Only	0.0230	**53.6%**
Attribution Only	0.0207	48.1%
Attribution + Relation	**0.0162**	52.5%

M Directed Relational Responses Are Distinct from Pairwise Interactions

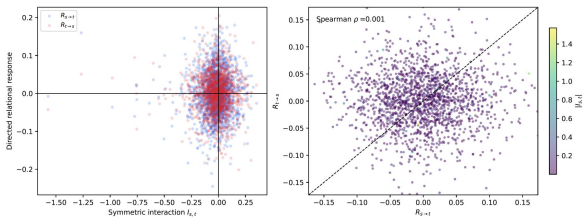
Pairwise interaction and relational response characterize different aspects of training-data influence. Conventional interaction measures whether two groups jointly affect the model beyond the combination of their individual effects, and therefore assigns a symmetric quantity to an unordered pair. In contrast, our relational response characterizes the directional dependency between attributed roles: $R_{s\to t}$ describes how the attribution of a target group changes with respect to a controlled change in a source group, and need not equal $R_{t\to s}$.

We compare these quantities empirically on CIFAR-10 using 30 training-data groups and four fixed queries. For each unordered group pair, we compute the inclusion–exclusion interaction

$$I_{s,t}(q) = F(\mathcal{D} \setminus G_s \setminus G_t, q) - F(\mathcal{D} \setminus G_s, q) - F(\mathcal{D} \setminus G_t, q) + F(\mathcal{D}, q), \tag{A26}$$

where F denotes the classification margin for query q. This interaction is compared with the two directed responses associated with the same pair.

Fig. A7 further illustrates the difference between conventional pairwise interaction and our directed relational response. For a pair of training-data groups, conventional interaction assigns a single symmetric value $I_{s,t}$, which measures whether jointly perturbing the two groups produces an additional prediction effect beyond the combination of their individual effects. As shown in Fig. A7(a), the same interaction value can correspond to two substantially different directional responses, indicating that a single symmetric interaction score does not determine how attribution changes in each direction. Fig. A7(b) makes this distinction more explicit by directly comparing $R_{s\to t}$ with $R_{t\to s}$. If the relationships were effectively symmetric, the points would concentrate around the diagonal; instead, they are widely dispersed and the two directions are nearly uncorrelated (Spearman $\rho = 0.001$). Consistently, the antisymmetric component accounts for 50.2% of the total response energy. These results show that pairwise interaction captures symmetric prediction-level non-additivity, whereas Relationship-Aware TDA captures directional changes in the attributed roles of training data, which cannot be recovered from a single interaction value.



(a) (b)

Figure A7: Symmetric pairwise interaction does not recover directed relational responses. (a) For each group pair, a single symmetric interaction value $I_{s,t}$ corresponds to two directional responses, which can differ substantially. (b) Forward and reverse responses are nearly uncorrelated, showing that exchanging the source and target generally changes the recovered relationship.