Structured Kernel Surrogate Models for Multimodal Attribution

Ziniu Zhang[*]
Northeastern University
Boston, MA

Zhenshuo Zhang[*]
Northeastern University
Boston, MA

Jianglin Lu
Northeastern University
Boston, MA

Ruoxuan Xiong
Emory University
Atlanta, GA

Yun Fu
Northeastern University
Boston, MA

Hongyang R. Zhang
Northeastern University
Boston, MA

Abstract

We consider the problem of *cross-sample modality attribution*, which involves measuring how one modality in a training sample affects the learning outcome in other multimodal samples. This problem is crucial because a modality may be redundant for one sample, but is uniquely informative for other samples, or useful only through synergy with other modalities. Existing methods from the literature of data attribution that focus on estimating sample-level influence remain insufficient in multimodal settings, without capturing nonlinear interactions across both samples and modalities. In this paper, we design structured kernel surrogate (SKS) models for *multimodal attribution*, by learning a mapping from both sample-level and modality-level selections to the prediction outcome. To approximate this mapping, SKS uses kernel ridge regression with a structured (radial basis function) kernel that captures nonlinear effects across both samples and modalities. In particular, the structured design applies a product composition of the kernel similarity within-sample across every pair of data modalities. Further, we design a gradient estimation algorithm to efficiently learn the kernel. We evaluate our approach on two multimodal benchmarks, spanning a range of data modalities including satellite images, road networks, texts, audio, and tabular data. We find that SKS achieves 57.8% higher correlation with the true prediction outcomes than existing methods. In downstream data selection evaluation, SKS improves the average accuracy by 1.93% over baselines. Models trained on data selected by SKS have lower Hessian trace values, indicating a flatter loss surface and better generalization. Finally, SKS also provides interpretable modality-level selections in spatiotemporal traffic accident prediction, preferentially retaining satellite tiles with visible road infrastructure over vegetation-dominated tiles.

1 Introduction

Multimodal data are increasingly available across domains and are often valuable for prediction. For example, transportation agencies can combine satellite images, traffic statistics, and weather records to predict accident risk [25, 44]. Autonomous self-driving systems integrate visual observations and audio instructions to support navigation [1]. As multimodal systems scale in both data size and model complexity, their internal mechanisms become highly complex [44] and susceptible to brittle features [41]. Understanding how different modalities interact and contribute to predictions is, therefore, central to interpreting multimodal models.

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

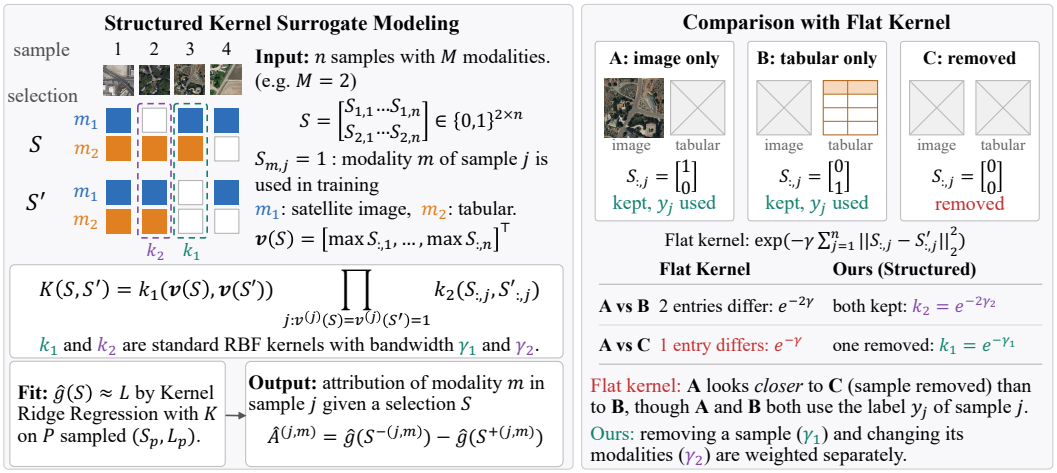


Figure 1: Overview of SKS. We fit a kernel ridge surrogate to predict the loss from the selection S, where $S_{m,j}$ indicates whether modality m of sample j is retained. We estimate its attribution by comparing the test losses under $S^{+(j,m)}$ and $S^{-(j,m)}$, which retain and remove this modality in the sample, respectively, with all other entries of S unchanged. We then illustrate the advantage of our structured kernel over a flat kernel.

A key challenge is that modality interactions can arise at multiple levels. Prior work studies such interactions through multi-view learning and information theory. Classical multi-view learning [6, 30] uses relationships estimated from unlabeled paired data to regularize predictors and encourage agreement across views. More recent work based on partial information decomposition characterizes whether modalities provide redundant, unique, or synergistic information that is useful only in combination [18, 4]. These approaches provide useful distribution-level descriptions of what information is present across modalities.

However, describing the information shared across modalities does not reveal how individual training samples and their modalities contribute to a trained model's performance. The effect of retaining or removing one modality from a training sample can depend on the other samples and modalities used during training. This motivates cross-sample modality attribution: measuring how retaining or removing a modality from an individual training sample affects the trained model's performance on the test set.

This is closely related to *data attribution*, which studies how training examples influence model predictions. Existing methods, such as influence functions [8], approximated data models [27], and linear and kernel surrogate methods [12, 43], provide useful tools for sample-level or task-level attribution. However, they are primarily designed for unimodal or task-level settings and do not capture the sample-modality structure of multimodal learning. In particular, removing an entire sample and removing one modality of that sample are *distinct perturbations*, and the effect of one modality can depend on how it interacts with other samples and modalities through a shared network or optimization procedure.

These gaps motivate us to design *Structured Kernel Surrogate* models, or SKS for short, for capturing cross-sample modality interaction with modality missingness and nonlinear relationships. The entire framework is illustrated in Figure 1. SKS considers perturbations of the training set in which some samples are included or removed, and some modalities are retained or masked within each included sample. For each such perturbation, the corresponding target is the generalization loss of the model trained on the perturbed multimodal dataset. SKS then learns this selection-to-loss map using kernel ridge regression with a structured radial basis function (RBF) kernel. The kernel separates sample-level participation from within-sample modality composition, which crucially allows for missing data modalities. Moreover, it captures nonlinear interactions across samples and modalities.

We evaluate SKS on the multimodal traffic accident prediction benchmark MMTRACE [44] and the standard multimodal benchmark MultiBench [20]. We first measure attribution quality using the linear data modeling score (LDS). Compared with the strongest baseline, SKS improves the per-task

relative LDS by 50.1% on MMTRACE and 65.6% on MultiBench. We then evaluate whether these attributions can guide downstream data selection. Under a fixed top-20% vision-modality selection budget, SKS improves the average accuracy by 1.52% on MMTRACE and by 2.33% on MultiBench over the strongest baseline. Similar gains are observed when the selected modality is changed to tabular information. Finally, we analyze the generalization behavior of models trained on data selected by SKS. The selected data lead to strong cross-state generalization and a smoother loss surface than competing attribution methods. This shows better robustness of the resulting models trained using SKS.

In summary, our contributions are threefold. First, we propose SKS, a *structured kernel surrogate* model that separates sample-level and modality-level selection to learn the nonlinear subset-to-loss mapping. We provide a sample complexity bound for the surrogate. Second, SKS achieves the highest LDS across six MultiBench tasks, improving over the strongest baseline by 65.6% in per-task relative LDS on average, with the largest gains on MOSI and Sarcasm. It also improves downstream test accuracy on MultiBench by 2.33% on average. Third, on spatiotemporal traffic accident prediction across six US states, SKS improves LDS by 50.1% over the strongest baseline. For downstream data selection, it improves average test accuracy by 1.52%. The learned attributions are also interpretable: SKS assigns higher scores to satellite images with visible road infrastructure than to vegetation-dominated tiles. Models trained on data selected by SKS show strong cross-state transferability. The loss sharpness analysis further shows that they achieve the lowest Hessian trace against baselines. Our code repository can be found at https://github.com/VirtuosoResearch/Multimodal-attribution-code.

2 Preliminaries

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote a multimodal training dataset sampled from distribution $\mathcal{D}$, where $\mathbf{x}_i$ is the multimodal input with M modalities, and y_i is the output. For exposition, we focus on the two-modality setting $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)})$. In traffic accident prediction, for example, each sample may represent a road segment, modeled as an edge in a road network graph. The two modalities can consist of a satellite image of the segment and tabular features such as traffic volume. We consider training a multimodal network f_W on selected samples and selected modalities within each sample. The global selection state is represented by the binary matrix

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,n} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,n} \end{bmatrix} \in \{0,1\}^{2\times n},$$

where $S_{m,i} = 1$ means that modality m of sample i is retained and $S_{m,i} = 0$ means that it is removed. Thus, column i of S indicates which modalities are retained in sample i, and row m indicates which samples retain modality m. The extension to more than two modalities is straightforward and is discussed in detail in Appendix A.3. We define $\mathbf{v}(S) = [\max(S_{:,1}), ..., \max(S_{:,n})]^\top$. $\mathbf{v}^{(i)}(S) = 1$ if at least one modality of sample i is retained, and $\mathbf{v}^{(i)}(S) = 0$ if the entire sample is removed.

For sample i, the masked training loss is $\ell\big(f_W(\mathbf{x}_i; S_{:,i}), y_i\big)$. We require at least one sample to retain at least one modality, so $\sum_{j=1}^n \mathbf{v}^{(j)}(S) > 0$. We define the empirical training loss under S as the average loss over all samples:

$$\hat{L}(f_W; S) = \left(\sum_{j=1}^n \mathbf{v}^{(j)}(S)\right)^{-1} \sum_{i=1}^n \mathbf{v}^{(i)}(S)\, \ell(f_W(\mathbf{x}_i; S_{:,i}), y_i).$$

The trained parameters W generally depend nonlinearly on S. Changing a modality in one sample changes the corresponding sample's contribution to the training objective and, through shared parameter updates, can further affect how the trained model uses information from other samples.

We characterize the effect of a selection S by the expected risk of the trained model under the test distribution $\mathcal{D}$, which is assumed to be the same as the training distribution.

$$L(f_W; S) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\ell(f_W(\mathbf{x}; S), y)\big].$$

Our goal is to characterize how perturbing a single entry of S, corresponding to retaining or removing one modality of a training sample, affects the resulting test loss $L(f_W; S)$.

Problem statement. We introduce an attribution score to measure how one modality of a training sample affects the test loss. For modality m in sample j, let $S^{+(j,m)}$ and $S^{-(j,m)}$ be obtained from S by setting $S_{m,j}$ to 1 and 0, respectively, keeping all other entries unchanged. We define the attribution of modality m in sample j at S as

$$A^{(j,m)}(S) = L(f_W; S^{-(j,m)}) - L(f_W; S^{+(j,m)}). \tag{1}$$

$A^{(j,m)}(S)$ measures the change in test loss caused by removing modality m from sample j and keeping other sample and modality choices unchanged. A positive value indicates that retaining the modality in this sample reduces the test loss. We note that this attribution is dependent on selection S. This formulation raises two questions:

- First, how can we model the nonlinear mapping from the selection of training samples and modalities specified by S to the test loss $L(f_W; S)$?
- Second, how can we use this mapping to efficiently estimate the attribution for each modality of each training sample?

The next section proposes a detailed methodology to answer the two questions.

3 Our Approach

We address the two questions in turn. We first introduce a structured kernel surrogate to model the mapping from S to $L(f_W; S)$, with a structured similarity that separately accounts for sample participation and modality composition. We then recover the attribution scores through finite differences of the fitted surrogate.

3.1 Designing Kernel Surrogate Models

To estimate the mapping from S to $L(f_W; S)$, we model how training samples and their modalities interact and jointly affect the test loss. These interactions arise at two levels: sample-level interaction and modality-level interaction. We model these two types of interactions separately using two RBF kernels. We then combine the two kernels to form a structured design that captures interactions among samples and modalities. We utilize the standard RBF kernel with squared Euclidean distance. We provide both a justification for the RBF kernel and an ablation of alternative kernel choices in Appendix B.1.

Structured design. Given two selections S, S', we first quantify their sample-level similarity by comparing $\mathbf{v}(S)$ and $\mathbf{v}(S')$ using an RBF kernel with bandwidth $\gamma_1 > 0$:

$$k_1(\mathbf{v}(S), \mathbf{v}(S')) = \exp\left(-\gamma_1 \|\mathbf{v}(S) - \mathbf{v}(S')\|_2^2\right).$$

Since $\mathbf{v}(S), \mathbf{v}(S')$ are binary, the squared distance counts the samples whose participation differs between the two states.

For each sample j that is selected in both S and S', we compare its modality selection $S_{:,j}$ and $S'_{:,j}$ to measure how similarly the modalities of this sample are selected across the two states. We define this modality-level similarity using an RBF kernel with bandwidth $\gamma_2 > 0$:

$$k_2(S_{:,j}, S'_{:,j}) = \exp\left(-\gamma_2 \|S_{:,j} - S'_{:,j}\|_2^2\right).$$

We then combine the sample-level and modality-level kernels to define the structured similarity between S and S':

$$K(S, S') = k_1(\mathbf{v}(S), \mathbf{v}(S')) \cdot \left[\prod_{j \in [n]: \mathbf{v}^{(j)}(S) = \mathbf{v}^{(j)}(S') = 1} k_2(S_{:,j}, S'_{:,j}) \right].$$

Restricting the product to samples that are present in both S and S' separates sample removal from modality changes within a retained sample. If a sample is absent in either S or S', its difference is captured by the sample-level component k_1 and does not contribute an additional modality-level term.

Algorithm 1 Structured Kernel Surrogate (SKS) Models for Multimodal Attribution

Input: P sampled selections and test loss on corresponding selections $\{(S_p, L(f_W; S_p))\}_{p=1}^{P}$
Require: γ_1, γ_2, S, ridge penalty λ
Output: Surrogate $\hat{g}$ and attributions $\{\widehat{A}^{(j,m)}(S) : j \in [n],\ m \in [M]\}$

/ Stage 1: build the structured Gram matrix and solve kernel ridge regression */*

for $p, q = 1, \ldots, P$ **do**
  $\mathcal{J}_{pq} \leftarrow \{j \in [n] : \mathbf{v}^{(j)}(S_p) = \mathbf{v}^{(j)}(S_q) = 1\}$
  $\mathbf{K}_{pq} \leftarrow k_1(\mathbf{v}(S_p), \mathbf{v}(S_q)) \prod_{j \in \mathcal{J}_{pq}} k_2\left(S^{(p)}_{:,j}, S^{(q)}_{:,j}\right)$
end for
$\mathbf{y} \leftarrow (L(f_W; S_1), \ldots, L(f_W; S_P))^\top$
$k_S \leftarrow (K(S, S_1), \ldots, K(S, S_P))^\top$
$\hat{g}(S) \leftarrow k_S^\top (\mathbf{K} + P\lambda \mathbf{I})^{-1}\mathbf{y}$

/ Stage 2: finite-difference attributions */*

for each $j \in [n]$ and $m \in [M]$ **do**
  Construct $S^{+(j,m)}, S^{-(j,m)}$ from S by setting $S_{m,j} = 1$ and 0 respectively
  $\widehat{A}^{(j,m)}(S) \leftarrow \hat{g}(S^{-(j,m)}) - \hat{g}(S^{+(j,m)})$
end for
return $\hat{g}$, $\{\widehat{A}^{(j,m)}(S) : j \in [n],\ m \in [M]\}$

This design allows K to capture interactions between samples. It also captures interactions between modalities within a sample and across samples. For two modalities, K is positive semidefinite if $\gamma_2 > 0$ and $\gamma_1 \geq \frac{1}{2}\log\left(\frac{3-e^{-\gamma_2}}{1+e^{-\gamma_2}}\right)$. Under these conditions, K is a valid kernel. We give the proposition and proof in Appendix A.1.

3.2 Surrogate Modeling of Sample-Modality Interaction

Since K defines a valid kernel, we use it to fit a surrogate model that predicts the test loss from S. We then estimate the attribution of each modality in each training sample using finite differences. Each difference compares the predicted losses with the modality retained and removed, where all other entries of S remain fixed.

Surrogate modeling. Let $\mathcal{H}_K$ denote the reproducing kernel Hilbert space (RKHS) induced by the structured kernel K. Given P sampled selection states $S_1, \ldots, S_P$ with measured test losses $L(f_W; S_1), \ldots, L(f_W; S_P)$, we fit a kernel ridge regression surrogate [43] by optimizing over candidate functions $g \in \mathcal{H}_K$, where $g(S)$ predicts the test loss associated with state S:

$$\hat{g} = \arg\min_{g \in \mathcal{H}_K} \frac{1}{P}\sum_{p=1}^{P}(g(S_p) - L(f_W; S_p))^2 + \lambda\|g\|^2_{\mathcal{H}_K}.$$

Let $\mathbf{K} \in \mathbb{R}^{P \times P}$ denote the Gram matrix over the P sampled selection states, whose (p, q)-th entry is $\mathbf{K}_{pq} = K(S_p, S_q)$. Let $\mathbf{y} = (L(f_W; S_1), \ldots, L(f_W; S_P))^\top$ and $k_S = (K(S, S_1), \ldots, K(S, S_P))^\top$. The closed-form solution is given by

$$\hat{g}(S) = k_S^\top (\mathbf{K} + P\lambda \mathbf{I})^{-1}\mathbf{y}.$$

Attribution from the surrogate model. Given a reference selection state S, we estimate the attribution using the fitted surrogate $\hat{g}$. For modality m in sample j, the estimated attribution is

$$\widehat{A}^{(j,m)}(S) = \hat{g}\left(S^{-(j,m)}\right) - \hat{g}\left(S^{+(j,m)}\right).$$

We compute the finite difference with all other entries fixed at their values in S. The attribution depends on which other samples and modalities are retained, and it can reflect interactions with those samples and modalities.

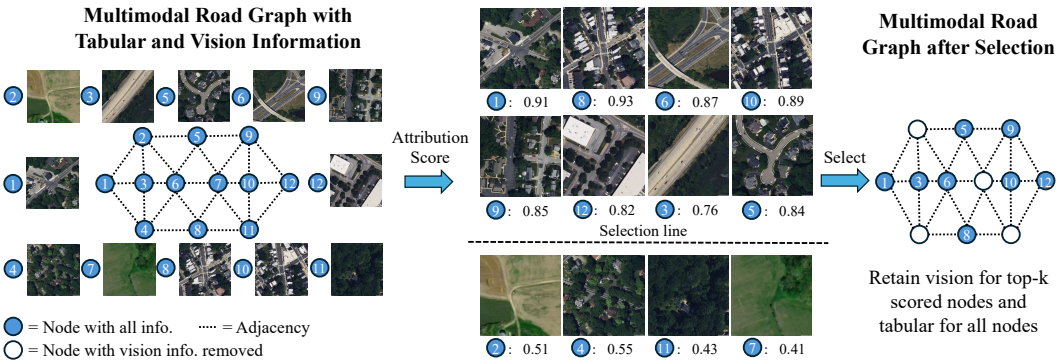


Figure 2: Overview of surrogate-guided data selection on MMTRACE. Each node represents a road-network sample with tabular and vision modalities. The structured kernel surrogate assigns an attribution score to each node image. We keep the vision information for the top-k scored nodes.

Sample complexity. We next analyze the statistical efficiency of the surrogate modeling. For the surrogate $\hat{g}$, the P training selections are drawn independently from the same selection procedure, and their observed losses equal $L(f_W; S)$. Assume that the function $S \mapsto L(f_W; S)$ belongs to $\mathcal{H}_K$ and satisfies $\|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho$.

Proposition 1. *For any $\lambda \in (0, 1]$ and $\delta \in (0, 1)$, there exist universal constants $C_0, C_1 > 0$ such that, if $P \geq \frac{C_0}{\lambda} \log\left(\frac{C_0}{\lambda\delta}\right)$, then, with probability at least $1 - \delta$ over the sampled selections,*

$$\mathbb{E}_S\left[\left(\hat{g}(S) - L(f_W; S)\right)^2\right] \leq C_1 \lambda \rho^2.$$

This result bounds the surrogate's prediction error in terms of the number of sampled selections and the regularization parameter. We defer the proof to Appendix A.2.

3.3 Efficient Estimation

For scalability, fitting the surrogate does not require retraining the model for every selection state. Inspired by the first-order approximation in previous works [24, 27, 15], we linearize the multimodal model around a checkpoint trained on all samples and modalities using a first-order approximation. Let $W^\star$ denote this checkpoint, and $f_W(\mathbf{x}_i; S_{:,i})$ denote the model output for sample i under the modality selection $S_{:,i}$. For the parameter W close to $W^\star$, we use the first-order Taylor approximation

$$f_W(\mathbf{x}_i; S_{:,i}) = f_{W^\star}(\mathbf{x}_i; S_{:,i}) + \nabla_W f_{W^\star}(\mathbf{x}_i; S_{:,i})(W - W^\star) + \epsilon, \tag{2}$$

where $\nabla_W f_{W^\star}(\mathbf{x}_i; S_{:,i})$ is the gradient with respect to W and ϵ denotes the approximation error. We compute and cache both the model output and gradient for each selection S. In Appendix B.3, we empirically evaluate ϵ and find it to be small in the estimation. The training objective under S can be approximated by

$$\frac{1}{\sum_{j=1}^n \mathbf{v}^{(j)}(S)} \sum_{i=1}^n \mathbf{v}^{(i)}(S)\, \ell(f_{W^\star}(\mathbf{x}_i; S_{:,i}) + \nabla_W f_{W^\star}(\mathbf{x}_i; S_{:,i})(W - W^\star), y_i)\,. \tag{3}$$

Thus, the loss associated with each sampled S can be estimated by optimizing the linearized objective, without retraining the original multimodal model from scratch. We report the computational speedup achieved by this approximation in Appendix B.3.

4 Experiments

Our experiments investigate two key questions. First, how does SKS compare with existing attribution methods and surrogate-based methods on multimodal attribution benchmarks? Second, can the attributions produced by SKS be used for downstream data selection? We evaluate these questions on MultiBench and MMTRACE. The results show that SKS achieves the best attribution performance across multiple tasks, and that data selected by SKS leads to the highest downstream accuracy under a fixed selection budget. We further show through ablation studies that SKS is robust to kernel bandwidth choices and generalizes across different modalities and selection rates.

Table 1: Results on multimodal attribution across MMTRACE and MultiBench. We report LDS scores for six MMTRACE states and six MultiBench tasks. All results are reported as the mean and standard deviation over three independent runs.

Methods	DE	MA	MD	NV	MT	IA
Influence functions	$0.22_{\pm 0.03}$	$0.12_{\pm 0.05}$	$0.17_{\pm 0.03}$	$0.10_{\pm 0.02}$	$0.23_{\pm 0.05}$	$0.07_{\pm 0.01}$
TracIn	$0.29_{\pm 0.05}$	$0.11_{\pm 0.08}$	$0.21_{\pm 0.04}$	$0.25_{\pm 0.05}$	$0.16_{\pm 0.07}$	$0.12_{\pm 0.01}$
Trak	$0.26_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.14_{\pm 0.02}$	$0.19_{\pm 0.03}$	$0.24_{\pm 0.01}$	$0.16_{\pm 0.07}$
BIF	$0.31_{\pm 0.04}$	$0.18_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.18_{\pm 0.03}$	$0.17_{\pm 0.04}$	$0.21_{\pm 0.03}$
Linear surrogates	$0.46_{\pm 0.06}$	$0.51_{\pm 0.02}$	$0.41_{\pm 0.01}$	$0.42_{\pm 0.10}$	$0.24_{\pm 0.06}$	$0.58_{\pm 0.02}$
SKS (Algorithm 1)	$\mathbf{0.64}_{\pm 0.03}$	$\mathbf{0.62}_{\pm 0.05}$	$\mathbf{0.59}_{\pm 0.10}$	$\mathbf{0.58}_{\pm 0.02}$	$\mathbf{0.57}_{\pm 0.06}$	$\mathbf{0.70}_{\pm 0.03}$
Methods	Avmnist	Humor	Imdb	Mosei	Mosi	Sarcasm
Influence functions	$0.14_{\pm 0.14}$	$0.11_{\pm 0.21}$	$0.23_{\pm 0.16}$	$0.27_{\pm 0.25}$	$0.17_{\pm 0.11}$	$0.17_{\pm 0.22}$
TracIn	$0.05_{\pm 0.07}$	$0.12_{\pm 0.18}$	$0.20_{\pm 0.18}$	$0.04_{\pm 0.13}$	$0.17_{\pm 0.17}$	$0.12_{\pm 0.10}$
Trak	$0.28_{\pm 0.16}$	$0.21_{\pm 0.14}$	$0.33_{\pm 0.15}$	$0.04_{\pm 0.08}$	$0.07_{\pm 0.20}$	$0.05_{\pm 0.11}$
BIF	$0.30_{\pm 0.12}$	$0.19_{\pm 0.11}$	$0.32_{\pm 0.09}$	$0.14_{\pm 0.07}$	$0.16_{\pm 0.11}$	$0.20_{\pm 0.13}$
Linear surrogates	$0.56_{\pm 0.10}$	$0.47_{\pm 0.05}$	$0.27_{\pm 0.07}$	$0.66_{\pm 0.05}$	$0.21_{\pm 0.17}$	$0.30_{\pm 0.24}$
SKS (Algorithm 1)	$\mathbf{0.73}_{\pm 0.04}$	$\mathbf{0.69}_{\pm 0.06}$	$\mathbf{0.43}_{\pm 0.05}$	$\mathbf{0.73}_{\pm 0.03}$	$\mathbf{0.62}_{\pm 0.01}$	$\mathbf{0.54}_{\pm 0.06}$

4.1 Experimental Setup

Datasets. We first evaluate our framework on the spatiotemporal traffic accident prediction task MMTRACE [44]. It contains a US state-level road network, accompanied by weather, traffic volume, satellite images, and longitudinal accident records from six states. It covers tabular and vision modalities, and the objective is to predict accident occurrences on a target edge. Then, we evaluate on six real-world tasks in MultiBench [20]: UR-FUNNY (Humor), MUStARD (Sarcasm), CMU-MOSI (Mosi), CMU-MOSEI (Mosei), AV-MNIST (Avmnist), MM-IMDB (Imdb). These datasets cover diverse modalities, including text, audio, vision, and tabular data.

Baselines. We compare our method with data attribution methods including Influence functions [8], TracIn [29], Trak [27], DSIR [36], LESS [35], and BIF [9]. We also compare with linear surrogate models [12].

Evaluation. We denote the evaluated selection states by $\{S_p\}_{p=1}^{P}$, with corresponding test losses y_p. The predicted test loss from an attribution method is defined as $\hat{g}(S_p)$. We first evaluate the LDS score for all methods. It is defined as the Spearman correlation coefficient between the predicted ranking and the real ranking: $\mathrm{LDS} = \mathrm{Spearman}\left(\{\hat{g}(S_p)\}_{p=1}^{P}, \{y_p\}_{p=1}^{P}\right)$. A higher LDS score indicates that the attribution method better recovers the ranking of subset-level test losses.

Then, we frame data selection as a downstream evaluation of attribution methods under an asymmetric multimodal training setting. The model uses all data from the first modality and only a 20% subset from the second modality. Each attribution method scores samples in the second modality, and we select the top 20% as the retained subset. We then train the multimodal model with the complete first modality and the selected second-modality subset, and evaluate it on a standard test set. We add an illustrative example in Figure 2 for better understanding.

4.2 Experimental Results

Attribution results. As shown in Table 1, SKS achieves the highest LDS score across both MMTRACE and MultiBench. Compared with the strongest baseline, SKS improves the per-task relative LDS by 50.1% on MMTRACE and by 65.6% on MultiBench. The gains are especially large on more challenging multimodal tasks, such as Mosi and Sarcasm, where linear attribution is less able to capture nonlinear subset-level effects. These results show that structured nonlinear surrogate modeling is important for recovering the ranking of subset-level test losses in multimodal prediction.

Data selection results. Given the strong attribution performance, we further evaluate SKS on downstream data selection. Under the fixed top-20% vision-modality selection budget, SKS achieves the best test accuracy on both MMTRACE and MultiBench, as shown in Table 2. Compared with the strongest baseline for each state, SKS improves the average accuracy by 1.52% on MMTRACE. On MultiBench, SKS also improves over the strongest baseline for each task by 2.33% on average, with

Table 2: Test accuracy of data selection methods on MMTRACE [44] and MultiBench [20]. Each approach selects the top 20% vision-modality samples under a fixed selection budget, and we report the mean and standard deviation over three independent runs.

Methods	DE	MA	MD	NV	MT	IA
Influence functions	$71.0_{\pm 0.1}$	$77.3_{\pm 0.8}$	$78.1_{\pm 0.3}$	$85.9_{\pm 0.1}$	$78.7_{\pm 0.2}$	$75.4_{\pm 0.1}$
DSIR	$70.2_{\pm 0.8}$	$76.1_{\pm 0.5}$	$77.9_{\pm 0.7}$	$85.2_{\pm 0.8}$	$77.5_{\pm 0.7}$	$75.7_{\pm 0.4}$
TracIn	$72.1_{\pm 0.4}$	$77.9_{\pm 0.3}$	$79.5_{\pm 0.4}$	$85.4_{\pm 0.1}$	$79.5_{\pm 0.1}$	$76.4_{\pm 0.2}$
Trak	$71.3_{\pm 0.5}$	$78.2_{\pm 0.2}$	$79.5_{\pm 0.2}$	$86.1_{\pm 0.0}$	$78.5_{\pm 0.2}$	$76.5_{\pm 0.1}$
BIF	$71.4_{\pm 0.3}$	$78.0_{\pm 0.3}$	$79.7_{\pm 0.3}$	$85.8_{\pm 0.2}$	$78.9_{\pm 0.3}$	$76.6_{\pm 0.2}$
LESS	$71.4_{\pm 0.0}$	$78.0_{\pm 0.3}$	$79.6_{\pm 0.2}$	$86.4_{\pm 0.2}$	$79.5_{\pm 1.2}$	$76.9_{\pm 0.8}$
Linear surrogates	$71.8_{\pm 0.4}$	$78.2_{\pm 0.1}$	$79.9_{\pm 0.1}$	$86.1_{\pm 0.1}$	$79.1_{\pm 0.3}$	$76.6_{\pm 0.1}$
SKS (Algorithm 1)	$\mathbf{73.6}_{\pm 0.3}$	$\mathbf{79.6}_{\pm 0.1}$	$\mathbf{81.3}_{\pm 0.1}$	$\mathbf{88.1}_{\pm 0.2}$	$\mathbf{81.6}_{\pm 0.2}$	$\mathbf{77.9}_{\pm 0.3}$
Methods	Avmnist	Humor	Imdb	Mosei	Mosi	Sarcasm
Influence functions	$34.9_{\pm 0.9}$	$56.6_{\pm 1.1}$	$52.8_{\pm 1.1}$	$62.9_{\pm 1.2}$	$67.8_{\pm 1.0}$	$56.5_{\pm 1.9}$
DSIR	$39.7_{\pm 0.4}$	$53.1_{\pm 0.7}$	$51.9_{\pm 0.8}$	$60.9_{\pm 0.9}$	$62.3_{\pm 1.1}$	$56.8_{\pm 1.5}$
TracIn	$52.9_{\pm 0.6}$	$56.6_{\pm 1.1}$	$51.9_{\pm 1.4}$	$63.2_{\pm 0.8}$	$60.9_{\pm 1.5}$	$53.6_{\pm 2.1}$
Trak	$53.9_{\pm 0.8}$	$56.4_{\pm 0.9}$	$52.8_{\pm 0.9}$	$60.6_{\pm 1.1}$	$66.2_{\pm 0.8}$	$58.0_{\pm 1.5}$
BIF	$53.5_{\pm 1.0}$	$56.7_{\pm 0.7}$	$52.5_{\pm 1.1}$	$61.7_{\pm 0.9}$	$67.5_{\pm 0.9}$	$57.8_{\pm 1.8}$
LESS	$53.3_{\pm 0.6}$	$56.4_{\pm 0.8}$	$52.8_{\pm 1.0}$	$60.8_{\pm 0.9}$	$66.3_{\pm 1.2}$	$57.6_{\pm 1.2}$
Linear surrogates	$56.5_{\pm 1.2}$	$56.6_{\pm 0.8}$	$52.9_{\pm 1.0}$	$63.3_{\pm 0.8}$	$68.0_{\pm 0.9}$	$57.2_{\pm 1.0}$
SKS (Algorithm 1)	$\mathbf{61.5}_{\pm 0.9}$	$\mathbf{58.5}_{\pm 0.5}$	$\mathbf{53.9}_{\pm 1.3}$	$\mathbf{65.4}_{\pm 1.4}$	$\mathbf{69.2}_{\pm 1.1}$	$\mathbf{60.9}_{\pm 1.5}$

the largest gain on AV-MNIST. These results show that the learned sample-modality attributions can effectively guide downstream data selection across both traffic prediction and general multimodal benchmarks.

4.3 Ablation Studies

Ablation for the data attribution. We ablate the surrogate modeling methods in SKS. We test the performance of the flat surrogate kernel instead of the structured kernel. As shown in Figure 3a, SKS performs better than the flat surrogate kernel on DE, MA, and MD, with a 9.6% LDS improvement. Similar results are found on MultiBench, and we defer the results to Figure 5a. We then ablate the sample-level bandwidth γ_1 and the modality-level bandwidth γ_2 of the structured surrogate on MMTRACE. Across a broad range of bandwidth values, the LDS score remains stable.

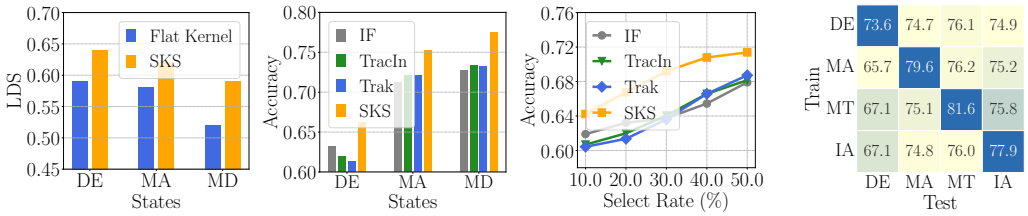
Ablation for the data selection. For data selection, we conduct two ablation studies on MMTRACE. First, we change the selected modality from vision to tabular information while keeping the same top-20% selection budget. As shown in Figure 3b, the accuracy drops noticeably compared with vision selection in Table 2. This indicates that tabular information is more critical for traffic accident prediction than satellite vision features. SKS still performs the best on DE, MA, and MD, with a 3.4% accuracy improvement. This shows that the learned attribution remains effective across different modalities. We defer the results for the rest of the states to Figure 5b.

Second, we vary the selection rate for the selected modality. Figure 3c shows that accuracy improves as the selection rate increases, but the gain gradually saturates at larger budgets. This suggests that the most useful samples are selected early, and that additional samples yield diminishing returns.

4.4 Local Geometric Measures

We empirically evaluate the cross-state generalization of SKS on MMTRACE. Specifically, we train the model on one source state using the downsampled data selected by SKS and test it on each target state. As shown in Figure 3d, the selected subsets generalize well across different source-target pairs. Among the source states, models trained on MT achieve stronger average performance than those trained on DE or MA, with an average accuracy of 74.9%. This shows that the selected MT subset may contain more transferable information for cross-state accident prediction.

Then, we assess model generalization through loss sharpness. A flatter loss landscape suggests that the selected subset induces a model less sensitive to small input perturbations and more likely to



(a) Compositional kernel (b) Data modality (c) Selection rate (d) Cross-state generalization

Figure 3: Ablation studies on the MMTRACE dataset [44]. Figure 3a evaluates surrogate modeling by comparing the LDS of SKS with the flat kernel on Delaware, Massachusetts, and Maryland. Figure 3b studies the effect of data modalities. We retain all visual data, select 20% of the tabular data using each method, and report the resulting test accuracy. Figure 3c evaluates selection rates ranging from 10% to 50% and reports the test accuracy on Delaware. Figure 3d evaluates cross-state generalization by training the model on one state and directly testing it on another state.

generalize across states. Since directly computing the Hessian trace is expensive, we estimate it by adding small isotropic Gaussian noise and measuring the expected loss increase.

We use the trace of the Hessian matrix of the trained model for evaluating the generalization ability [5, 40]. We compare the model trained on different states in MMTraCE with the training data selected by different attribution methods. As shown in Table 3, SKS not only achieves the lowest test loss, but also has the lowest Hessian trace value. This shows that the model on data selected by SKS generalizes better than models trained on data selected by other attribution methods.

Table 3: The Hessian trace and test loss of the model trained on Delaware, Massachusetts, Maryland, and Nevada. We compare SKS against several attribution methods including influence functions [8], TracIn [29], Trak [27] and BIF [9].

Methods	DE		MA		MD		NV	
Metric	Trace	Loss	Trace	Loss	Trace	Loss	Trace	Loss
Influence functions	3.88×10^2	0.61	8.75×10^2	0.84	6.17×10^3	1.04	2.53×10^2	0.70
TracIn	4.69×10^2	0.62	3.63×10^2	0.90	1.19×10^4	1.14	2.74×10^2	0.74
Trak	5.45×10^2	0.60	3.27×10^2	0.87	1.90×10^3	1.11	3.95×10^2	0.71
BIF	3.76×10^2	0.61	1.47×10^3	0.81	7.84×10^3	1.11	4.70×10^2	0.74
SKS (Algorithm 1)	$\mathbf{2.46 \times 10^2}$	**0.59**	$\mathbf{2.68 \times 10^2}$	**0.79**	$\mathbf{1.59 \times 10^3}$	**0.98**	$\mathbf{7.35 \times 10^1}$	**0.67**

5 Related Work

Multimodal representation learning. Multimodal learning relies on relationships between information sources to improve prediction. Different modalities can compensate for missing or unreliable information [39] and provide context needed to interpret one another [38]. Learning these relationships provides a basis for constructing useful representations and reducing the number of training samples required [6]. The reliability of these relationships also matters. Sun et al. [31] model uncertainty in image-text alignment to capture whether the available information supports a connection between their meanings.

These relationships guide representation learning by identifying which information should be preserved and how it should be organized. For example, semantic correspondence between language and vision guides representations to emphasize visual content relevant to the text and bring their meanings into alignment [7, 2]. To connect representations to prediction, deep supervision makes representations themselves predictive rather than relying solely on the final output [10], and knowledge transfer through representations supplies localization information missing from target labels [32].

Missing information and temporal changes also shape what representations need to capture. A modality's absence can itself carry information, so representations should account for what is missing as well as what is observed [21]. For evolving states, this missingness information helps refine estimates based on past observations, so that representations can track changes at each time step [22].

Beyond using observations to track changing states, representations can model how feedback drives subsequent changes [26].

Algorithms for influence estimation. The problem of influence estimation aims to quantify how individual examples affect predictions. Classical influence functions [8] estimate the effect of upweighting or removing a training point through a local second-order approximation. However, estimating inverse Hessian-vector products is still expensive for large models. To reduce computation, gradient similarity across training checkpoints avoids explicit Hessian computation [29], and model linearization with random projections enables attribution at larger scales [27]. Other than computational efficiency, attribution also raises the question of whether estimated importance rankings are reliable. For feature attribution, Goldwasser and Hooker [3] address this question by testing whether differences in estimated scores support the reported rankings despite sampling uncertainty.

Attribution also extends from individual examples to entire tasks. In multi-task learning, a task's contribution depends on the other tasks trained with it. Surrogate models estimate how task combinations affect performance, revealing positive and negative transfer [12]. Since a task's contribution depends on its training partners, estimating interactions helps determine which tasks should share a model [11, 13]. The cost of comparing combinations motivates first-order approximations of fine-tuning performance [16, 45]. These estimates inform both which tasks to group during optimization [42, 17] and where they should share parameters within a model [14, 46]. Theoretical analysis explains how input distributions and task relationships influence the benefits of sharing [37].

Information-theoretic perspectives on multimodal learning. Information-theoretic analyses relate the benefits of multimodal learning to how predictive information is distributed across modalities. Shared information allows unlabeled paired data to constrain learning [30], and bounds on information available only through joint observation quantify the potential gains from combining modalities [19]. Beyond information content, analyses of data augmentation distinguish gains from expanded data coverage and regularization [34]. Complementary optimization work addresses the tradeoff between data reuse and the distribution mismatch induced by repeated updates [33]. These perspectives characterize conditions for effective data use, but do not directly quantify how individual samples and their modalities affect test performance.

6 Conclusion

We studied cross-sample modality attribution in multimodal learning. We proposed SKS, a structured kernel surrogate that estimates the subset-to-loss response while preserving the hierarchy between sample participation and within-sample modality composition. We further applied the learned sample-modality attributions to downstream data selection. Across MMTRACE and MultiBench, SKS improves over prior work in terms of attribution quality and data selection accuracy. Further geometric analysis shows that models trained on data selected by SKS have a lower Hessian trace, suggesting a flatter loss surface and better generalization.

Broader Impact and Ethics Statement

This paper studies the connection between influence estimation and multimodal learning. This work can help us understand how the multimodal model learns from different modalities in different samples. There are no negative social impacts in this work.

Acknowledgment

This work received financial support from NSF award IIS-2412008 and a startup fund from Northeastern University. Ziniu Zhang and Zhenshuo Zhang received additional financial support from Northeastern University's Khoury PhD fellowships. R. Xiong is supported in part by the URC award at Emory University.

References

[1] A. Balasingam, J. Chandler, C. Li, Z. Zhang, and H. Balakrishnan. "Drivetrack: A benchmark for long-range point tracking in real-world videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 22488–22497.

[2] Z. Chen, X. Zhang, H. Xu, J. Xie, and Z. Tu. "Cvp: Central-peripheral vision-inspired multimodal model for spatial reasoning". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2026, pp. 2295–2305.

[3] J. Goldwasser and G. Hooker. "Statistical Significance of Feature Importance Rankings". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2025, pp. 1476–1496.

[4] S. Gupta, S. Sundaram, C. Wang, S. Jegelka, and P. Isola. "Better together: Leveraging unpaired multimodal data for stronger unimodal models". In: *arXiv preprint arXiv:2510.08492* (2025).

[5] H. Ju, D. Li, and H. R. Zhang. "Robust fine-tuning of deep neural networks with hessian-based generalization guarantees". In: *International conference on machine learning*. PMLR. 2022, pp. 10431–10461.

[6] S. M. Kakade and D. P. Foster. "Multi-view regression via canonical correlation analysis". In: *International Conference on Computational Learning Theory*. Springer. 2007, pp. 82–96.

[7] Z. Khan and Y. Fu. "Contrastive Alignment of Vision to Language Through Parameter-Efficient Transfer Learning". In: *International Conference on Learning Representations*. 2023.

[8] P. W. Koh and P. Liang. "Understanding black-box predictions via influence functions". In: *International conference on machine learning*. PMLR. 2017, pp. 1885–1894.

[9] P. A. Kreer, W. Wu, M. Adam, Z. Furman, and J. Hoogland. "Bayesian Influence Functions for Hessian-Free Data Attribution". In: *International Conference on Learning Representations (ICLR)*. 2026.

[10] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. "Deeply-supervised nets". In: *Artificial Intelligence and Statistics*. Pmlr. 2015, pp. 562–570.

[11] D. Li, H. Ju, A. Sharma, and H. R. Zhang. "Boosting multitask learning on graphs through higher-order task affinities". In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2023, pp. 1213–1222.

[12] D. Li, H. Nguyen, and H. R. Zhang. "Identification of Negative Transfers in Multitask Learning Using Surrogate Models". In: *Transactions on Machine Learning Research* (2023).

[13] D. Li, A. Sharma, and H. R. Zhang. "Scalable multitask learning using gradient-based estimation of task affinity". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 1542–1553.

[14] D. Li, Z. Zhang, M. Duan, E. Dobriban, and H. R. Zhang. "Efficiently Learning Branching Networks for Multitask Algorithmic Reasoning". In: *arXiv preprint arXiv:2512.01113* (2025).

[15] D. Li, Z. Zhang, L. Wang, and H. R. Zhang. "Scalable fine-tuning from multiple data sources: A first-order approximation approach". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024, pp. 5608–5623.

[16] D. Li, Z. Zhang, L. Wang, and H. R. Zhang. "Efficient ensemble for fine-tuning language models on multiple datasets". In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2025, pp. 25347–25364.

[17] D. Li, Z. Zhang, L. Wang, and H. R. Zhang. "MoPLEx: Estimating Plackett-Luce Mixture Models for Multi-Objective Alignment". In: *Conference on Empirical Methods in Natural Language Processing*. 2026.

[18] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L.-P. Morency. "Quantifying & modeling multimodal interactions: An information decomposition framework". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 27351–27393.

[19] P. P. Liang, C. K. Ling, Y. Cheng, A. Obolenskiy, Y. Liu, R. Pandey, A. Wilf, L.-P. Morency, and R. Salakhutdinov. "Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications". In: *The Twelfth International Conference on Learning Representations*. 2024.

[20] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L.-P. Morency. "Multibench: Multiscale benchmarks for multimodal representation learning". In: *Advances in neural information processing systems* 2021.DB1 (2021), p. 1.

[21] Z. Liang, Z. Pan, and R. Xiong. "Causal representation learning from multimodal clinical records under non-random modality missingness". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 28779–28796.

[22] Z. Liang, Z. Pan, and R. Xiong. "Learning Dynamic Representations and Policies from Multimodal Clinical Time-Series with Informative Missingness". In: *Findings of the Association for Computational Linguistics: ACL 2026*. 2026, pp. 26363–26392.

[23] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. "Learn to explain: Multimodal reasoning via thought chains for science question answering". In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 2507–2521.

[24] S. Malladi, A. Wettig, D. Yu, D. Chen, and S. Arora. "A kernel-based view of language model fine-tuning". In: *International Conference on Machine Learning (ICML)*. PMLR. 2023, pp. 23610–23641.

[25] A. Nippani, D. Li, H. Ju, H. Koutsopoulos, and H. Zhang. "Graph neural networks for road safety modeling: Datasets and evaluations for accident analysis". In: *Advances in neural information processing systems* 36 (2023), pp. 52009–52032.

[26] Z. Pan, Z. Liang, and R. Xiong. "Two-Sided State-Space Models for Sequential Recommendation with Non-Random Multimodal Review Feedback". In: *arXiv preprint arXiv:2609.00165* (2026).

[27] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Mądry. "TRAK: attributing model behavior at scale". In: *Proceedings of the 40th International Conference on Machine Learning*. 2023, pp. 27074–27113.

[28] T. Poggio and F. Girosi. "Networks for approximation and learning". In: *Proceedings of the IEEE* 78.9 (1990), pp. 1481–1497.

[29] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. "Estimating training data influence by tracing gradient descent". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19920–19930.

[30] K. Sridharan and S. M. Kakade. "An information theoretic framework for multi-view learning". In: *COLT*. 114. 2008, pp. 403–414.

[31] G. Sun, Y. Bai, X. Yang, Y. Fang, Y. Fu, and Z. Tao. "Aligning out-of-distribution web images and caption semantics via evidential learning". In: *Proceedings of the ACM Web Conference 2024*. 2024, pp. 2271–2281.

[32] Y. Wang, R. Guerrero, and V. Pavlovic. "D2F2WOD: Learning Object Proposals for Weakly-Supervised Object Detection via Progressive Domain Adaptation". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2023, pp. 22–31.

[33] Z. Wang, Z. Wang, J. Fu, X. Qu, Q. Cheng, S. Tang, M. Zhang, and X. Huo. "Slow-fast policy optimization: Reposition-before-update for llm reasoning". In: *arXiv preprint arXiv:2510.04072* (2025).

[34] S. Wu, H. Zhang, G. Valiant, and C. Ré. "On the generalization effects of linear transformations in data augmentation". In: *International conference on machine learning*. PMLR. 2020, pp. 10410–10420.

[35] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen. "Less: Selecting influential data for targeted instruction tuning". In: *International Conference on Machine Learning (ICML)*. PMLR. 2024.

[36] S. M. Xie, S. Santurkar, T. Ma, and P. S. Liang. "Data selection for language models via importance resampling". In: *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023), pp. 34201–34227.

[37] F. Yang, H. R. Zhang, S. Wu, C. Ré, and W. J. Su. "Precise high-dimensional asymptotics for quantifying heterogeneous transfers". In: *Journal of Machine Learning Research (JMLR)* 26.113 (2025), pp. 1–88.

[38] J. Yang, J. Gao, Y. Zhang, X. Chen, and A. Waibel. "An automatic sign recognition and translation system". In: *Proceedings of the 2001 workshop on Perceptive user interfaces*. 2001, pp. 1–8.

[39] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. "Multimodal people ID for a multimedia meeting browser". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. 1999, pp. 159–168.

[40] H. R. Zhang, D. Li, and H. Ju. "Noise Stability Optimization for Finding Flat Minima: A Hessian-based Regularization Approach". In: *Transactions on Machine Learning Research* (2024).

[41] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Re. "Correct-N-Contrast: a Contrastive Approach for Improving Robustness to Spurious Correlations". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. 2022, pp. 26484–26516.

[42] Z. Zhang, M. Duan, Y. Ye, and H. R. Zhang. "Scalable multi-objective and meta reinforcement learning via gradient estimation". In: *arXiv preprint arXiv:2511.12779* (2025).

[43] Z. Zhang, M. Duan, and H. Zhang. "Efficient estimation of kernel surrogate models for task attribution". In: *International Conference on Learning Representations (ICLR)*. Vol. 2026. 2026, pp. 61822–61848.

[44] Z. Zhang, M. Duan, H. Koutsopoulos, and H. R. Zhang. "Learning Multimodal Embeddings for Traffic Accident Prediction and Causal Estimation". In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2026, pp. 2901–2912.

[45] Z. Zhang, Z. Zhang, D. Li, L. Wang, J. Dy, and H. R. Zhang. "Linear-Time Demonstration Selection for In-Context Learning via Gradient Estimation". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 16470–16488.

[46] Z. Zhang, Z. Zhang, R. Xiong, G. Cooperman, and H. R. Zhang. "Long-Context Demonstration Selection Using State Space Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2026*. 2026.

A Omitted Materials from Section 3

We first introduce the notations. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote a training dataset of n samples, where $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(M)})$ contains M input modalities and y_i is the corresponding output. We use S to denote the selection state of the samples and their corresponding modalities. $\mathbf{v}(S)$ indicates which samples retain at least one modality under S. Let f_W be a multimodal model parameterized by W, let ℓ be the per-sample loss, and $L(f_W; S)$ be the test loss of the model trained under S. The RBF kernels k_1 and k_2 compare which samples are retained and which modalities are retained within each sample, with bandwidths γ_1 and γ_2, respectively. Their composition defines $K(S, S')$ for two different selections. Let $\mathcal{H}_K$ be the reproducing kernel Hilbert space induced by K, we fit a surrogate function $g \in \mathcal{H}_K$ using P sampled combinations of retained samples and modalities. For kernel ridge regression, $\mathbf{K}$ denotes the Gram matrix with entries $K_{pq} = K(S_p, S_q)$, $\mathbf{k}_S$ denotes the kernel-similarity vector between a query selection state S and the P sampled states, and λ is the ridge regularization parameter.

Given a selection state S, $A^{(j,m)}(S)$ denotes the attribution of modality m in sample j. These attribution values are used to rank and select sample-modality pairs for downstream model training.

We use B_M as the Gram matrix over its nonzero M-modality states and ρ to bound the RKHS norm of the mapping $S \mapsto L(f_W; S)$, so that $\|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho$. We further use ϵ to denote the error introduced by the first-order approximation.

A.1 Positive Semidefiniteness of K

We use K to measure similarity between different selections and fit a loss predictor through kernel ridge regression. To establish its validity as a kernel, for a sample j, let $B_M \in \mathbb{R}^{(2^M-1)\times(2^M-1)}$ be the Gram matrix indexed by the nonzero modality states $S_{:,j}, S'_{:,j} \in \{0,1\}^M \setminus \{\mathbf{0}\}$, with entries

$$(B_M)_{S_{:,j}, S'_{:,j}} = \exp\left(-\gamma_2 \|S_{:,j} - S'_{:,j}\|_2^2\right)$$

Let $\mathbf{1} \in \mathbb{R}^{2^M-1}$ denote the all-ones vector.

Proposition 2 (Positive semidefiniteness of K)**.** *The structured design K is positive semidefinite over all sample and modality selections if*

$$\gamma_2 > 0 \quad \textit{and} \quad \gamma_1 \geq \frac{1}{2} \log\left(\mathbf{1}^\top B_M^{-1} \mathbf{1}\right).$$

Proof. To analyze the positive semidefiniteness of K, we first rewrite K into an equivalent product form over the samples. Since the first term is

$$\exp\left(-\gamma_1 \|\mathbf{v}(S) - \mathbf{v}(S')\|_2^2\right) = \prod_j \exp\left(-\gamma_1 \|\mathbf{v}^{(j)}(S) - \mathbf{v}^{(j)}(S')\|_2^2\right),$$

and the second term is

$$\prod_{j \in [n]: \mathbf{v}^{(j)}(S) = \mathbf{v}^{(j)}(S') = 1} \exp\left(-\gamma_2 \|S_{:,j} - S'_{:,j}\|_2^2\right) = \prod_j \exp(-\gamma_2 \mathbf{v}^{(j)}(S) \mathbf{v}^{(j)}(S') \|S_{:,j} - S'_{:,j}\|_2^2),$$

we have the alternative formulation of K

$$K(S, S') = \prod_{j=1}^n \exp\left(-\gamma_1 \|\mathbf{v}^{(j)}(S) - \mathbf{v}^{(j)}(S')\|_2^2 - \gamma_2 \mathbf{v}^{(j)}(S) \mathbf{v}^{(j)}(S') \|S_{:,j} - S'_{:,j}\|_2^2\right).$$

Since $K(S, S')$ is a product over samples, K is positive semidefinite if the factor corresponding to each sample is positive semidefinite. Fix a sample j. The j-th factor in the product measures the similarity between the modality selections $S_{:,j}$ and $S'_{:,j}$ for this sample. Ordering the possible columns with the zero vector first, followed by all nonzero binary vectors, gives the Gram matrix

$$\begin{pmatrix} 1 & e^{-\gamma_1} \mathbf{1}^\top \\ e^{-\gamma_1} \mathbf{1} & B_M \end{pmatrix}.$$

where $\mathbf{1} \in \mathbb{R}^{2^M-1}$ is the all-ones vector and B_M is indexed by the nonzero columns $S_{:,j}, S'_{:,j} \in \{0,1\}^M$, with entries

$$(B_M)_{S_{:,j},S'_{:,j}} = \exp\left(-\gamma_2 \|S_{:,j} - S'_{:,j}\|_2^2\right).$$

Since B_M is the Gram matrix of an RBF kernel on distinct points, it is positive definite for $\gamma_2 > 0$. By the Schur complement, the block matrix above is positive semidefinite if and only if

$$1 - e^{-2\gamma_1} \mathbf{1}^\top B_M^{-1} \mathbf{1} \geq 0,$$

which is equivalent to

$$\gamma_1 \geq \frac{1}{2} \log\left(\mathbf{1}^\top B_M^{-1} \mathbf{1}\right).$$

Under this condition, the factor corresponding to sample j is positive semidefinite over all possible columns $S_{:,j}$. Since K is a product of these factors, it is also positive semidefinite.

■

Thus, kernel ridge regression has a unique solution for any finite collection of sampled selections and any positive regularization parameter.

A.2 Proof of Proposition 1

Proof. Since $\hat{g}$ minimizes the kernel ridge regression objective over $\mathcal{H}_K$, its objective value cannot exceed that obtained by choosing $g = L(f_W; \cdot) \in \mathcal{H}_K$. Thus,

$$\begin{aligned}
&\frac{1}{P} \sum_{p=1}^{P} \left(\hat{g}(S_p) - L(f_W; S_p)\right)^2 + \lambda \|\hat{g}\|_{\mathcal{H}_K}^2 \\
&\leq \frac{1}{P} \sum_{p=1}^{P} \left(L(f_W; S_p) - L(f_W; S_p)\right)^2 + \lambda \|L(f_W; \cdot)\|_{\mathcal{H}_K}^2 \\
&= \lambda \|L(f_W; \cdot)\|_{\mathcal{H}_K}^2 \leq \lambda \rho^2,
\end{aligned}$$

where the last inequality uses the assumption $\|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho$. Thus,

$$\frac{1}{P} \sum_{p=1}^{P} \left(\hat{g}(S_p) - L(f_W; S_p)\right)^2 \leq \lambda \rho^2, \qquad \|\hat{g}(\cdot)\|_{\mathcal{H}_K} \leq \rho.$$

By the triangle inequality, together with $\|\hat{g}(\cdot)\|_{\mathcal{H}_K} \leq \rho$ and the assumption $\|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho$, we obtain

$$\|\hat{g}(\cdot) - L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \|\hat{g}(\cdot)\|_{\mathcal{H}_K} + \|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho + \rho = 2\rho.$$

Since $S_1, \ldots, S_P$ are independent, $K(S, S) = 1$, and

$$P \geq \frac{C_0}{\lambda} \log\left(\frac{C_0}{\lambda \delta}\right)$$

for a sufficiently large universal constant C_0, the operator Bernstein inequality bounds the expected squared error by the empirical squared error and a regularization term. Specifically, with probability at least $1 - \delta$, simultaneously for every $g \in \mathcal{H}_K$,

$$\mathbb{E}_S\left[\left(g(S) - L(f_W; S)\right)^2\right] \leq \frac{2}{P} \sum_{p=1}^{P} \left(g(S_p) - L(f_W; S_p)\right)^2 + \lambda \|g(\cdot) - L(f_W; \cdot)\|_{\mathcal{H}_K}^2.$$

Applying this inequality to $\hat{g}(\cdot)$ and using the bounds above yields

$$\mathbb{E}_S\left[\left(\hat{g}(S) - L(f_W; S)\right)^2\right] \leq 2\lambda\rho^2 + 4\lambda\rho^2 = 6\lambda\rho^2.$$

This proves the claim with $C_1 = 6$. ■

A.3 Extensions of Algorithm 1

Multiple modalities. Our structured kernel surrogate naturally extends from two modalities to $M > 2$ modalities. In this case, we define the global sample-modality selection state as

$$S = \begin{bmatrix} S_{1,1} & \cdots & S_{1,n} \\ \vdots & \ddots & \vdots \\ S_{M,1} & \cdots & S_{M,n} \end{bmatrix} \in \{0,1\}^{M \times n},$$

where $S_{m,j} = 1$ denotes that modality m of sample j is retained, and $S_{m,j} = 0$ otherwise. Accordingly, the j-th column $S_{:,j} \in \{0,1\}^M$ represents which modalities are retained in sample j, and $\mathbf{v}^{(j)}(S) = \max(S_{:,j})$.

The structured kernel retains the same form as in the two-modality case. The sample-level kernel measures the distance between $\mathbf{v}(S)$ and $\mathbf{v}(S')$, while the modality-level kernel measures the distance between the corresponding M-dimensional states $S_{:,j}$ and $S'_{:,j}$. The positive semidefiniteness condition changes only through the local state space. Specifically, let B_M denote the Gram matrix over the $2^M - 1$ nonzero modality states in $\{0,1\}^M$. Then the structured kernel is positive semidefinite if

$$\gamma_2 > 0 \quad \text{and} \quad \gamma_1 \geq \frac{1}{2} \log\left(\mathbf{1}^\top B_M^{-1} \mathbf{1}\right).$$

Thus, SKS extends directly by replacing the two-dimensional sample-modality state with $S_{:,j} \in \{0,1\}^M$.

B Omitted Experiments

B.1 Implementation Details

Models and hyperparameters. In the MultiBench experiments, we use the cross-attention network as the fusion model. The model directly takes the pre-computed modality embeddings from the datasets as inputs. We report the hyperparameters in Table 4. In the MMTRACE experiments, we use the standard three-layer GNN as the encoder and leverage an MLP as the predictor. The vision embedding is fetched by a pretrained CLIP and ViT model. We report the hyperparameters in Table 5. The models we use for validating the first-order approximation are Qwen2-VL-2B, Qwen3-VL-8B, Qwen3-VL-30B-A3B, and Llama-3.2-11B-Vision. All the experiments are conducted on one NVIDIA RTX A6000 GPU and one NVIDIA B200 GPU. The number P of $\{S_p, L(f_W; S_p)\}_{p=1}^P$ used for training the surrogate model is 100 on both MMTRACE and MultiBench.

Table 4: Hyperparameters of the Cross-Attention model.

Hyperparameter	**Value**
Projection dimension	256
Number of attention heads	4
Number of layers	2
Batch size	4096
Dropout probability	0.1
Number of epochs	15
Learning rate	5×10^{-4}
Optimizer	AdamW
Weight decay	1×10^{-4}

Since instance-level selections introduce high variance and computational overhead, we design the surrogate sampling at the group level. We partition the training data into 25 clusters using K-means. To fit the surrogate model, we generate selections by masking each cluster with probability 0.8 and compute the loss for each selection. We use 100 selections for each of MMTraCE and MultiBench.

Baselines. For all baselines, each input sample initially contains information from all available modalities. For example, for one modality in a sample, we mask the selected modality of this sample

Table 5: Hyperparameters of the GNN model used in MMTRACE.

Hyperparameter	Value
Number of GNN layers	3
Number of Classifier layers	2
Batch size	4096
Hidden dimension	256
Dropout probability	0.1
Number of epochs	15
Learning rate	5×10^{-4}
Optimizer	AdamW
Weight decay	1×10^{-4}

(with zeros), and keep the remaining modalities of the sample unchanged. Then, we apply the baseline procedure.

Influence functions [8] estimate how upweighting one masked sample affects the loss of another masked sample. We flatten the modalities across samples into individual training items, treating modality m of sample i ($\mathbf{x}_i^{(m)}$) as a separate item with label y_i. We compute the loss gradient for each item and the gradient of the average test loss. Their negative inner product, weighted by the inverse Hessian of the training objective, gives the influence score for that item.

DSIR [36] estimates an importance weight for $\mathbf{x}_i^{(m)}$ by taking the ratio of its likelihood under the target distribution to its likelihood under the training distribution in a reduced feature space. A larger ratio indicates that $\mathbf{x}_i^{(m)}$ is more representative of the target distribution. We use this weight to rank and select sample-modality pairs under the fixed budget.

TracIn [29] estimates pairwise influence by accumulating gradient similarity over multiple training checkpoints. Treating each $\mathbf{x}_i^{(m)}$ as an individual training item with label y_i, we compute its loss gradient and the gradient of the average test loss at each saved checkpoint. We then take the inner product between these gradients, weight it by the corresponding learning rate, and sum across checkpoints to obtain the attribution score for $\mathbf{x}_i^{(m)}$. A positive score indicates that training on this item is estimated to reduce the average test loss.

TRAK [27] estimates how each training item contributes to a target prediction. Treating each $\mathbf{x}_i^{(m)}$ as an individual training item with label y_i, we compute gradients of the log-odds of the correct-class probability at $W^\star$ for both training items and target examples. We randomly project these gradients into a lower-dimensional space and compute their similarity using the inverse Gram matrix of the projected training gradients. Each training item's score is then weighted by one minus the probability assigned to y_i. Averaging over target examples gives the attribution score for $\mathbf{x}_i^{(m)}$.

BIF [9] estimates influence through loss covariance over a localized parameter distribution, avoiding the inverse-Hessian computation used by classical influence functions. Starting from $W^\star$, we use stochastic gradient Langevin dynamics to sample model parameters from a local posterior defined by the training loss and a Gaussian prior centered at $W^\star$. At each sampled parameter value W, we evaluate $\ell(f_W(\mathbf{x}_i^{(m)}), y_i)$ for each individual training item and the average test loss. We compute their negative empirical covariance across parameter draws as the attribution score for $\mathbf{x}_i^{(m)}$. We use the resulting scalar as the attribution score between the two sample-modality pairs.

LESS [35] estimates influence using optimizer-aware gradient similarity. After a short LoRA warm-up, it records model checkpoints and Adam optimizer states. For the masked candidate and target inputs, LESS computes an Adam-aware update feature for the candidate and a loss gradient for the target, projects both into a shared low-dimensional space, and accumulates their learning-rate-weighted cosine similarity across checkpoints. We use the resulting value as the attribution score between the two sample-modality pairs.

Choice of the RBF Kernel. We use the RBF kernel because it provides expressivity for any attribution function theoretically. Zhang et al. [43] show that the Gaussian RBF kernel is strictly positive definite on any finite set of distinct vectors in $\mathbb{R}^d$. This implies that its Gram matrix is of

Table 6: Ablation study of nonlinear kernel choices on MMTraCE. Each entry reports the mean and standard deviation of the LDS score over three runs.

Kernel	DE	MA	MD	NV	MT	IA
ℓ_2-Laplacian	$0.58_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.50_{\pm 0.02}$	$0.48_{\pm 0.01}$	$0.29_{\pm 0.03}$	$0.65_{\pm 0.04}$
Quadratic	$0.50_{\pm 0.02}$	$0.47_{\pm 0.03}$	$0.52_{\pm 0.01}$	$0.40_{\pm 0.04}$	$0.33_{\pm 0.01}$	$0.56_{\pm 0.02}$
RBF	$0.64_{\pm 0.03}$	$0.62_{\pm 0.05}$	$0.59_{\pm 0.10}$	$0.58_{\pm 0.02}$	$0.57_{\pm 0.06}$	$0.70_{\pm 0.03}$



Reference subset



Positive interaction with reference subset



Negative interaction with reference subset

Figure 4: Examples of cross-sample modality interactions on MMTraCE. The top row shows the reference subset, while the middle and bottom rows show satellite images with positive and negative interactions with the reference subset, respectively. Positive interactions generally share similar road structures, and negative interactions exhibit substantially different spatial patterns.

full rank. Therefore, for any function F, there exists a kernel mapping that matches F on the input vectors. Classical results [28] guarantee that on this discrete domain, the RBF kernel can approximate any attribution function without loss of expressivity [43].

Moreover, the RBF kernel encodes an inductive bias that matches the geometry of the subset space. On $\{0,1\}^d$, two subsets that differ on only a few coordinates are close in Hamming distance, and intuitively their performances tend to be similar. The RBF kernel directly reflects this structure. By contrast, polynomial kernels do not encode distance-based locality; two subsets with small Hamming distance are not necessarily mapped to nearby points in the feature space.

While we utilize the squared Euclidean distance function within the RBF kernel, the distance measure can be set to other types of measures depending on the application scenario. We conducted an ablation analysis comparing the RBF kernel to using an ℓ_2-Laplacian kernel ($\exp(-\gamma\|S-S'\|_2)$), and another quadratic kernel. The results are reported in Table 6, which show that the RBF kernel gives a 31.8% average advantage over the other two kernels in LDS scores on the MMTraCE benchmark.

B.2 Cross-sample Modality Interaction

We examine satellite image pairs from MMTraCE to understand how visual information relates to interactions between samples. As shown in Figure 4, positive interactions are observed when both images contain similar complex road structures, such as dense residential networks, curved or wide roads, and intersections. This shows that they reinforce each other. In contrast, negative interactions occur when the images differ substantially, such as when a dense urban scene is paired with a sparse, low-density, or rural scene. This demonstrates that the latter may weaken the evidence provided by the former.

B.3 First-order Approximation in Multimodal Models

To evaluate the accuracy of the first-order approximation, we measure the relative mean squared error (MSE) between the output of the real trained model and the approximated output using the validation set:

$$\text{Relative MSE}(S) = \frac{\sum_{j \in \mathcal{D}_{\text{val}}} \| f_W(\mathbf{x}_j; S_{:,j}) - f_{W^\star}(\mathbf{x}_j; S_{:,j}) - \nabla f_{W^\star}(\mathbf{x}_j; S_{:,j})(W - W^\star) \|_2^2}{\sum_{j \in \mathcal{D}_{\text{val}}} \| f_W(\mathbf{x}_j; S_{:,j}) \|_2^2}. \tag{4}$$

We empirically evaluate the accuracy and efficiency of the first-order approximation on MMTraCE in Table 7, where efficiency is measured by FLOPs. Across all six states, the approximation error remains lower than 1.2%. The linearized objective provides a speedup up to $18.9\times$ over exact retraining. These results show that the first-order approximation closely estimates the state-specific loss while substantially reducing the computational cost of constructing state-loss pairs.

Table 7: First-order approximation error and speedup over exact retraining on MMTraCE. Errors are reported as the mean and standard deviation over three runs.

	DE	MA	MD	NV	MT	IA
Relative MSE	$1.1_{\pm 0.2}\%$	$1.0_{\pm 0.4}\%$	$1.2_{\pm 0.1}\%$	$0.8_{\pm 0.2}\%$	$1.1_{\pm 0.1}\%$	$1.1_{\pm 0.5}\%$
Speed up	$15.39\times$	$4.07\times$	$4.74\times$	$5.62\times$	$18.9\times$	$14.7\times$

The complexity of first-order approximation is linear in the size of the model, thus, we apply it to pretrained VLMs to show the scalability. We evaluate the accuracy of this estimation procedure on the multimodal ScienceQA benchmark [23], with model architectures ranging from 2B to 30B parameters. As shown in Table 8, the error is less than 1.1% on all models.

Table 8: First-order approximation error on ScienceQA across pretrained vision-language models ranging from 2B to 30B parameters. Results report the mean and standard deviation of the relative MSE over three runs.

	Qwen2-VL-2B	Qwen3-VL-8B	Qwen3-VL-30B-A3B	Llama-3.2-11B-Vision
Relative MSE	$3.6_{\pm 0.5} \times 10^{-3}$	$2.9_{\pm 0.4} \times 10^{-3}$	$1.1_{\pm 0.76} \times 10^{-2}$	$1.1_{\pm 0.2} \times 10^{-2}$

B.4 Computational Costs

By using the first-order approximation, SKS avoids retraining the multimodal model for each sampled state, allowing the cost of constructing state-loss pairs to scale linearly with the input size. We report the total computational cost, measured in FLOPs, on MMTRACE in Table 9. Overall, SKS has a FLOP cost comparable to the baselines and is nearly identical to TracIn and LESS across all six tasks.

Table 9: Computational cost of SKS and baseline attribution methods on MMTRACE, measured by the total number of FLOPs.

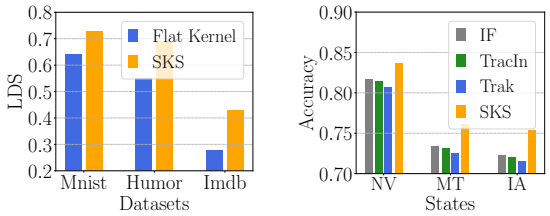
Method	DE	MA	MD	NV	MT	IA
Influence functions	6.09×10^{12}	4.36×10^{14}	1.49×10^{14}	5.86×10^{13}	1.80×10^{13}	7.75×10^{13}
DSIR	4.28×10^{12}	3.16×10^{14}	9.23×10^{13}	3.81×10^{13}	9.70×10^{12}	4.85×10^{13}
TracIn	5.26×10^{12}	4.34×10^{14}	1.47×10^{14}	5.65×10^{13}	1.60×10^{13}	7.54×10^{13}
LESS	5.17×10^{12}	4.34×10^{14}	1.47×10^{14}	5.65×10^{13}	1.60×10^{13}	7.54×10^{13}
BIF	5.73×10^{12}	4.89×10^{14}	1.64×10^{14}	6.38×10^{13}	1.68×10^{13}	8.52×10^{13}
SKS	5.15×10^{12}	4.34×10^{14}	1.47×10^{14}	5.67×10^{13}	1.59×10^{13}	7.55×10^{13}

B.5 Omitted Ablation Analysis

In Section 4.3, we study how the kernel construction and the selected modality affect the performance of SKS. Here, we report the remaining results on additional MultiBench datasets and MMTRACE states.

For the kernel-structure ablation, the flat kernel vectorizes each sample-modality selection matrix S into a single binary vector and applies one RBF kernel to the entire vector. It treats all sample-modality decisions as flat coordinates and does not distinguish sample-level similarity from modality-level similarity. We fit the same kernel ridge regression surrogate using each kernel and evaluate its LDS. Figure 5a reports the comparison on AV-MNIST, Humor, and MM-IMDb.

For the modality ablation, we change which modality is selected under the fixed data budget. In the ablation, we retain all visual data and select the top 20% of the tabular data according to each attribution method. We train the multimodal model on the resulting data subset and report its test accuracy. Figure 5b presents the remaining results on Nevada, Montana, and Iowa.



(a) Compositional kernel (b) Data modality

Figure 5: Additional ablation results on MultiBench and MMTRACE. Figure 5a compares the LDS of SKS with that of the flat surrogate kernel on AV-MNIST, Humor, and MM-IMDb. Figure 5b studies tabular-data selection on Nevada, Montana, and Iowa. We retain all visual data, select the top 20% of the tabular data using each method, and report the resulting test accuracy.

B.6 Discussions

Our current empirical evaluation focuses on standard multimodal fusion models. This establishes a baseline for the efficacy of SKS but bounds our current scope. Investigating the scalability and behavior of SKS on large-scale Vision-Language Models (VLMs) is a promising direction for future research.