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 study the problem of *cross-sample modality attribution*, which involves measuring how one modality in one sample affects the learning outcome in other multimodal samples. This problem is crucial in a multimodal system because a data modality (such as a text description) may be redundant for one sample (a self-explanatory image), but the same data modality would otherwise be uniquely informative for another sample, or informative when combined with other modalities. Prior work focuses on influence estimation at the level of one sample, and it remains an open question how one should capture the interactions across both samples and modalities. To capture scenarios like the above, we propose *structured kernel surrogate* models, which are surrogate mappings from both samples and modalities to the prediction outcome when they are trained together. We design an algorithm, SKS, which applies kernel ridge regression with a structured, compositional kernel on the selection states of individual samples and modalities. Specifically, the kernel function multiplies kernel-based similarity scores between every pair of data modalities for one sample and a pair of two samples. Additionally, we design a gradient-based estimation algorithm to efficiently estimate the structured kernel surrogate model. We evaluate our approach on two multimodal benchmarks, spanning a range of data modalities including satellite images, road networks, text, 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, SKS improves the average accuracy by 1.93% over baselines. Models with data selected by SKS have lower Hessian traces, indicating a flatter loss surface. Using SKS to select satellite images identifies road segments, where accident scenes commonly occur, as opposed to background images.

1 Introduction

Data in the form of multiple modalities are widely available in many domains. For example, in transportation, satellite images capturing road segments, traffic statistics such as traffic volume, and weather statistics can jointly be used to make accurate predictions on accident risks [27, 47]. Autonomous self-driving vehicles integrate visual cues and sounds from the surrounding environment to provide safety-critical navigation [1]. While multimodal systems have scaled up in terms of both model size and complexity, their internal mechanisms remain highly complex [46] (one concrete example where susceptible models may be brittle to feature representations is spurious correlations

[*]Equal contribution.

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

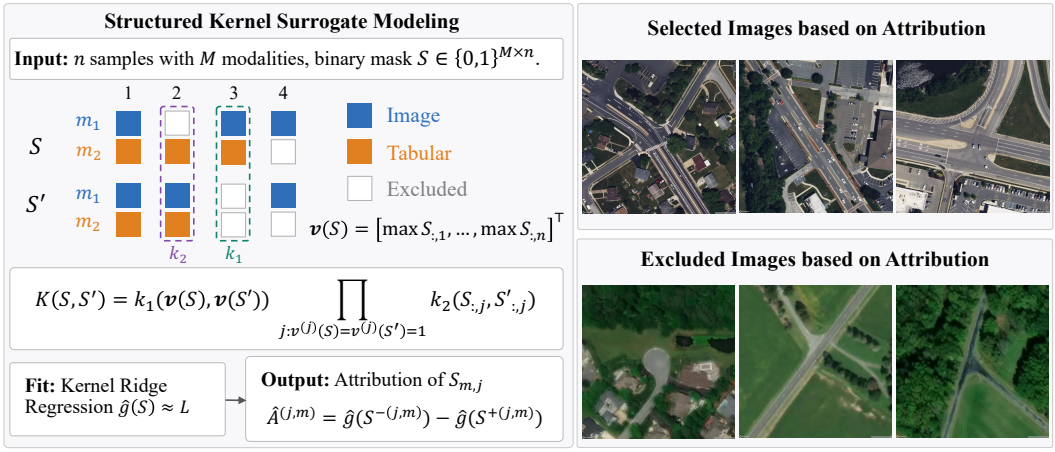


Figure 1: Overview of our proposed approach. We design a structured compositional kernel by multiplying both sample-level and modality-level RBF kernel similarity scores, given by k_1 and k_2 (with bandwidth parameters γ_1 and γ_2, respectively). Then, we fit a surrogate model mapping the selection states to their loss outcomes using kernel ridge regression on N pairs of randomly sampled selections and their corresponding loss outcomes (after fitting a model). Lastly, we estimate the influence score of one sample-modality pair (m, j) under a binary mask $S_{m,j}$ by subtracting the test losses under $S^{-(j,m)}$ and $S^{+(j,m)}$, which retain and remove this modality in the sample, with all other entries of S fixed. A positive score indicates that removing this pair (m, j) is detrimental. On the right, we give examples of satellite images retained and excluded based on their influence scores. Our approach correctly identifies the road segments from the top panel, where accident scenes often occur, and estimates a higher influence score for these segments compared to the background images from the bottom panel.

[44]). Thus, understanding how different modalities interact and impact predictions is central to providing meaningful interpretations to multimodal systems.

Compared to a unimodal system, data influences a multimodal system at different granularities. In this vein, the problem of cross-sample modality interaction can be related to earlier literature on multi-view learning. For instance, Kakade and Foster [7] use relationships estimated from unlabeled paired data to regularize predictors and encourage agreement across views (see also a follow-up work [32]). Another related concept built on partial information decomposition characterizes redundant, unique, or synergistic information in the presence of heterogeneous data modalities, under different combinations [19, 5]. While these existing works describe what information is present across modalities, the question of how individual samples and their modalities affect prediction remains open: is there an efficient algorithm for measuring the effect of adding or removing one modality from an individual sample on model performance?

Prior work in the literature of influence estimation studies the effect of adding or removing an individual example on model predictions. Existing methods, including influence functions [9], datamodels [29], and linear and kernel surrogate methods [13, 46], have considered the estimation problem both for individual samples and for groups (or sets) of samples. These existing works primarily focus on unimodal settings. In particular, the effect of adding or removing one modality from a sample can depend on how it interacts with other samples and modalities through a shared network or optimization procedure. The problem of estimating the influence of individual samples in multimodal learning remains open.

These gaps motivate us to introduce *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 dataset in which some samples are included or removed, and some modalities are retained or masked within each included sample. For each such sampling, the corresponding target is the loss of the model based on the perturbed 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. An implication of this work is thus connecting the literature on influence estimation in multimodal learning to concepts in causal inference (such as difference-in-differences and modality missingness [22, 23]).

We evaluate SKS on the multimodal traffic accident prediction benchmark MMTRACE [47] and the standard multimodal benchmark MultiBench [21]. 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 from 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 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. The code repository for reproducing the empirical findings can be found at https://github.com/VirtuosoResearch/Multimodal-attribution-SKS-code.

2 Preliminaries

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote a multimodal dataset sampled from an unknown 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 learn 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

$$\mathbf{v}(S) = [\max(S_{:,1}), ..., \max(S_{:,n})]^\top.$$

In particular, $\mathbf{v}^{(i)}(S) = 1$ 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 loss is denoted as $\ell\big(f_W(\mathbf{x}_i; S_{:,i}), y_i\big)$. 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 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 learned parameters W generally depend nonlinearly on S. Changing a modality in one sample changes the corresponding sample's contribution to the objective and, through shared parameter updates, can further affect how the model uses information from other samples.

We characterize the effect of a selection S by the expected risk of the resulting model under the test distribution $\mathcal{D}$:

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

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

Problem statement. We introduce an attribution score to measure how one modality of an individual 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 influence score 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 is dependent on the binary mask matrix S. This formulation raises two questions:

- First, how can we estimate the nonlinear mapping from the binary masks of samples and modalities specified by S to the loss outcome $L(f_W; S)$?
- Second, how can we use the surrogate model to efficiently estimate the influence score for each modality of an individual 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 individual 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.[2]

A compositional kernel. Given two binary mask matrices 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 $S_{:,j}$ 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 $\gamma_2 > 0$:

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

[2]We provide both a justification for the RBF kernel and an ablation of alternative kernel choices in Appendix B.4.

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

Input: N randomly sampled selections $\{S_1, S_2, \ldots, S_N\}$ and their evaluation loss outcomes based on corresponding selections $\{(S_p, \tilde{L}(f_W; S_p))\}_{p=1}^{N}$
Require: Sample-level RBF bandwidth γ_1, modality-level RBF bandwidth γ_2, reference selection matrix S, ridge regression penalty λ
Output: Surrogate model and estimated influence scores

/ Stage 1: calculate the kernel and solve a kernel ridge regression problem */*
for $p, q = 1, \ldots, N$ **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
$\mathbf{y} \leftarrow (\tilde{L}(f_W; S_1), \ldots, \tilde{L}(f_W; S_N))^\top$
$k_S \leftarrow (K(S, S_1), \ldots, K(S, S_N))^\top$
$\hat{g}(S) \leftarrow k_S^\top (\mathbf{K} + N\lambda \mathbf{I})^{-1} \mathbf{y}$
/ Stage 2: calculate the estimated influence scores */*
for each $j \in [n]$ and $m \in [M]$ **do**
  $S^{+(j,m)}, S^{-(j,m)} \leftarrow$ modify 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 $\hat{g}$, $\{\widehat{A}^{(j,m)}(S) : j \in [n],\ m \in [M]\}$

We then combine the sample-level and modality-level kernel similarity scores 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.

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.2.

3.2 Surrogate Modeling of Sample-Modality Interaction

Given a valid kernel, we then fit a surrogate model that predicts the test loss from S. We then estimate the attribution of each modality in each sample using finite differences. Each difference compares the predicted losses with the modality added 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 N sampled selection states $S_1, \ldots, S_N$ with measured test losses $L(f_W; S_1), \ldots, L(f_W; S_N)$, we fit a kernel ridge regression surrogate [46] by optimizing over candidate functions $g \in \mathcal{H}_K$, where $g(S)$ predicts the test loss associated with state S:

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

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

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

Influence estimation. 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

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

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.

For the surrogate $\hat{g}$, the N samples 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 $\|L(f_W; \cdot)\|_{\mathcal{H}_K} \leq \rho$. We show the following statement.

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 $N \geq \frac{C_0}{\lambda} \log\left(\frac{C_0}{\lambda\delta}\right)$, 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.1.

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 [25, 29, 16], we linearize the multimodal model around a checkpoint learned 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.2, we empirically evaluate ϵ and find it to be small in the estimation. The 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.2.

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.

4.1 Experimental Setup

Datasets. We first evaluate our framework on the spatiotemporal traffic accident prediction task MMTRACE [47]. 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 [21]: UR-FUNNY (Humor), MUStARD (Sarcasm), CMU-MOSI

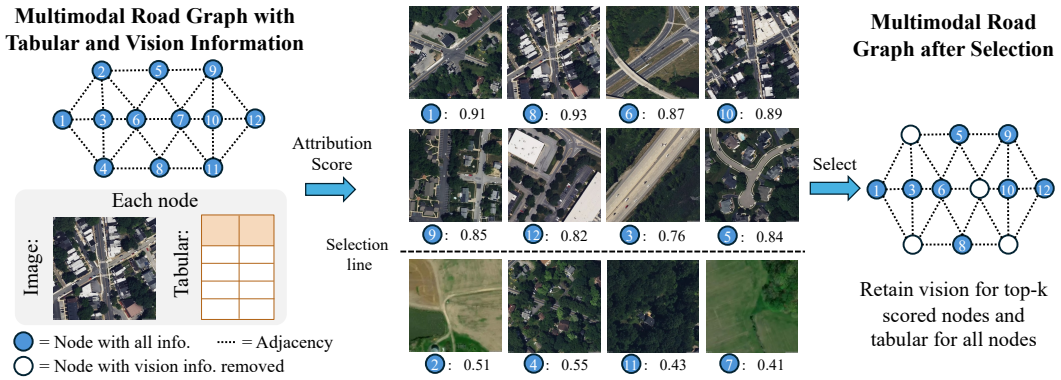


Figure 2: Overview of surrogate-guided data selection on MMTRACE. Each node represents a road-network sample with tabular and vision modalities. The structured kernel surrogate assigns an attribution score to each node image. We keep the vision information for the top-k scored nodes.

(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 [9], TracIn [31], Trak [29], DSIR [39], LESS [38], and BIF [10]. Lastly, we compare our approach with linear surrogate models [13].

Evaluation. We denote the evaluated sampling states by $\{S_p\}_{p=1}^{N}$, 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: $\text{LDS} = \text{Spearman}\left(\{\hat{g}(S_p)\}_{p=1}^{N}, \{y_p\}_{p=1}^{N}\right)$. 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 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 fit 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 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.

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)	$\mathbf{0.73}_{\pm 0.04}$	$\mathbf{0.69}_{\pm 0.06}$	$\mathbf{0.43}_{\pm 0.05}$	$\mathbf{0.73}_{\pm 0.03}$	$\mathbf{0.62}_{\pm 0.01}$	$\mathbf{0.54}_{\pm 0.06}$

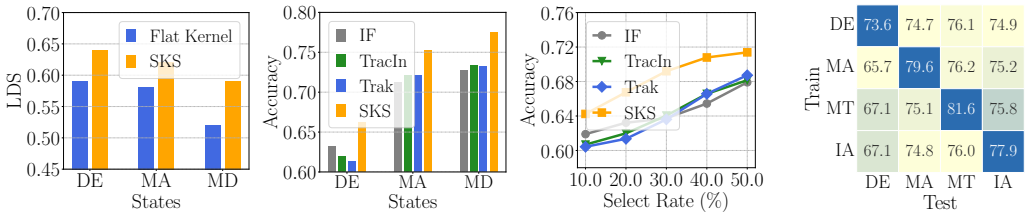
Table 2: Test accuracy of data selection methods on MMTRACE [47] and MultiBench [21]. Each approach selects the top 20% vision-modality samples under a fixed selection budget, and we report the mean and standard deviation over three independent runs.

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 surrogate models	$71.8_{\pm 0.4}$	$78.2_{\pm 0.1}$	$79.9_{\pm 0.1}$	$86.1_{\pm 0.1}$	$79.1_{\pm 0.3}$	$76.6_{\pm 0.1}$
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 surrogate models	$56.5_{\pm 1.2}$	$56.6_{\pm 0.8}$	$52.9_{\pm 1.0}$	$63.3_{\pm 0.8}$	$68.0_{\pm 0.9}$	$57.2_{\pm 1.0}$
SKS (Algorithm 1)	$\mathbf{61.5}_{\pm 0.9}$	$\mathbf{58.5}_{\pm 0.5}$	$\mathbf{53.9}_{\pm 1.3}$	$\mathbf{65.4}_{\pm 1.4}$	$\mathbf{69.2}_{\pm 1.1}$	$\mathbf{60.9}_{\pm 1.5}$

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 analysis on data selection. For data selection, we conduct two ablation studies on the MMTRACE dataset. 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 ablation 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.



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

Figure 3: Ablation studies on the MMTRACE dataset [47]. 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 fitting the model on one state and directly testing it on another state.

4.4 Local Geometric Measures

Next, we evaluate the cross-state generalization results of applying SKS on MMTRACE. Specifically, we learn 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 based on MT data achieve stronger average performance than those based on DE or MA data, 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 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 model for evaluating the generalization ability [6, 43]. We compare the model from different states in MMTraCE with the 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 on data selected by other attribution methods.

Table 3: The Hessian trace and test loss of the model based on Delaware, Massachusetts, Maryland, and Nevada. We compare SKS against several attribution methods including influence functions [9], TracIn [31], Trak [29] and BIF [10].

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 [42] and provide context needed to interpret one another [41, 50]. Learning these relationships provides a basis for constructing useful representations and reducing the number of samples required [7]. Bei et al. [2] show that answering questions about past conversations requires connecting visual details with facts introduced in other dialogue turns. It is important for multi-step search [26], since evidence from one modality can guide what to retrieve from another. The reliability

of these relationships also matters. Sun et al. [33] 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 [8, 3]. To connect representations to prediction, deep supervision makes representations themselves predictive rather than relying solely on the final output [11], and knowledge transfer through representations supplies localization information missing from target labels [34].

Missing information and temporal changes also shape what representations need to capture. The absence of one modality can itself carry information, so representations should account for what is missing as well as what is observed [22]. For evolving states, this missingness information helps refine estimates based on past observations, so that representations can track changes at each time step [23]. Beyond using observations to track changing states, representations can model how feedback drives subsequent changes [28].

Algorithms for influence estimation. The problem of influence estimation aims to quantify how individual examples affect predictions. Classical influence functions [9] estimate the effect of upweighting or removing one sample through a local second-order approximation. However, estimating inverse Hessian-vector products is still expensive for large models. To reduce computation, gradient similarity across checkpoints avoids explicit Hessian computation [31], and model linearization with random projections enables attribution at larger scales [29]. Other than computational efficiency, attribution also raises the question of whether estimated importance rankings are reliable. For feature attribution, Goldwasser and Hooker [4] address this question by testing whether differences in estimated scores support the reported rankings despite sampling uncertainty.

This type of influence analysis also extends naturally from quantifying individual examples to entire evaluation tasks. In multi-task learning, joint training can improve or hurt a task's performance depending on the tasks' input distributions and sample sizes [37, 40]. Li et al. [13] use surrogate models to estimate the effects of different task combinations on performance, identifying which combinations lead to positive or negative transfer. Then, Li et al. [12, 14] use estimated transfer relationships to identify task groups that benefit from joint training. The cost of comparing combinations motivates first-order approximations of fine-tuning performance [17, 48]. These estimates inform both which tasks to group during optimization [45, 18] and where they should share parameters within a model [15, 49]. Theoretical analysis in the setting of high-dimensional linear regression shows when learning a shared linear model on source and target data improves target-task prediction compared with fitting the model on target data alone [40].

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 [32], and bounds on information available only through joint observation quantify the potential gains from combining modalities [20]. Beyond information content, analyses of data augmentation distinguish gains from expanded data coverage and regularization [36]. Complementary optimization work addresses the tradeoff between data reuse and the distribution mismatch induced by repeated updates [35]. 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 study cross-sample modality attribution in multimodal learning. We propose a new algorithm that estimates the subset-to-loss response while preserving the hierarchy between sample participation and within-sample modality composition. We further apply the learned sample-modality attributions to downstream data selection. Across MMTRACE and MultiBench, our approach consistently improves over prior work in terms of attribution quality and data selection accuracy. Further geometric analysis shows that models trained on data selected by our proposed algorithm have a lower Hessian trace, suggesting a flatter loss surface and better generalization.

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] Y. Bei, T. Wei, X. Ning, Y. Zhao, Z. Liu, X. Lin, Y. Zhu, H. Hamann, J. He, and H. Tong. "Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents". In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2026, pp. 40750–40784.

[3] 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.

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

[5] 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).

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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).

[14] 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.

[15] 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).

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] X. Ning, D. Fu, T. Wei, M. Ai, J. Zou, T.-W. Li, H. Tong, Y. Zhu, H. Hamann, and J. He. "MC-Search: Evaluating and Enhancing Multimodal Agentic Search with Structured Long Reasoning Chains". In: *International Conference on Learning Representations*. 2026.

[27] 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.

[28] 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).

[29] 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.

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

[31] 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.

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

[33] 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.

[34] 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.

[35] 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).

[36] 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.

[37] S. Wu, H. R. Zhang, and C. Ré. "Understanding and improving information transfer in multi-task learning". In: *ICLR* (2020).

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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).

[44] 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.

[45] 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).

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] Y. Zhao, T. Wei, J. Zou, X. Ning, Y. Bei, L. Chen, S. Rana, W. H. Yang, H. Tong, and J. He. "PAPERMIND: Benchmarking Agentic Reasoning and Critique over Scientific Papers in Multimodal LLMs". In: *Findings of the Association for Computational Linguistics: ACL 2026*. 2026, pp. 10457–10474.

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. The selection state matrix S specifies which modalities are retained in each sample, and $\mathbf{v}(S)$ indicates which samples retain at least one modality under S. We write f_W for the multimodal model with parameters W, ℓ for the per-sample loss, and $L(f_W; S)$ for the test loss of the model trained under S.

The RBF kernels k_1 and k_2, with bandwidths γ_1 and γ_2, measure similarity in sample participation and within-sample modality selection, respectively. Combining these kernels gives the structured kernel $K(S, S')$ 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 N sampled combinations of retained samples and modalities. In kernel ridge regression, $\mathbf{K}$ is the Gram matrix with entries $K_{pq} = K(S_p, S_q)$, $\mathbf{k}_S$ contains the kernel similarity between a query selection state S and the N sampled states, and λ controls the ridge regularization.

For 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.

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

A.1 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,

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

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

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

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

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

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 $g \in \mathcal{H}_K$,

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

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.2 Positive Semidefiniteness of the Structured Kernel

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. *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.3 Extensions

Our structured kernel surrogate naturally extends from two modalities to $M > 2$ modalities by defining 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 $S_{m,j} = 1$ if modality m of sample j is retained, and $S_{m,j} = 0$ otherwise. The column $S_{:,j} \in \{0,1\}^M$ specifies the modalities retained in sample j, and $\mathbf{v}^{(j)}(S) = \max(S_{:,j})$ indicates whether this sample retains at least one modality.

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 $\mathbf{v}(S')$, and 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, B_M now covers the $2^M - 1$ nonzero modality states in $\{0,1\}^M$. The structured kernel is positive semidefinite if $\gamma_2 > 0$, and $\gamma_1 \geq \frac{1}{2}\log\left(\mathbf{1}^\top B_M^{-1}\mathbf{1}\right)$. Under this condition, SKS extends directly to M modalities using the same kernel ridge regression and attribution procedures.

B Omitted Experimental Details and Results

Models and hyperparameters. For MultiBench, we use the cross-attention network to fuse the pre-computed modality embeddings provided by the datasets. The hyperparameters are listed in Table 4. For MMTRACE, we use a three-layer GNN encoder and an MLP predictor. The vision embedding is fetched by pretrained CLIP and ViT models. 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. For both MMTRACE and MultiBench, we fit the surrogate model using $N = 100$ pairs of selections and crossponding losses $\{S_p, L(f_W; S_p)\}_{p=1}^N$.

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, we start with samples containing all available modalities. To evaluate a particular modality in a sample, we replace its input with zeros, keep the remaining modalities of the sample unchanged, and then apply the corresponding baseline method.

Influence functions [9] 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

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}

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 [39] 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 [31] 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 [29] 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 [10] 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 [38] 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.

B.1 Cross-sample Modality Interaction

We examine pairs of satellite images 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,



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.

low-density, or rural scene. This demonstrates that the latter may weaken the evidence provided by the former.

B.2 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:

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

Table 6 reports the approximation error and computational savings on MMTRACE, with computational cost measured in 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 loss at a substantially lower computational cost.

Table 6: 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 [24], with model architectures ranging from 2B to 30B parameters. As shown in Table 7, the error is less than 1.1% on all models.

Table 7: 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.3 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 8. Overall, SKS has a FLOP cost comparable to the baselines and is nearly identical to TracIn and LESS across all six tasks.

Table 8: 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.4 Omitted Ablation Analysis

Choice of the RBF Kernel. We use the RBF kernel because it provides expressivity for any attribution function theoretically. Zhang et al. [46] 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 full rank. Therefore, for any function F, there exists a kernel mapping that matches F on the input vectors. Classical results [30] guarantee that on this discrete domain, the RBF kernel can approximate any attribution function without loss of expressivity [46].

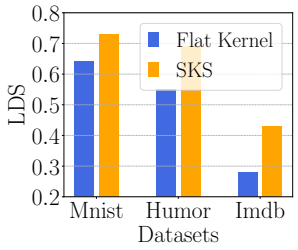
Moreover, the RBF kernel encodes an inductive bias that matches the geometry of the subset space. On $\{0,1\}^d$, two subsets that differ on only a few coordinates are close in Hamming distance, and intuitively their performances tend to be similar. The RBF kernel directly reflects this structure. By 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 9, which show that the RBF kernel gives a 31.8% average advantage over the other two kernels in LDS scores on the MMTraCE benchmark.

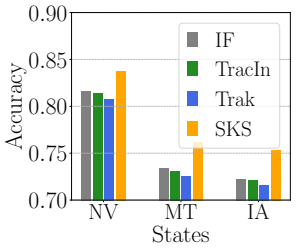
Table 9: Ablation study of nonlinear kernel choices on MMTraCE. Each entry reports the mean and standard deviation of the LDS score over three runs.

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	$0.64_{\pm 0.03}$	$0.62_{\pm 0.05}$	$0.59_{\pm 0.10}$	$0.58_{\pm 0.02}$	$0.57_{\pm 0.06}$	$0.70_{\pm 0.03}$

Comparing the compositional kernel to a flat kernel design. For the kernel-structure ablation analysis, 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



(a) Compositional kernel



(b) Data modality

Figure 5: Ablation analysis 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.

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.

Next, 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.

B.5 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1 summarizes the contribution and the scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix B.5 describes the limitations of this work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3.1 and Appendix A provide the assumptions and the complete proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix B disclose the information needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Section 1 includes the link to the entire code repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix B describe the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4 and Appendix B report the error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B provides the information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Due to the technical nature, this paper presents no ethical concerns or issues.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: 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.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We pose no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 4 includes the citations to the datasets that we have used in our experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines: This paper doesn't introduce or release new datasets or models.

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research only utilizes publicly available datasets and does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We did not use large language models as a core component of our methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.