

Select the Key, Then Generate the Rest: Improving Multi-Modal Learning with Limited Data Budget

Xinyu Zhao¹, Kaishen Wang², Jiaye Yang¹, Jie Peng¹,
Lirong Xiang³, Faizan Siddiqui⁴, Kwonjoon Lee⁴, Behzad Dariush⁴, Tianlong Chen^{1,*}

¹UNC of North Carolina at Chapel Hill, ²University of Maryland,

³Cornell University, ⁴Honda Research Institute

*Corresponding author

Abstract

Multimodal learning serves as a promising approach for applications with diverse information sources. However, there are numerous challenges when scaling up multimodal learning data from all modalities due to availability or varied cost of data collection. We formulate and study multimodal learning under limited data collection budgets as a modality selection problem, demonstrating that selectively collecting a subset of modalities and synthesizing the missing ones can match or outperform models trained with uniformly augmented full modalities. Our research problem is formulated as: *given a limited data collection budget, how to find the appropriate modalities to collect new data and generate for the rest to maximize model performance gain?* To answer this, we propose a new paradigm - Select the Key modality, then generate the rest to enable learning with limited data (SK-11). SK-11 contains two essential components: (1) *Select the key*. We propose a modality importance indicator to find the optimal modalities by assessing their single modal marginal contribution and cross-modal interactions. (2) *Generate the rest*. We substitute with generated embeddings for the rest of modalities. We conducted extensive experiments applying SK-11 across affective computing, healthcare with diverse multimodal learning backbones, obtaining average accuracy gains of {2.37%, 6.55%, 6.73%} on {MOSI, MOSEI, ADNI} respectively. Meanwhile, we present interesting empirical insights such as the data efficiency. Codes are provided in <https://github.com/UNITES-Lab/SK-11>.

1 Introduction

Multimodal learning (MML) integrates information from different modalities, such as text, visual, sensor, biomarkers to learn implicit representation, which has been applied to diverse areas including visual question answering (Ilievski and Feng, 2017; Ding et al., 2022; Lu et al., 2023),

sentiment analysis (Soleymani et al., 2017; Chen et al., 2017; Gandhi et al., 2023), robotics (Sun et al., 2021a), and healthcare application (Yun et al., 2024). Despite its great potential in these applications, collecting data with full modalities in real-world scenarios can be challenging and costly. This is due to restrictions in real life, for example in biomedical settings, measurement devices would destroy paired samples (Xi et al., 2024). In addition, the cost of collecting different modalities varies significantly, with easily accessible image-text data being far more abundant than more complex modalities such as depth or thermal maps (Zhu et al., 2024; Girdhar et al., 2023), tactile data requires specialized sensors (Zou et al., 2017; Yang et al., 2017; Meribout et al., 2024). Previous studies have devised various approach on multimodal learning with missing modalities (Qiu et al., 2023; Wang et al., 2023a; Wu et al., 2024; Lee et al., 2023a; Guo et al., 2024). These methods simulate missing modalities by randomly dropping observations from fully available data, sometimes using the dropped modality for reconstruction learning (Wang et al., 2023a). Our setting addresses prospective acquisition, where modality scarcity is present before training and a fixed collection budget determines which new observations can be obtained. This budgeted acquisition setting remains underexplored in MML.

We conducted preliminary experiments over a few affective computing datasets (Zadeh et al., 2016, 2018) to show the importance of multimodal learning, while the contribution of different modality combinations differs substantially. We first investigate the efficiency of data utilization of multimodal learning compared to unimodal learning. This is experimented with evaluating the performance of a multimodal transformer (Yu et al., 2019) across different training data ratios against the performance of the best unimodal transformer model

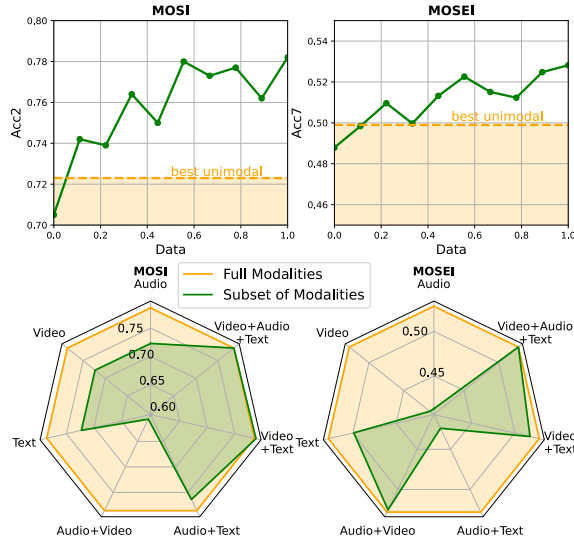


Figure 1: **Top:** Accuracy of multimodal model at different training data ratios (0.1 to 1.0) across affective computing datasets, compared to best unimodal learning (orange area) at full training data. **Bottom:** Test accuracy of multimodal models learned on different modality combination (green line) compared to the model trained on full modalities (orange line) across 2 affective computing datasets. Acc2 denotes binary accuracy, while Acc7 denotes 7-class accuracy.

trained with the complete dataset. As illustrated in Figure 1 (Upper) for both the MOSI and MOSEI affective computing datasets, the models trained with full multimodality surpass the best unimodal baseline even when using as little as 10% of the total training data. However, not all modalities or their combinations contribute equally to the performance of multimodal learning. To further assess the potential of utilizing a subset of modalities, we trained multimodal models with various combinations of the available modalities. Their test accuracies were then compared to a model trained on the full set of modalities. As depicted in Figure 1 (Bottom), specific subsets of modalities (indicated by the green line) can achieve strong performance relative to other subsets, and in some cases, can approach the performance of the model trained with all modalities (indicated by the orange line). For example, in the MOSI dataset, the Audio+Text combination is a notably high-performing subset. For the MOSEI dataset, combinations such as Video+Text and Audio+Text also yield strong results. The observation that the most effective modality subsets can vary by dataset highlights that a strategic selection of modalities is crucial. This approach can lead to superior or more efficient performance, particularly when data acquisition or processing resources are constrained, without needing a complete set of all available modalities.

Inspired by the findings from the preliminary experiment and data collection difficulty in reality, we introduce a new research problem: *given a certain limited data collection budget, how to determine which modalities should be collected and the rest should be generated for improved multimodal learning?* To address this question, we propose a novel multimodal paradigm, **Select the Key**, then generate the rest to enable learning with limited data (SK-11). Specifically, SK-11 consists of two key components: the modality importance indicator and the missing modality generation.

Firstly, the modality importance indicator selects the key modality combination by evaluating both the marginal contribution of each modality and their cross-modality interactions. Given a multimodal dataset, we apply our proposed Step-wise Maximization of Modality Selection algorithm to identify the most informative modality combination. We then allocate the data collection budget to the key combination to ensure that the most informative modalities are prioritized. Secondly, we generate the missing modalities based on the combination to ensure the completion of multimodal data while mitigating modality interference. This approach enables more effective model training even with a limited data collection budget. To validate the effectiveness of SK-11, we conduct experiments across MML backbones and applications. The contributions of this paper are:

- We investigate multimodal learning under limited data collection budgets, addressing the practical constraint of varying modality acquisition costs. We reveal that MML is more data-efficient than unimodal approaches, while performance varies significantly across modality subsets, necessitating principled modality selection for achieving optimal MML performance under budget constraints.
- We introduce a novel framework that first Select the Key modality subsets for new data collection, then generate for the rest modalities, thus enable Learning with Limited new data. Then, we develop an indicator for assessing the relative importance of each modality subset, which guides modality selection in SK-11. Experiments demonstrate a strong correlation between the rankings of this indicator and the accuracy of actual MML task across different subsets of modality.
- Our framework offers flexibility by adapting to any available data at any new data budget

constraint, enhancing performance across diverse MML tasks. We validate the effectiveness of SK-11 through extensive experiments across multiple multimodal backbones. Our approach demonstrates consistent performance improvements across affective computing and healthcare. For example, SK-11 obtains average accuracy gains of $\{2.37\%, 6.55\%, 6.73\%, 0.59\%\}$ on $\{\text{MOSI}, \text{MOSEI}, \text{ADNI}, \text{MIMIC}\}$ compared to the baseline approach.

2 Related Work

Multimodal Learning with Missing Modality.

Real-world multimodal systems often face the challenge of uncertain modality missingness due to various factors such as environmental interference, sensor malfunctions, and privacy concerns, which can significantly degrade model performance (Ma et al., 2022; Wei et al., 2023; Lee et al., 2023b; Qiu et al., 2023; Zhang et al., 2023b; Wu et al., 2024). Recent studies have explored modality robustness and unified frameworks to mitigate these effects: Hazarika et al. (2022) and Lin and Hu (2023) analyze robustness against modality drops, while UniMF (Huan et al., 2023) proposes a unified framework for unaligned and missing sequences. Techniques range from simple imputation (Tran et al., 2017; Pham et al., 2019; Wang et al., 2023b) to noise imitation (Yuan et al., 2023) and multimodal mixup (Lin and Hu, 2024), with other approaches using deep generative models to synthesize missing data or latent representations (Hoffman et al., 2016; Zheng et al., 2021; Zhou et al., 2021; Zhi et al., 2024). Huang et al. (2025) further explore out-of-modal generalization without instance-level modal correspondence. However, these studies focus on handling missing values within fixed, pre-existing datasets. In contrast, we address the prospective challenge of prioritizing which modalities to collect under a limited budget to maximize performance gain.

Modality Imbalance and Selection in MML.

While MML leverages complementary information from multiple sources, some modalities are more dominant than others (Du et al., 2023; Peng et al., 2022), and different modalities overfit and converge at different rates (Wang et al., 2020), leading to modality imbalance and counterproductive MML performance (Ismail et al., 2020; Sun et al., 2021b; Fan et al., 2023). Methods to address this include modulating learning pace (Zhang et al., 2023a), alternating unimodal learning (Zhang et al., 2024), sparse mixture-of-experts (Peng et al.,

2023), Pareto-optimal unimodal assistance (Wei and Hu, 2024), and dynamic modality gap bridging (Yang et al., 2024). Gradient conflicts have also been studied to mitigate modality collapse (Javaloy et al., 2022). Most related to ours, He et al. (2024) select modalities via Shapley values to improve inference efficiency by pruning modalities at test time. Our work instead targets the *data acquisition* stage: we select synergistic modalities for collection and generate the rest, optimizing both acquisition cost and modality interference.

3 Methodology

3.1 Task Formulation of Multimodal Learning under Limited Data Budget

We study resource-constrained multimodal learning with an available paired seed dataset and a limited budget for collecting additional observations. Let $\mathcal{M} = \{M_1, \dots, M_m\}$ denote the modality set and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ the available training data, where $x_i = (x_i^1, \dots, x_i^m)$ and x_i^j is the observation of modality M_j for sample i . Each acquisition of one modality for one new sample consumes one modality-sample unit. An allocation $\beta = (\beta_1, \dots, \beta_m) \in \mathbb{N}_0^m$, where β_j is the number of newly acquired observations of M_j , is feasible when

$$\sum_{j=1}^m \beta_j \leq B. \quad (1)$$

Let $\mathcal{D}'(\beta)$ denote the observations acquired under this allocation, and define $J(\beta) = \text{Perf}(\text{Train}(\mathcal{D}, \mathcal{D}'(\beta)))$ for a fixed training and evaluation procedure. The budgeted acquisition problem is

$$\beta^* \in \arg \max_{\beta \in \mathbb{N}_0^m} J(\beta) \quad \text{s.t.} \quad \sum_{j=1}^m \beta_j \leq B. \quad (2)$$

When n aligned samples are collected for every modality in a selected subset S of size k , $\beta_j = n$ for $M_j \in S$ and 0 otherwise, so the acquisition cost is $B = kn$. Our experiments therefore study finite-sample allocation at equal acquisition cost, complementing population-level analyses of the information available from the full modality set.

3.2 Select the Key Modality

The introduction of a new modality usually could bring a non-negative effect, as prior work

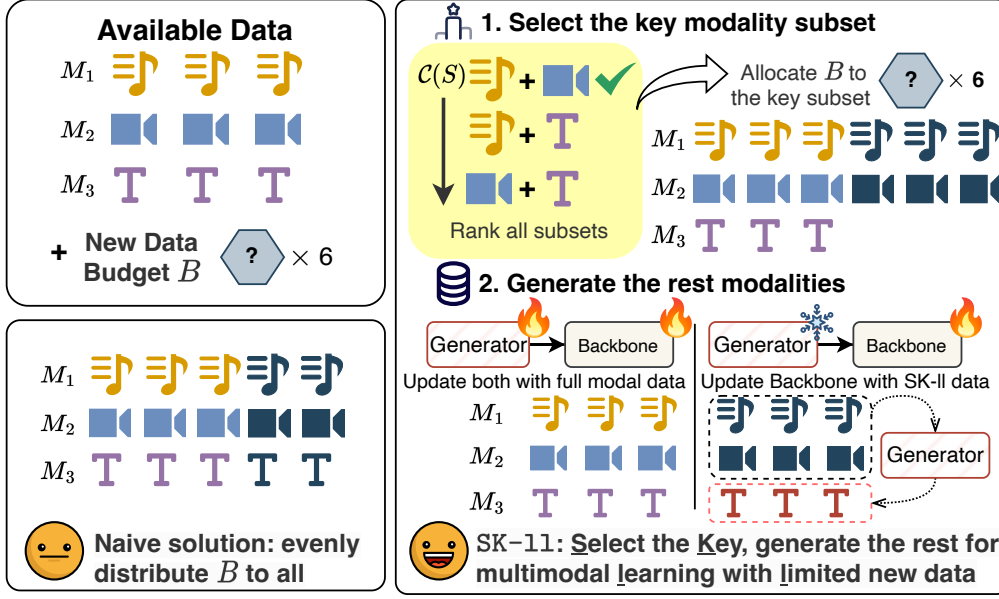


Figure 2: Overview of SK-11. *Top-left* shows the task setting, where there are some available multimodal data along with a limited new data budget B . *Right* illustrates SK-11 framework approaching MML under limited new data with two steps. (1) We first select the key modality subset by evaluating the cooperation indicator $C(S)$, ranking all possible subsets, and allocating the data budget B to the top subset. (2) We then synthesize missing modalities corresponding to increased sample amounts through generators. In the second step, we start model training by updating both the generator and multimodal backbone with available full modal data, then continue improving the multimodal backbone with new data from SK-11. *Bottom-left* contrasts with a naive solution that evenly distributes the new data budget to all modalities.

shows that learning with more modalities provably achieves a smaller population risk (Huang et al., 2021). However, the marginal benefit from new modalities may also decrease as more modalities are included. Therefore, being able to proactively select the most useful modalities can help reduce the cost of collecting and maintaining weak modalities, and improve computational efficiency.

There is a line of works that explore how to quantify the contributions of modality and modalities interaction in multimodal learning. (Gat et al., 2021) introduce the perceptual score that assesses the degree to which a model relies on the different subsets of the input features. (Hu et al., 2022) use the Shapley value to evaluate the cross-modal cooperation for the whole dataset. (Wei et al., 2024) further improve the multimodal cooperation at the sample level using the introduced Shapley-based metric. (Wenderoth et al., 2024) introduces InterSHAP that dissects cross-modal interaction uses the Shapley interaction index.

Our work focuses on measuring the importance of individual modalities and cooperation in a data-limited scenario. We formalize this intuition by defining an indicator of importance for the combination of modality that captures both the contribu-

tion of the individual modalities and the degree of complementarity between the modalities.

Inspired by the cooperative game theory, the Shapley value (Roth, 1988) $\phi(\cdot, \cdot)$ measures each player’s contribution to overall performance via the weighted average of their marginal contributions to all possible coalitions. Multimodal learning is considered a coalition game of $|\mathcal{M}|$ modalities, and the Shapley value of modality M_j is defined as:

$$\phi_{\mathcal{M}_j}(\mathcal{M}, f_u) := \sum_{S \subseteq \mathcal{M} \setminus \{j\}} \frac{|S|!(|\mathcal{M}| - |S| - 1)!}{|\mathcal{M}|!} \cdot (f_u(S \cup \{M_j\}) - f_u(S)) \quad (3)$$

where S is a modality subset of $\mathcal{M}$, and $j \notin S$. The contribution of each modality (subset) is measured by the utility function $f_u(\cdot) : 2^{\mathcal{M}} \rightarrow \mathbb{R}$.

We quantify the utility of an input-modality subset S by the reduction in the best achievable expected loss relative to a constant predictor. Let $G(S)$ be the predictor class that observes X_S , and let $G(\emptyset)$ contain constant predictors. For the classification experiments, we instantiate the loss as $\ell_{0-1}(y, \hat{y}) = 1[y \neq \hat{y}]$.

$$f_u(S) = \min_{c \in G(\emptyset)} \mathbb{E}[\ell_{0-1}(Y, c)] - \min_{g \in G(S)} \mathbb{E}[\ell_{0-1}(Y, g(X_S))] \quad (4)$$

Under 0–1 loss, the optimal constant is the majority-class predictor and the second term equals one minus the best achievable classification accuracy. Thus, $f_u(S)$ equals the accuracy gain from observing S over the majority-class predictor, directly connecting the loss-based definition to the accuracy used in our experiments. Empirically, we fit both the subset model and the majority-class predictor on the training portion of the seed set and evaluate both on its held-out validation split. Individual marginal contributions characterize the additive component of joint utility, and the interaction score below captures its non-additive component.

Concretely, we use a scaled Shapley value to measure each modality’s individual contribution, where Z is the total utility gain of the full model, acting as a dataset-level normalization factor (Gat et al., 2021):

$$\hat{\phi}_{M_j}(\mathcal{M}; f_u) = \frac{1}{Z} \phi_{M_j}(\mathcal{M}; f_u) \quad (5)$$

The Shapley value allocates the total payoff to individual modalities. To additionally represent the interaction structure of a particular subset, we define **Interaction Synergy** as the departure of joint utility from an additive unimodal reference. Specifically, for $S \subseteq \mathcal{M}$, the normalized cooperation score $\mathcal{C}(S)$ is

$$\mathcal{C}(S) = \frac{1}{Z_S} \left(f_u(S) - \sum_{M_j \in S} f_u(\{M_j\}) \right), \quad (6)$$

where $f_u(S)$ is the joint utility and $\sum_{M_j \in S} f_u(\{M_j\})$ is an additive reference built from the unimodal utilities. This reference is only a computational baseline; it does not assume the modalities are statistically independent. $\mathcal{C}(S)$ measures how far the joint utility departs from it. We normalize by $Z_S = f_u(S)$ to compare subsets at different utility scales. A positive $\mathcal{C}(S)$ indicates synergy beyond the additive reference, a negative value indicates redundancy or interference, and zero indicates approximately additive utility.

Table 1 presents the modality importance scores over a few affective computing datasets under limited data availability, using classification accuracy as the utility function. We take the majority-class

accuracy as a baseline when certain modalities are absent. All possible permutations are evaluated so that both unimodal contributions and cross-modal interactions are thoroughly assessed.

3.3 Modality-Interference-aware Incomplete Modality Generation

Informed by SK-11, we collect a selected subset $S_{\text{ava}} \subset \mathcal{M}$ and synthesize the unobserved modalities $S_{\text{miss}} = \mathcal{M} \setminus S_{\text{ava}}$. The incomplete-modality generation process comprises feature extraction and missing-modality generation. **(1) Shallow feature extraction.** For each sample, we extract features $X^{(M_i)}$ for every $M_i \in S_{\text{ava}}$ and project them into a shared-dimensional space. **(2) Missing modality generation.** For the classification tasks, we use class-conditional normalizing flows to transfer the available representations into each missing-modality space (Wang et al., 2023a). The class condition shapes a discriminative latent distribution during training, while the available features provide the instance-specific input. Given an available sample of class c , the flows are optimized by the cross-modal distribution transfer loss

$$l_{\text{gen}} = - \sum_{M_i \in S_{\text{ava}}} \left[\log p_{Z^{(M_i)}}(\mathbf{Z}^{(M_i)} | y = c) + \log \left| \det \left(\frac{\partial \mathbf{Z}^{(M_i)}}{\partial \mathbf{X}^{(M_i)}} \right) \right| \right] \quad (7)$$

where the first term is the class-conditional latent log-density and the second is the flow log-Jacobian determinant. In particular, $Z^{(M_i)} = \mathcal{F}^{(M_i)}(X^{(M_i)}) \sim \mathcal{N}(\mu_c, \Sigma_c)$ is a deterministic function of that sample’s available feature. To generate a missing target modality M_t , the cross-modal transfer module decodes the sample-derived latent through the target inverse flow, $\hat{X}^{(M_t)} = (\mathcal{F}^{(M_t)})^{-1}(Z^{(M_i)})$. The class label parameterizes the training prior but is not the input to this reverse mapping. Consequently, two samples in the same class with different available features produce different missing-modality representations: the objective is class-conditional, whereas generation is instance-conditional.

The entire model is trained end to end by jointly optimizing multimodal fusion and incomplete-modality generation: $\mathcal{L} = l_{\text{task}} + \alpha l_{\text{gen}}$, where l_{task} is the appropriate classification loss (cross-entropy or binary cross-entropy) and α controls the weight of the generation objective.

Table 1: Accuracy (Acc.) of MuLT on MOSI and MOSEI, all with 40% training data. The cooperation $\mathcal{C}(\cdot)$ scores over each modality subset S , our proposed indicator to measure the importance of modality in multimodal learning, and the rank of each subset in $\mathcal{C}(S)$.

Datasets	S	Acc.	$\mathcal{C}(S)$	Rank.
MOSI	T, A	75.02	-0.015	2
	T, V	77.52	0.014	1
	A, V	55.63	-0.066	3
MOSEI	T, A	70.92	0.037	1
	T, V	70.35	0.029	2
	A, V	52.57	-0.101	3

The overview of our method is illustrated in Figure 2. The complete algorithm for crafting a recipe for dataset in the limited data regime is summarized in Algorithm 1. Algorithm 1 starts with an empty set, and subsequently adds to the current set the new modality that maximizes the marginal gain at each iteration. For each candidate modality M_i , we (1) train two models on S and $S \cup M_i$ respectively until training losses converge, and (2) record the difference between two sets of performance metrics, and compute the importance scores; (3) select the top p modalities (where p is the cardinal constraint of modalities that can be selected) based on the ranked importance score, and add the selected modality to S to construct S_{i+1} . Normally we end up with more complementary modalities, ensuring that the modality importance score (i.e., $\hat{\phi}$ or $\mathcal{C}(S)$ values) selected in each iteration is a positive value. The modality selection process terminates when no modality with a positive effect on the current combination can be found in any iteration. Given the optimal modality combination, we next utilize available data modalities to generate missing modality, to form full modal representations.

4 Experiment

4.1 Implementation Details

Backbone Models and Benchmarks. We evaluate our proposed SK-11 across below multimodal learning backbones and application:

- **Affective Computing:** We apply SK-11 on Multimodal Transformer (Yu et al., 2019) with Flow-based generation (Wang et al., 2023a) on MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018), which provide Text (T), Video (V), and Audio (A) modalities aligned at the clip level. Each sample is labeled with sentiment valence from -3 to +3. We also apply SK-11 on CoMM (Dufumier et al., 2025) with UR-Funny (Hasan et al., 2019)

and MUSTARD (Castro et al., 2019). Accuracy2 is used for evaluation.

- **Healthcare Application:** We apply SK-11 on Flex-MoE (Yun et al., 2024) with ADNI (Weiner et al., 2010) (four modalities for AD stage prediction) and MIMIC-IV (Johnson et al., 2019) (ICD-9 codes, clinical text, and vital values). Only samples with all modalities are kept for training. F1 is used for evaluation.

Baselines. The baseline allocates the data collection budget evenly across all modalities. In SK-11, the modality importance indicator selects the key modality combination, and the budget is distributed across this combination, with missing modalities generated to complete the data. The **upper bound** allocates a separate, dedicated budget to each modality.

Training and Evaluation Details. We conduct four groups of experiments per dataset with initial-data ratios of 10%, 20%, 30%, and 40%, each repeated with five random seeds (mean $\pm$ standard deviation). The new-data acquisition budget is fixed at $B = 10\%$ of the full training-set size in modality-sample units. Detailed hyperparameters are provided in Appendix A.3.

We equalize acquisition cost across methods, while the number of completed training instances follows from how many modalities are acquired per new sample. For a dataset with m modalities, the Baseline spends B/m units on each modality for aligned samples, adding B/m complete real instances. If SK-11 selects a subset of k modalities, it spends B/k units on each selected modality for aligned samples and generates the remaining modalities, adding B/k complete instances. Thus, with $m = 3$, $k = 2$, and $B = 10\%$, the Baseline adds 3.33% complete real instances, whereas SK-11 adds 5% completed instances from two acquired and one generated modality. At an initial-data ratio r , their respective training-instance ratios are $r + B/m$ and $r + B/k$. The upper bound acquires all modalities for an additional B complete instances and consequently uses acquisition cost mB ; the Baseline and SK-11 each use B .

4.2 Superior Performance of SK-11

Naively allocating new data budget to all modalities results in suboptimal performance gain and even decrease. Table 2 shows that the return from uniform allocation depends on the dataset and data

Table 2: Performance of SK-11 across applied to classification tasks and traditional multimodal model. 4 groups of experiments are shown in rows with same color (, , ,), with upper and lower adjacent experiments trained on full modality data as upper and lower bound. Best results are **bolden**. “Ava. $\mathcal{D}$ ” stand for available data ratio. “Gen.” stand for whether we generate data embedding for missing modalities. “Baseline” method simply allocate quota to all modalities evenly. For Baseline and SK-11 we select data budget B as 10%.

Ava. $\mathcal{D}$	Gen.	Method	MulT		CoMM		Flex-MoE	
			MOSI	MOSEI	UR-Funny	MUSTARD	ADNI	MIMIC
10%	-	-	72.01 $\pm$ 2.54	69.83 $\pm$ 1.43	51.79 $\pm$ 2.02	59.07 $\pm$ 0.32	30.61 $\pm$ 7.15	57.34 $\pm$ 0.85
	✗	Baseline	70.52 $\pm$ 3.83	71.94 $\pm$ 0.33	51.10 $\pm$ 2.82	59.60$\pm$0.47	32.34 $\pm$ 4.95	57.70 $\pm$ 0.61
	✓	SK-11	75.18$\pm$1.36	76.82$\pm$6.67	51.62$\pm$1.11	59.19 $\pm$ 0.04	48.87$\pm$3.65	58.41$\pm$1.05
20%	-	-	71.78 $\pm$ 2.32	71.07 $\pm$ 0.67	50.16 $\pm$ 2.65	59.67 $\pm$ 0.82	40.90 $\pm$ 7.24	59.16 $\pm$ 0.55
	✗	Baseline	76.24 $\pm$ 0.77	70.63 $\pm$ 1.28	57.72$\pm$1.11	59.73 $\pm$ 0.15	44.44 $\pm$ 6.91	59.78 $\pm$ 0.75
	✓	SK-11	77.93$\pm$0.60	80.81$\pm$0.37	52.06 $\pm$ 3.20	61.61$\pm$1.02	54.06$\pm$5.49	60.13$\pm$0.81
30%	-	-	76.24 $\pm$ 0.72	71.29 $\pm$ 0.65	52.99 $\pm$ 0.97	59.27 $\pm$ 0.37	49.69 $\pm$ 4.55	60.37 $\pm$ 0.73
	✗	Baseline	76.53 $\pm$ 1.42	71.71 $\pm$ 0.89	54.28 $\pm$ 3.15	61.61 $\pm$ 0.20	54.11 $\pm$ 2.12	61.52 $\pm$ 0.83
	✓	SK-11	79.18$\pm$0.56	80.36$\pm$0.50	59.41$\pm$1.29	61.87$\pm$0.07	54.83$\pm$5.99	61.83$\pm$0.56
40%	-	-	77.26 $\pm$ 0.82	71.21 $\pm$ 1.14	53.86 $\pm$ 1.52	60.37 $\pm$ 0.22	53.70 $\pm$ 4.32	60.48 $\pm$ 0.82
	✗	Baseline	77.11 $\pm$ 1.70	71.10 $\pm$ 0.84	55.93 $\pm$ 3.36	61.51 $\pm$ 0.29	54.28 $\pm$ 2.23	61.30 $\pm$ 1.03
	✓	SK-11	79.94$\pm$0.87	80.16$\pm$0.25	57.70$\pm$2.16	62.24$\pm$0.19	61.76$\pm$0.43	62.87$\pm$0.80
50%	-	-	76.82 $\pm$ 1.93	71.84 $\pm$ 0.52	56.82 $\pm$ 1.58	61.40 $\pm$ 0.48	63.96 $\pm$ 3.20	62.19 $\pm$ 1.78

Table 3: Performance of SK-11 on MOSI (Acc) under three new-data budgets (+10%, +30%, +50%), each evaluated at 10% - 50% available data. For a budget i , the per-modality acquisition quota is $i/|\mathcal{M}|$, where $|\mathcal{M}|$ is the number of modalities.

Ava. $\mathcal{D}$	+10% Data		+30% Data		+50% Data	
	Baseline	SK-11	Baseline	SK-11	Baseline	SK-11
10%	70.52 $\pm$ 3.83	75.18$\pm$1.36	71.78 $\pm$ 2.32	75.83$\pm$1.42	73.61 $\pm$ 4.84	76.12$\pm$0.71
30%	76.53 $\pm$ 1.42	79.18$\pm$0.56	77.26 $\pm$ 0.82	77.67$\pm$0.95	77.63 $\pm$ 1.58	79.59$\pm$0.39
50%	77.93 $\pm$ 0.99	79.24$\pm$0.70	78.68 $\pm$ 2.56	79.49$\pm$0.68	79.13 $\pm$ 1.94	80.10$\pm$0.44

regime. On MOSI at 10% and 40% available data, the Baseline falls below training on the initial data alone. The same pattern occurs on UR-Funny at 10% and MUSTARD at 40%, whereas uniform acquisition improves performance on ADNI. These results motivate directing new observations toward modalities that contribute most effectively under the available-data regime.

SK-11 surpasses Baseline with significant margins. Our proposed method SK-11, which first determines key modalities for budget allocation and generates missing data, demonstrates substantial performance improvements across the all benchmarks. It consistently outperforms the initial available data setting and, in most cases, surpasses the Baseline approach, often by significant margins. **① Generalization across tasks and backbones:** The effectiveness of SK-11 is not confined to a specific task or model architecture. On affective computing, we experiment with both fusion-based (MulT on MOSI/MOSEI) and self-supervised (CoMM on UR-Funny/MUSTARD) multimodal mod-

els, where SK-11 consistently yields significant gains over Baseline. On healthcare tasks with MoE-based Flex-MoE (ADNI/MIMIC), SK-11 also delivers strong results, achieving a 6.73% average improvement over Baseline on ADNI. These consistent improvements across different task domains and architectures highlight the broad applicability of our selective budget allocation strategy. **② Generalization across data ratios:** The advantages of SK-11 hold across the spectrum of initial data availability from 10% to 40%. At low ratios, SK-11 provides substantial benefits even with minimal data — the improvement on MOSEI at 10% (76.82 vs. 71.94 Baseline) exemplifies this. At higher ratios, SK-11 maintains its edge, as seen on MIMIC at 40% (62.87 vs. 61.30) and MOSEI at 40% (80.16 vs. 71.10).

4.3 Extra Studies

SK-11 Performs Consistently Under Varied New Data Budgets In our primary experiments, we fixed the new data budget to 10% across varying availability levels. To examine the generalizability of SK-11 across different data budget scenarios, we

Table 4: Ablation study on the Budget Allocation Strategy. We fix the modality subset from the “Selection” step and vary the budget ratio assigned to the identified “Primary Modality.” The baseline is a uniform split (50% for 2 modalities). The dataset is MOSI and metrics is Acc2.

Ava. $\mathcal{D}$	Modality Subset		Ratio for Primary Modality			
	Selected	Primary	50%	60%	70%	80%
10%	T, A	T	75.18 ± 1.36	74.17 ± 0.92	73.89 ± 1.70	74.05 ± 1.09
30%	T, V	T	79.18 ± 0.56	78.20 ± 1.36	77.86 ± 0.61	77.67 ± 0.60
50%	T, V	T	79.24 ± 0.70	78.24 ± 0.70	77.90 ± 0.58	77.36 ± 0.76

Table 5: The stability of our modality importance indicator $\mathcal{C}(S)$ on MOSI. The indicator’s ranking aligns with the actual subset performance (Acc2) as the amount of available data changes. Best performing subset at each data ratio is bolded.

Ava. $\mathcal{D}$	Accuracy (Acc)			Indicator $\mathcal{C}(S)$ Value		
	T+A	T+V	A+V	T+A	T+V	A+V
10%	71.05 ± 1.16	68.48 ± 2.26	48.57 ± 4.63	0.120	0.035	0.078
20%	76.33 ± 0.48	73.00 ± 1.16	52.65 ± 2.05	0.092	0.035	-0.010
30%	75.54 ± 1.44	76.70 ± 1.39	55.02 ± 1.50	-0.008	0.041	0.006
40%	75.02 ± 1.31	77.52 ± 0.64	55.63 ± 1.29	-0.015	0.014	-0.066

Table 6: Model performance (Accuracy) by subgroups on the ADNI dataset.

Model and Ava. $\mathcal{D}$	Male	Female	Gender Gap	Young	Old	Age Gap
No-Gen (40%)	0.489	0.485	0.004	0.457	0.492	0.035
Baseline (40%+10%)	0.552	0.549	0.003	0.457	0.551	0.094
SK-11 (40%+10%)	0.631	0.655	0.024	0.571	0.638	0.067
No-Gen (50%)	0.659	0.645	0.014	0.714	0.647	0.067

conduct additional experiments using three different budget ratios (10%, 30%, and 50%), representing multimodal learning contexts ranging from low to high resource availability. The results presented in Table 3 demonstrate consistent and significant performance improvements by SK-11 across all tested budget settings, highlighting its robustness and flexibility in data resource constraints.

Balanced Data Allocation Strategy Improves Efficiency and Stability We investigate the effectiveness of different budget allocation strategies for newly collected data. Specifically, we compare a weighted allocation strategy, where modalities within a subset are ranked according to their single-modal Shapley values, prioritizing modalities with higher contributions (denoted as the “Primary Modality”) against a balanced allocation. Results summarized in Table 4 indicate that performance differences among various allocation strategies are marginal, with excessively emphasizing the primary modality potentially harming overall performance. Consequently, for simplicity and pipeline efficiency, we uniformly distribute the data budget among modalities.

Separating Subset Selection from Generation and Sample Count. Table 7 holds the acquisition budget, generator, and training-instance count fixed while varying only the acquired pair. The pair selected by our indicator performs best at every data ratio, exceeds the worst pair by 10.62–13.17 Acc2 points, and remains within 0.28 points of the corresponding SK-11 result in Table 2. Table 8 controls for sample count: the Generation baseline

creates B new instances versus $B/2$ for SK-11, yet SK-11 scores 1.68–2.13 points higher. These controls isolate the benefit of subset choice and conditioning generation on two real modalities.

Generator Choice. On MOSI with 20% available data and a 10% new-data budget, the normalizing flow achieves 77.93 ± 0.60 Acc2, compared with 75.49 ± 1.28 for diffusion, while reducing average training time per epoch from approximately 15 to 3 seconds. The flow therefore provides both higher completion accuracy and a $5\times$ training-speed advantage in this setting.

Generation Cost. The flow generator adds 8.40M parameters (7.13%), 12.8% forward-pass overhead for one generated modality (26.4% for two), and 2.14 GiB peak memory; Appendix Tables 10 and 11 report the full breakdown.

Evaluating the Modality Importance Indicator. To evaluate our modality importance indicator, we compare its ranking against actual downstream accuracy for all modality pair combinations on MOSI. With a 20% lower bound, Algorithm 1 orders the pairs as $\{T, A\} \succ \{T, V\} \succ \{A, V\}$, matching the experimental ranking (details in Appendix A.1). We further test stability by varying the available data ratio from 10% to 40%. As shown in Table 5, the optimal subset shifts with data availability—from $\{T, A\}$ at 10–20% to $\{T, V\}$ at 30–40%—and $\mathcal{C}(S)$ correctly tracks this shift at every stage, confirming the indicator’s robustness across data-limited regimes.

Table 7: Controlled subset selection on MOSI. Each column acquires two modalities and generates the third with the same budget, generator, and new-instance count. Bold marks the selected subset.

Ava. $\mathcal{D}$	$T+A \rightarrow V$	$T+V \rightarrow A$	$A+V \rightarrow T$
10%	75.32$\pm$0.62	71.40 $\pm$ 0.61	62.15 $\pm$ 0.58
20%	77.65$\pm$0.58	74.50 $\pm$ 0.50	65.20 $\pm$ 0.90
30%	76.20 $\pm$ 0.90	79.40$\pm$0.64	67.80 $\pm$ 0.29
40%	78.50 $\pm$ 0.42	79.72$\pm$0.73	69.10 $\pm$ 0.73

Comparison with Alternative Indicators. We compare our cooperation indicator with a Shapley-only score (He et al., 2024), which captures individual modality contributions, and the SHAPe interaction index (Hu et al., 2022), which measures pairwise interactions. Across MOSI and MOSEI at four data ratios, our indicator identifies the best subset in 12 of 16 settings, compared with 10 for Shapley-only and 9 for SHAPe. It also achieves the highest average Kendall τ with downstream subset performance (0.750 vs. 0.667 and 0.417), supporting the use of both individual importance and cross-modal synergy for subset selection.

Scope of Subset Selection. Our results characterize two regimes that the indicator identifies before any budget is spent. On MOSI, the $\{A, V\}$ pair has $\mathcal{C}(S) = -0.066$ at 40% available data and substantially lower subset accuracy than $\{T, V\}$ (Table 5); selecting a stronger pair yields gains of 10.62–13.17 points under the controlled generation setting (Table 7). On MIMIC, all three pairwise scores are positive (0.141, 0.118, and 0.063), which the indicator reads as broadly complementary modalities with little interference to avoid. The minimum interaction score thus distinguishes an interference regime with substantial selection headroom from a broad-complementarity regime, letting practitioners prioritize subset selection precisely when the weakest interaction is negative.

Fairness Audit in Healthcare Application To evaluate the fairness and potential biases of our method in sensitive domains, we conducted a fairness audit on the ADNI healthcare dataset. We disaggregated the model’s performance by gender and age subgroups, comparing our SK-11 framework against the uniform-allocation baseline and no-generation models under a 40% available data and 10% new data budget setting. As shown in Table 6, SK-11 mitigates the age-related accuracy gap observed in the baseline (0.067 vs. 0.094) while maintaining a comparable gender gap. This analysis suggests that our strategic data selection and

Table 8: Sample-count control on MOSI. At the same budget, Generation generates two modalities per instance and adds twice as many new instances as SK-11.

Method	Budget	10%	20%	30%	40%
Baseline	$B/3$	70.52 $\pm$ 3.83	76.24 $\pm$ 0.77	76.53 $\pm$ 1.42	77.11 $\pm$ 1.70
Generation	B	73.20 $\pm$ 0.64	75.80 $\pm$ 1.23	77.50 $\pm$ 0.40	78.10 $\pm$ 0.93
SK-11	$B/2$	75.18$\pm$1.36	77.93$\pm$0.60	79.18$\pm$0.56	79.94$\pm$0.87

generation process not only enhances overall accuracy but also promotes more equitable performance across different patient populations.

5 Conclusion

In this paper, we propose SK-11, which identifies the most synergistic modality subset via a cooperation score $\mathcal{C}(S)$, allocates the data acquisition budget to it, and synthesizes the missing modalities. On affective computing and healthcare benchmarks, SK-11 consistently outperforms uniform allocation across data availability ratios, budget sizes, and multimodal backbones. A fairness audit on ADNI further shows that SK-11 narrows the age-related accuracy gap, suggesting that strategic data acquisition can improve both utility and subgroup equity. We hope this work motivates resource-aware data collection as an important consideration in multimodal learning.

Limitations

The selection and generation pipeline relies on an initially paired seed set. Its quality and representativeness influence the accuracy of the importance assessment, and biases in demographic groups, domains, or recording conditions can propagate to the selected acquisition strategy. Future work can combine the importance assessment with bias detection and evaluate selection robustness under deliberately imbalanced seed data.

Our experiments assume equal acquisition costs across modalities. In practice, these costs may vary due to different sensors, procedures, or annotation requirements. The selector can account for such differences by prioritizing modality subsets that provide stronger cooperation relative to their collection cost. Future work can evaluate this extension using application-specific costs.

Acknowledgment

This project is supported by the Honda Research Institute USA.

References

- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an obviously perfect paper). In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4619–4629.
- Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 163–171.
- Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5089–5098.
- Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In International Conference on Machine Learning, pages 8632–8656. PMLR.
- Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. 2025. What to align in multimodal contrastive learning? In International Conference on Learning Representations.
- Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20029–20038.
- Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91:424–444.
- Itai Gat, Idan Schwartz, and Alex Schwing. 2021. Perceptual score: What data modalities does your model perceive? Advances in Neural Information Processing Systems, 34:21630–21643.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Manat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023, pages 15180–15190. IEEE.
- Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1726–1736.
- Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and 1 others. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2046–2056.
- Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. arXiv preprint arXiv:2205.15465.
- Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient modality selection in multimodal learning. Journal of Machine Learning Research, 25(47):1–39.
- Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning with side information through modality hallucination. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 826–834.
- Pengbo Hu, Xingyu Li, and Yi Zhou. 2022. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. arXiv preprint arXiv:2205.00302.
- Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. 2023. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. IEEE Transactions on Multimedia, 26:5753–5768.
- Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In Proceedings of the 35th International Conference on Neural Information Processing Systems, page 13.
- Zhuo Huang, Gang Niu, Bo Han, Masashi Sugiyama, and Tongliang Liu. 2025. Towards out-of-modal generalization without instance-level modal correspondence. In The Thirteenth International Conference on Learning Representations.
- Ilija Ilievski and Jiashi Feng. 2017. Multimodal learning and reasoning for visual question answering. Advances in neural information processing systems, 30.
- Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ishtiaq. 2020. Improving multimodal accuracy through modality pre-training and attention. arXiv preprint arXiv:2011.06102.
- Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal vaes via impartial optimization. In International Conference on Machine Learning.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6.

- Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023a. Multimodal prompting with missing modalities for visual recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023b. Multimodal prompting with missing modalities for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14943–14952.
- Ronghao Lin and Haifeng Hu. 2023. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. Transactions of the Association for Computational Linguistics, 11:1686–1702.
- Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. Information Fusion, 105:102216.
- Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. 2023. The multi-modal fusion in visual question answering: a review of attention mechanisms. PeerJ Computer Science, 9:e1400.
- Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18177–18186.
- Mahmoud Meribout, Natnael Abule Takele, Olyad Derege, Nidal Rifiki, Mohamed El Khalil, Varun Tiwari, and Jing Zhong. 2024. Tactile sensors: A review. Measurement, page 115332.
- Jie Peng, Kaixiong Zhou, Ruida Zhou, Thomas Hartvigsen, Yanyong Zhang, Zhangyang Wang, and Tianlong Chen. 2023. Sparse moe as a new treatment: Addressing forgetting, fitting, learning issues in multimodal multi-task learning.
- Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8238–8247.
- Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI conference on artificial intelligence, pages 6892–6899.
- Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, and Zheng Wang. 2023. Scratch each other’s back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21317–21326.
- Alvin E Roth. 1988. Introduction to the shapley value. The Shapley value, 1.
- Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. Image and Vision Computing, 65:3–14.
- Charles Sun, Jkdrzej Orbik, Coline Devin, Brian Yang, Abhishek Gupta, Glen Berseth, and Sergey Levine. 2021a. Fully autonomous real-world reinforcement learning with applications to mobile manipulation. In Conference on Robot Learning.
- Ya Sun, Sijie Mai, and Haifeng Hu. 2021b. Learning to balance the learning rates between various modalities via adaptive tracking factor. IEEE Signal Processing Letters, 28:1650–1654.
- Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1405–1414.
- Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. Journal of Machine Learning Research, 9:2579–2605.
- Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12695–12705.
- Yuanzhi Wang, Zhen Cui, and Yong Li. 2023a. Distribution-consistent modal recovering for incomplete multimodal learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22025–22034.
- Yuanzhi Wang, Yong Li, and Zhen Cui. 2023b. Incomplete multimodality-diffused emotion recognition. Advances in Neural Information Processing Systems, 36:17117–17128.
- Shicai Wei, Chunbo Luo, and Yang Luo. 2023. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20039–20049.
- Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 27338–27347.
- Yake Wei and Di Hu. 2024. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. arXiv preprint arXiv:2405.17730.
- Michael W. Weiner, Paul S. Aisen, Clifford R. Jack, William J. Jagust, John Q. Trojanowski, Les Shaw, Andrew J. Saykin, John C. Morris, Nigel J. Cairns, Laurel A Beckett, Arthur W. Toga, Robert C. Green, Sarah Walter, Holly D. Soares, Peter J. Snyder, Eric R. Siemers, William Z. Potter, Patricia E. Cole, and Mark E. Schmidt. 2010. The alzheimer’s disease neuroimaging initiative: Progress report and future plans. Alzheimer’s & Dementia, 6:202–211.e7.

- Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. 2024. Measuring cross-modal interactions in multimodal models. [arXiv preprint arXiv:2412.15828](#).
- Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.
- Johnny Xi, Jana Osea, Zuheng Xu, and Jason S Hartford. 2024. Propensity score alignment of unpaired multimodal data. *Advances in Neural Information Processing Systems*, 37:141103–141128.
- Tingting Yang, Dan Xie, Zhihong Li, and Hongwei Zhu. 2017. Recent advances in wearable tactile sensors: Materials, sensing mechanisms, and device performance. *Materials Science and Engineering: R: Reports*, 115:1–37.
- Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems*, 37:62108–62122.
- Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.
- Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.
- Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. *Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts*. Preprint, arXiv:2410.08245.
- Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. [arXiv preprint arXiv:1606.06259](#).
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023a. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR.
- Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466.
- Yunhua Zhang, Hazel Doughty, and Cees Snoek. 2023b. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36:54716–54726.
- Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. 2021. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3529–3537.
- Zhuo Zhi, Ziquan Liu, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. 2024. Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity. [arXiv preprint arXiv:2403.09428](#).
- Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. 2021. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274.
- Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Liang Zou, Chang Ge, Z Jane Wang, Edmond Cretu, and Xiaoou Li. 2017. Novel tactile sensor technology and smart tactile sensing systems: A review. *Sensors*, 17(11):2653.

A Appendix

A.1 Extra Experiments

Evaluating Modality Importance Indicator As shown in Table 9, we allocate extra data collection budget to all possible modality combinations and then report corresponding results. When setting lower bound to 10%, our modality importance indicator orders the modality combinations as T+A, followed by T+V, then A+V. The results shown in Table 9 is consistent with the order obtained by indicator, proving the effectiveness of our proposed modality importance indicator.

Evaluating SK-11 under Different Testing Conditions. To assess the robustness of SK-11, we take MOSI as an example, evaluating it under various testing conditions using combinations of T , A , and V modalities: $\{T, A\}$, $\{T, V\}$, $\{A, V\}$, and $\{T, A, V\}$. For instance, the T+A condition uses only the Text and Audio modalities during testing. All results are based on the best models validated with the full validation set. As shown in Table 12, the $\{T, A\}$ and $\{T, V\}$ conditions perform similarly to the full testing set $\{T, A, V\}$, demonstrating the model’s robustness. In contrast, $\{A, V\}$ yields lower results, which aligns with the indicator’s findings that $\{A, V\}$ is less suitable for budget allocation.

Computational Cost and Scalability Analysis

Computing the modality importance indicator requires training or evaluating a model for each candidate subset. Table 13 reports this one-time wall-clock cost on a single NVIDIA A6000 GPU, with early stopping used throughout. We also estimate the indicator on MOSI from five non-overlapping subsets of 100 samples. Although the absolute $\mathcal{C}(S)$ values change, Table 14 shows that every subset preserves the full-data ordering $\{T, V\} > \{T, A\} > \{A, V\}$. The selection decision can therefore be estimated from a compact paired seed set, substantially reducing indicator-computation cost.

A.2 Overview of Our framework

A.3 Extra Implementation Details

Hyperparameter Details. For unimodal training on affective computing datasets, the maximum number of epochs is 100, learning rate is 1e-4, batch size is 128, weight decay is 0.005, and early stopping patience is 8. For multimodal training,

Algorithm 1 Step-wise Maximization of Modality Selection

```

1: Data: Full modality set  $M = \{M_1, \dots, M_{\mathcal{M}}\}$ , fixed sampling budget  $B$ 
2: Input: Unimodal contribution  $\hat{\phi}$  (Eq. 5), cross-modal cooperation  $\mathcal{C}$  (Eq. 6), max number of modalities to select  $k$ 
3: Output:  $\mathcal{S}_k$ 
4:  $\mathcal{S}_0 = \emptyset$ 
5: for  $i = 0, 1, \dots, k - 1$  do
6:   if  $i = 0$  then
7:      $M^i = \arg \max_{M_j \in M} \hat{\phi}(M_j)$ 
8:   else
9:      $M^i = \arg \max_{M_j \in M \setminus \mathcal{S}_i} \mathcal{C}(\mathcal{S}_i \cup \{M_j\})$ 
10:  end if
11:   $\mathcal{S}_{i+1} = \mathcal{S}_i \cup \{M^i\}$ 
12: end for

```

the same settings apply except patience is set to 10. The MOSI dataset contains 1,284 training, 229 validation, and 686 testing samples; MOSEI includes 16,326 training, 1,871 validation, and 4,659 testing samples. The ADNI and MIMIC settings follow (Yun et al., 2024).

Generation Model Details For the MulT and CoMM backbones, we employ a class-conditional Normalizing Flow to synthesize missing modalities. The architecture comprises a shallow feature extractor (kernel size 5), a cross-modal distribution transfer module consisting of 32 invertible layers, and a reconstruction decoder utilizing 20 Residual Channel Attention Blocks. To accommodate varying dataset complexities, the hidden dimensions are set to 64 for CMU-MOSI and 128 for CMU-MOSEI. For Flex-MoE, its learnable missing-modality bank supplies the generated modality representation. This framework is configured with a hidden dimension of 128 across tasks; specifically, on the ADNI dataset, it employs 16 experts with a Top-4 gating strategy, while on MIMIC-IV, it scales to 32 experts with a Top-3 gating strategy to effectively model diverse modality combinations.

A.4 Visualization of Generated Data

To qualitatively evaluate feature generation quality, we visualize the distributions of real and generated data using t-SNE (van der Maaten and Hinton, 2008). We randomly select 100 testing samples from MOSI, projecting their multimodal features onto a 2D plane. As illustrated in Figure 3, the generated features reside in the same latent manifold as the authentic data, preserving modality-specific

Table 9: Results of different modality combination for generation.

Combination	Acc2	F1 Score	Acc7
T+A	77.29±0.78	77.33±0.72	29.26±0.60
T+V	75.88±2.00	75.91±2.06	28.69±1.92
A+V	63.14±11.72	59.11±15.78	20.44±6.89

Table 10: Parameter cost of the MOSI generator.

Component	Parameters	Share
Generator (3 flows, 3 CATs, 7 priors)	8.40M	7.13%
Backbone (including BERT)	109.48M	92.87%
Total	117.88M	100%

Table 11: Forward-pass cost on MOSI with batch size 32. Peak GPU memory is 2.14 GiB.

Setting	ms/batch	Overhead
Full modalities, no generation	12.50	–
2 real modalities, generate 1 (SK-11)	14.10	+12.8%
1 real modality, generate 2 (Uniform+Gen)	15.80	+26.4%

Table 12: Ablation results under different testing conditions across different available data (Ava. $\mathcal{D}$) on MOSI.

Ava. $\mathcal{D}$	Testing Conditions			
	$\{T, A\}$	$\{T, V\}$	$\{A, V\}$	$\{T, A, V\}$
0.1	72.26	73.72	44.34	75.65
0.2	73.96	76.44	47.82	77.95
0.3	72.77	79.19	36.84	79.08
0.4	74.26	72.81	44.84	80.07

structures. This confirms that SK-11 effectively synthesizes realistic representations without additional data collection. Furthermore, comparing the visualization across different data settings reveals that the feature clusters become more distinct and structured with increased availability of initial training data (from 10% to 50%). This suggests that SK-11 effectively leverages richer data contexts to refine the separability and quality of the synthetic features.

Table 13: One-Time Indicator Computation and Model Training Time. We report the total wall-clock time (in minutes) for the modality selection process on a single NVIDIA A6000 GPU across different datasets and data availability ratios.

Dataset	Modalities	Available Data	SK-11 Indicator Computation	SK-11 Model Training	Baseline Model Training
MOSI	3	10%	8.13	~2.5	~1.8
		50%	13.99	~3.5	~2.5
MOSEI	3	10%	12.07	~3.8	~3.0
		50%	22.2	~6.5	~5.0
ADNI	4	10%	73	~12	~9
		50%	147	~25	~18
MIMIC	3	10%	39	~8	~6
		50%	137	~28	~22

Table 14: Indicator Scores C(S) on MOSI computed from all available data versus 5 non-overlap small random subset of 100 samples

. The relative ranking of subsets (T,V > T,A > A,V) is preserved, demonstrating the feasibility of using a small sample for efficient estimation.

Modality Subset	Acc. with All Data	C(S) with All Data	C(S) with 100 Samples
T, V	77.52	0.014	0.121 ± 0.008
T, A	75.02	-0.015	0.062 ± 0.007
A, V	55.63	-0.066	-0.094 ± 0.005

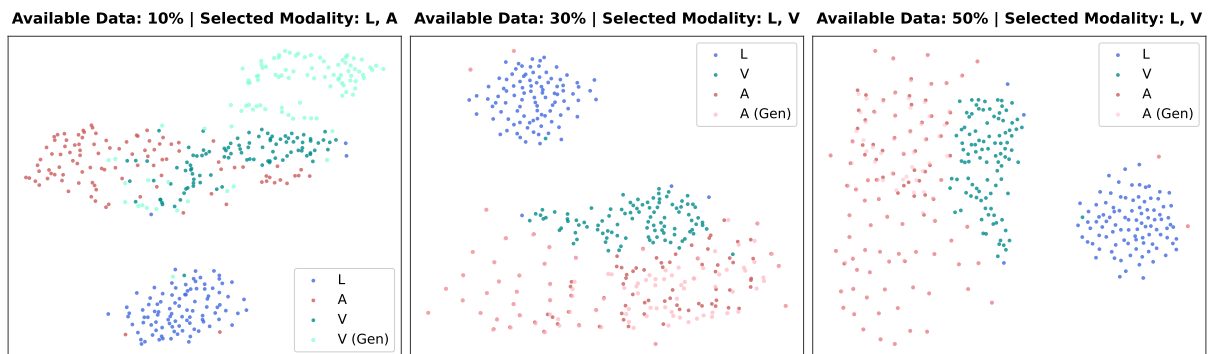


Figure 3: Visualization of real and synthetically generated multimodal features (MOSI), at different availability levels (10%, 30%, and 50%) with an additional 10% data budget. Modality abbreviations: Text (T), Video (V), Audio (A).