SMURF: Statistical Modality Uniqueness and Redundancy Factorization

Torsten Wörtwein[*]
twoertwein@ets.org
Educational Testing Service
Pittsburgh, PA, USA

Nicholas B. Allen
nallen3@uoregon.edu
University of Oregon
Eugene, OR, USA

Jeffrey F. Cohn
jeffcohn@pitt.edu
University of Pittsburgh
Pittsburgh, PA, USA

Louis-Philippe Morency
morency@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Multimodal late fusion is a well-performing fusion method that sums the outputs of separately processed modalities, so-called modality contributions, to create a prediction; for example, summing contributions from vision, acoustic, and language to predict affective states. In this paper, our primary goal is to improve the interpretability of what modalities contribute to the prediction in late fusion models. More specifically, we want to factorize modality contributions into what is consistently shared by at least two modalities (pairwise redundant contributions) and what the remaining modality-specific contributions are (unique contributions). Our secondary goal is to improve robustness to missing modalities by encouraging the model to learn redundant contributions. To achieve our two goals, we propose SMURF (Statistical Modality Uniqueness and Redundancy Factorization), a late fusion method that factorizes its outputs into a) unique contributions that are uncorrelated with all other modalities and b) pairwise redundant contributions that are maximally correlated between two modalities. For our primary goal, we 1) verify SMURF's factorization on a synthetic dataset, 2) ensure that its factorization does not degrade predictive performance on eight affective datasets, and 3) observe significant relationships between its factorization and human judgments on three datasets. For our secondary goal, we demonstrate that SMURF leads to more robustness to missing modalities at test time compared to three late fusion baselines.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Multimodal, Machine Learning, Unique, Redundant

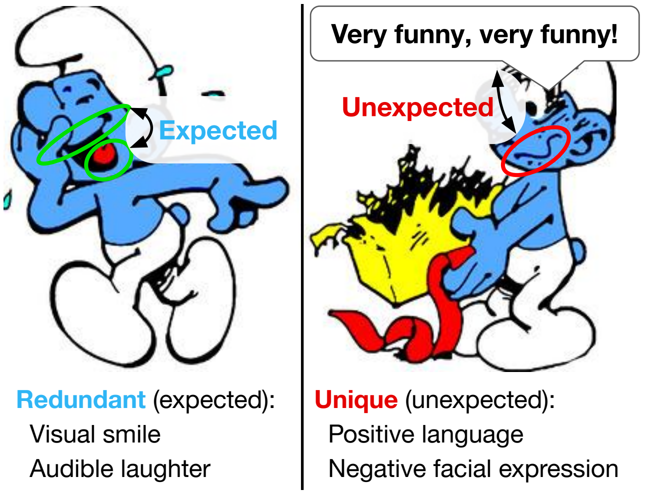


Figure 1: Left: Modalities can provide *redundant* information for a task, such as visually smiling while audible laughing, where both might indicate a positive state. If two modalities consistently contribute redundant information, we might expect what the other modality contributes while only observing one of them, such as expecting a smile when observing laughter. Right: Any remaining contributions we cannot consistently expect are *unique*. The image is created from an outline on oncoloring.com, accessed on December 20th, 2023.

ACM Reference Format:
Torsten Wörtwein, Nicholas B. Allen, Jeffrey F. Cohn, and Louis-Philippe Morency. 2024. SMURF: Statistical Modality Uniqueness and Redundancy Factorization. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3678957.3685716

1 INTRODUCTION

Multimodal late fusion is a common multimodal fusion method that sums the outputs of separately processed modalities, so-called modality contributions, to create the prediction; for example, summing the contributions of vision, acoustic, and language to predict affective states. Knowing what modalities contribute to the prediction can improve the interpretability of such models. For example, it

[*] This work was conducted while the first author was at Carnegie Mellon University.

ICMI '24, November 4–8, 2024, San Jose, Costa Rica
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685716

might be reassuring if a model's modality contributions align with human expectations. Humans often express affective states through multiple modalities, such as visually smiling while audibly laughing, as shown in Figure 1 on the left. To provide a more detailed interpretation, we aim to separate what a modality uniquely contributes and what is redundantly contributed by a pair of modalities. Figure 1 illustrates the concept of pairwise redundant contributions on the left at the example of smiling and laughing: they often co-occur and are likely to indicate the same affective state. The right side of Figure 1 illustrates the remaining unique contributions that are exclusively expressed by one modality. While unique contributions are primarily relevant for interpretability, encouraging pairwise redundant contributions might improve robustness to missing modalities, as even a completely redundant modality will be used, which otherwise might be ignored [1, 34, 37].

This paper is motivated by two research questions:

RQ1: Can we factorize unique and pairwise redundant contributions in late fusion models to improve interpretability?

RQ2: Does encouraging late fusion models to learn redundancies improve their robustness to missing modalities?

The main challenge for the two research questions is that we need to mathematically operationalize unique and pairwise redundant contributions in the context of late fusion models for multiple modalities. Since late fusion models already provide a coarse separation of modality contributions, they are well suited to further factorize them into unique and pairwise redundant contributions. However, late fusion models also make this factorization difficult, as only the final summing operation simultaneously observes all modalities. This means we have to predict from one modality which pairwise redundancies we expect it to have with other modalities and what we expect to remain unique without observing the other modalities. While challenging, we expect this to capture frequently co-occurring patterns such as visual smiles and audible laughter.

In this paper, we propose SMURF (Statistical Modality Uniqueness and Redundancy Factorization) to learn late fusion models that factorize their outputs into the sum of a) unique contributions that are uncorrelated with all other modalities and b) pairwise redundant contributions that are maximally correlated between pairs of modalities. SMURF achieves its factorization through two auxiliary loss terms adapted from the interpretable factor analysis in statistics [33]: the first term maximizes the covariance between pairwise redundant contributions and the other term minimizes the absolute value of the covariance between a modality's unique contributions and its pairwise redundant contributions.

For RQ1, our primary goal, we first verify that SMURF achieves its intended factorization on synthetic data in both a bimodal and trimodal context, we then evaluate that SMURF does not degrade predictive performance on eight affective datasets, and finally, we compare SMURF's factorization to human judgments studies on three datasets. For RQ2, our secondary goal, we test whether SMURF is more robust to missing modalities compared to three late fusion models by reconstructing the fully observed predictions using only one available modality.

2 RELATED WORK

We cover three related topics: approaches that quantify the amount of unique and redundant information, coordinated representations to learn redundancy, and multimodal collinearity.

Unique and Redundant Information: Statistical and information-theoretical approaches, such as factor analysis [33], mutual information [15], and partial information decomposition [14], provide statistics about redundant information, for example, in terms of variance or bits. Such statistics have been predicted using neural networks [7, 17, 39], for example, by training a multi-task neural network for a primary task and adding a secondary task to predict the redundant information in bits between two modalities [7]. These two tasks are, however, only linked through the shared parameters of the neural network, meaning it is unclear if the predicted redundancy is actually used to predict the primary task. SMURF addresses this issue by directly factorizing its prediction into unique and pairwise redundant contributions. SMURF combines neural networks and ideas of the factor analysis to define pairwise redundant contributions as maximally correlated contributions and unique contributions as uncorrelated contributions.

Coordinated Representations for Redundancy: Coordinated representation learning tries to learn a representation by maximizing a similarity measure between two modalities [3], meaning this representation focuses on redundant information simultaneously present in both modalities. Many similarity measures have been proposed to learn such coordinated representations [2, 18, 27], and previous work also learned representations that focus on what other modalities do not contain [10, 16, 32, 38]. The main differences are that SMURF is applied to the output of a model instead of a representation space and that SMURF independently processes modalities (late fusion) instead of combining modalities as done, for example, in cross-modal attention. SMURF's approach has three potential advantages: 1) it ensures that the unique and pairwise redundant contributions impact the prediction (representation spaces undergo further layers which might learn not to use, for example, the correlated information); 2) it might be easier to inspect the low-dimensional contributions for a machine learning practitioners than to inspect the high-dimensional embedding spaces; and 3) it ensures that the unique contributions can never contain unique contributions from other modalities as they are independently processed.

Multimodal Collinearity: Multimodal models can learn to ignore a modality even though the modality contains predictive information when another modality provides the same and more information. This has been observed in multimodal machine translation [31] and multimodal sentiment recognition [34, 37]. Ignoring an informative modality is related to collinearity in statistics, where at least one feature is a linear combination of the remaining features [1]. In such a situation, a model might use the redundant feature to some degree or ignore it in the extreme case. SMURF tries to avoid this ambiguity by maximally relying on pairwise redundant contributions, meaning a redundant modality will be used by SMURF. This can potentially improve robustness to missing modalities, for example, when only a completely redundant modality is available at test time.

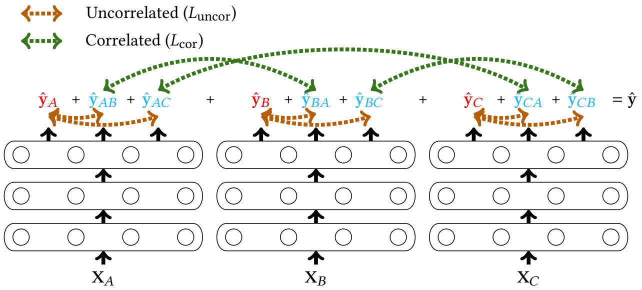

Uncorrelated (L_{uncor})
Correlated (L_{cor})
$\hat{\mathbf{y}}_A$ + $\hat{\mathbf{y}}_{AB}$ + $\hat{\mathbf{y}}_{AC}$ + $\hat{\mathbf{y}}_B$ + $\hat{\mathbf{y}}_{BA}$ + $\hat{\mathbf{y}}_{BC}$ + $\hat{\mathbf{y}}_C$ + $\hat{\mathbf{y}}_{CA}$ + $\hat{\mathbf{y}}_{CB}$ = $\hat{\mathbf{y}}$
$\mathbf{X}_A$
$\mathbf{X}_B$
$\mathbf{X}_C$

Figure 2: Illustration of SMURF for three modalities where L_{uncor} leads to uncorrelated unique contributions and L_{cor} correlates pairwise redundant contributions.

3 SMURF

We first describe the SMURF model, then detail how it is learned, and lastly, explain its design decisions[1].

3.1 Bimodal SMURF

For simplicity, we start by focusing on regression in the bimodal case with two modalities A and B, where the model input for a dataset of N samples are two matrices $\mathbf{X}_A \in \mathbb{R}^{N\times|A|}$ and $\mathbf{X}_B \in \mathbb{R}^{N\times|B|}$, and the model output is a vector $\hat{\mathbf{y}} \in \mathbb{R}^N$ predicting the ground truth labels $\mathbf{y} \in \mathbb{R}^N$.

SMURF is a late fusion model, meaning it processes modalities A and B separately using two neural network f_{θ_A} and f_{θ_B} that each have their own learnable parameters θ_A and θ_B. Unlike most late fusion models that output one contribution for each modality, SMURF separates unique and pairwise redundant contributions and outputs, therefore, two contributions per modality in the bimodal case:

$$[\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB}] = f_{\theta_A}(\mathbf{X}_A) \tag{1}$$
$$[\hat{\mathbf{y}}_B,\ \hat{\mathbf{y}}_{BA}] = f_{\theta_B}(\mathbf{X}_B) \tag{2}$$

where we use a single-letter subscript to denote the unique contributions ($\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_B$) and a two-letter subscript to denote the pairwise redundant contributions ($\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$). $\hat{\mathbf{y}}_{AB}$ is the redundant contribution as predicted from A, while $\hat{\mathbf{y}}_{BA}$ is the redundant contribution as predicted from B. We want these two contributions, $\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$, to be the same, or at least, be very similar to each other.

Finally, SMURF creates the predictions $\hat{\mathbf{y}}$ by summing all the contributions:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_A + \hat{\mathbf{y}}_B + \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA}\ . \tag{3}$$

Figure 2 illustrates the SMURF model for three modalities.

The novelty of SMURF comes from its two auxiliary loss terms L_{uncor} and L_{cor} that encourage factorization of the unique and pairwise redundant contributions. The SMURF model is optimized with the following joint loss function

$$L(\mathbf{y}, \hat{\mathbf{y}}) + \lambda(L_{\text{uncor}} + L_{\text{cor}}) \tag{4}$$

where $L(\mathbf{y}, \hat{\mathbf{y}})$ is a downstream loss (for example, the mean squared error to predict emotional valence) and λ is a hyper-parameter determining the trade-off between it and SMURF's factorization. As illustrated in Figure 2, the goal of L_{uncor} is to uncorrelate a modality's unique contribution from its pairwise redundant contributions

$$\min |r(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB})| + |r(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{BA})|, \tag{5}$$

where r is Pearson's correlation coefficient. L_{uncor} encourages the unique contribution to learn what we expect is uncorrelated from what the other modalities learn. As Pearson's r is scale-invariant, it fluctuates widely when the contributions are almost zero, making the optimization unstable. We, instead, use the sample covariance

$$\text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) = \frac{\sum_{i=1}^{N}(\hat{\mathbf{y}}_A^i - \bar{\hat{\mathbf{y}}}_A)(\hat{\mathbf{y}}_{AB}^i - \bar{\hat{\mathbf{y}}}_{AB})}{N-1}, \tag{6}$$

where $\bar{\hat{\mathbf{y}}}_A$ is the mean over the N samples in $\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_A^i$ refers to the i-th element in the vector $\hat{\mathbf{y}}_A$. Using the sample covariance, we implement L_{uncor} as

$$L_{\text{uncor}} = \frac{1}{2}\left(|\text{cov}(\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB})| + |\text{cov}(\hat{\mathbf{y}}_B,\ \hat{\mathbf{y}}_{BA})|\right)\ . \tag{7}$$

The goal of the second loss term, L_{cor}, is to learn highly correlated pairwise contributions so that $\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$ become similar to each other. We operationalize L_{cor} again with the sample covariance

$$L_{\text{cor}} = -\text{cov}(\hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_{BA}) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{AB})\text{var}(\hat{\mathbf{y}}_{BA}) \tag{8}$$

where var is the sample variance

$$\text{var}(\hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N}(\hat{\mathbf{y}}^i - \bar{\hat{\mathbf{y}}})}{N-1}. \tag{9}$$

The first term of L_{cor} maximizes the covariance between the pairwise redundant contributions. The second term limits the individual variances. While the second term might not seem intuitive, it is needed as otherwise contributions can increase their covariance by increasing only their variances without increasing their correlation[2]. L_{cor} is known as an Hirschfeld-Gebelein-Ŕenyi (HGR) correlation [12] approximation proposed to learn maximally correlated representation in neural networks and was demonstrated to perform better than maximizing Pearson's r, the sample covariance, and canonical correlation analysis [18].

The design of L_{uncor} and L_{cor} is inspired by the interpretable factor analysis in statistics [33], which expresses a variable as the sum of uncorrelated factors. In our case, we express the predictions $\hat{\mathbf{y}}$ as the sum of three uncorrelated "factors" $\hat{\mathbf{y}}_A + \hat{\mathbf{y}}_B + (\hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA})$ where we combine the two highly correlated pairwise redundant contributions $\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$. We can now analyze the unique and pairwise redundant contributions, similar to how we would analyze the factors of the factor analysis. Extracting the pairwise redundant contributions from two modalities, e.g., $\hat{\mathbf{y}}_{AB}$ from A and $\hat{\mathbf{y}}_{BA}$ from B, has the advantage that even if one modality is missing, we still have an expectation of what a modality has in common with a missing one. This might allow us to recover the redundant information present in the missing modalities to improve robustness.

[1] The supplementary material describes one way how SMURF can be implemented beyond late fusion models.

[2] In practise, we also observed infinite values without the variance term.

3.2 m-modal SMURF

In the case of m modalities (A, B, ..., M), we learn again the unique contributions but also all pairwise redundant contributions as already illustrated in Figure 2 for $m = 3$. We focus on all pairwise redundancies, as this entails all possible redundancies, allowing us to define the remaining contributions as unique. We do not separate pairwise redundancies from tertiary redundancies—redundancies between three modalities—as this is not necessary to determine the unique contributions, and having additional loss terms might degrade predictive performance. The model architecture becomes

$$[\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB},\ \ldots,\ \hat{\mathbf{y}}_{AM}] = f_{\theta_A}(\mathbf{x}_A) \tag{10}$$

$$\cdots$$

$$[\hat{\mathbf{y}}_M,\ \hat{\mathbf{y}}_{MA},\ \ldots,\ \hat{\mathbf{y}}_{MN}] = f_{\theta_M}(\mathbf{x}_M)\ . \tag{11}$$

L_{uncor} uncorrelates the unique contribution of modality I from all the pairwise redundant contributions with modality J

$$L_{\text{uncor}} = \alpha \sum_{(I,J),I\neq J} |\text{cov}(\hat{\mathbf{y}}_I, \hat{\mathbf{y}}_{IJ})| \tag{12}$$

where $\alpha = \frac{1}{m^2-m}$ is a normalization term to average over all the covariance terms so that the same hyper-parameter λ can be used across bimodal and m-modal experiments. Similarly, L_{cor} maximizes the covariance between pairwise redundant contributions of all modality pairs (I, J)

$$L_{\text{cor}} = \beta \sum_{(I,J),I<J} -\text{cov}(\hat{\mathbf{y}}_{IJ},\ \hat{\mathbf{y}}_{JI}) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{IJ})\text{var}(\hat{\mathbf{y}}_{JI})\ . \tag{13}$$

where $\beta = \frac{2}{m^2-m}$ is again a normalization term to average over the covariance terms.

3.3 Classification SMURF

To extend SMURF to classification tasks where we predict one out of c classes labels, we represent $\hat{\mathbf{y}} \in \mathbb{R}^{N\times c}$ as a matrix containing the c logits for each sample. We generalize the single-output regression case to the multiple-output classification case by applying the loss terms separately to each output: the unique contribution for the i-th class should be uncorrelated of the pairwise redundant contributions of i-th class. The two loss terms in case of c classes become in the bimodal case with modalities A and B

$$L_{\text{uncor}} = \frac{1}{2c} \sum_{i\in[1,c]} |\text{cov}(\hat{\mathbf{y}}_A^{*,i},\ \hat{\mathbf{y}}_{AB}^{*,i})| + |\text{cov}(\hat{\mathbf{y}}_B^{*,i},\ \hat{\mathbf{y}}_{BA}^{*,i}])| \tag{14}$$

$$L_{\text{cor}} = \frac{1}{c} \sum_{i\in[1,c]} -\text{cov}(\hat{\mathbf{y}}_{AB}^{*,i},\ \hat{\mathbf{y}}_{BA}^{*,i}) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{AB}^{*,i})\text{var}(\hat{\mathbf{y}}_{BA}^{*,i}) \tag{15}$$

where $\hat{\mathbf{y}}^{*,i}$ represents the i-th column in the matrix $\hat{\mathbf{y}}$.

4 EXPERIMENTAL SETUP

For RQ1, we first describe the synthetic dataset used to verify SMURF's factorization and summarize the eight multimodal datasets that are used to test whether SMURF impacts predictive performance. Three of those eight datasets have human judgments, which we will use later for analysis. For RQ2, we compare SMURF against three baseline models on how well each model can recover their multimodal predictions using only one modality to evaluate their robustness to missing modalities.

4.1 Datasets

To evaluate SMURF's factorization, we create a synthetic dataset with a ground truth of the unique and pairwise redundant contributions. As affective states are often expressed through multiple modalities [5, 26], we focus on eight affective datasets that include sentiment and emotion annotations. See Table 1 for a summary.

Synthetic: To test whether SMURF recovers the intended unique and pairwise redundant contributions, we create a synthetic dataset. We define $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{u}_A + \mathbf{u}_B + \mathbf{u}_C + \mathbf{r}_{AB} + \mathbf{r}_{AC} + \mathbf{r}_{BC} \tag{16}$$

where $\mathbf{u}_A, \ldots, \mathbf{r}_{BC} \sim \mathbb{N}(0,1)$ are randomly sampled. We define the three modalities A, B, and C as containing three features each

$$A = [\mathbf{u}_A, \mathbf{r}_{AB}, \mathbf{r}_{AC}], \tag{17}$$

$$B = [\mathbf{u}_B, \mathbf{r}_{AB}, \mathbf{r}_{BC}], \tag{18}$$

$$C = [\mathbf{u}_C, \mathbf{r}_{AC}, \mathbf{r}_{BC}], \tag{19}$$

where [] is the concatenation operator. $\mathbf{u}_A, \mathbf{u}_B$, and $\mathbf{u}_C$ are features with unique contributions that are in only one modality, and $\mathbf{r}_{AB}, \mathbf{r}_{AC}$, and $\mathbf{r}_{BC}$ are the pairwise redundant contributions that are in multiple modalities, for example, $\mathbf{r}_{AB}$ is in A and B. We use this synthetic dataset in two settings: in the introduced trimodal setting where the model has access to modalities A, B, and C and in a bimodal setting, where the model has access to only modalities A and B. In both settings, $\mathbf{y}$ is defined as in Equation 16 even when modality C is not available in the bimodal setting.

MOSI [35] and MOSEI [36]: These two datasets consist of single-person YouTube videos where the person expresses an opinion, e.g., about a movie. In both cases, we predict the continuous sentiment ratings (MOSI-S and MOSEI-S) and also the happiness intensity ratings on MOSEI (MOSEI-H).

IEMOCAP [6]: We use the improvised dyadic interactions of IEMOCAP and predict their continuous arousal (IEMOCAP-A) and valence (IEMOCAP-V) ratings separately for each person and utterance.

RECOLA [23]: This dataset consists of French-speaking dyadic interactions. Similar to IEMOCAP, we predict arousal (RECOLA-A) and valence (RECOLA-V) ratings for each person and utterance.

SEWA [25]: This dataset consists of German-speaking dyadic interactions. As previously, we predict arousal (SEWA-A) and valence (SEWA-V) ratings for each person and utterance.

UMEME [21]: The UMEME dataset contains a set of sentences enacted in different emotional settings. We predict arousal (UMEME-A) and valence (UMEME-V) separately for each enacted sentence. UMEME has further combinations of mismatched audio and video, e.g., the video from a positive enactment but the audio from a negative enactment. As we focus on more natural interactions, we exclude those mismatched combinations.

TPOT [19]: The TPOT dataset contains video recordings of dyadic interactions between mothers and their adolescents. These interactions consist of segments annotated for four affective states (other, aggressive, dysphoric, and positive). We classify these segments for each person independently of the previous and following segments.

Table 1: Dataset characteristics.

Dataset	Tasks	Samples	Modalities (abbreviations)
MOSEI [36]	Sentiment and happiness (regression)	23.3k	audio (A), text (T), video (V)
MOSI [35]	Sentiment (regression)	2.2k	audio (A), text (T), video (V)
IEMOCAP [6]	Arousal and valence (regression)	4.8k	audio (A), text (T), video (V)
RECOLA [23]	Arousal and valence (regression)	1.0k	audio (A), ECG (E), video (V)
SEWA [25]	Arousal and valence (regression)	2.2k	audio (A), text (T), video (V)
UMEME [21]	Arousal and valence (regression)	1.6k	audio (A), text (T), video (V)
TPOT [19]	Four affective states (multiclass classification)	15.2k	audio (A), text (T), video (V)
VREED [24]	Arousal-valence quadrants (multiclass classification)	312	ECG (E), GSR (G), gaze (V)

VREED [24]: VREED is a virtual reality dataset of people watching emotion-eliciting 360-degree videos. We predict the four quadrants of the arousal-valence space (the four combinations of low/high arousal and low/high valence) separately for each person and video.

4.2 Multimodal Features

We use the same features for all modality: MiniLM-L12-v2's sentence embedding [28] for text, openSMILE's eGeMaPs [8] features for audio, and OpenFace 2.2 [4] features for video. In all cases, we use statistics, such as mean and standard deviation, to aggregate the extracted features at the labeled utterance level.

RECOLA does not provide transcripts of the spoken text: we use the dataset author-provided heart rate related features (ECG) as a third modality instead of the text modality. VREED does not share the raw audio-video recordings and has no transcripts. We use the author-provided features for eye-gaze, skin-conductance (GSR), and heart rate (ECG).

4.3 Baseline Models

We compare SMURF to three models. To better evaluate the impact of the two auxiliary loss terms, we focus mainly on models with the same architecture and otherwise change only the auxiliary loss terms, meaning except for the normal late fusion model, all other models have multiple outputs per modality as in Equation 3.

Late fusion: A normal late fusion model [9] that outputs one contribution per modality and has no additional loss terms. This model makes its prediction as following $\hat{\mathbf{y}} = \hat{\mathbf{y}}_A + \hat{\mathbf{y}}_B = f_{\theta_A}(\mathbf{X}_A) + f_{\theta_B}(\mathbf{X}_B)$ and its loss function is only the downstream loss $L(\mathbf{y}, \hat{\mathbf{y}})$. We choose this standard baseline to evaluate the impact of having multiple outputs per modality as in Equation 3.

SMURF w/o $L_{\text{cor}}+L_{\text{uncor}}$: Uses the same architecture as SMURF, but it has no auxiliary loss terms. This helps us to isolate how the two auxiliary loss terms L_{cor} and L_{uncor} impact performance of SMURF. Compared to the normal late fusion model, SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$ has multiple outputs per modality as in Equation 3.

E-HGR [18]: We refer to previous work that also uses the Hirschfeld-Gebelein-Renyi (HGR) correlation approximation to maximize redundacy as E-HGR [18]. The E in E-HGR stands for embedding, as E-HGR maximizes redundancy in an embedding space (SMURF maximizes redundancies at the modality contribution level). While SMURF maximizes all pairwise redundancies, E-HGR maximizes only what all modalities simultaneously have in common, meaning it might learn fewer redundancies than SMURF. To focus the comparison on the effect of E-HGR's and SMURF's auxiliary losses, E-HGR uses the same architecture as SMURF, meaning it also has multiple modality output as in Equation 3.

4.4 Evaluation Methodology

We evaluate SMURF using our two research questions. Our primary RQ1 has three parts: 1) verify SMURF's factorization, 2) test whether SMURF maintains predictive performance, and 3) analyze whether SMURF's factorization is related to human judgments. Our secondary RQ2 explores whether SMURF is more robust to missing modalities.

RQ1: Factorization, Performance, and Analysis: We evaluate SMURF's factorization on the synthetic dataset by verifying that it recovers the known ground truth of the unique and pairwise redundant contributions, e.g., SMURF's $\hat{\mathbf{y}}_A$ should correspond to the feature $\mathbf{u}_A$ on the synthetic dataset, meaning $r(\hat{\mathbf{y}}_A, \mathbf{u}_A)$ should be high. On all nine datasets (which includes the synthetic dataset), we further report Pearson's r for regression tasks and accuracy for classification tasks to evaluate whether SMURF impacts predictive performance.

The UMEME, IEMOCAP, and TPOT datasets have additional human judgments related to the predictive task that enable us to quantitatively compare SMURF's factorization with them. Specifically, we test whether SMURF's unique contributions have a relationship to two types of human judgments: a) partial arousal and valence judgments based on a subset of the modalities, e.g., assessing valence based on only textual transcripts; and b) judgment ratings of how informative a modality is, e.g., a single modality alone might provide sufficient information to determine an affective state.

UMEME and IEMOCAP have partial judgments of arousal and valence, where humans are, for example, given only the muted video to rate valence. All samples of UMEME have partial judgments for the muted video ($\mathbf{y}_V$) and the original audio ($\mathbf{y}_{A+T}$; including the spoken texts). A subset of 100 IEMOCAP samples has partial judgments for all combinations of acoustic, text, and vision modalities, which include bimodal judgments based on the low-pass filtered audio with the transcripts ($\mathbf{y}_{A+T}$), the low-pass filtered audio with the muted video ($\mathbf{y}_{A+V}$), and the transcripts with the muted video ($\mathbf{y}_{T+V}$) [29]. As the original dataset labels are based on all three available modalities, we can define what is unique to modality m by subtracting the judgments when modality m is unavailable, e.g., we define the unique human vision contributions as $\mathbf{y} - \mathbf{y}_{A+T}$. We can now use the correlation between human unique contributions and the learned unique contributions from SMURF to test whether they are similar.

TPOT has judgments of how informative modalities appear to humans when confirming its four affective states [30] ranging from *no*, *relevant*, and *sufficient* information. We hypothesize larger unique contributions for samples where the modality is judged as *relevant* or *sufficient* compared to different samples of the same modality that are judged as providing *no* informative. As we classify TPOT's four discrete affective states, we operationalize "larger unique contributions" as the absolute probability change when setting a modality's unique contributions to 0.

Lastly, we qualitatively inspect samples on the IEMOCAP that have large unique contributions and large pairwise redundant contributions.

RQ2: Redundancy and Robustness: SMURF's covariance maximization in L_{cor} explicitly encourages it to derive the same contributions from modality pairs. This might make SMURF more robust to missing modalities than the other models, which might learn pairwise redundancies to a lesser degree. We evaluate how well models perform when only one modality is present at test time, for example, only A, which means we have only $[\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}]$ remaining. To best recover the original fully-observed predictions $\hat{\mathbf{y}}$ from the remaining contributions, we train a linear model that takes the remaining contributions as input to predict $\hat{\mathbf{y}}$. This step is necessary to re-scale the contributions. If the original model did not extract all the pairwise redundant information from a modality (e.g., if a model ignored audible laughter and relied only on visual smiles) this linear model will perform poorly on the downstream task. This performance quantifies whether the explicitly encouraged pairwise redundancy in SMURF improves robustness to missing modalities.

4.5 Implementation Details

The models (f_{θ_A}, f_{θ_B}, and f_{θ_C}) are instantiated as multi-layer perceptions (MLP) using PyTorch [20]. All models are learned with the optimizer Adam [13], a batch size of 256, and have their hyperparameters validated on the validation sets. Hyper-parameters include $\lambda \in [0.1, 1]$ (for both SMURF and E-HGR; SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$ and the standard late fusion use $\lambda = 0$), the number of layers of the MLP and their number of neurons, the learning rate, and the strength of L2 weight decay.

Early stopping is performed on the loss values on the validation set, which includes auxiliary loss terms for SMURF and E-HGR. The predictive performance metric on the validation set (Pearson's r for regression tasks and accuracy for classification tasks) determines the best model of the hyperparameter search. We use a 5-fold testing for all datasets. These folds are person-independent except for MOSI and MOSEI for which we use the official test set.

5 RESULTS AND DISCUSSION

5.1 RQ1: Factorization and Performance

The primary research question (RQ1) aims to evaluate whether SMURF 1) achieves its factorization, 2) maintains predictive performance, and 3) whether its unique contribution correlate with human judgements.

Table 2: Performance of the trimodal models. Higher is better in all cases and bold indicates best numeric performance. $^{\downarrow}$ (and $^{\uparrow}$) indicates when a model performs significantly worse (or better) than SMURF at $\alpha = 0.05$.

	Late Fusion	E-HGR [18]	SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$	**SMURF** (proposed)
Pearson's r (regression)				
Synthetic	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000
MOSEI-S	0.716 ± 0.004	0.714 ± 0.006	0.715 ± 0.006	**0.717** ± 0.002
MOSEI-H	**0.638** ± 0.003	0.635 ± 0.005	0.634 ± 0.005	**0.638** ± 0.007
MOSI-S	0.688 ± 0.012	0.693 ± 0.013	0.694 ± 0.014	**0.698** ± 0.013
IEMOCAP-A	0.664 ± 0.050	0.646 ± 0.061$^{\downarrow}$	0.663 ± 0.048	**0.665** ± 0.048
IEMOCAP-V	**0.671** ± 0.073	0.664 ± 0.078	0.662 ± 0.085	0.667 ± 0.079
RECOLA-A	0.609 ± 0.045$^{\downarrow}$	**0.623** ± 0.047	0.623 ± 0.046	**0.623** ± 0.043
RECOLA-V	**0.495** ± 0.064	0.490 ± 0.073	0.485 ± 0.067$^{\downarrow}$	**0.495** ± 0.065
SEWA-A	0.497 ± 0.041	0.499 ± 0.055	**0.525** ± 0.029	0.509 ± 0.048
SEWA-V	**0.473** ± 0.031	0.467 ± 0.046	0.467 ± 0.023	0.465 ± 0.024
UMEME-A	0.710 ± 0.064	0.711 ± 0.078	0.712 ± 0.071	**0.713** ± 0.066
UMEME-V	0.725 ± 0.046$^{\downarrow}$	0.740 ± 0.053$^{\downarrow}$	0.745 ± 0.042	**0.750** ± 0.043
Accuracy (classification)				
TPOT	0.533 ± 0.008	0.531 ± 0.011	**0.535** ± 0.008	**0.535** ± 0.010
VREED	0.515 ± 0.061$^{\downarrow}$	0.551 ± 0.073$^{\downarrow}$	0.602 ± 0.083	**0.608** ± 0.041

Table 3: Pearson's r between the human and SMURF's unique contributions. ** indicates $p < 0.01$ on UMEME.

Unique Contributions		Arousal	Valence
Vision	$r(\hat{\mathbf{y}}_V, \mathbf{y} - \mathbf{y}_{A+T})$	0.373^{**}	0.578^{**}
Acoustic+Text	$r(\hat{\mathbf{y}}_{A+T}, \mathbf{y} - \mathbf{y}_V)$	0.392^{**}	0.295^{**}

Achieving factorization: We have a ground truth of the unique and pairwise redundant contributions on the synthetic dataset to evaluate SMURF's factorization; for example, $\hat{\mathbf{y}}_A$ should correlate highly with the unique contribution $\mathbf{u}_A$ of modality A. The average correlation between the ground truth contributions and the actual contributions learned by SMURF is $r = 0.964$ in the bimodal case and $r = 0.922$ in the trimodal case. These high correlations indicate that SMURF achieved its factorization.

Maintaining performance: SMURF often maintains similar predictive performance, sometimes statistically significantly improves performance, and never significantly decreases performance, see Table 2. Two reasons why SMURF might lead numerically to better performance are that 1) its auxiliary loss terms might act as an regularization making overfitting less likely and 2) L_{cor}'s covariance maximization across modalities encourages the model to use all modalities, which might make SMURF more robust to noisy and missing modalities. We explore this second explanation in RQ2.

Human partial judgments: We test whether the human unique contributions on UMEME and IEMOCAP correlate with SMURF's unique contributions, e.g., are the human unique vision contributions $\mathbf{y} - \mathbf{y}_{A+T}$ correlated with SMURF's unique vision contributions $\mathbf{y}_V$.

On UMEME, we train a bimodal SMURF model by collapsing acoustic and language features into one modality to be compatible with UMEME's human judgments which combine acoustic and language. Figure 3 hints at a linear relationship between the human unique vision contributions and SMURF's unique vision contribution $\hat{\mathbf{y}}_V$ on the test set. We quantify this relationship in Table 3 through Pearson's r and observe a moderate correlation.

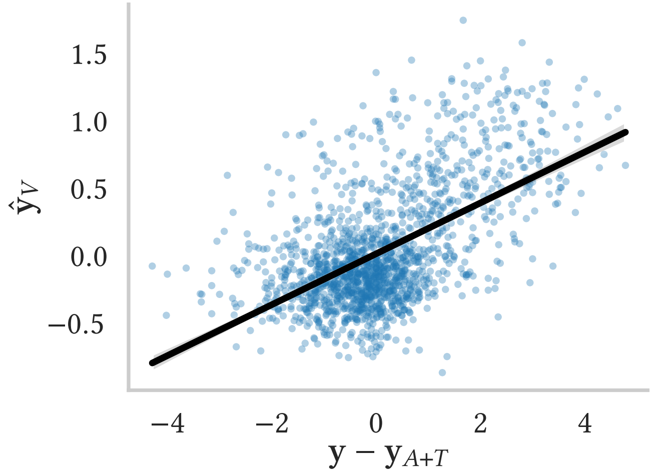


Figure 3: Scatterplot visualizing the relation between the human unique vision contributions ($y - y_{A+T}$) and SMURF's unique vision contributions ($\hat{y}_V$) on the test set when predicting valence on UMEME.

Table 4: Pearson's r between the human and SMURF's unique contributions. * and ** indicate $p < 0.05$ and $p < 0.01$ respectively on IEMOCAP.

Unique Contributions		Arousal	Valence
Acoustic	$r(\hat{y}_A, y - y_{T+V})$	0.297^{**}	0.197^{*}
Text	$r(\hat{y}_T, y - y_{A+V})$	0.269^{**}	0.228^{*}
Vision	$r(\hat{y}_V, y - y_{A+T})$	0.325^{**}	0.468^{**}

On IEMOCAP, we use our trimodal SMURF model as we can define human unique contributions for all modalities using the bimodal human judgments. Like for UMEME, we observe significant correlations for the subset of IEMOCAP in Table 4. The results on UMEME and IEMOCAP indicate that SMURF's factorization has similarities to partial human judgements.

Human informativeness judgments: We expect that a modality has larger unique contributions when humans rate the informativenss of this modality as *relevant* or *sufficient*. As we classify TPOT's four discrete affective states, we operationalize "larger unique contributions" as the absolute probability change when setting a modality's unique contributions to 0. We use Wilcoxon's unpaired ranksums test to compare whether the absolute probability change is lower for samples with *no* information compared to samples that are judged as having *relevant* or *sufficient* information within each modality. We observe that this is the case for all three modalities of our trimodal SMURF on the test set: acoustic ($T = 3.691, p < 0.001$), text ($T = 7.043, p < 0.001$), video ($T = 6.075, p < 0.001$).

Qualitative Inspection: We present samples for the SMURF valence model on IEMOCAP in Table 6 to qualitatively inspect the learned factorization. The top row in Table 6 shows examples of large unique contributions where other modalities are either not providing similar information, such as not audibly laughing while only slightly smiling, or that might be ambiguous without more context, such as saying "What do you mean?". The second row of examples depicts large pairwise redundant contributions between two modalities where both modalities contribute similarly. We observe similar pairwise redundant contributions during common behavior co-occurrences, such as laughing while smiling. The last row focuses on examples with large predicted pairwise redundant contributions in one modality but not the other modality (unexpected pairwise redundant contributions). These examples exemplify that unexpected pairwise redundancies are likely during infrequent behavior patterns, such as smiling while saying thanks

Table 5: Performance of the trimodal models when recovering the performance from only one modality. Higher is better in all cases and bold indicates best numeric performance. $^{\downarrow}$ (and $^{\uparrow}$) indicates when a model performs significantly worse (or better) than SMURF at $\alpha = 0.05$.

Available		Approach			
Modality		Late Fusion	E-HGR [18]	SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$	**SMURF** (proposed)
Pearson's r (regression)					
Synthetic	A	0.633 ± 0.001$^{\downarrow}$	0.644 ± 0.033$^{\downarrow}$	0.693 ± 0.006$^{\downarrow}$	**0.702** ± 0.003
	B	0.674 ± 0.001$^{\downarrow}$	0.578 ± 0.037$^{\downarrow}$	0.693 ± 0.008$^{\downarrow}$	**0.703** ± 0.004
	C	0.672 ± 0.000$^{\downarrow}$	0.399 ± 0.044$^{\downarrow}$	0.646 ± 0.004$^{\downarrow}$	**0.685** ± 0.004
MOSEI-S	A	0.314 ± 0.009$^{\downarrow}$	0.321 ± 0.014	0.313 ± 0.012$^{\downarrow}$	**0.324** ± 0.012
	T	**0.695** ± 0.003	0.690 ± 0.006	0.691 ± 0.007	**0.695** ± 0.017
	V	0.246 ± 0.004$^{\downarrow}$	0.253 ± 0.009	0.252 ± 0.007	**0.256** ± 0.007
MOSEI-H	A	0.304 ± 0.006$^{\downarrow}$	0.313 ± 0.007$^{\downarrow}$	0.304 ± 0.004$^{\downarrow}$	**0.319** ± 0.011
	T	**0.368** ± 0.003	0.366 ± 0.009	0.362 ± 0.009	0.365 ± 0.004
	V	**0.551** ± 0.002	**0.551** ± 0.003	0.550 ± 0.001	**0.551** ± 0.003
MOSI-S	A	0.004 ± 0.063$^{\downarrow}$	-0.067 ± 0.056$^{\downarrow}$	-0.050 ± 0.051$^{\downarrow}$	**0.034** ± 0.069
	T	0.701 ± 0.014$^{\downarrow}$	0.707 ± 0.021	0.707 ± 0.017	**0.710** ± 0.013
	V	0.084 ± 0.037	0.073 ± 0.030	0.082 ± 0.030	**0.088** ± 0.039
IEMOCAP-A	A	0.639 ± 0.060	0.632 ± 0.048$^{\downarrow}$	0.644 ± 0.052	**0.652** ± 0.059
	T	0.302 ± 0.038$^{\downarrow}$	0.322 ± 0.053	0.301 ± 0.029$^{\downarrow}$	**0.334** ± 0.020
	V	0.355 ± 0.150	0.357 ± 0.138	0.348 ± 0.163	**0.358** ± 0.133
IEMOCAP-V	A	0.356 ± 0.116$^{\downarrow}$	0.444 ± 0.099	0.425 ± 0.085$^{\downarrow}$	**0.448** ± 0.093
	T	**0.557** ± 0.054	0.554 ± 0.041	0.541 ± 0.046$^{\downarrow}$	0.552 ± 0.049
	V	**0.416** ± 0.167	**0.416** ± 0.176	0.408 ± 0.182	**0.416** ± 0.168
RECOLA-A	A	0.562 ± 0.045	0.530 ± 0.044$^{\downarrow}$	0.544 ± 0.047$^{\downarrow}$	**0.569** ± 0.057
	E	0.223 ± 0.083	0.187 ± 0.125	**0.248** ± 0.077	0.213 ± 0.144
	V	0.324 ± 0.098	0.277 ± 0.106	0.290 ± 0.144	**0.326** ± 0.106
RECOLA-V	A	0.212 ± 0.086	0.210 ± 0.070	**0.213** ± 0.072	**0.213** ± 0.070
	E	0.234 ± 0.134	0.227 ± 0.122$^{\downarrow}$	0.230 ± 0.122	**0.253** ± 0.078
	V	0.457 ± 0.126	0.451 ± 0.126	0.452 ± 0.106	**0.460** ± 0.119
SEWA-A	A	0.251 ± 0.053	0.145 ± 0.106$^{\downarrow}$	0.172 ± 0.108$^{\downarrow}$	**0.258** ± 0.055
	T	0.093 ± 0.054	0.049 ± 0.067$^{\downarrow}$	0.090 ± 0.067	**0.123** ± 0.044
	V	0.522 ± 0.009	0.499 ± 0.049$^{\downarrow}$	0.521 ± 0.011	**0.532** ± 0.038
SEWA-V	A	0.181 ± 0.031	0.182 ± 0.047	0.192 ± 0.042	**0.199** ± 0.025
	T	0.032 ± 0.035$^{\downarrow}$	0.081 ± 0.024$^{\downarrow}$	0.026 ± 0.029$^{\downarrow}$	**0.101** ± 0.037
	V	**0.536** ± 0.025	0.524 ± 0.028	0.532 ± 0.017	0.526 ± 0.028
UMEME-A	A	0.496 ± 0.101	0.460 ± 0.115	0.490 ± 0.076	**0.504** ± 0.104
	T	0.138 ± 0.052$^{\downarrow}$	0.137 ± 0.092$^{\downarrow}$	0.152 ± 0.047	**0.189** ± 0.067
	V	**0.526** ± 0.084	0.509 ± 0.101	0.509 ± 0.099	0.518 ± 0.084
UMEME-V	A	0.097 ± 0.071$^{\downarrow}$	0.105 ± 0.056$^{\downarrow}$	0.109 ± 0.079$^{\downarrow}$	**0.155** ± 0.060
	T	0.265 ± 0.050	0.278 ± 0.053	0.274 ± 0.048	**0.286** ± 0.050
	V	0.652 ± 0.071	0.673 ± 0.045	0.672 ± 0.043	**0.677** ± 0.045
Accuracy (classification)					
TPOT	A	0.368 ± 0.014	**0.380** ± 0.012	0.374 ± 0.011	0.376 ± 0.012
	T	0.349 ± 0.011	0.353 ± 0.011	0.352 ± 0.012	**0.355** ± 0.005
	V	0.519 ± 0.006	0.518 ± 0.015	**0.520** ± 0.006	**0.520** ± 0.005
VREED	E	0.276 ± 0.044	0.282 ± 0.030	0.273 ± 0.054	**0.288** ± 0.018
	G	0.313 ± 0.074	0.332 ± 0.047	**0.333** ± 0.036	0.325 ± 0.051
	V	0.531 ± 0.050	0.551 ± 0.063	0.544 ± 0.059	**0.559** ± 0.057

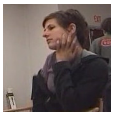
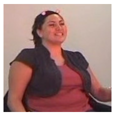
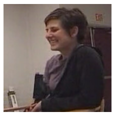
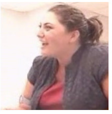
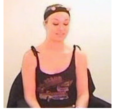
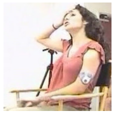
		Large unique contributions	
Video			
Voice	Bored	Controlled, neutral	Amused, not laughing
Utterance	What do you mean?	You're not making this any easier for me, you know.	She needs a new name each state that we go into.
Unique Contributions	$\hat{y}_A = \mathbf{-0.75}, \hat{y}_L = 0.08, \hat{y}_V = 0.03$	$\hat{y}_A = -0.01, \hat{y}_L = \mathbf{-1.33}, \hat{y}_V = 0.00$	$\hat{y}_A = -0.02, \hat{y}_L = 0.25, \hat{y}_V = \mathbf{0.74}$
		Large pairwise redundant contributions	
Video			
Voice	Surprised, excited	Excited	Joking, laughing
Utterance	Are you serious? You're getting married.	U.S.C.	Yeah. It's a joke.
Redundant Contributions	$\hat{y}_{AL} = 0.17, \hat{y}_{LA} = 0.17$	$\hat{y}_{AV} = 0.31, \hat{y}_{VA} = 0.30$	$\hat{y}_{LV} = 0.46, \hat{y}_{VL} = 0.47$
		Large unexpected pairwise redundant contributions	
Video			
Voice	Loud	Quiet	Loud
Utterance	That is incredible.	Thanks.	This has been going on for the past two weeks.
Redundant Contributions	$\hat{y}_{AL} = -0.11, \hat{y}_{LA} = 0.33$	$\hat{y}_{AV} = -0.66, \hat{y}_{VA} = 0.17$	$\hat{y}_{LV} = -0.66, \hat{y}_{VL} = 0.29$

Table 6: Examples from the SMURF valence model on IEMOCAP.

despite expressing sadness through the voice, or during ambiguous expression; for example, speaking loudly might be either very positive or very negative.

5.2 RQ2: Redundancy and Robustness

The goal of our secondary research question is to evaluate whether SMURF's maximization of pairwise redundant contributions is beneficial for robustness to missing modalities.

Robustness to missing modalities: To evaluate how robust SMURF and the other models are to missing modalities, we evaluate the performance of using the learned contributions from just one modality to simulate missing modalities. We reconstruct $\hat{\mathbf{y}}$ using the contributions from only one modality, e.g., $[\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}]$, with a linear model. The original trimodal models should always derive unique contributions as the model would otherwise perform worse, so the main performance difference between the models should reflect the degree to which the trimodal model extracted redundant contributions from multiple modalities. In Table 5, we observe that SMURF's contributions often lead numerically to a better performance (bold) than for its baselines, showing that SMURF is more robust to missing modalities. This indicates that explicitly learning redundant contributions is a step towards making models more robust to missing modalities. Similar to previous work [31], we observe that SMURF is mainly beneficial when the most predictive modality, such as language for sentiment, is missing.

6 CONCLUSION

Our primary goal of SMURF was to make multimodal late fusion models more interpretable by separating what modalities consistently have in common (redundant contribution) from what remains specific to each modality (unique contributions). This is especially interesting for human behavior, as we often express ourselves redundantly through multiple modalities [5, 26], for example, when expressing affective states. We first verified that SMURF achieved its factorization on a synthetic dataset, then demonstrated that despite its additional factorization, SMURF maintains predictive performance, and finally we observed significant relationships between the learned factorization and human judgments. As models can, hypothetically, ignore a completely redundant modality [1, 34, 37], our secondary goal was to explore whether encouraging a model to learn redundancies might make it more robust when a modality is missing at test time. We investigated this hypothesis by measuring how well we can reconstruct multimodal predictions using the modality contributions from just one modality and observed that SMURF tended to enable better reconstructions.

ACKNOWLEDGMENTS

This material is based upon work partially supported by the National Science Foundation (Awards #1722822 and #1750439), and National Institutes of Health (Awards #R01MH125740, #R01MH096951,

#U01MH116925, and #U01MH116923). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

REFERENCES

[1] Aylin Alin. 2010. Multicollinearity. *Wiley interdisciplinary reviews: computational statistics* 2, 3 (2010), 370–374.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*. PMLR, 1247–1255.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[5] Kenton L Burns and Ernst G Beier. 1973. Significance of vocal and visual channels in the decoding of emotional meaning. *Journal of Communication* 23, 1 (1973), 118–130.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359.

[7] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922* (2021).

[8] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[9] Trevor Hastie and Robert Tibshirani. 1986. Generalized Additive Models. *Statist. Sci.* 1, 3 (1986), 297–310.

[10] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.

[11] Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 861–877. https://doi.org/10.18653/v1/2020.emnlp-main.62

[12] Hermann O Hirschfeld. 1935. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 31. Cambridge University Press, 520–524.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Artemy Kolchinsky. 2022. A novel approach to the partial information decomposition. *Entropy* 24, 3 (2022), 403.

[15] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[16] Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6631–6640.

[17] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023. Quantifying & modeling feature interactions: An information decomposition framework. *arXiv preprint arXiv:2302.12247* (2023).

[18] Fei Ma, Shao-Lun Huang, and Lin Zhang. 2021. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *2021 IEEE international conference on multimedia and Expo (ICME)*. IEEE, 1–6.

[19] Benjamin W Nelson, Lisa Sheeber, Jennifer Pfeifer, and Nicholas B Allen. 2021. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry* 62, 2 (2021), 199–211.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[21] Emily Mower Provost, Yuan Shangguan, and Carlos Busso. 2015. UMEME: University of Michigan emotional McGurk effect data set. *IEEE Transactions on Affective Computing* 6, 4 (2015), 395–409.

[22] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).

[23] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2018. Summary for AVEC 2018: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 26th ACM international conference on Multimedia*. 2111–2112.

[24] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 4 (2021), 1–20.

[25] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.

[26] Harald G Wallbott and Klaus R Scherer. 1986. Cues and channels in emotion recognition. *Journal of personality and social psychology* 51, 4 (1986), 690.

[27] Tinghua Wang, Xiaolu Dai, and Yuze Liu. 2021. Learning with Hilbert–Schmidt independence criterion: A review and new perspectives. *Knowledge-based systems* 234 (2021), 107567.

[28] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33 (2020), 5776–5788.

[29] Torsten Wörtwein, Lisa Sheeber, Nicholas Allen, Jeffrey Cohn, and Louis-Philippe Morency. 2022. Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 4681–4696.

[30] Torsten Wörtwein, Lisa B Sheeber, Nicholas Allen, Jeffrey F Cohn, and Louis-Philippe Morency. 2021. Human-guided modality informativeness for affective states. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 728–734.

[31] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462* (2021).

[32] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1642–1651.

[33] An Gie Yong, Sean Pearce, et al. 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology* 9, 2 (2013), 79–94.

[34] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826* (2019).

[35] 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* (2016).

[36] 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)*. 2236–2246.

[37] Yufei Zeng, Zhixin Li, Zhenjun Tang, Zhenbin Chen, and Huifang Ma. 2023. Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Systems with Applications* 213 (2023), 119240.

[38] Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. 2022. Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[39] Jiahao Zheng, Sen Zhang, Xiaoping Wang, and Zhigang Zeng. 2022. Multimodal representations learning based on mutual information maximization and minimization and identity embedding for multimodal sentiment analysis. *arXiv preprint arXiv:2201.03969* (2022).

A SMURF BEYOND LATE-FUSION

While the focus of this paper is SMURF for late-fusion models, we want to highlight that SMURF can also be extended to non-late fusion to learn more complex interactions between modalities, so-called non-additive interactions [11]. A non-additive model with modalities A, B, and C can learn non-additive interactions between pairs of modalities and between the triplet of modalities. Multimodal Residual Optimization (MRO) [29] was proposed to separate additive (unimodal), pairwise non-additive (bimodal), and triplet non-additive (trimodal) interactions from each other. We can extend MRO by applying SMURF within the additive and within the bimodal interactions. This allows us to explore a) whether there are non-additive interactions between A and B that derive the same contributions as non-additive interactions between other modality pairs (redundant bimodal interactions) and b) what the unique non-additive interactions between modalities contribute to the prediction.

We have the following seven models when combining SMURF and MRO to factorize non-additive interactions for three modalities A, B, and C as illustrated in Figure 4: the three unimodal models

$$[\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB},\ \hat{\mathbf{y}}_{AC}] = f_{\theta_A}(\mathbf{x}_A) \tag{20}$$

$$[\hat{\mathbf{y}}_B,\ \hat{\mathbf{y}}_{BA},\ \hat{\mathbf{y}}_{BC}] = f_{\theta_B}(\mathbf{x}_B) \tag{21}$$

$$[\hat{\mathbf{y}}_C,\ \hat{\mathbf{y}}_{CA},\ \hat{\mathbf{y}}_{CB}] = f_{\theta_C}(\mathbf{x}_C)\ ; \tag{22}$$

the three bimodal models

$$[\hat{\mathbf{y}}_{(AB)},\ \hat{\mathbf{y}}_{(AB)(AC)},\ \hat{\mathbf{y}}_{(AB)(BC)}] = f_{\theta_{AB}}(\mathbf{x}_A, \mathbf{x}_B) \tag{23}$$

$$[\hat{\mathbf{y}}_{(AC)},\ \hat{\mathbf{y}}_{(AC)(AB)},\ \hat{\mathbf{y}}_{(AC)(BC)}] = f_{\theta_{AC}}(\mathbf{x}_A, \mathbf{x}_C) \tag{24}$$

$$[\hat{\mathbf{y}}_{(BC)},\ \hat{\mathbf{y}}_{(BC)(AB)},\ \hat{\mathbf{y}}_{(BC)(AC)}] = f_{\theta_{BC}}(\mathbf{x}_B, \mathbf{x}_C)\ , \tag{25}$$

where $\hat{\mathbf{y}}_{(AB)}$ are the unique non-additive contributions from the modality pair AB, $\hat{\mathbf{y}}_{(AB)(AC)}$ are the redundant non-additive contributions that can be derived from the pair AB and also from the pair AC; and the trimodal model

$$\hat{\mathbf{y}}_{(ABC)} = f_{\theta_{ABC}}(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C) \tag{26}$$

which learns the unique trimodal interactions $\hat{\mathbf{y}}_{(ABC)}$.

To define the loss function of MRO, we define $\hat{\mathbf{y}}_{\text{uni}}$ as the sum of all unimodal contributions (the sum over all outputs from the three unimodal models)

$$\begin{aligned}\hat{\mathbf{y}}_{\text{uni}} =& \hat{\mathbf{y}}_A + \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{AC} + \hat{\mathbf{y}}_B + \hat{\mathbf{y}}_{BA} + \hat{\mathbf{y}}_{BC} \\ &+ \hat{\mathbf{y}}_C + \hat{\mathbf{y}}_{CA} + \hat{\mathbf{y}}_{CB}\end{aligned} \tag{27}$$

and $\hat{\mathbf{y}}_{\text{bi}}$ as the sum of all bimodal contributions (the sum over all outputs from the three bimodal models)

$$\begin{aligned}\hat{\mathbf{y}}_{\text{bi}} =& \hat{\mathbf{y}}_{(AB)} + \hat{\mathbf{y}}_{(AB)(AC)} + \hat{\mathbf{y}}_{(AB)(BC)} \\ &+ \hat{\mathbf{y}}_{(AC)} + \hat{\mathbf{y}}_{(AC)(AB)} + \hat{\mathbf{y}}_{(AC)(BC)} \\ &+ \hat{\mathbf{y}}_{(BC)} + \hat{\mathbf{y}}_{(BC)(AB)} + \hat{\mathbf{y}}_{(BC)(AC)}\ .\end{aligned} \tag{28}$$

To encourage that $\hat{\mathbf{y}}_{\text{bi}}$ contains only bimodal non-additive interactions, meaning no unimodal additive interactions, and $\hat{\mathbf{y}}_{\text{tri}}$ only trimodal non-additive interactions, we use the MRO loss formulation which prioritizes the unimodal contributions $\hat{\mathbf{y}}_{\text{uni}}$, falls back on the bimodal contributions $\hat{\mathbf{y}}_{\text{bi}}$ to correct the unimodal mistakes (residuals), and only then uses the trimodal contributions $\hat{\mathbf{y}}_{(ABC)}$. We use the MRO loss as our downstream loss L

$$\begin{aligned}L(\mathbf{y}, \hat{\mathbf{y}}) =& L(\mathbf{y}, \hat{\mathbf{y}}_{\text{uni}}) + L(\mathbf{y}, sg(\hat{\mathbf{y}}_{\text{uni}}) + \hat{\mathbf{y}}_{\text{bi}}) \\ &+ L(\mathbf{y}, sg(\hat{\mathbf{y}}_{\text{uni}} + \hat{\mathbf{y}}_{\text{bi}}) + \hat{\mathbf{y}}_{(ABC)})\end{aligned} \tag{29}$$

where sg means stop gradient [22] which prevents back-propagation through sg's arguments.

To achieve the factorization constraints from SMURF within the unimodal and within the bimodal contributions, we define the two auxiliary loss terms L_{uncor} and L_{cor} as following

$$\begin{aligned}L_{\text{uncor}} = \frac{1}{12}\big(&\text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) + \text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) \\ &+ \text{cov}(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{BA}) + \text{cov}(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{BC}) \\ &+ \text{cov}(\hat{\mathbf{y}}_C, \hat{\mathbf{y}}_{CA}) + \text{cov}(\hat{\mathbf{y}}_C, \hat{\mathbf{y}}_{CB}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(AB)}, \hat{\mathbf{y}}_{(AB)(AC)}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(AB)}, \hat{\mathbf{y}}_{(AB)(BC)}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(AC)}, \hat{\mathbf{y}}_{(AC)(AB)}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(AC)}, \hat{\mathbf{y}}_{(AC)(BC)}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(BC)}, \hat{\mathbf{y}}_{(BC)(AB)}) \\ &+ \text{cov}(\hat{\mathbf{y}}_{(BC)}, \hat{\mathbf{y}}_{(BC)(AC)})\big)\end{aligned} \tag{30}$$

$$\begin{aligned}L_{\text{cor}} = \frac{1}{6}\big(&\text{HGR}(\hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_{BA}) \\ &+ \text{HGR}(\hat{\mathbf{y}}_{AC}, \hat{\mathbf{y}}_{CA}) \\ &+ \text{HGR}(\hat{\mathbf{y}}_{BC}, \hat{\mathbf{y}}_{CB}) \\ &+ \text{HGR}(\hat{\mathbf{y}}_{(AB)(AC)}, \hat{\mathbf{y}}_{(AC)(AB)}) \\ &+ \text{HGR}(\hat{\mathbf{y}}_{(AB)(BC)}, \hat{\mathbf{y}}_{(BC)(AB)}) \\ &+ \text{HGR}(\hat{\mathbf{y}}_{(AC)(BC)}, \hat{\mathbf{y}}_{(BC)(AC)})\big)\end{aligned} \tag{31}$$

where for brevity we use $\text{HGR}(\mathbf{a}, \mathbf{b})$ to refer to $-\text{cov}(\mathbf{a}, \mathbf{b}) + \frac{1}{2}\text{var}(\mathbf{a})\text{var}(\mathbf{b})$. As illustrated in Figure 4, we apply the same loss terms as for the late-fusion trimodal SMURF, but also use them for the bimodal non-additive contributions.

Synthetic (Non-Additive): We validate MRO-SMURF on a synthetic dataset that has two non-additive interactions in the form of two multiplications between $\mathbf{u}_A$ and $\mathbf{u}_B$ and between $\mathbf{u}_C$ and $\mathbf{r}_{AB}$

$$\mathbf{y} = \mathbf{u}_A \mathbf{u}_B + \mathbf{u}_C \mathbf{r}_{AB} \tag{32}$$

where $\mathbf{u}_A, \ldots, \mathbf{r}_{AB} \sim \mathbb{N}(0, 1)$ are randomly sampled. We define the three modalities A, B, and C as

$$A = [\mathbf{u}_A, \mathbf{r}_{AB}],\ B = [\mathbf{u}_B, \mathbf{r}_{AB}],\ \text{and } C = [\mathbf{u}_C]\ . \tag{33}$$

$\mathbf{u}_A \mathbf{u}_B$ is a unique non-additive interaction present only in the modality pair (A, B), while $\mathbf{u}_C \mathbf{r}_{AB}$ is a pairwise redundant non-additive interaction between modality pairs (A, C) and (B, C).

Factorizing non-additive interactions: We test whether MRO-SMURF is able to reconstruct the one unique ($\mathbf{u}_A \mathbf{u}_B$) and the one pairwise redundant ($\mathbf{u}_C \mathbf{r}_{AB}$) non-additive interaction. MRO-SMURF closely reconstructs the unique non-additive interaction, i.e., $r(\hat{\mathbf{y}}_{(AB)}, \mathbf{u}_A \mathbf{u}_B) = 0.944$, and also the pairwise redundant non-additive interaction, i.e., $r(\hat{\mathbf{y}}_{(AC)(BC)}, \mathbf{u}_C \mathbf{r}_{AB}) = 0.999$ and $r(\hat{\mathbf{y}}_{(BC)(AC)}, \mathbf{u}_C \mathbf{r}_{AB}) = 0.998$, indicating that MRO-SMURF can conceptually learn to factorize non-additive interactions.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

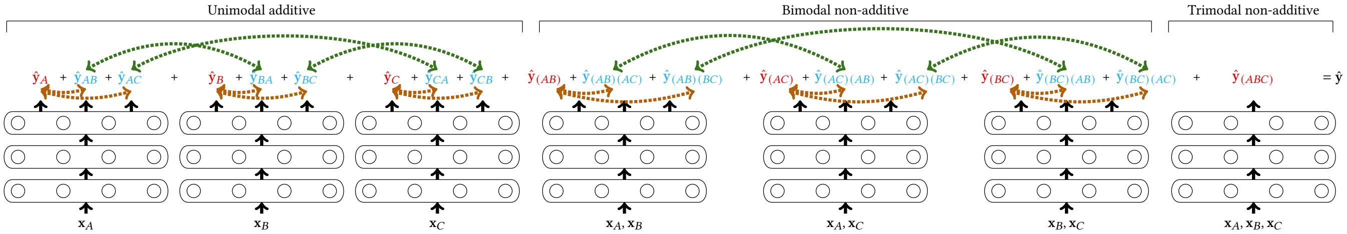

Unimodal additive
Bimodal non-additive
Trimodal non-additive
$\hat{y}_A + \hat{y}_{AB} + \hat{y}_{AC} + \hat{y}_B + \hat{y}_{BA} + \hat{y}_{BC} + \hat{y}_C + \hat{y}_{CA} + \hat{y}_{CB} + \hat{y}_{(AB)} + \hat{y}_{(AB)(AC)} + \hat{y}_{(AB)(BC)} + \hat{y}_{(AC)} + \hat{y}_{(AC)(AB)} + \hat{y}_{(AC)(BC)} + \hat{y}_{(BC)} + \hat{y}_{(BC)(AB)} + \hat{y}_{(BC)(AC)} + \hat{y}_{(ABC)} = \hat{y}$
$\mathbf{x}_A$
$\mathbf{x}_B$
$\mathbf{x}_C$
$\mathbf{x}_A, \mathbf{x}_B$
$\mathbf{x}_A, \mathbf{x}_C$
$\mathbf{x}_B, \mathbf{x}_C$
$\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C$

Figure 4: Illustration of combining SMURF and MRO for three modalities. MRO factorizes unimodal additive, bimodal non-additive and trimodal non-additive interactions and SMURF further factorizes the additive and bimodal non-additive interactions into unique and redundant contributions.