Balanced Multimodal Representation Learning

Qing-Yuan Jiang

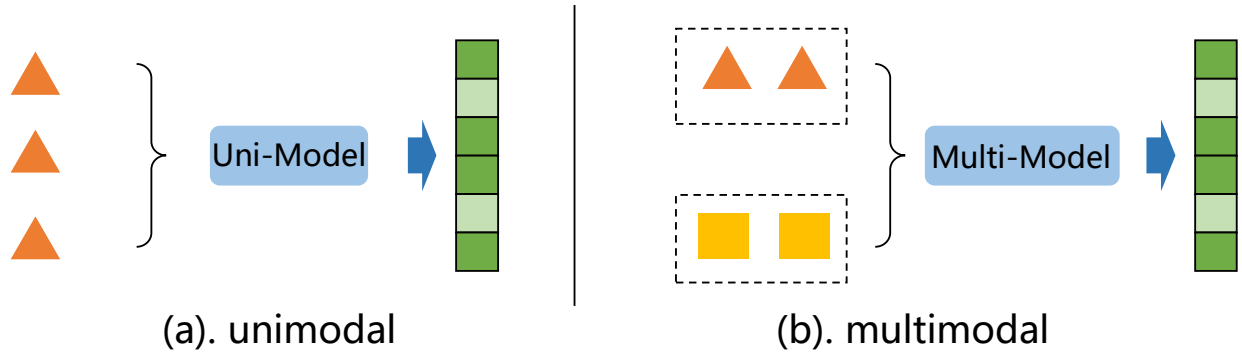
NJUST-KMG group,
The School of Computer Science and Engineering, NJUST

March 26, 2026

Multimodal Representation Learning

Multimodal Learning:

- **Goal:** fuse multimodal data to boost model performance.
- **Optimal Scenario:** maximize information extraction from multiple modalities for **better performance**.
- Taking multimodal data as inputs, and then predicting.

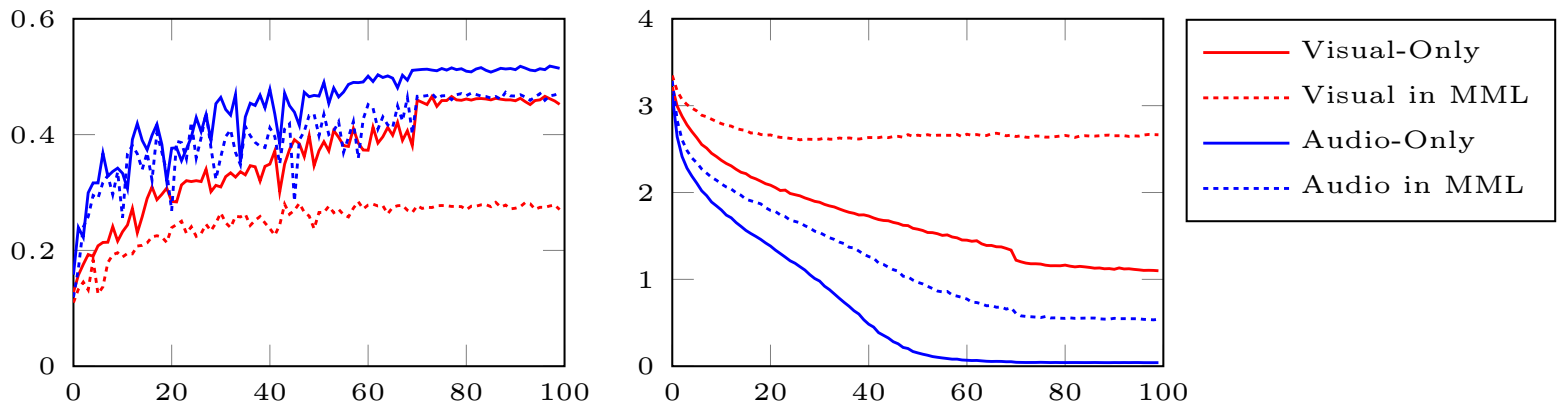


(a). unimodal (b). multimodal

Imbalance in Multimodal Representation Learning

Modality Imbalance:

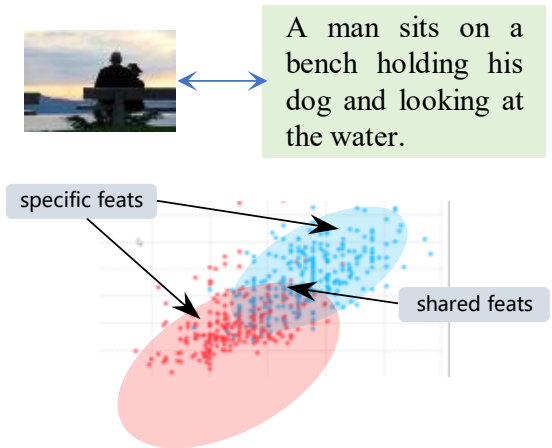
- **Phenomenon:** MML underperforms single-modality models.
- Strong-Weak Modality $\mapsto$ Modality Imbalance



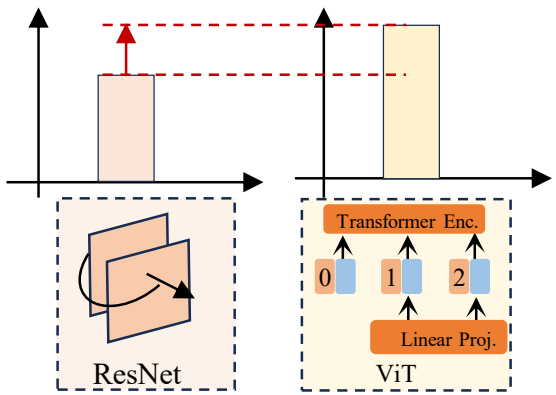
(a). Training Accuracy. (b). Training Loss.

Multi-dimensional Modality Imbalance

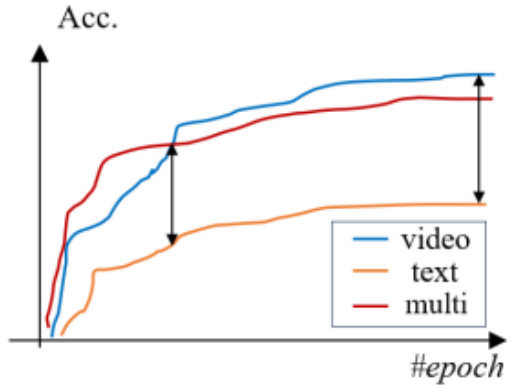
- Modality imbalance arises throughout the entire process of multimodal learning, including the data, model, and learning levels.



(a). data



(b). model



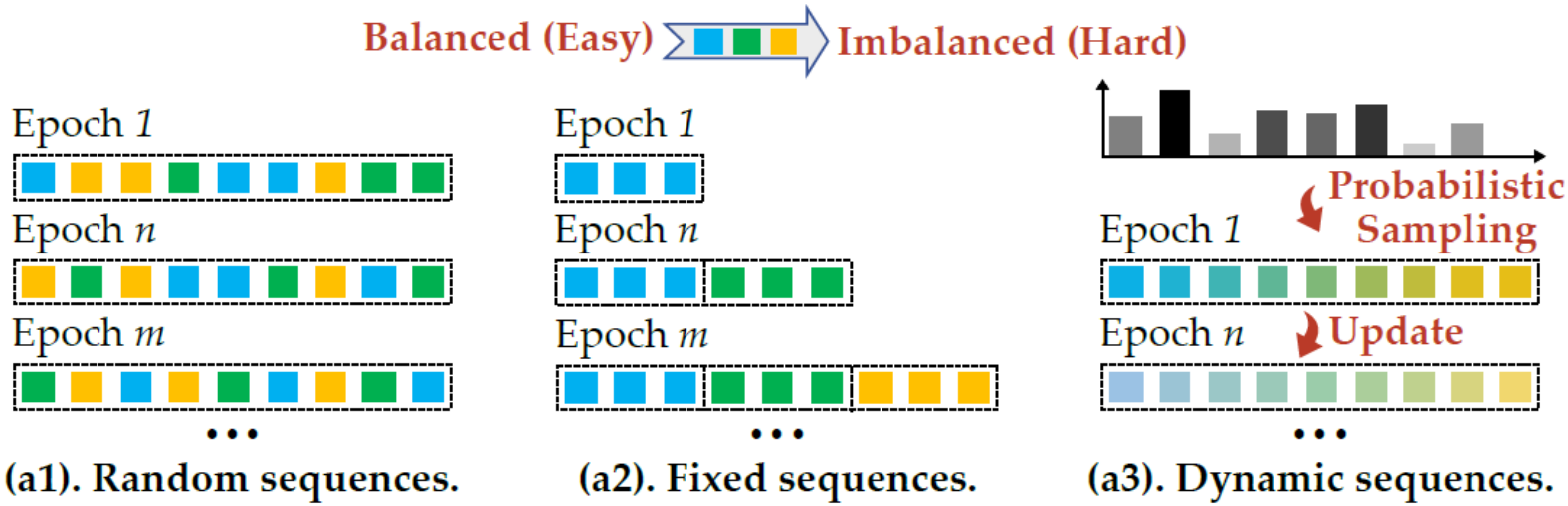
(c). learning

Our Works

- **Data-Level**: Balance-Aware Sequence Sampling [IJCAI'25]
- **Model-Level**: Adaptive Unimodal Gradient Boosting [NeurIPS'25], Adaptively Masked Subnetworks Considering Modal Significance [TPAMI'25]
- **Learning-Level**: Multimodal Learning with Fusion and Alignment [NeurIPS'24], Interactive MML via Gradient Modification [IJCAI'25], Instantaneous Probe-and-Rebalance Multimodal Learning [IJCAI'25]

Data-Level Modality Imbalance

- The training of multimodal models is sensitive to **task difficulty** across **different learning stages**.



Challenge of Data-Level Modality Imbalance

How to measure the *task difficulty*?

- **Correlation criterion:**

$$\mathrm{sim}(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v) = \frac{[\hat{\boldsymbol{y}}_i^a]^\top \hat{\boldsymbol{y}}_i^v}{\|\hat{\boldsymbol{y}}_i^a\|_2 \|\hat{\boldsymbol{y}}_i^v\|_2}$$

- **Information criterion:** modality-aware training loss.

How to use this property to achieve *balanced multimodal learning*?

- **Curriculum learning:** learning from balanced to imbalanced.
- **Fine-grained sampling:** learning a sampling probability *w.r.t.* a balance-oriented measure.

Balance-Aware Sequence Sampling

Balance Score:

- Combining the correlation and information criterion:

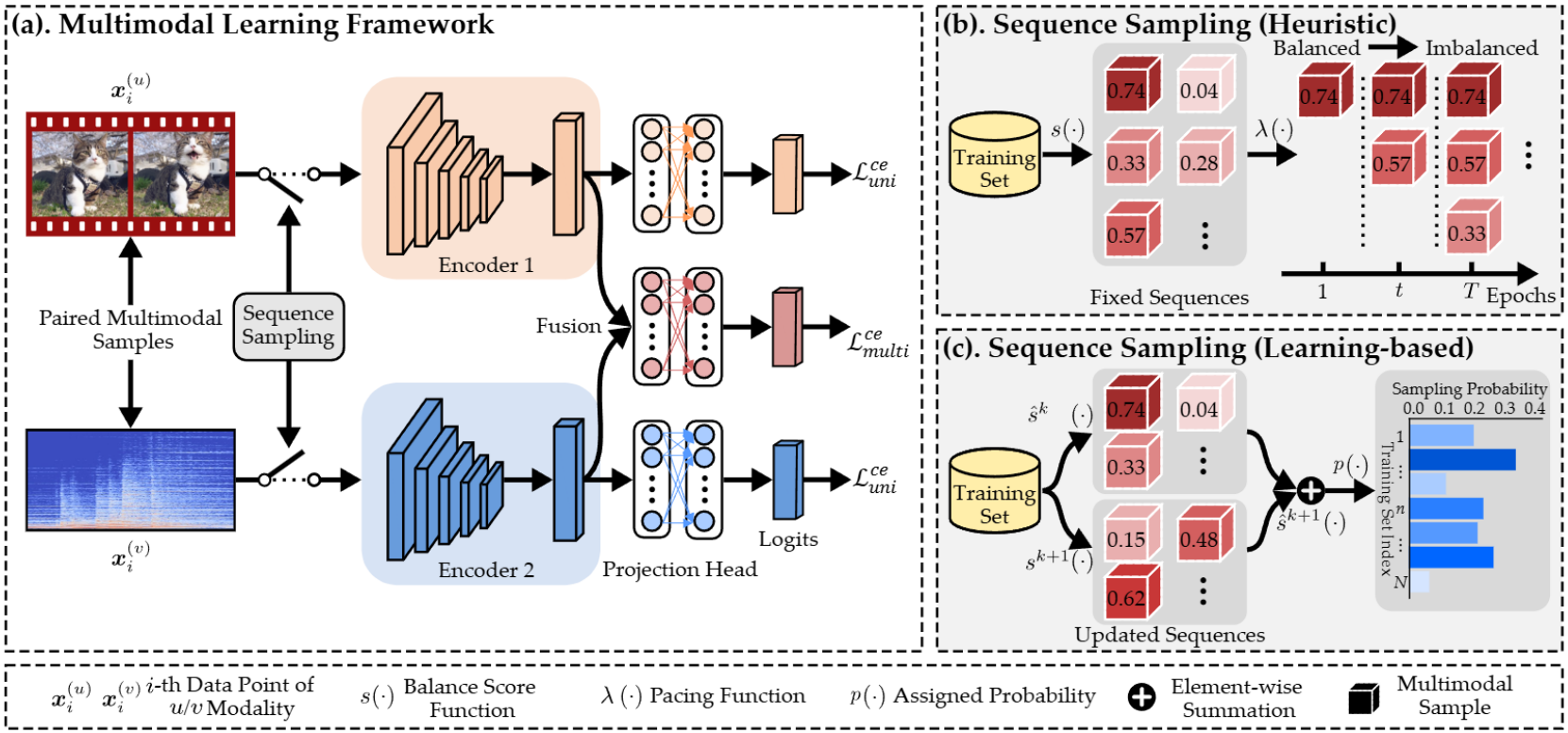
$$s(\boldsymbol{x}_i) = \frac{\mathrm{sim}(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v) - \min(\mathcal{S})}{\max(\mathcal{S}) - \min(\mathcal{S})} - \frac{\ell_{total}(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, \boldsymbol{y}_i) - \min(\mathcal{L})}{\max(\mathcal{L}) - \min(\mathcal{L})}$$

Training scheduler:

- **Heuristic scheduler:** spliting training set as multiple blocks based on pace function. Then sample batch from each block.
- **Learning-based scheduler:** sampling batch based on sampling probability:

$$p(\boldsymbol{x}_i) = \frac{e^{\hat{s}^{k+1}(\boldsymbol{x}_i)}}{\sum_{j=1}^{n} e^{\hat{s}^{k+1}(\boldsymbol{x}_j)}}.$$

The Architecture of BSS



The BSS Algorithm

Algorithm 1 Multimodal Learning with Balance-Aware Sequence Sampling (BSS)

Input: Training set $\boldsymbol{X}$, category labels $\boldsymbol{Y}$.
Output: Learned params θ of all models.

Initialize params θ^0, maximum epochs T, training set for ranking $\boldsymbol{X}_{\text{rank}} = \varnothing$, curriculum period T_{grow}, initial proportion λ_0, epoch interval E.
for each sample $\boldsymbol{x}_i$ in $\boldsymbol{X}$ **do**
    Obtain balance score $s(\boldsymbol{x}_i)$.
    Add $\boldsymbol{x}_i$ to $\boldsymbol{X}_{\text{rank}}$ in descending order of $s(\boldsymbol{x}_i)$.
for $t = 0$ to $T - 1$ **do**
    if $scheduler ==' \text{heuristic}'$ **then**
        Calculate the proportion of the training samples $\lambda(t)$.
        Obtain current batch data $\boldsymbol{X}_{\text{data}}$ from $\boldsymbol{X}_{\text{rank}}$.
    else if $scheduler ==' \text{learning-based}'$ **then**
        Update $s(\boldsymbol{x}_i)$ every E epochs.
        Assign sampling probability $p(\boldsymbol{x}_i)$ based on $s(\boldsymbol{x}_i)$.
        Obtain current batch data $\boldsymbol{X}_{\text{batch}}$.
    Train model with $\boldsymbol{X}_{batch}$ and update parameters θ.
    Update $t = t + 1$.

Experiments for BSS

- Main results compared with SOTA:

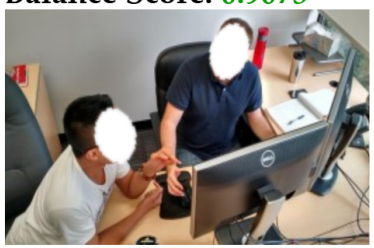
Method	CREMA-D		Kinetics-Sounds		Twitter2015		Sarcasm		NVGesture	
	ACC	MAP	ACC	MAP	ACC	F1	ACC	F1	ACC	F1
Audio/Text/RGB	63.17	68.61	54.12	56.69	73.67	68.49	81.36	80.65	78.22	78.33
Video/Image/OF	45.83	58.79	55.62	58.37	58.63	43.33	71.81	70.73	78.63	78.65
Depth	-	-	-	-	-	-	-	-	81.54	81.83
Concat	63.31	68.31	64.55	71.31	70.11	63.86	82.86	82.43	81.33	81.47
Affine	66.26	71.93	64.24	69.31	72.03	59.92	82.47	81.88	82.78	82.81
Channel	66.13	71.75	63.51	68.66	-	-	-	-	81.54	81.57
ML-LSTM	62.94	64.73	63.84	69.02	70.68	65.64	82.05	70.73	83.20	83.30
Sum	63.44	69.08	64.97	71.03	73.12	66.61	82.94	82.47	82.99	83.05
Weight	66.53	73.26	65.33	71.33	72.42	65.16	82.65	82.19	83.42	83.57
ETMC	65.86	71.34	65.67	71.19	73.96	67.39	83.69	83.23	83.61	83.69
MSES	61.56	68.83	64.71	70.63	71.84	66.55	84.18	83.60	81.12	81.47
OGR-GB	64.65	84.54	67.10	71.39	74.35	68.69	83.35	82.71	82.99	83.05
DOMFN	67.34	85.72	66.25	72.44	74.45	68.57	83.56	82.62	-	-
OGM	66.94	71.73	66.06	71.44	74.92	68.74	83.23	82.66	-	-
MSLR	65.46	71.38	65.91	71.96	72.52	64.39	84.23	83.69	82.86	82.92
AGM	67.07	73.58	66.02	72.52	74.83	69.11	84.02	83.44	82.78	82.82
PMR	66.59	70.30	66.56	71.93	74.25	68.60	83.60	82.49	-	-
ReconBoost	74.84	81.24	70.85	74.24	74.42	68.34	84.37	83.17	84.13	86.32
MMPareto	74.87	85.35	70.00	78.50	73.58	67.29	83.48	82.48	83.82	84.24
SMV	78.72	84.17	69.00	74.26	74.28	68.17	84.18	83.68	83.52	83.41
MLA	79.43	85.72	70.04	74.13	73.52	67.13	84.26	83.48	83.40	83.72
AMSS	70.30	76.14	72.25	79.13	75.12	69.23	84.35	83.77	84.64	84.94
BSS-H	80.78	87.86	72.67	78.61	74.73	68.67	84.41	83.86	85.06	85.15
BSS-L	**82.80**	**88.61**	**73.95**	**79.43**	**75.22**	**69.51**	**85.01**	**84.62**	**86.72**	**87.04**

Further Analysis for BSS

- Ablation study and visualization:

Criterion		ACC (%) / MAP (%)		
PreSim	Loss	Audio	Video	Multi
✗	✗	49.37/51.07	54.03/57.48	70.44/76.62
✗	✓	52.11/54.40	54.23/57.91	72.44/79.41
✓	✗	52.38/54.32	**54.93/58.52**	73.25/78.98
✓	✓	**52.73/54.43**	54.74/58.46	**73.95/79.43**

Label: Positive
Balance Score: 0.9073



Congratulations to South Greene freshman Taylor Lamb for earning state honors from the TSWA.

Label: Neutral
Balance Score: 0.4788

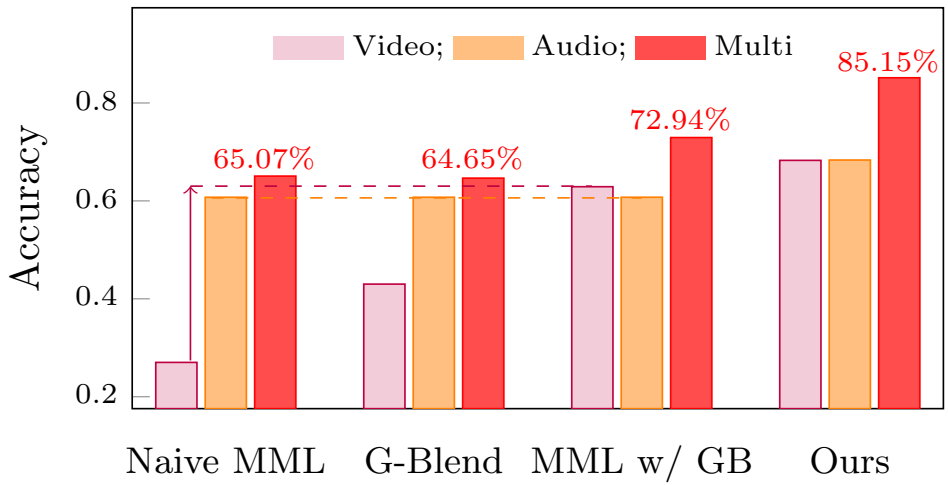
Of course Tesla has a solar camera battery to document their buildout.

Label: Neutral
Balance Score: 0.0641

Good morning, church! Grab a coffee from Elevate Cafe and join us at 9 or 11 am.

Model-Level Modality Imbalance

- Capability discrepancies across **heterogeneous models** play a crucial role in causing modality imbalance.



How to balance the *model capability* for multimodal learning?

- **Enhancing the learning capabilities of weak modality:** classifier boosting.

Adaptive Unimodal Gradient Boosting

How to enhance the **learning capability** of weaker modalities?

- Evaluating current learning status: OGM score [1], etc.
- Enhancing the learning capability: **Sustained Boosting**.

The objective function of multimodal learning:

$$\mathcal{L}^o(\boldsymbol{X}^o, \boldsymbol{Y}) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i^\top \log(\boldsymbol{p}_i^o).$$

Residual objective at t-th iteration:

$$\hat{\boldsymbol{y}}_{it}^o = \boldsymbol{y}_i - \lambda \sum_{j=1}^{t-1} \boldsymbol{y}_i \odot \boldsymbol{p}_{ij}^o; \quad \epsilon(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) = \ell\left(\boldsymbol{p}_{it}^o, \hat{\boldsymbol{y}}_{it}^o\right).$$

Here, λ is used to soften hard labels.

Adaptive Unimodal Gradient Boosting (cont'd)

Since we utilize a shared encoder, we have to ensure that all classifiers are updated simultaneously:

$$\epsilon_o(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) = \ell\left(\boldsymbol{p}_{it}^o + \sum_{j=1}^{t-1} \boldsymbol{p}_{ij}^o, \boldsymbol{y}_i\right) = \ell\left(\sum_{j=1}^{t} \boldsymbol{p}_{ij}^o, \boldsymbol{y}_i\right).$$

Furthermore, we have to ensure that the $t-1$ classifiers are well-trained:

$$\epsilon_p(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) = \ell\left(\sum_{j=1}^{t-1} \boldsymbol{p}_{ij}^o, \boldsymbol{y}_i\right).$$

In summary, we have:

$$L(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) = \epsilon(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) + \epsilon_o(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t) + \epsilon_p(\boldsymbol{x}_i^o, \boldsymbol{y}_i, t).$$

Adaptive Unimodal Gradient Boosting (cont'd)

A strategy to add the classifier adaptively:

- Monitor the learning status:

$$\forall o \in \{a, v\}, s_t^o = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i^\top \left[\sum_{j=1}^{n^o} \boldsymbol{p}_{ij}^o \right].$$

- If $s_t^a - \sigma s_t^v > \tau$, add a classifier for video, and verse vice.

Adaptive Unimodal Gradient Boosting (cont'd)

Define the gap function:

$$\mathcal{G}(\Phi) = \mathcal{L}^a(\Phi^a) - \mathcal{L}^v(\Phi^v),$$

We have:

Theorem: Convergence of Gap Loss, Informal

Under some assumptions for the loss function and the effectiveness of sustained boosting algorithm, we have:

$$\mathcal{G}(\Phi(T)) \leq \frac{1}{1 + \frac{\nu^2\kappa^2}{2L_a\beta^2} T\mathcal{G}(\Phi(0))}\mathcal{G}(\Phi(0))$$

where ν, κ, L_a and β are constant.

The AUG Algorithm

Algorithm 2 Multimodal Learning with Adaptive Unimodal Gradient Boosting

Input: Training set $\boldsymbol{X}$, category labels $\boldsymbol{Y}$.
Output: Learned DNN models for all modalities.

Initialize the number of classifier $n^a = 1, n^v = 1$. Initialize iteration $t = 1$. Initialize the params of DNN models Φ_t^a, Φ_t^v.
for $t = 1 \mapsto \#iteration$ **do**
  Sample a mini-batch $\boldsymbol{X}_t = \{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v)\}_{i=1}^{n_b}$.
  $\forall \boldsymbol{x}_i^a, \boldsymbol{x}_i^v \in \boldsymbol{X}_t$, calculate features $\boldsymbol{u}_i^a$ and $\boldsymbol{u}_i^v$.
  Calculate predictions $\{\boldsymbol{p}_{ij}^a\}_{j=1}^{n^a}$ and $\{\boldsymbol{p}_{ij}^v\}_{j=1}^{n^v}$.
  Calculate loss based on predictions;
  Update DNN parameters Φ_t^a and Φ_t^v based on SGD;
  if $\mod(t, t_N) = 0$ **then**
    Calculate confident score $\{s_t^a, s_t^v\}$ based on predictions;
    if $s_t^a - \sigma s_t^v > \tau$ **then**
      Add a classifier for audio modality;
      $n^a = n^a + 1$;
    else if $s_t^a - \sigma s_t^v < \tau$ **then**
      Add a classifier for video modality;
      $n^v = n^v + 1$;

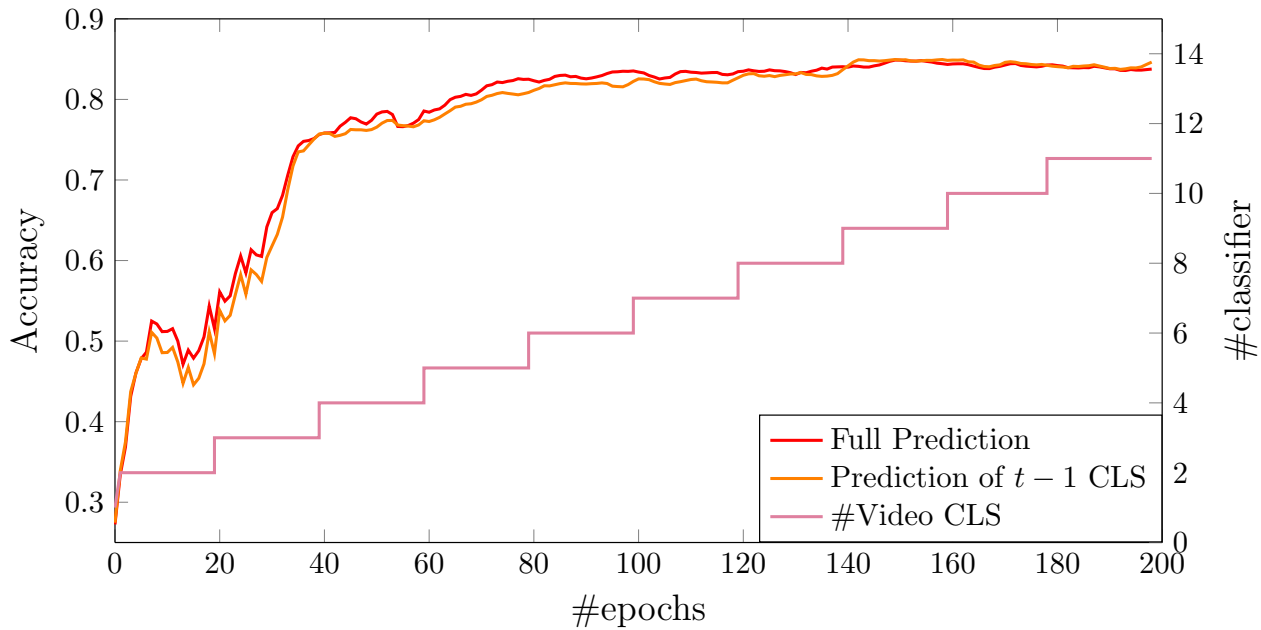
Experiments for AUG

- Main results compared with SOTA:

Method	CREMA-D		KSounds		VGGSound		Twitter2015		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	MAP	Acc.	F1	Acc.	F1	Acc.	F1
Unimodal-1	.4583	.5879	.5412	.5669	.4655	.4701	.5863	.4333	.7181	.7073	.7822	.7833
Unimodal-2	.6317	.6861	.5562	.5837	.3494	.3478	.7367	.6849	.8136	.8056	.7863	.7865
Unimodal-3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.8154	.8183
Concat	.6361	.6841†	.6455	.7130	.5116	.5352	.7011†	.6386	.8286	.8240	.8237	.8270
Affine	.6626	.7193	.6424	.6931	.5001	.5155	.7203†	.5992†	.8240	.8188	.8278	.8281
ML-LSTM	.6290†	.6473†	.6394	.6902	.4966	.5139	.7068†	.6564†	.8277	.8205	.8320	.8330
Sum	.6344	.6908	.6490	.7103	.5136	.5338	.7300†	.6661†	.8294	.8247	.8050†	.8067†
Weight	.6653	.7134	.6533	.7110	.5144	.5300	.7242†	.6516†	.8265	.8219	.7842†	.7939†
MSES	.6546	.7138	.6591	.7196	.4891	.5429	.7252†	.6439†	.8423	.8369	.8112†	.8147†
G-blend	.6465	.7392	.6722	.7274	.5086	.5555	.7309†	.6799†	.8286	.8215	.8299	.8305
MSLR	.6868	.7412	.6756	.7282	.4987	.5415	.7232†	.6382†	.8439	.8378	.8237	.8284
OGM	.6612	.7372	.6582	.7159	.4829	.4978	.7058†	.6435†	.8360	.8293	—	—
PMR	.6659	.7058	.6675	.7274	.4647	.4866	.7357†	.6636†	.8310	.8256	—	—
AGM	.6733	.7807	.6791	.7388	.4711	.5198	.7261†	.6502†	.8306	.8293	.8279	.8284
MMParato	.7487	.8535	.7000	.7850	.5125	.5473	.7358†	.6729†	.8348	.8284	.8382	.8424
SMV	.7872	.8417	.6900	.7426	.5031	.5362	.7428	.6817†	.8418	.8368	.8352	.8341
MLA	.7943	.8572	.7004	.7945	.5165	.5473	.7352†	.6713†	.8426	.8348	.8340	.8372
DI-MML	.8158	.8592	.7203	.7426	.5173	.5479	.7248†	.6686†	.8411	.8315	—	—
LFM	.8362	.9006	.7253	.7897	.5274	.5694	.7501	**.7057**	.8497	.8457	.8436	.8468
ReconBoost	.7557	.8140	.6855	.7662	.5097	.5387	.7442	.6832†	.8437	.8317	.8386	.8434
Ours	**.8515**	**.9103**	**.7263**	**.7901**	**.5301**	**.5826**	**.7512**	.6962	**.8510**	**.8458**	**.8501**	**.8533**

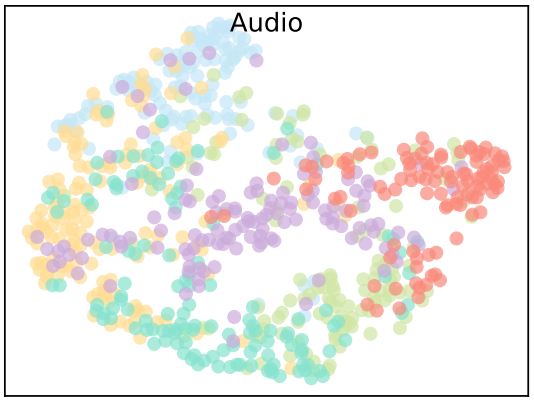
Further Analysis for AUG

- **Effectiveness:** the impact of residual learning on classification performance.

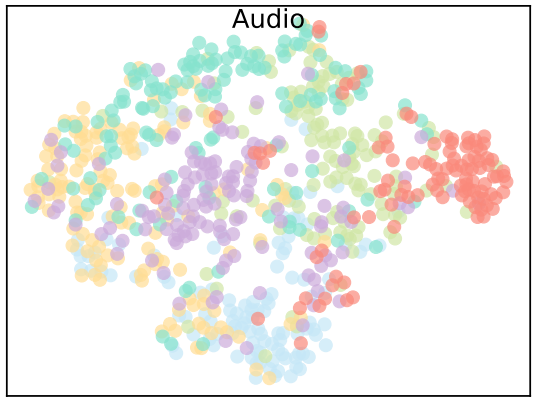


Further Analysis for AUG (cont'd)

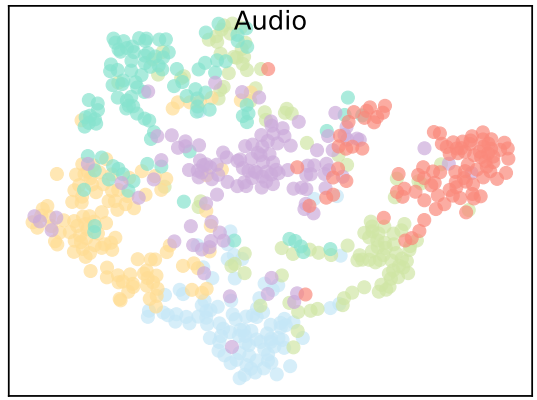
- **Visualization:** the improvement of weak modality.



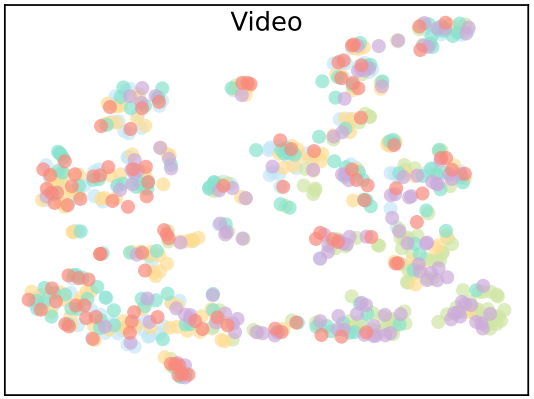
(a). naive MML@Audio.



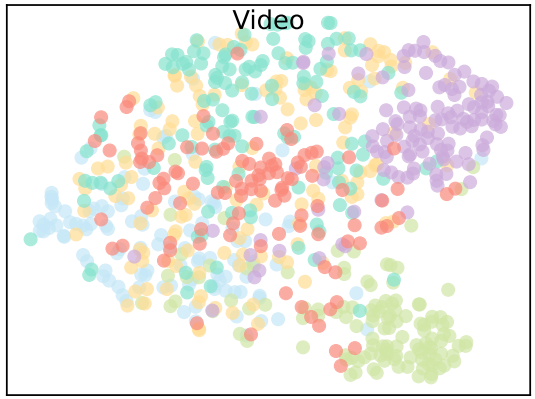
(b). ReconBoost@Audio.



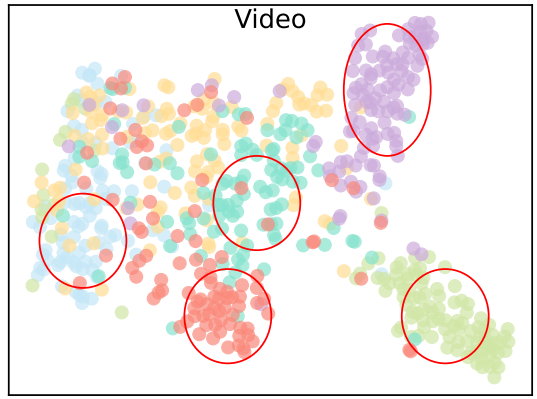
(c). Ours@Audio.



(d). naive MML@Video.



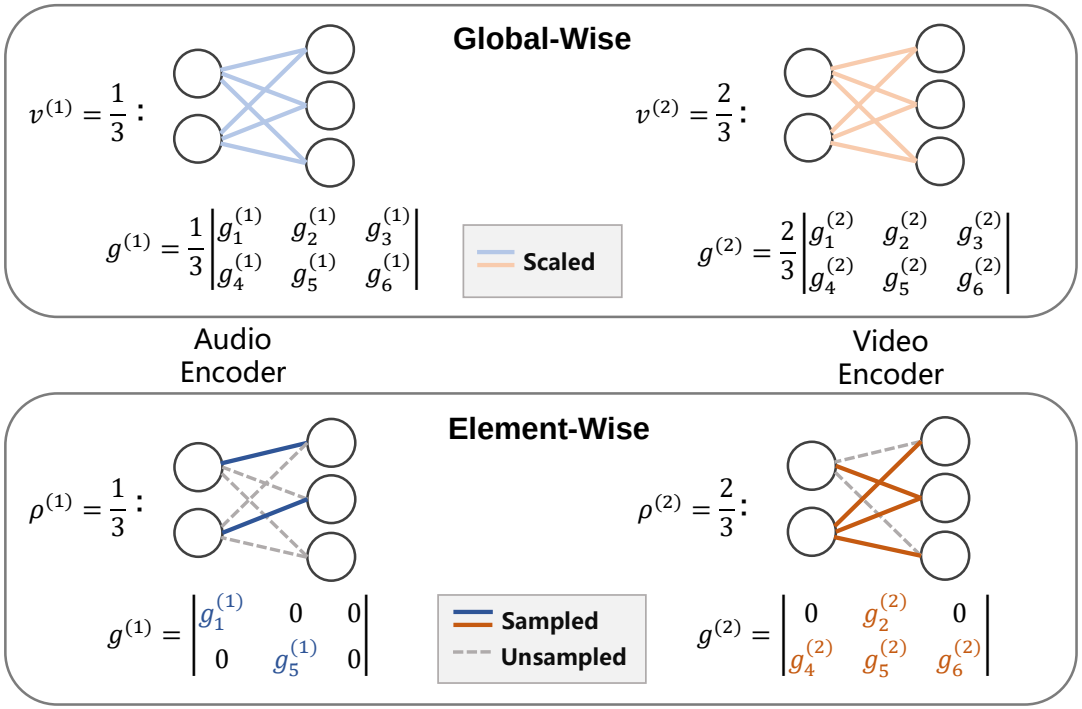
(e). ReconBoost@Video.



(f). Ours@Video.

Adaptively Mask Subnetworks Considering Modal Significance

- **Fine-grained** gradient adjustment strategy.



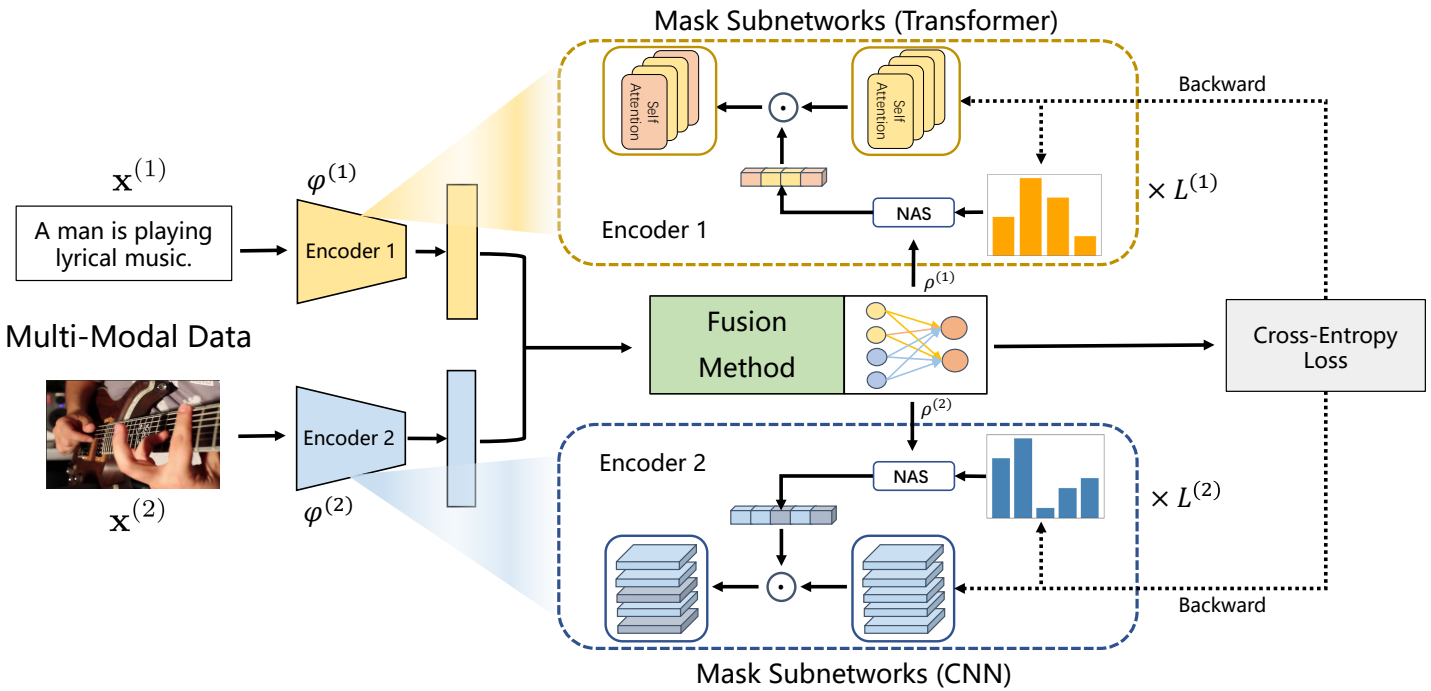
What is the key point of this matter?

- Adjusting the learning weights of modality-specific models
- Fine-grained adjustment: searching modality-specific subnetwork

Adaptively Mask Subnetworks Considering Modal Significance (cont'd)

How to "search" a modality-specific subnetwork?

- Evaluating the **modality significance**
- Masking **modality-specific models** during learning



Adaptively Mask Subnetworks Considering Modal Significance (cont'd)

Learning procedure:

- **Learning objective:**

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i^{\top} \log(q(\boldsymbol{x}_i)).$$

- Updating rule with adjustment weight $v(t)$:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \cdot v(t) \cdot \nabla \mathcal{L}(\boldsymbol{w}(t)).$$

- Fine-grained updating rule with masking $\boldsymbol{m}(t)$:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \cdot \boldsymbol{m}(t) \odot \nabla \mathcal{L}(\boldsymbol{w}(t)).$$

Modality Significance Evaluation

- **Update ratio:** the quantity of masking parameter

$$\bar{u}^{(k)} = \frac{\mathbb{I}(\boldsymbol{X}^{(k)}; \boldsymbol{Y})}{\mathbb{H}(\boldsymbol{X}^{(k)})},$$
$$u^{(k)} = \lambda u^{(k)} + (1-\lambda)\bar{u}^{(k)},$$
$$\rho^{(k)} = 1 - \frac{\exp(u^{(k)}/\tau)}{\sum_{n=1}^{K} \exp(u^{(n)}/\tau}.$$

- **Masking ratio:** the weight of masking parameter

$$\boldsymbol{F}(\boldsymbol{w}^{(k)}) = \mathbb{E}[(\frac{\partial \log(p(\bar{\boldsymbol{y}} \mid \boldsymbol{x}^{(k)}; \boldsymbol{w}^{(k)}))}{\partial \boldsymbol{w}^{(k)}})(\frac{\partial \log(p(\bar{\boldsymbol{y}} \mid \boldsymbol{x}^{(k)}; \boldsymbol{w}^{(k)}))}{\partial \boldsymbol{w}^{(k)}})^{\top}],$$

$$p_j^{(k)} = \frac{F_j(\boldsymbol{w}^{(k)})}{\sum_{j=1}^{|\boldsymbol{w}^{(k)}|} F_j(\boldsymbol{w}^{(k)})}$$

Parameter Sampling

- **Parameter sampling:** based on update ratio and sampling probability

$$\boldsymbol{P}^{(k)} = \{p_1^{(k)}, p_2^{(k)}, \cdots, p_{|\boldsymbol{w}^{(k)}|}^{(k)}\}$$

- Based on $\boldsymbol{P}^{(k)}$, we obtain the selected sampling sequence.

$$\mathcal{S}^{(k)} = \{s_1^{(k)}, s_2^{(k)}, \cdots, s_{\lceil \rho^{(k)}|\boldsymbol{w}^{(k)}| \rceil}^{(k)}\}$$

- Constructing masking:

$$\bar{m}_j(t) = \begin{cases} \dfrac{1}{p_j}, & \text{if } w_j^{(k)}(t) \in \mathcal{S}^{(k)}(t), \\ 0, & \text{otherwise.} \end{cases}$$

Theoretical Analysis

Theorem: Convergence, Informal

Under some assumptions for the stochastic gradient $\nabla \ell(\boldsymbol{w}^{(k)}) \odot \boldsymbol{m}^{(k)}(t)$, we have:

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \ell(\boldsymbol{w}^{(k)})\|^2 \leq \mathcal{O}\left(\frac{1+(1+\nu)^2}{\sqrt{T}(1+\nu)(1-\delta^2)}\right)$$

where $\delta \in (0,1)$ and $\nu \geq 0$ are two constants.

Experiments for AMSS

- Main results compared with SOTA:

Methods		KSounds		CREMA-D		Sarcasm		Twitter2015	
		ACC	mAP	ACC	mAP	ACC	Mac-F1	ACC	Mac-F1
Uni-modal	Audio/Text	54.12	56.69	63.17	68.61	81.36	80.65	73.67	68.49
	Video/Image	55.62	58.37	45.83	58.79	71.81	70.73	58.63	43.33
Multi-modal Fusion	Concat	64.55	71.30	63.31	68.41	82.86	82.40	70.11	63.86
	Affine	64.24	69.31	66.26	71.93	82.40	81.88	72.03	59.92
	Channel	63.51	68.66	66.13	71.70	-	-	-	-
	ML-LSTM	63.94	69.02	62.90	64.73	82.77	82.05	70.68	65.64
	Sum	64.90	71.03	63.44	69.08	82.94	82.47	73.00	66.61
	Weight	65.33	71.10	66.53	73.26	82.65	82.19	72.42	65.16
	ETMC	65.67	72.50	65.86	71.34	83.69	83.23	73.96	67.39
Multi-modal Rebalance Fusion	MSES	64.71	70.63	61.56	66.83	84.18	83.60	71.84	66.55
	OGR-GB	67.10	71.39	64.65	68.54	83.35	82.71	74.35	68.69
	OGM-GE	66.06	71.44	66.94	71.73	83.23	82.66	74.92	68.74
	Greedy	66.52	72.81	66.64	72.64	-	-	-	-
	DOMFN	66.25	72.44	67.34	73.72	83.56	82.62	74.45	68.57
	MSLR	65.91	71.96	65.46	71.38	84.23	83.69	72.52	64.39
	PMR	66.56	71.93	66.59	70.30	83.60	82.49	74.25	68.60
	AGM	66.02	72.52	67.07	73.58	84.02	83.44	74.83	69.11
	AMSS	68.96	74.89	67.61	73.97	84.14	83.69	**75.89**	**69.81**
	AMSS+	**72.25**	**79.13**	**70.30**	**76.14**	**84.35**	**83.77**	75.12	69.23

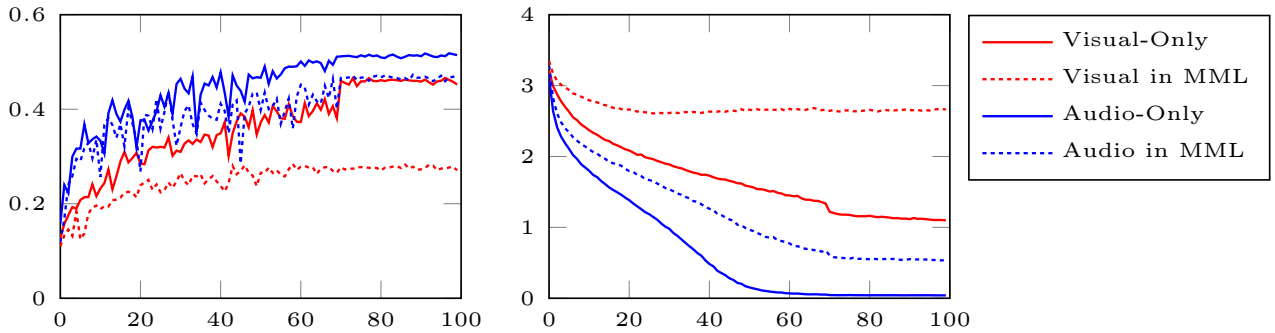
Ablation Study for AMSS

- Ablation study:

Methods	Kinetics-Sound		Sarcasm-Detection	
	ACC	mAP	ACC	Mac-F1
Baseline	64.55	71.30	82.86	82.40
w/ Classifier Mask	65.37	70.87	83.40	83.05
w/ Backbone Mask	66.68	72.23	83.81	83.17
AMSS	**68.96**	**74.89**	**84.14**	**83.69**
w/ Classifier Mask+	66.80	73.80	83.44	82.93
w/ Backbone Mask+	69.15	76.13	83.77	83.10
AMSS+	**72.25**	**79.13**	**84.35**	**83.77**

Learning-Level Modality Imbalance

During training, strong modality tends to dominate the learning process.



(a). Training Accuracy. (b). Training Loss.

How to adjust modal learning to achieve balance during training?

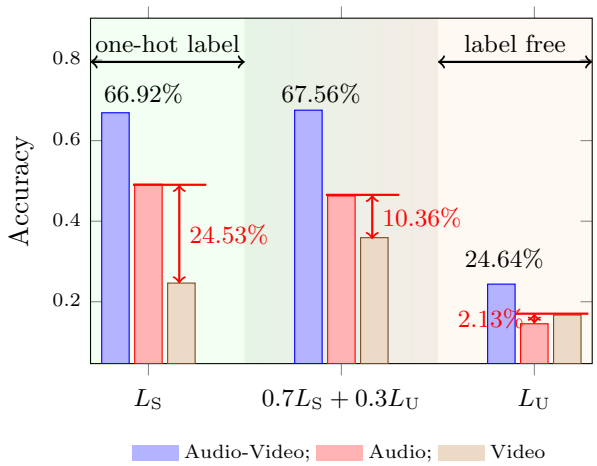
- **Considering fusion and alignment:** dynamically integrating fusion and alignment simultaneously.

Multimodal Learning with Fusion and Alignment

Multimodal learning: *fitting labels*.

$$L_{CLS}(\boldsymbol{X}, \boldsymbol{Y}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^{\top} \log(\hat{\boldsymbol{y}}_i).$$

The performance gap widens as the model fits the labels.



- Introduce contrastive learning to bridge the heterogeneous data in feature space.

Multimodal Learning with Fusion and Alignment (cont'd)

Definition of pair similarity:

$$s(\boldsymbol{x}_i^a, \boldsymbol{x}_k^v) = \frac{\left[\boldsymbol{z}_i^a\right]^\top \boldsymbol{z}_k^v}{\|\boldsymbol{z}_i^a\|_2 \|\boldsymbol{z}_k^v\|_2}.$$

Loss function for contrastive learning:

$$L_{MM}(\boldsymbol{X}) = -\frac{1}{2n_b} \sum_i^{n_b} \left[\log \left(\frac{e^{s(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v)/\tau}}{\sum_k e^{s(\boldsymbol{x}_i^a, \boldsymbol{x}_k^v)/\tau}} \right) + \log \left(\frac{e^{s(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v)/\tau}}{\sum_k e^{s(\boldsymbol{x}_k^a, \boldsymbol{x}_i^v)/\tau}} \right) \right].$$

Total loss:

$$L_{Total} = (1-\alpha) L_{CLS}(\boldsymbol{X}, \boldsymbol{Y}) + \alpha L_{MM}(\boldsymbol{X}).$$

Multimodal Learning with Fusion and Alignment (cont'd)

Dynamically integration:

$$\min_{0\leq\alpha\leq 1} L_{CLS}(\theta^*(\alpha))$$
$$s.t. \quad \theta^*(\alpha) \in \operatorname{argmin}_{\theta} \big\{(1-\alpha)L_{CLS}(\theta) + \alpha L_{MM}(\theta)\big\}.$$

Solution through bi-level optimization:

$$\nabla L_{CLS}(\theta(\alpha)) = -\nabla^2_{\alpha,\theta} L_{Total} [\nabla^2_{\theta,\theta} L_{Total}]^{-1} \nabla_\theta L_{CLS}(\boldsymbol{X}, \boldsymbol{Y}).$$

The LFA Algorithm

Algorithm 3 Multimodal Learning with Label Fitting and Alignment

Input: Training set $\mathcal{X}$, labels $\mathcal{Y}$, method.
Output: Learned parameters $\{\theta\}$ of all models.

Initialize parameters θ, parameter α, maximum iterations T, learning rate η_α.
for $t = 1 \mapsto T$ **do**
  for $i = 1 \mapsto$ Inner_Iters **do**
    Calculate total loss L_{Total} by forward phase.
    Update parameters θ according to its gradient.
    if method $==$ 'learning-based' **then**
      Calculate gradient approximation: $\nabla L_{\text{CLS}}(\theta(\alpha)) = -\nabla^2_{\alpha,\theta} L_{\text{Total}} \cdot [\nabla^2_{\theta,\theta} L_{\text{Total}}]^{-1} \cdot \nabla_\theta L_{\text{CLS}}(\boldsymbol{X}, \boldsymbol{Y})$.
      Update α according to: $\alpha = \alpha - \eta_\alpha \nabla L_{\text{CLS}}(\theta(\alpha))$.
      Clip α into $[0, 1]$: $\alpha := \max(0, \min(1, \alpha))$.
    else if method $==$ 'heuristic' **then**
      Update α according to: $\alpha = 1 - e^{-1/t}$.

Experiments for LFA

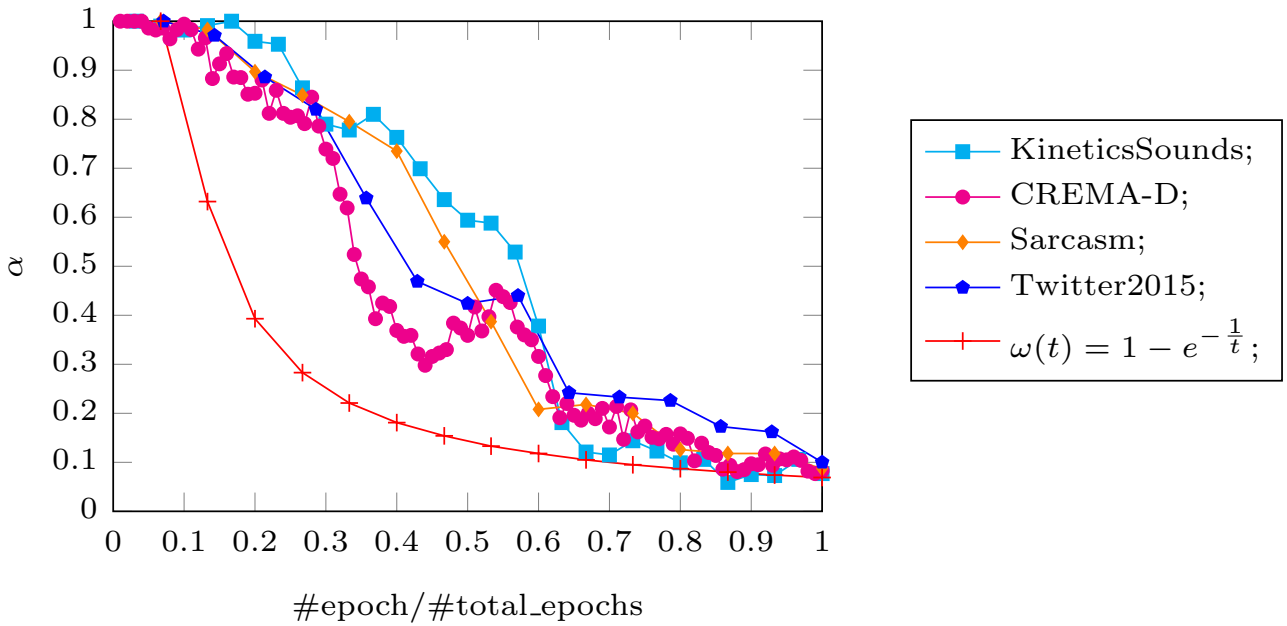
- Main comparison with SOTA:

Method	KSounds		CREMA-D		Sarcasm		Twitter2015		NVGesture	
	ACC	MAP	ACC	MAP	ACC	F1	ACC	F1	ACC	F1
Unimodal-1	54.12%	56.69%	63.17%	68.61%	81.36%	80.65%	73.67%	68.49%	78.22%	78.33%
Unimodal-2	55.62%	58.37%	45.83%	58.79%	71.81%	70.73%	58.63%	43.33%	78.63%	78.65%
Unimodal-3	—	—	—	—	—	—	—	—	81.54%	81.83%
Concat	64.55%	71.31%	63.31%	68.41%†	82.86%	82.43%	70.11%†	63.86%†	81.33%†	81.47%†
Affine	64.24%	69.31%	66.26%	71.93%	82.47%	81.88%	72.03%†	59.92%†	82.78%	82.81%
Channel	63.51%	68.66%	66.13%	71.75%	—	—	—	—	81.54%	81.57%†
ML-LSTM	63.84%	69.02%	62.94%†	64.73%†	82.05%	70.73%†	70.68%†	65.64%†	83.20%	83.30%
Sum	64.97%	71.03%	63.44%	69.08%	82.94%	82.47%	73.12%†	66.61%†	82.99%	83.05%
Weight	65.33%	71.33%	66.53%	73.26%	82.65%	82.19%	72.42%†	65.16%†	83.42%	83.57%
ETMC	65.67%	71.19%	65.86%	71.34%	83.69%	83.23%	73.96%	67.39%†	83.61%	83.69%
MSES	64.71%	72.52%	61.56%†	66.83%†	84.18%	83.60%	71.84%†	66.55%†	81.12%†	81.47%†
G-Blend	67.12%	71.39%	64.65%	68.54%	83.35%	82.71%	74.35%	68.69%	82.99%	83.05%
OGM	66.06%	71.44%	66.94%	71.73%	83.23%	82.66%	74.92%	68.74%	—	—
Greedy	66.52%	72.81%	66.64%	72.64%	—	—	—	—	82.74%	82.69%
DOMFN	66.25%	72.44%	67.34%	73.72%	83.56%	82.62%	74.45%	68.57%	—	—
MSLR	65.91%	71.96%	65.46%	71.38%	84.23%	83.69%	72.52%†	64.39%†	82.86%	82.92%
PMR	66.56%	71.93%	66.59%	70.36%	83.61%	82.49%	74.25%	68.62%	—	—
AGM	66.02%	72.52%	67.07%	73.58%	84.28%	83.44%	74.83%	69.11%	82.78%	82.82%
MLA	70.04%	74.13%	79.43%	85.72%	84.26%	83.48%	73.52%†	67.13%†	83.73%	83.87%
ReconBoost	70.85%	74.24%	74.84%	81.24%	84.37%	83.17%	74.42%	68.34%†	84.13%	**86.32%**
MMPareto	70.00%	**78.50%**	74.87%	75.15%	83.48%	82.84%	73.58%	†67.29%†	83.82%	84.24%
LFA-H	69.05%	72.97%	72.15%	80.45%	84.12%	83.98%	73.87%	69.17%	83.24%	83.87%
LFA-LB	**72.53%**	78.38%	**83.62%**	**90.06%**	**84.97%**	**84.57%**	**75.01%**	**70.57%**	**84.36%**	84.68%

Further Analysis for LFA

- Learning strategy and visualization:

Dataset	Modality	Constant			Stepwise				Dynamic	
		0	0.5	1	$h(0)$	$h(1)$	$h(0.05)$	$h(0.95)$	Ours-H	Ours-LB
KSounds	Multiple	64.55%	64.70%	28.67%	65.17%	66.92%	66.01%	67.41%	69.32%	72.89%
	Audio	49.17%	46.30%	34.11%	51.12%	52.34%	52.21%	53.41%	53.89%	54.32%
	Video	24.64%	44.02%	28.41%	41.21%	41.45%	42.31%	46.72%	49.18%	54.17%
CREMA-D	Multiple	63.31%	70.45%	26.49%	66.45%	70.24%	69.11%	71.45%	72.39%	84.11%
	Audio	55.65%	60.17%	33.15%	56.19%	57.38%	58.09%	60.18%	61.89%	65.13%
	Video	18.68%	42.54%	20.42%	45.14%	49.97%	46.41%	55.32%	57.14%	64.89%



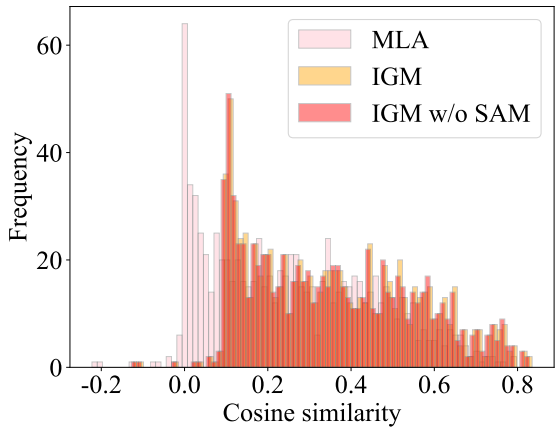
Interactive Multimodal Learning with Gradient Modification

Alternating Learning: Learning each modality one-by-one

Poor Plasticity:

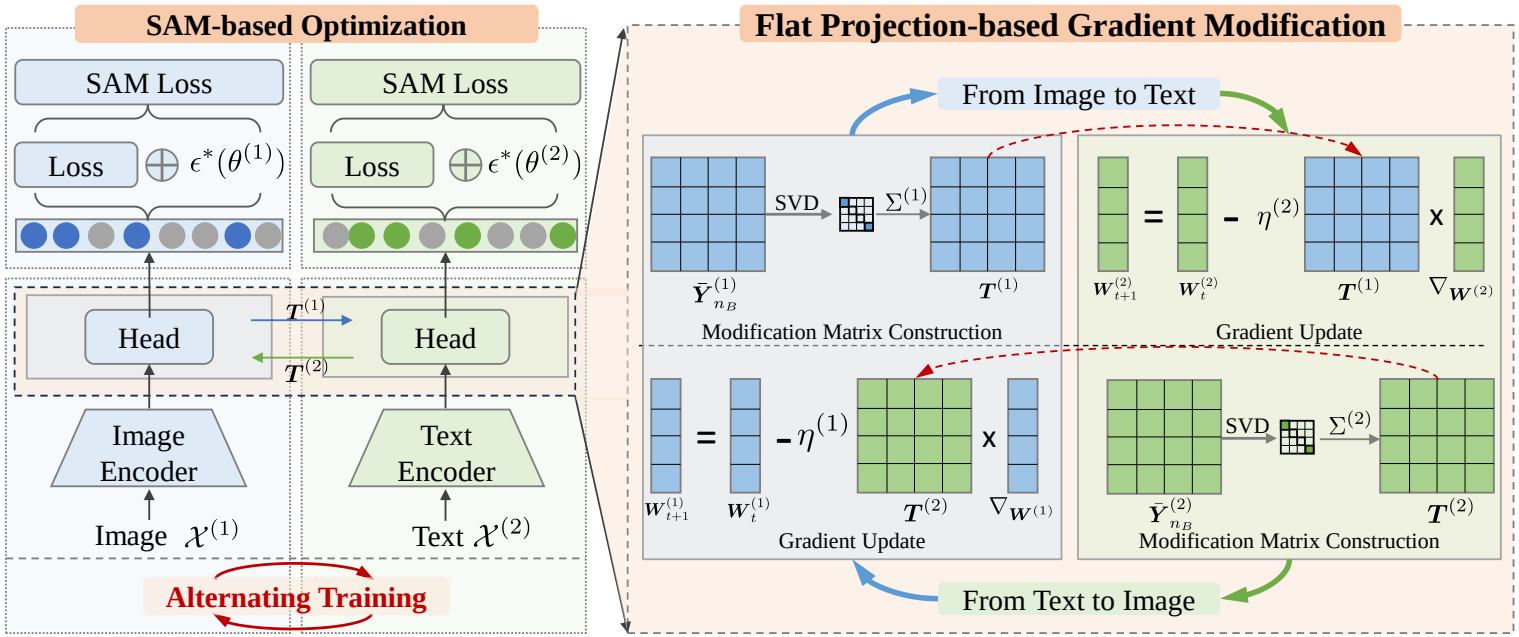
- The orthogonal projection suffers from **poor plasticity** problem, i.e., leading to **feasible gradient direction becomes narrow**.
- The poor plasticity problem results in **suboptimal solution**.

Gradient Change:



Interactive Multimodal Learning via Gradient Modification

- **Flat Projection Gradient Modification**: Project updating direction along **flat direction**.
- **SAM-based Optimization**: Smooth the objective function, enhancing the **flatness**.



- **Find Flat Direction**: $\boldsymbol{U}^v \Lambda^v [\boldsymbol{V}^v]^\top = svd(\boldsymbol{Y}^v)$
- **Relationship of Flatness and Singular Value**: Small $\lambda^v \mapsto$ Flat Area
- **Conduct Projection Matrix**: $\Sigma^v = \exp(-\frac{\tau}{\lambda^v_{\max} - \lambda^v_{\min}}(\Lambda^v - \lambda^v_{\min} \boldsymbol{I}))$
- **Project during Updating**:

$$\boldsymbol{T}^v = \boldsymbol{U}^v \Sigma^v [\boldsymbol{V}^v]^\top$$

$$\boldsymbol{\omega}^a_{t+1} = \boldsymbol{\omega}^a_t - \eta \cdot \boldsymbol{T}^v \cdot \nabla_{\boldsymbol{\omega}} \mathcal{L}$$

SAM-based Optimization

- **Perturb the Loss**: $\mathcal{L}^{\text{SAM}}(\omega) = \max_{\epsilon:||\epsilon||\leq\rho} \mathcal{L}(\omega+\epsilon)$.
- **Optimal perturbation**: $\epsilon^*(\omega) = \text{argmax}_{||\epsilon||\leq\rho} \mathcal{L}(\omega+\epsilon)$.
- **Gradient of SAM Loss**: $\nabla_\omega \mathcal{L}^{\text{SAM}}(\omega) = \nabla_\omega \mathcal{L}(\omega)|_{\omega+\epsilon^*(\omega)}$.

The IGM Algorithm

Updating rule: $\boldsymbol{\omega}_{t+1}^a = \boldsymbol{\omega}_t^a - \eta \cdot \boldsymbol{T}^v \cdot \nabla_{\omega^a} \mathcal{L}^{\text{SAM}}(\omega^a)$

Algorithm 4 Multimodal Learning with Label Fitting and Alignment

Input: Training set $\mathcal{D}$ and labels $\boldsymbol{Y}$;

Output: The learned parameters $\{\theta^{(j)}\}_{j=1}^{(m)}$;

Initialize gradient modification matrix, $\{\boldsymbol{T}^{(k)}\}_{j=1}^{(m)}$: $\forall k \in \{1, \cdots, m\}$, $\boldsymbol{T}^{(k)} = \boldsymbol{I}$;
for $i = 1 \to$ *Out_Iters* **do**
  for $j = 1 \to m$ **do**
    for $t = 1 \to$ *Inner_Iters* **do**
      Randomly construct a mini-batch $\mathcal{X}_t^{(j)}$.
      Calculate loss $L(\theta^{(j)})$ for data in $\mathcal{X}_t^{(j)}$.
      Calculate $\epsilon^*(\theta^{(j)})$ and $\nabla_{\theta^{(j)}} L^{\text{SAM}}$.
      Calculate modality index: $k = \texttt{mod}(j + m - 2, m) + 1$.
      Update $\theta^{(j)}$: $\theta_{t+1}^{(j)} = \theta_t^{(j)} - \eta^{(j)} \boldsymbol{T}^{(k)} \nabla_{\theta^{(j)}} L^{\text{SAM}}$.
    for $j = 1 \to n_B$ **do**
      Update cumulative variance.
    Update $\boldsymbol{T}^{(j)}$.

Experiments for IGM

- Main comparison with SOTA:

Method	CREMA-D		KSounds		Twitter2015		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1
Unimodal-1	.6317	.6861	.5312	.5669	.7367	.6849	.8136	.8065	.7822	.7833
Unimodal-2	.4583	.5879	.5462	.5837	.5863	.4333	.7181	.7073	.7863	.7865
Unimodal-3	-	-	-	-	-	-	-	-	.8154	.8183
OGR-GB	.6465	.6854†	.6710	.7139	.7435	.6869	.8335	.8271	.8299	.8305
OGM	.6694	.7173	.6606	.7144	**.7492**	.6874	.8323	.8266	-	-
DOMFN	.6734	.7372	.6625	.7244	.7445	.6857	.8356	.8262	-	-
MSES	.6156†	.6683†	.6471	.7063	.7184†	.6655†	.8418	.8360	.8112†	.8147†
PMR	.6659	.7030	.6656	.7193	.7425	.6860	.8360	.8249	-	-
AGM	.6707	.7358	.6602	.7252	.7483	.6911	.8402	.8344	.8278	.8282
MSLR	.6546	.7138	.6591	.7196	.7252†	.6439†	.8423	.8369	.8286	.8292
ReconBoost	.7484	.8124	.7085	.7424	.7442	.6834	.8437	.8317	.8413	.8632
SMV	.7872	.8417	.6900	.7426	.7428	.6817	.8418	.8368	.8352	.8341
MMPareto	.7487	.8535	.7000	.7850	.7358	.6729	.8348	.8284	.8382	.8424
MLA	.7943	.8572	.7004	.7413	.7352†	.6713†	.8426	.8348	.8373	.8387
IGM w/o SAM	.8026	.8830	.7159	.7623	.7395	.6912	.8455	.8390	.8487	.8634
IGM	**.8105**	**.8948**	**.7403**	**.7855**	.7489	**.6917**	**.8468**	**.8392**	**.8693**	**.8703**

Further Analysis of IGM

- Ablation study.

Ablation study on CREMA-D dataset.

SAM	GM	Audio	Video	Multi
✘	✘	45.83%	63.17%	64.52%
✔	✘	58.60%	64.79%	73.42%
✘	✔	60.13%	65.06%	80.26%
✔	✔	**61.16%**	**67.82%**	**81.05%**

Interactive enhancement analysis.

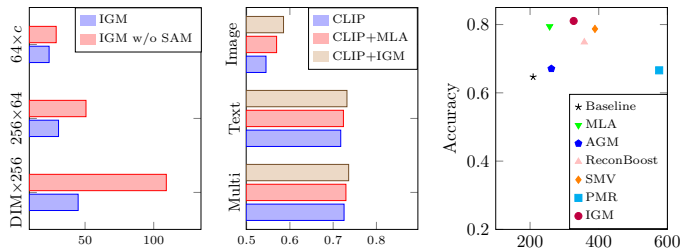
Method	*Initial*	*Out_Iters*=1		*Out_Iters*=2	
		Audio	Video	Audio	Video
w/o a-GM	.0325	.5312	.6803	.7231	.7482
w/o v-GM	.0325	.5312	.7023	.7472	.7646
IGM	.0325	.5312	.7023	**.7557**	**.8105**

- Impact of different scope of GM, singular values, the pretrained models, and training time.

Results with different scope of GM.

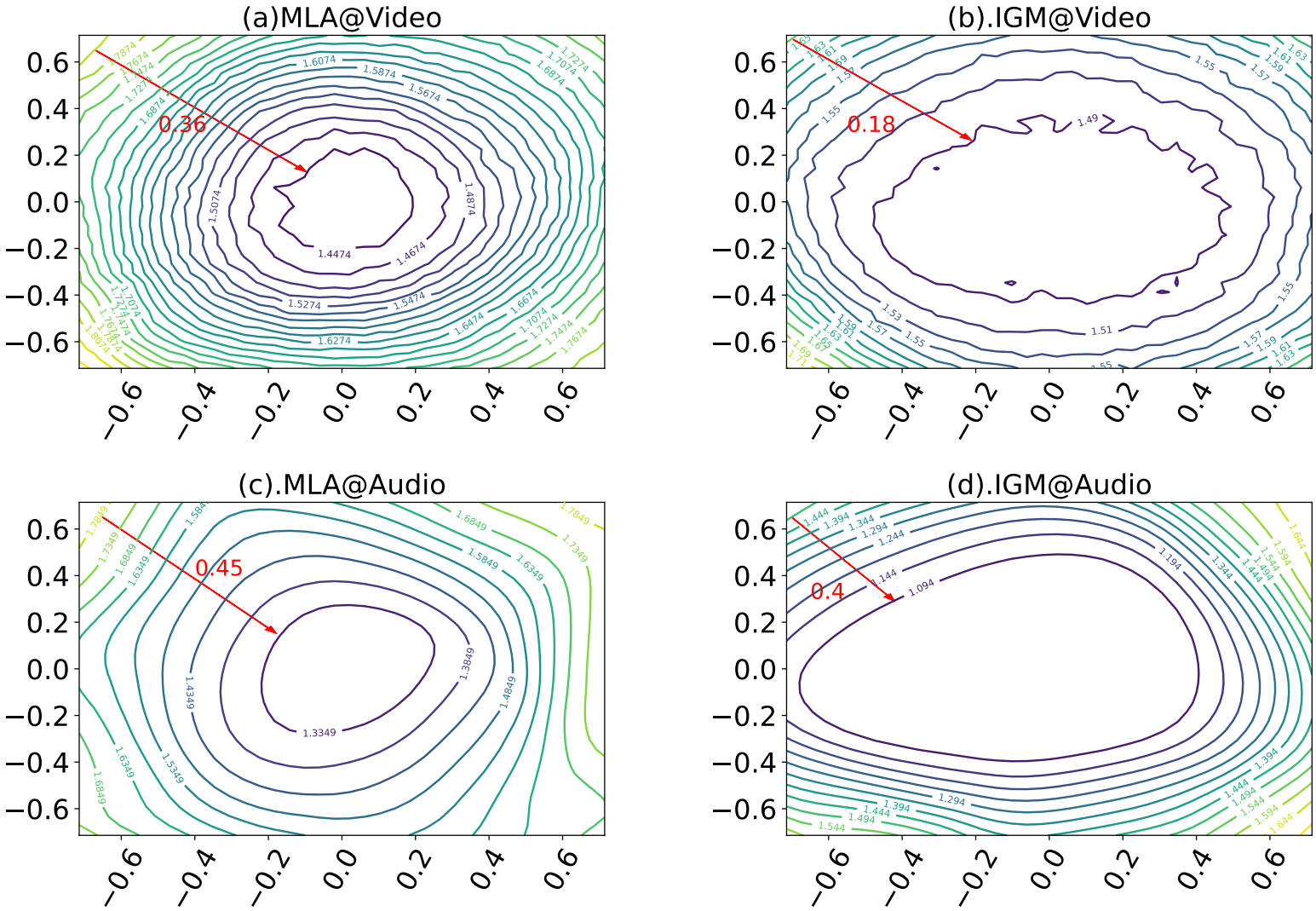
Scope of GM	Accuracy	MAP
100%	75.34%	81.23%
50%	78.97%	85.58%
30%	82.97%	90.15%
1.3% (Classification head)	81.05%	89.48%
0% (w/o GM)	73.42%	81.77%

Impact of singular values, CLIP, and training time.



(a). Singular values. (b). Pretrained Model. (c). Training time.

Loss Landscape Visualization



Instantaneous Probe-and-Rebalance Multimodal Learning

Existing Methods: Post-hoc Balance Learning After **"Imbalance"**.

Gradient Modulation-based methods:

- Adjust updating direction during backward.
- G-blend [2], OGM [3], CGGM [4].

Interaction-based methods:

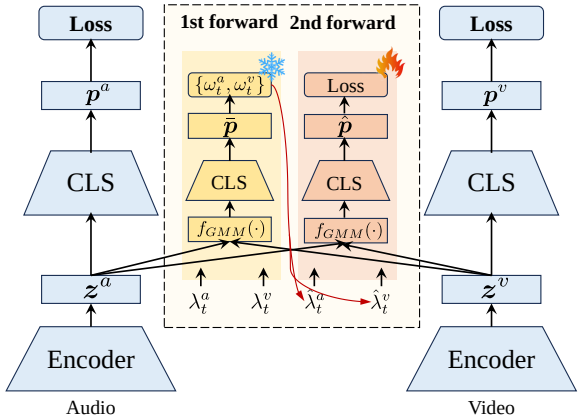
- enhance the interaction using features, logits after obtained.
- MLA [5], DI-MML [6], OLM [7]

Instantaneous Probe-and-Rebalance Multimodal Learning

IPRM: A Probe-then-Rebalance Strategy.

Two-pass forward strategy:

- Probes the strength of modality imbalance firstly.
- Facilitates multimodal learning under the balanced status secondly.



Instantaneous Probe-and-Rebalance Multimodal Learning

The 1st Forward Probing Phase:

- **Feature Extraction**: $\forall o \in \{a, v\},\ \boldsymbol{z}_i^o = g_o(\boldsymbol{x}_i^o; \Theta_o).$
- **Fusion with GMM**:
  $\bar{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \lambda_t^a),\quad \bar{\boldsymbol{p}}_i = \mathrm{softmax}(h(\bar{\boldsymbol{z}}_i)).$
- **Evaluate Balancing Instantaneous**:
  $\omega_t^a \triangleq \frac{\mathcal{D}_{\mathsf{KL}}(\mathcal{P}^v|\mathcal{P};\mathcal{T}_t)}{\mathcal{D}_{\mathsf{KL}}(\mathcal{P}^a|\mathcal{P};\mathcal{T}_t)+\mathcal{D}_{\mathsf{KL}}(\mathcal{P}^v|\mathcal{P};\mathcal{T}_t)}$

The 2nd Forward Rebalancing Phase:

- **Reassign Wight**: $\forall o \in \{a, v\}, \hat{\lambda}_t^o = \omega_t^o.$
- **Fusion with Balanced Status**:
  $\hat{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \hat{\lambda}_t^a),\quad \hat{\boldsymbol{p}}_i = \mathrm{softmax}(h(\hat{\boldsymbol{z}}_i)).$

The IPRM Algorithm

Algorithm 5 IPRM Algorithm

Input: Training set $\mathcal{D}$ and labels $\boldsymbol{Y}$;

Output: The learned parameters $\{\theta^{(j)}\}_{j=1}^{(m)}$;

1: Initialize parameters $\{\Theta_a, \Theta_v, \Phi, \Phi_a, \Phi_v\}$, maximum iterations M_t, learning rate η_α, modality weight $\omega_0^a = \omega_0^v = 0.5$, λ_1^a and λ_1^v;
2: **for** $t = 1 \mapsto M_t$ **do**
3: Sample a mini-batch data samples $\mathcal{T}_t$.
4: Calculate features based on $g_a(\cdot)$ and $g_v(\cdot)$.
5: Calculate fused feature $\bar{\boldsymbol{z}}$ the by the first forward phase.
6: Calculate the instantaneous strength score.
7: Calculate the balanced weight.
8: Calculate the prediction $\bar{\boldsymbol{p}}$ by the second forward phase.
9: Calculate the gradients based on backward phase.
10: Update the network parameters based on SGD.
11: Update λ_t^a and λ_t^v.

Experiments for IPRM

- Main comparison with SOTA:

Dataset	Metric	Unimodal			Naive Fusion			IPRM
		A/A/R/A/I	V/V/O/V/T	D/T	Concat	Sum	Weight	
CREMA-D	Accuracy	63.17%	45.83%	N/A	63.61%	63.44%	66.53%	**84.27%**
	MAP	68.61%	58.79%	N/A	68.41%†	69.08%	71.34%	**90.66%**
KSounds	Accuracy	54.12%	55.62%	N/A	64.55%	64.90%	65.33%	**74.37%**
	MAP	56.69%	58.37%	N/A	71.30%	71.03%	71.10%	**80.63%**
NVGesture	Accuracy	78.22%	78.63%	81.54%	82.37%	80.50%†	78.42%†	**85.89%**
	Macro-F1	78.33%	78.65%	81.83%	82.70%	80.67%†	79.39%†	**86.34%**
IEMOCAP	Accuracy	58.45%	30.71%	70.55%	75.97%	76.06%	69.29%†	**80.22%**
	Macro-F1	58.29%	11.75%	69.93%	75.88%	76.03%	68.91%†	**80.63%**
Sarcasm	Accuracy	71.81%	81.36%	N/A	82.86%	82.94%	82.65%	**85.14%**
	Macro-F1	70.73%	80.56%	N/A	82.40%	82.47%	82.19%	**84.41%**

Experiments for IPRM (cont'd)

- Main comparison with SOTA:

Dataset	Metric	OGR-GB	MSLR	OGM	AGM	MMPareto	ReconBoost	MLA	LFM	IPRM
CREMA-D	Accuracy	64.65%	68.68%	66.12%	67.33%	74.87%	75.57%	79.43%	83.62%	**84.27%**
	MAP	73.92%	74.12%	73.72%	78.07%	85.35%	81.40%	85.72%	90.06%	**90.66%**
KSounds	Accuracy	67.22%	67.56%	65.82%	67.91%	70.00%	68.55%	70.04%	72.53%	**74.37%**
	MAP	72.74%	72.82%	71.59%	73.88%	78.50%	76.62%	79.45%	78.97%	**80.63%**
NVGesture	Accuracy	82.99%	82.37%	N/A	82.79%	83.82%	83.86%	83.40%	84.36%	**85.89%**
	Macro-F1	83.05%	82.84%	N/A	82.84%	84.24%	84.34%	83.72%	84.68%	**86.34%**
IEMOCAP	Accuracy	70.10%	76.69%	N/A	77.51%	77.69%	76.87%	79.31%	78.41%	**80.22%**
	Macro-F1	69.90%	76.77%	N/A	77.29%	77.89%	77.08%	79.73%	78.51%	**80.63%**
Sarcasm	Accuracy	82.86%	84.39%	83.60%	83.06%	83.48%	84.37%	84.26%	84.97%	**85.14%**
	Macro-F1	82.15%	83.78%	82.93%	82.93%	82.84%	83.17%	83.48%	**84.57%**	84.41%

Further Analysis

- Ablation Study:

Dataset	w/ L-Mixup	w/o EMA	One-Pass	IPRM
CREMA-D	75.53%	83.06%	83.47%	**84.27%**
KSounds	71.94%	73.91%	73.64%	**74.37%**
NVGesture	84.85%	85.27%	84.44%	**85.89%**
IEMOCAP	75.79%	78.05%	77.60%	**80.22%**
Sarcasm	84.52%	84.81%	84.10%	**85.14%**

Summary

Data-Level Modality Imbalance:

- Zhi-Hao Guan, **Qing-Yuan Jiang***, Yang Yang. Balance-aware Sequence Sampling Makes Multimodal Learning Better. *IJCAI*, 2025.

Model-Level Modality Imbalance:

- **Qing-Yuan Jiang**, Longfei Huang, Yang Yang. Rethinking Multimodal Learning from the Perspective of Mitigating Classification Ability Disproportion. *NeurIPS*, 2025 (Oral, 1.46%).
- Yang Yang, Hongpeng Pan, **Qing-Yuan Jiang**, Yi Xu, and Jinhui Tang. Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks. *TPAMI*, 2025.

Learning-Level Modality Imbalance:

- Yang Yang, Fengqiang Wan, **Qing-Yuan Jiang***, Yi Xu. Facilitating Multimodal Classification via Dynamically Learning Modality Gap. *NeurIPS*, 2024.
- **Qing-Yuan Jiang**, Zhouyang Chi, Yang Yang. Interactive Multimodal Learning via Flat Gradient Modification. *IJCAI*, 2025.
- Yang Yang, Xixian Wu, **Qing-Yuan Jiang***. Towards Equilibrium: An Instantaneous Probe-and-Rebalance Multimodal Learning Approach. *IJCAI*, 2025.

Take Home Message

Robust multimodal representation learning:

- Key Points of Efficiently Mining Multimodal Information for Large Models
- Multi-dimensional Challenges: data-, model-, and learning-level

Our contributions: A Series of Multidimensional RMML Solutions

- Data-Level: sampling sequence
- Model-Level: weak modality capability enhancement; strong modality masking
- Learning-Level: fusion and alignment simultaneously; instantaneous probe-and-rebalance fusion

Thanks
Q&A

References I

[1] Y. Wei, D. Hu, H. Du, and J. Wen, "On-the-fly modulation for balanced multimodal learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 469–485, 2025.

[2] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" In *CVPR*, 2020, pp. 12 692–12 702.

[3] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8228–8237.

[4] Z. Guo, T. Jin, J. Chen, and Z. Zhao, "Classifier-guided gradient modulation for enhanced multimodal learning," in *NeurIPS*, 2024.

References II

[5] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," in *CVPR*, 2024, pp. 27 446–27 456.

[6] Y. Fan, W. Xu, H. Wang, J. Liu, and S. Guo, "Detached and interactive multimodal learning," in *ACM MM*, 2024, pp. 5470–5478.

[7] D. Zong, C. Ding, B. Li, J. Li, and K. Zheng, "Balancing multimodal learning via online logit modulation," in *IJCAI*, 2024, pp. 5753–5761.