

Balance-aware Sequence Sampling Makes Multimodal Learning Better

Zhi-Hao Guan, Qing-Yuan Jiang*, Yang Yang*

Nanjing University of Science and Technology, Nanjing, China

2025.08.30

Background

Method

Experiments

Conclusion

Background

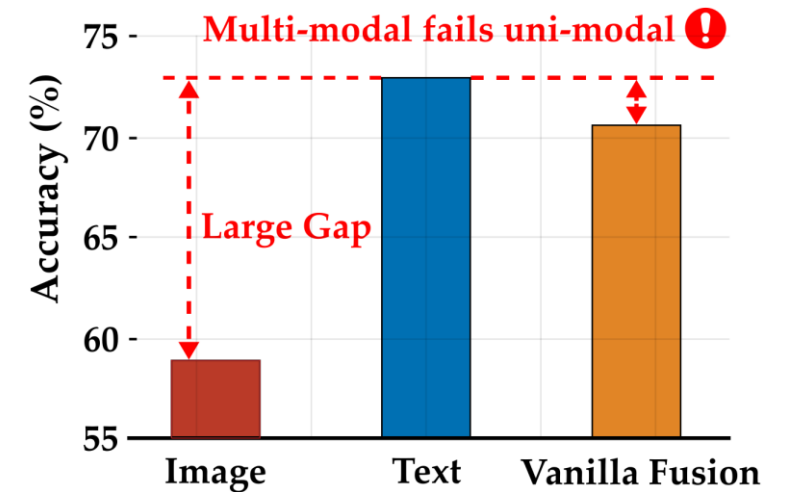
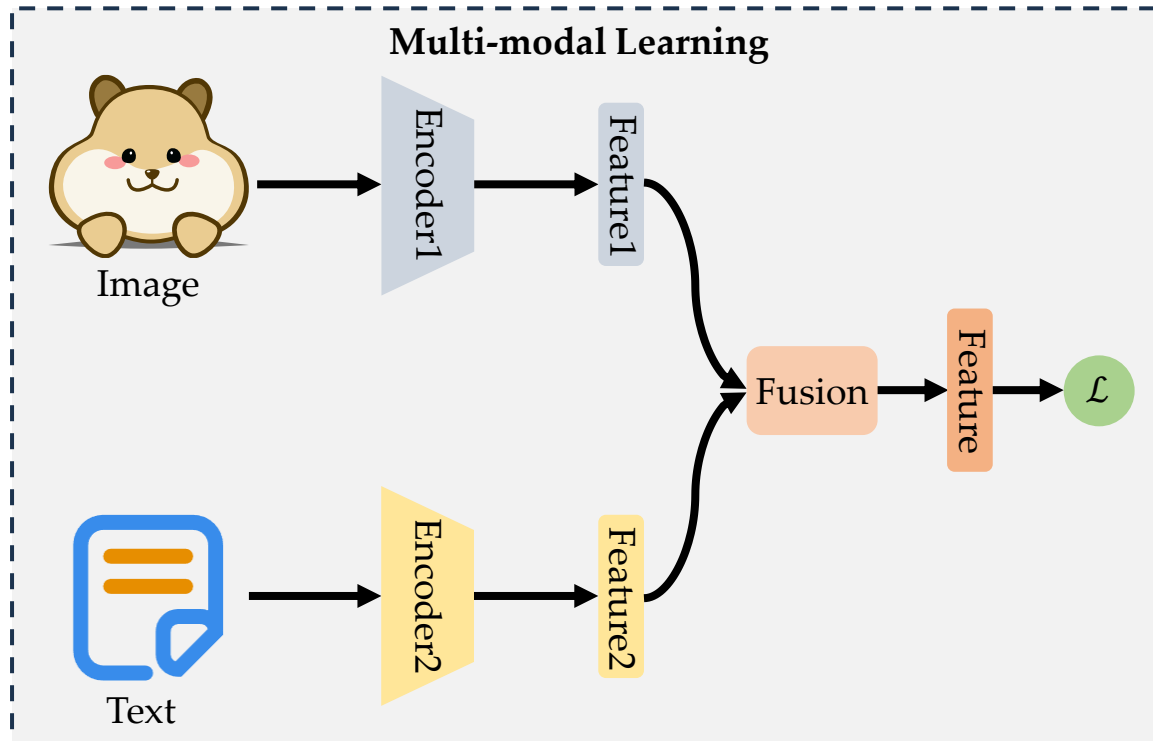
Method

Experiments

Conclusion

□ Modality Imbalance

- Due to **modality heterogeneity**, multi-modal learning (MML) is often dominated by stronger modalities, which leads to insufficient learning of weaker ones and ultimately suboptimal overall performance.

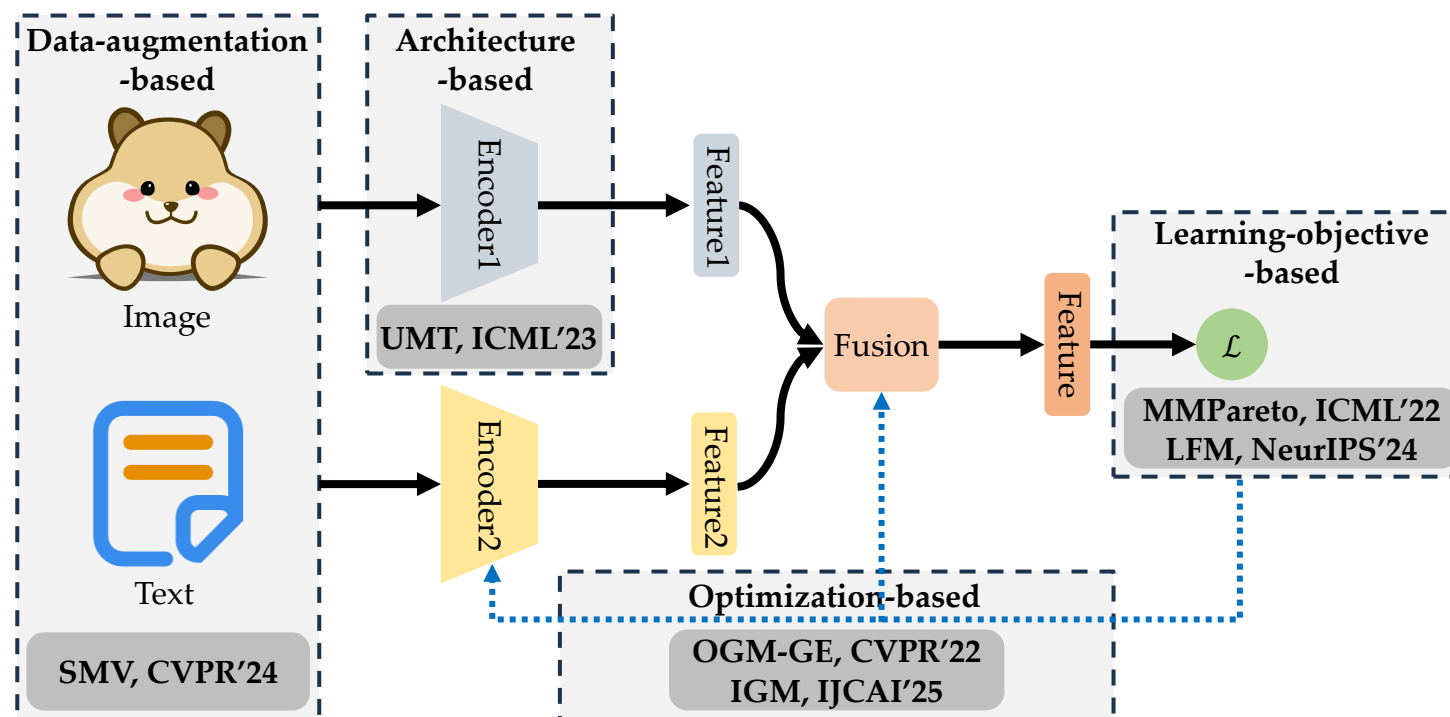


Modality Rebalance Method

- Learning-objective-based: **MMPareto**_[ICML'22] , **LFM**_[NeurIPS'24]
- Optimization-based: **OGM-GE**_[CVPR'22] , **IGM**_[IJCAI'25]
- Architecture-based: **UMT**_[ICML'23]
- Data-augmentation-based: **SMV**_[CVPR'24]

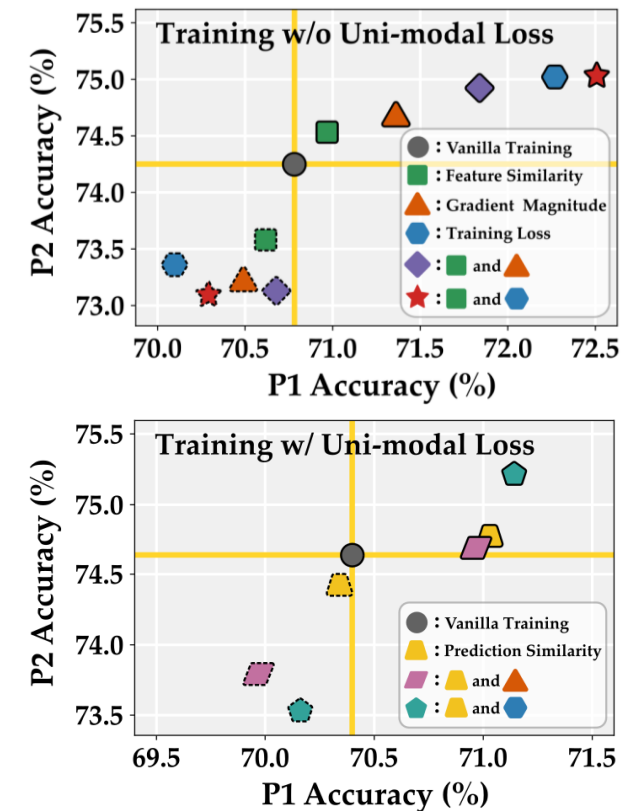
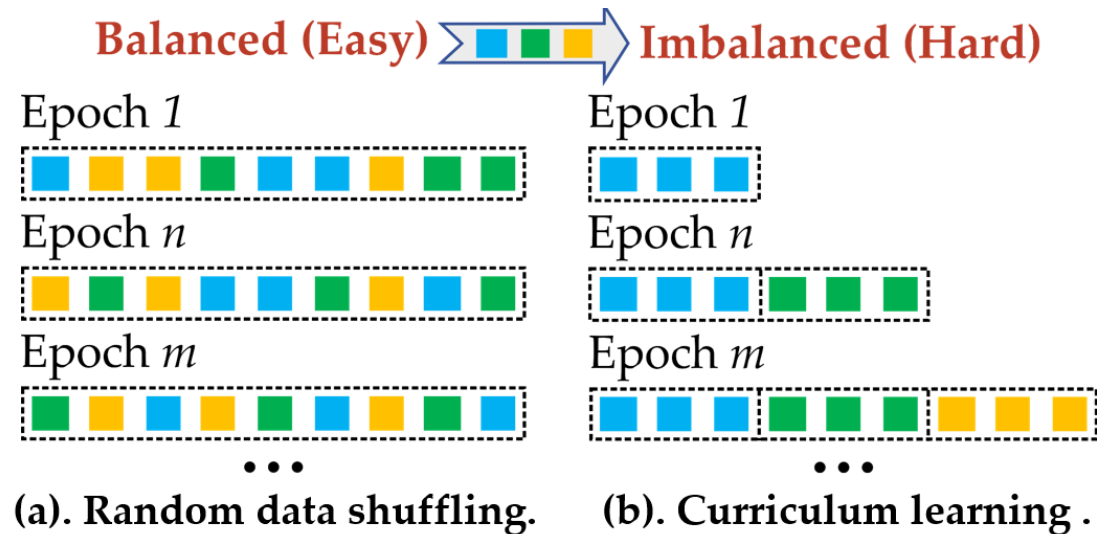


Outstanding Performance!



□ Motivation

- Although existing methods have shown promising results, they generally overlook a key aspect: **MML can be highly sensitive to the training sequence.**



CL effectively boosts MML performance, whereas anti-CL degrades it!

Background

Method

Experiments

Conclusion

□ Multi-perspective Measurer

- To construct well-structured training sequences that address modality imbalance, we first measure the **balance degree** of a multimodal sample from the following perspectives:

◆ Correlation Criterion

$$\text{sim}(\mathbf{x}_i^{(u)}, \mathbf{x}_i^{(v)}) = \frac{[\hat{\mathbf{y}}_i^{(u)}]^\top \hat{\mathbf{y}}_i^{(v)}}{\|\hat{\mathbf{y}}_i^{(u)}\|_2 \|\hat{\mathbf{y}}_i^{(v)}\|_2}.$$

◆ Information Criterion

$$\ell_{total} = \ell_{multi}^{ce}(\mathbf{x}_i, \mathbf{y}_i) + \sum_{j \in \{u, v\}} \ell_{uni}^{ce}(\mathbf{x}_i^{(j)}, \mathbf{y}_i).$$

◆ Balance Score

The balance score of a sample $\mathbf{x}_i = \{\mathbf{x}_i^{(u)}, \mathbf{x}_i^{(v)}\}$ can be formulated as:

$$s(\mathbf{x}_i) = \frac{\text{sim}(\mathbf{x}_i^{(u)}, \mathbf{x}_i^{(v)}) - \min(\mathcal{S})}{\max(\mathcal{S}) - \min(\mathcal{S})} - \frac{\ell_{total}(\mathbf{x}_i^{(u)}, \mathbf{x}_i^{(v)}, \mathbf{y}_i) - \min(\mathcal{L})}{\max(\mathcal{L}) - \min(\mathcal{L})}.$$

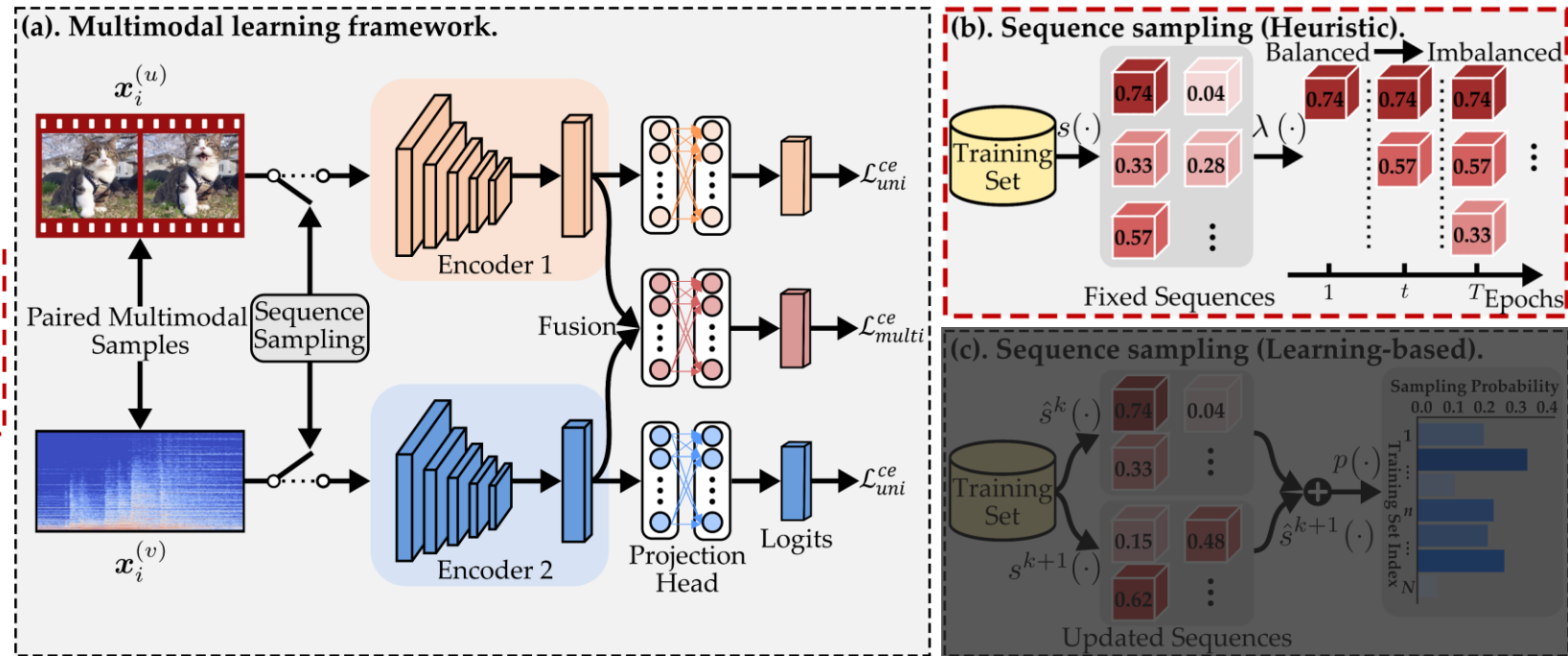
Training Scheduler

- After evaluating the balance score of each sample, we then proceed to control the **presentation order** of training data from balanced to imbalanced samples:

Heuristic Scheduler

$$\lambda(t) = \min \left(1, \sqrt{\frac{1 - \lambda_0^2}{T_{grow}} \cdot t + \lambda_0^2} \right).$$

$\lambda(t)$ maps the training epoch t to an interval $\lambda \in (0, 1]$.



At epoch t , the current batch data X_{batch} is randomly sampled from the top λ proportion of the training data in the entire ranked sequence X_{rank} :

$$X_{batch}(t) = \text{Sampling}(\{x_i | x_i \in X_{rank}, i < \lfloor n \cdot \lambda(t) \rfloor\}).$$

Training Scheduler

Learning-based Scheduler

We update the balance score in a certain epoch E . The $k + 1$ -th balance score can be denoted as:

$$\hat{s}^{k+1}(\mathbf{x}_i) = \begin{cases} s^{k+1}(\mathbf{x}_i), & \text{if } k = 0, \\ (1 - \beta)\hat{s}^k(\mathbf{x}_i) + \beta s^{k+1}(\mathbf{x}_i), & \text{otherwise,} \end{cases}$$

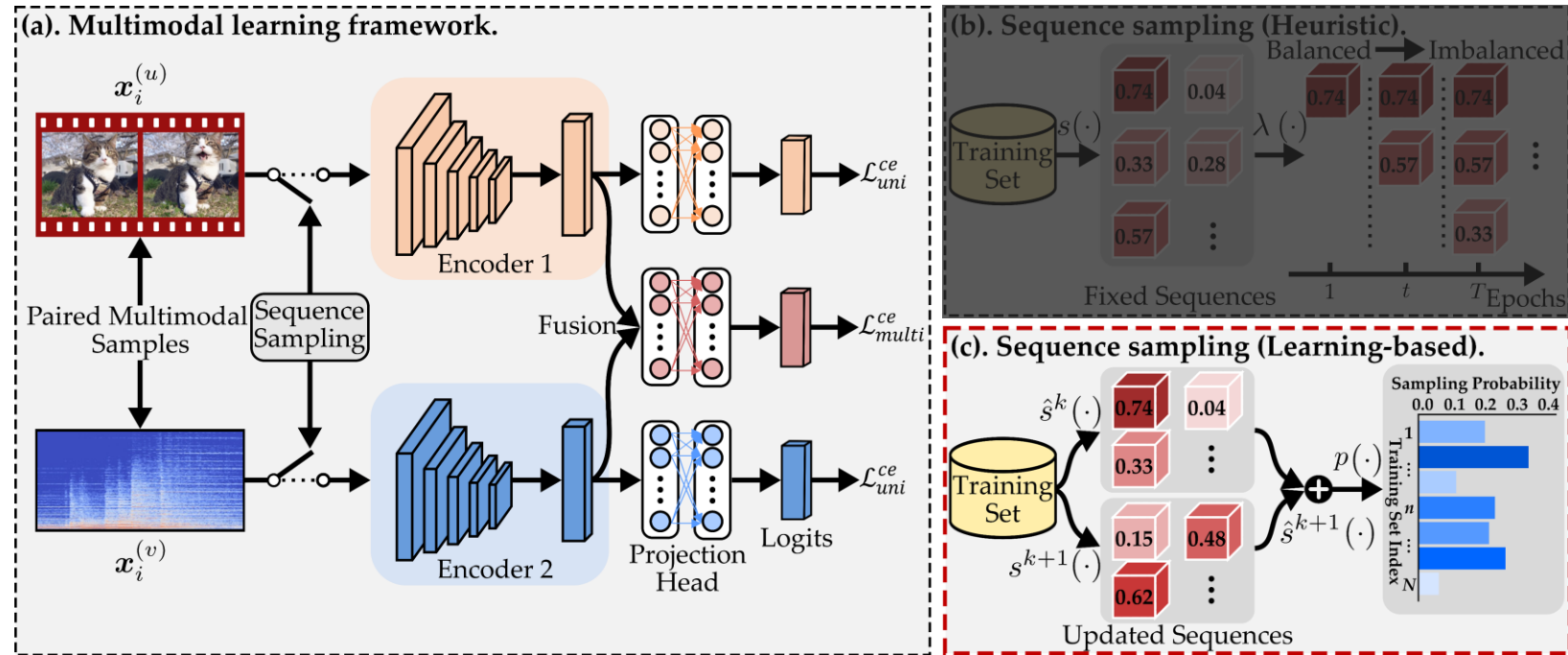
where $k = \lfloor t/E \rfloor$, t denotes the t -th epoch.

The sampling probability for each data point \mathbf{x}_i can be denoted as:

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

Finally, \mathbf{x}_i is sampled with p to construct the current batch data \mathbf{X}_{batch} :

$$\mathbf{X}_{batch}(t) = \text{Sampling}(\{p(\mathbf{x}_1), p(\mathbf{x}_2), \dots, p(\mathbf{x}_n)\}).$$



Background

Method

Experiments

Conclusion

Classification Results

Table 1: Comparison with SOTA multimodal learning methods. The best performances are highlighted in bold, and the second best is underlined. Higher ACC, MAP, or F1 scores indicate better performance.

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	<u>86.32</u>
MMPareto	74.87	85.35	70.00	78.50	73.58	67.29	83.48	82.48	83.82	<u>84.24</u>
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	<u>79.13</u>	<u>75.12</u>	<u>69.23</u>	84.35	83.77	84.64	84.94
BSS-H	<u>80.78</u>	<u>87.86</u>	<u>72.67</u>	78.61	74.73	68.67	<u>84.41</u>	<u>83.86</u>	<u>85.06</u>	85.15
BSS-L	82.80	88.61	73.95	79.43	75.22	69.51	85.01	84.62	86.72	87.04

Table 2: Performances on the VGGSound dataset.

Method	ACC (%)	MAP (%)
OGM	48.29	49.78
AGM	47.11	51.98
ReconBoost	50.97	53.87
MMPareto	51.25	54.73
SMV	50.31	53.62
MLA	<u>51.65</u>	54.73
BSS-H	51.61	<u>55.68</u>
BSS-L	52.80	56.61



Our method achieves **SOTA performance** across various datasets.

Further Analysis

- Both "PreSim" and "Loss", can boost classification performance.
- BSS is not sensitive to hyperparameters.
- BSS is robust to the large pre-trained model.

Table 3: Ablation study on the Kinetics-Sounds dataset under the learning-based setting.

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

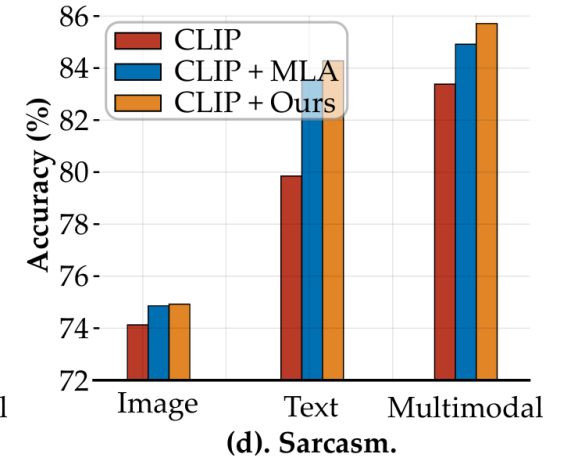
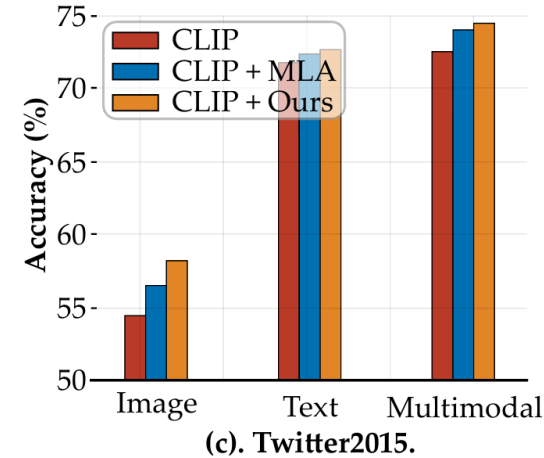
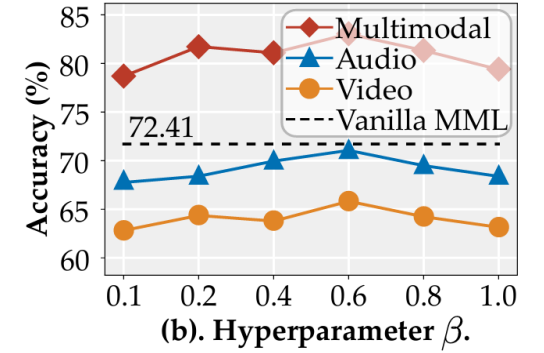
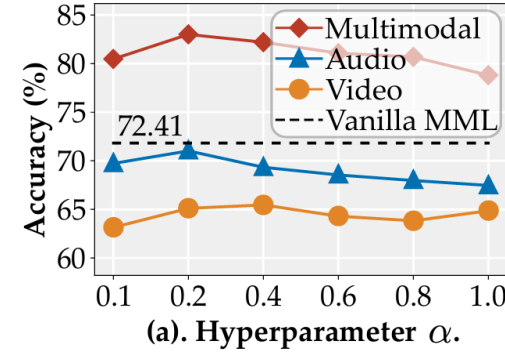








Figure 3: (a). and (b). Sensitivity to hyperparameters α and β on the CREMA-D dataset. (c). and (d). Robust performance achieved by using the CLIP pre-trained model as encoders.

Case Study

 <p>Congratulations to South Greene freshman Taylor Lamb for earning state honors from the TSWA.</p> <p>Label: Positive Balance Score: 0.8614 Balanced Sample</p>	 <p>Food event weekend include Taste of Toronto, Summerlicious amp free tacos!</p> <p>Label: Positive Balance Score: 0.7850 Balanced Sample</p>	 <p>Allegiant flights cancelled, delayed in Orlando.</p> <p>Label: Negative Balance Score: 0.6502 Semi-balanced Sample</p>
 <p>Lots of fun judging Santa Parade entries, and riding in parade afterward. Thanks KCBIA kamloops.</p> <p>Label: Positive Balance Score: 0.5267 Semi-balanced Sample</p>	 <p>At Costa Coffee in Edinburgh. Great coffee, great view, no WiFi.</p> <p>Label: Neutral Balance Score: 0.1623 Imbalanced Sample</p>	 <p>Helena Bonham Carter and Time Burton have split after 13 years.</p> <p>Label: Negative Balance Score: 0.1278 Imbalanced Sample</p>

(a). Sample with different balance scores

 <p>Once the 2020 Olympics are over, Tokyo faces a bleak future.</p> <p>Label: Negative Balance Score: 0.5302 1st Evaluation</p>	 <p>Once the 2020 Olympics are over, Tokyo faces a bleak future.</p> <p>Label: Negative Balance Score: 0.5873 2nd Evaluation</p>	 <p>Once the 2020 Olympics are over, Tokyo faces a bleak future.</p> <p>Label: Negative Balance Score: 0.6165 3rd Evaluation</p>
---	--	---

(b). Dynamic variation of sample scores

Figure C1: Qualitative results of sample evaluation. (a). Some representative samples selected from different segments on the Twitter2015 dataset, designated as balanced, semi-balanced, and imbalanced. (b). Dynamic variation of sample scores with the learning-based scheduler.

Background

Method

Experiments

Conclusion

□ Contributions

- We **highlight the critical role of training sequences** in addressing modality imbalance, and show that well-structured sequences can significantly improve MML performance.
- We **define a multi-perspective measurer** to quantify the balance degree of each sample. Based on the resulting balance scores, we then **propose both a heuristic and a learning-based sampling method** to adjust the training sequences.

□ Future Work

- Expect to extend BSS to other downstream tasks, such as cross-modal retrieval.
- ...

Thank you for your listening!



KMG Group



WeChat

