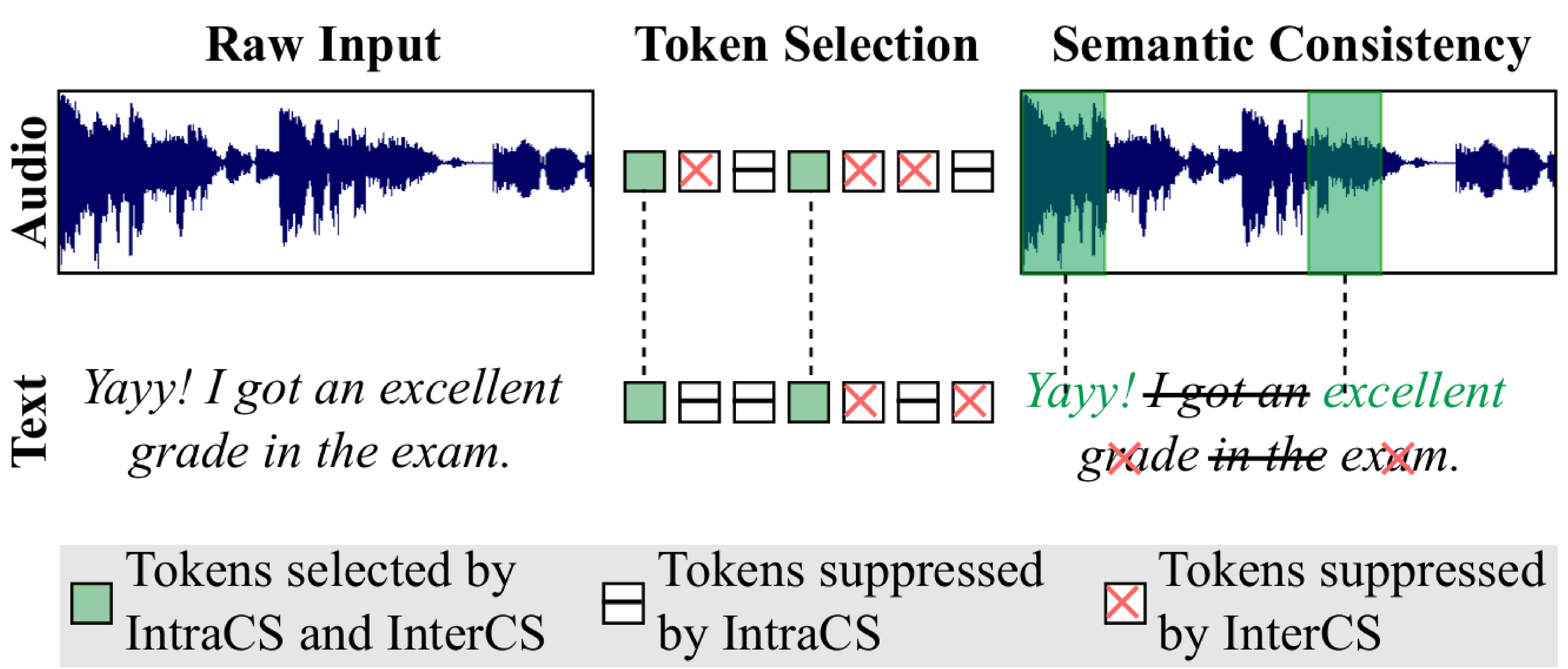
ConsMSA: Semantic Distribution Consistency Learning for Multimodal Sentiment Analysis

Pan Wang[1,2] Lipeng Ke[2] Huajun Ying[2] Pritish Mohapatra[2] Rohan Sarkar[2] Suresh Lakhani[2] Sankar Venkataraman[2] Jingtong Hu[1]

[1]University of Pittsburgh [2]Amazon

① Motivation: Multimodal Sentiment Analysis (MSA)



- Raw multimodal sequences suffer from semantic inconsistency:
  - **Intra-modality:** redundancy or neutral signals within a modality
  - **Inter-modality:** conflicts across modalities (e.g., sarcasm)
- We explicitly formalize semantic distribution consistency for robust and efficient MSA.

② Key Contribution

1. Holistic Consistency Paradigm
Formalizes intra- & inter-modality semantic distribution consistency for multimodal sentiment analysis.

2. I²CS Metric (via JS Divergence)
Information-theoretic metric quantifying semantic agreement within and across modalities.

3. Consistency-aware Importance
Couples predictive relevance with I²CS: rewards tokens that are both informative and coherent.

4. Three Synergistic Mechanisms
(i) Consistency regularization (ii) Adaptive token reweighting (iii) Token pruning criterion.

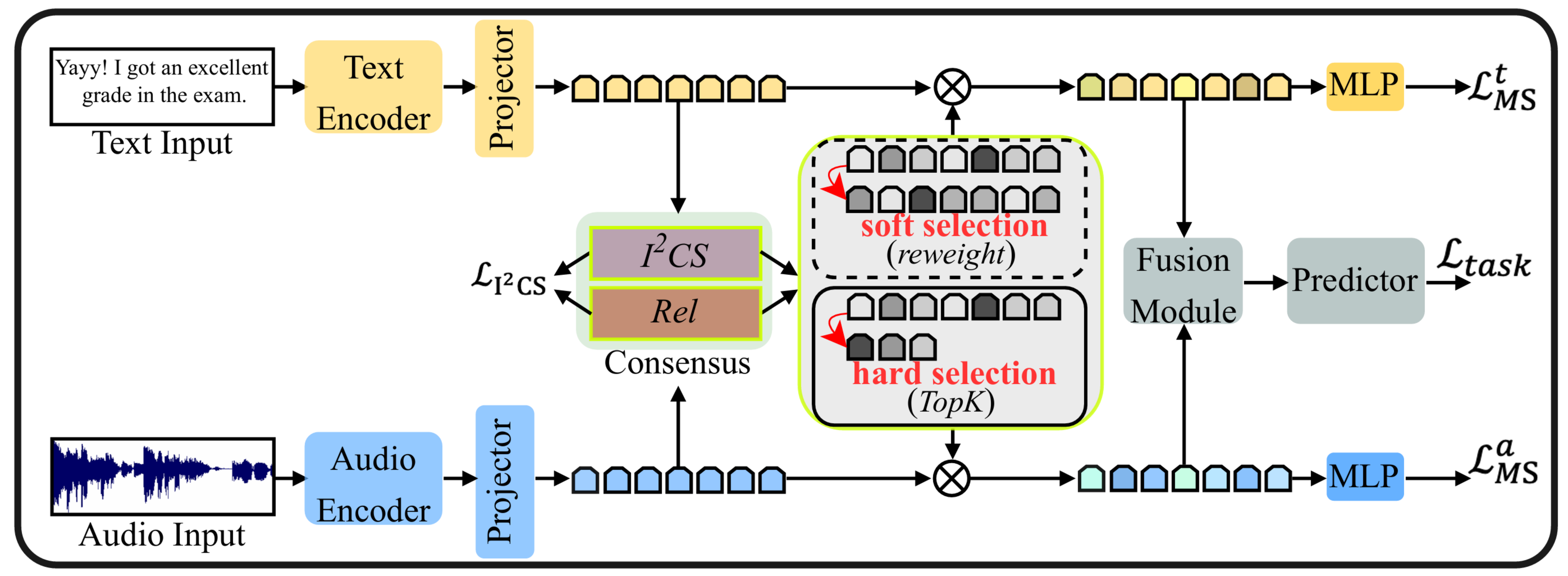
5. The State-of-the-art Performance
Our ConsMSA achieves the SOTA performance on three MSA benchmarks, including MOSEI, MOSI, and CH-SIMS.

Contact Information

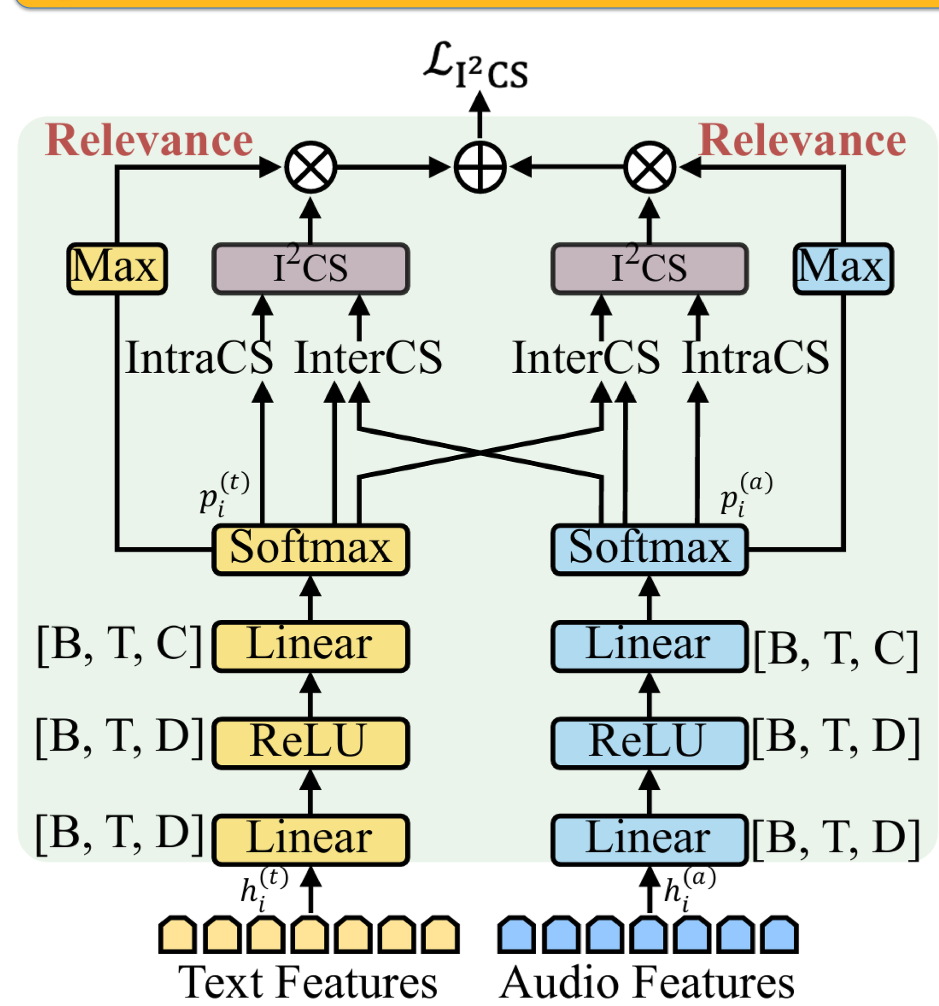
Email: pan.wang@pitt.edu

(Wechat) (Linkedin)

③ Proposed Framework



④ I²CS: Consistency Score



Intra-modality:

$$\mathrm{IntraCS}(h_i) = \frac{1}{T-1}\sum_{j\neq i} \mathrm{JS}(p_i \| p_j)$$

Inter-modality:

$$\mathrm{InterCS}(h_i) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T}\sum_{j=1}^{T} \mathrm{JS}(p_i^{\mathrm{src}} \| p_j^{\mathrm{align}})$$

Combined:

$$I^2CS(h_i) = \alpha \cdot \mathrm{IntraCS}(h_i) + \beta \cdot \mathrm{InterCS}(h_i)$$

Importance:

$$\mathrm{Signal}(h_i) = Rel(h_i) - I^2CS(h_i)$$

⑤ Token Optimization

Soft Selection

$$w_i = \sigma(\gamma \cdot \mathrm{Signal}(h_i)), \quad w_i \in (0,1)$$

Hard Selection

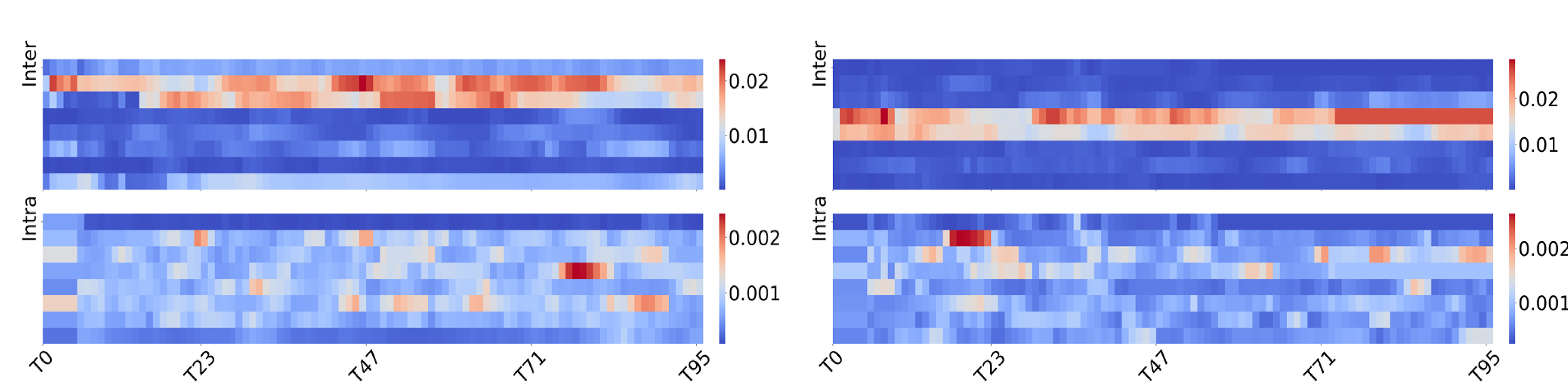
$$m_i = TopK[w_i]$$

Joint Learning Objective

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{task}} + \lambda_2 \mathcal{L}_{I^2CS} + \lambda_3 \mathcal{L}_{\text{MS}}$$

Task accuracy + Consistency regularization + Modality-specific Loss

⑥ Visualization of Token Consistency



(a) Before ConsMSA (b) After ConsMSA

⑦ Main Results

Table 1. Comparison of our method with prior SOTA methods on MOSI and MOSEI datasets. The best performance is highlighted in bold. T: Text, A: Audio, and V: Video. Note: † denotes the result from THUIAR's GitHub page, * denotes the result from (Hazarika et al., 2020), - denotes that the result from the original paper is not provided, and others are all from the original papers. For fair comparison, the BERT-based models utilize T+A+V following DLF, while the RoBERTa-based model utilizes T+A following MMML.

Method	Modality	CMU-MOSI				CMU-MOSEI			
		Acc-7(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Acc-7(↑)	Acc-2(↑)	F1(↑)	MAE(↓)
BERT-based									
TFN*	T+A+V	34.90	80.08	80.07	0.901	50.20	82.50	82.10	0.593
LMF*	T+A+V	33.20	82.50	82.40	0.917	48.00	82.00	82.10	0.623
EF-LSTM†	T+A+V	35.39	78.48	78.51	0.949	50.01	80.79	80.67	0.601
LF-DNN†	T+A+V	34.52	78.63	78.63	0.955	50.83	82.74	82.52	0.580
MFN†	T+A+V	35.83	78.87	78.90	0.927	51.34	82.85	82.85	0.575
Graph-MFN†	T+A+V	34.64	78.35	78.35	0.956	51.37	83.48	83.43	0.575
MulT	T+A+V	40.00	83.00	82.00	0.871	51.80	82.50	82.30	0.580
PMR	T+A+V	40.60	83.60	83.60	-	52.50	83.60	83.40	-
MISA†	T+A+V	41.37	83.54	83.58	0.777	52.05	84.67	84.66	0.558
MAG-BERT	T+A+V	43.62	84.43	84.61	0.727	52.67	84.82	84.71	0.543
FDMER	T+A+V	44.10	84.60	84.70	0.724	54.10	86.10	85.80	0.536
DMD	T+A+V	45.60	86.00	86.00	-	54.50	86.60	86.60	-
MMIM	T+A+V	46.65	86.06	85.98	0.700	54.24	85.97	85.94	0.526
Self-MM†	T+A+V	46.67	85.98	85.95	0.713	53.87	85.17	85.30	0.530
DLF	T+A+V	47.08	85.06	85.04	0.731	53.90	85.42	85.27	0.536
ULMD	T+A+V	47.81	85.82	85.71	0.700	53.81	85.95	85.91	0.531
EMOE	T+A+V	47.70	85.40	85.40	0.710	54.10	85.30	85.30	0.536
MDF	T+A+V	**48.25**	85.82	85.90	**0.692**	54.80	86.61	86.41	0.525
ConsMSA (Ours)	T+A+V	47.96	**86.06**	**86.06**	0.698	**54.92**	**86.67**	**86.65**	**0.525**
RoBERTa-based									
MMML	T+A	50.34	89.69	89.67	0.580	55.74	88.02	88.15	0.492
ConsMSA (Ours)	T+A	**51.46**	**90.09**	**90.08**	**0.557**	**56.00**	**88.41**	**88.39**	**0.485**

Table 2. Comparison on CH-SIMS dataset.

Method	Modality	Acc-2 ↑	Acc-3 ↑	Acc-5 ↑	F1 ↑	MAE ↓
MMML	T+A	82.93	69.37	49.38	82.90	0.332
Ours	**T+A**	**83.59**	**70.68**	**50.77**	**83.64**	**0.331**

⑧ Further Analysis

Table 3. Results of ablation studies on the MOSI benchmark.

Method	Acc-7(↑)	Acc-2(↑)	F1(↑)	MAE(↓)
ConsMSA (Ours)	**51.46**	**90.09**	**90.08**	**0.557**
w/o IntraCS	47.52	87.20	87.21	0.629
w/o InterCS	46.79	88.87	88.83	0.626
w/o *Rel* in Eq. (6)	45.48	87.96	87.99	0.668
w/o *Rel* in Eq. (9)	46.36	87.80	87.82	0.647
w/o Gating	45.77	87.20	87.21	0.653
w/o $\mathcal{L}_{MS}$	47.38	87.50	87.51	0.643
w/o $\mathcal{L}_{I^2CS}$	48.83	86.59	86.62	0.634

Table 4. Different token retention ratios on CH-SIMS and MOSI.

Ratio	CH-SIMS (Acc2/F1/MAE)	MOSI (Acc2/F1/MAE)
1.0	83.59/83.64/**0.331**	90.09/90.08/0.557
0.8	82.06/82.25/0.341	**90.85/90.84/0.553**
0.7	83.15/83.20/0.347	89.33/89.34/0.562
0.6	**83.59/83.70**/0.337	89.63/89.66/0.585
0.5	82.71/83.00/0.337	88.72/88.75/0.585
0.3	81.84/82.19/0.333	89.33/89.31/0.567
0.1	81.18/81.32/0.355	89.18/89.17/0.576

Key Takeaways:

Semantic distribution consistency learning is an efficient and effective strategy to mitigate the multimodal redundancy and conflicts, especially for MSA tasks.