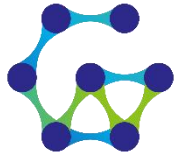
中国人民大學
RENMIN UNIVERSITY OF CHINA



GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

RollingQ: Reviving the Cooperation Dynamics in Multimodal Transformer

GeWu-lab

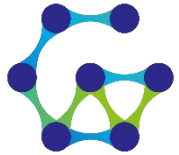
haotian_ni@buaa.edu.cn

hangliu@xmu.edu.cn

{yakewei; dihu}@ruc.edu.cn

2025-06-15

Content



GeWu-Lab

Gaoling School of Artificial Intelligence
Renmin University of China

- Introduction
- Method
- Experiments

Introduction

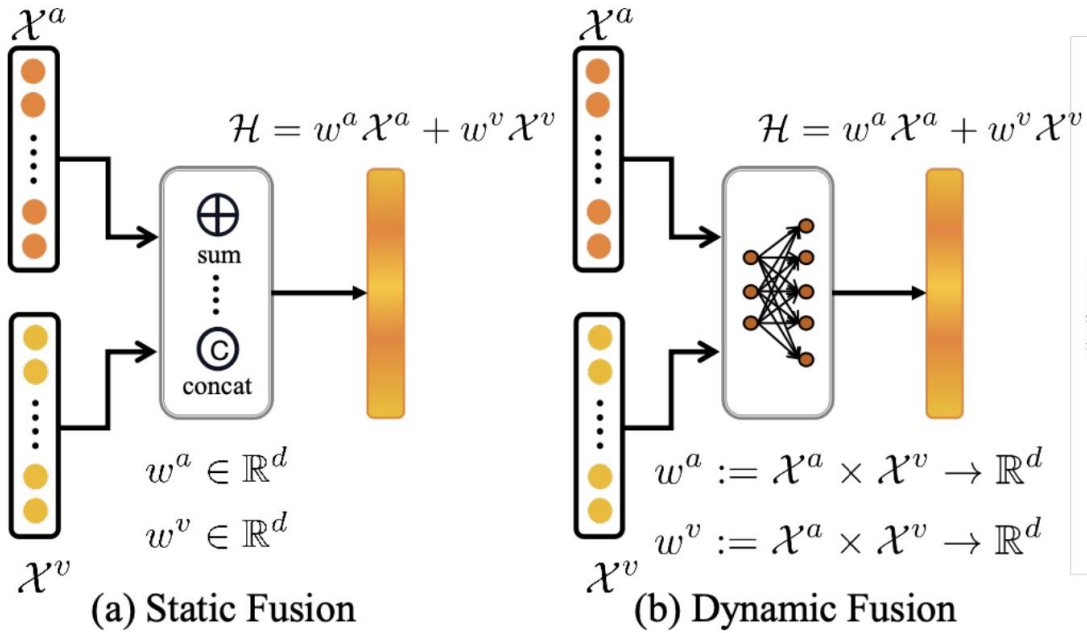
Fusion Paradigms

Multimodal Learning focus on effectively fuse information from diverse modalities.

Static fusion applies fixed weights to different modalities during inference.

However, modality quality may be varied.

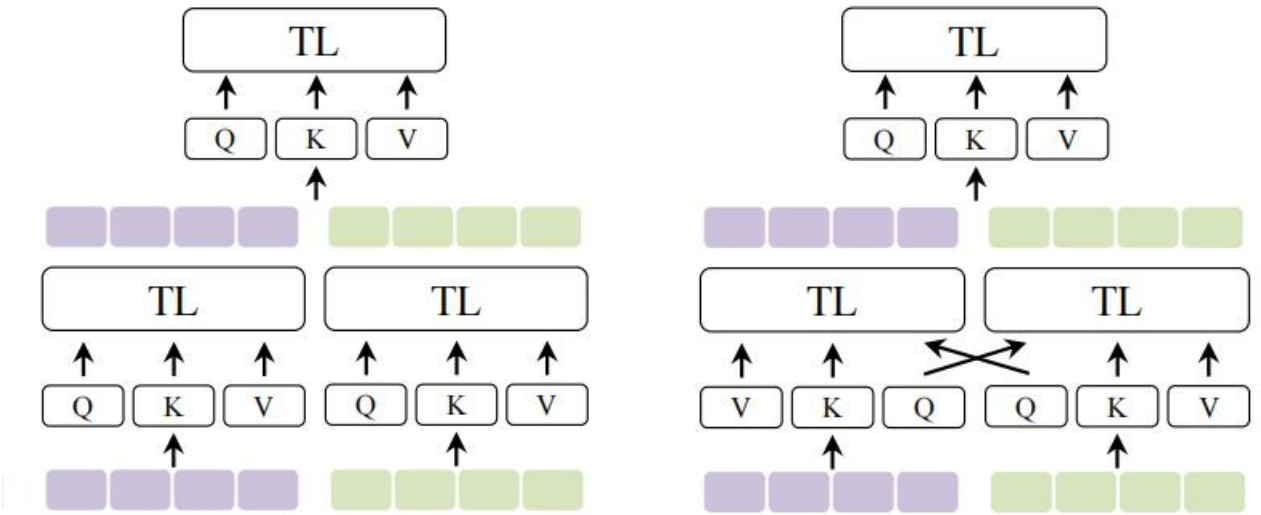
Dynamic fusion, by contrast, assigning weights dynamically to each modality based on the characteristics of input data.



[1] Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., et al. Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv:2404.18947, 2024.

■ Multimodal Transformer for Dynamic Fusion

To enable dynamic fusion, Multimodal Transformers have emerged as a powerful scheme, leveraging the attention mechanism to identify and focus on the informative and task-relevant tokens in the input.

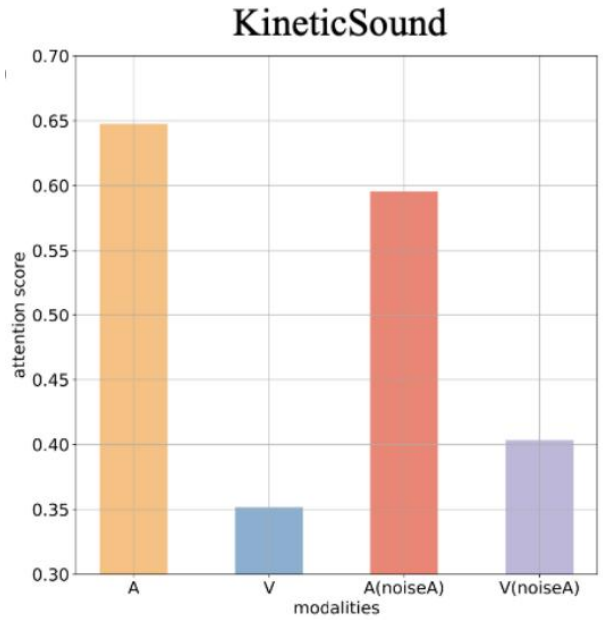


[1] Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., et al. Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv:2404.18947, 2024.

Deactivation of Cooperation Dynamics

Surprisingly, dynamic fusion achieves an accuracy of *67.0*, which **underperforms** the static fusion's accuracy of *68.0*.

Under noisy input test, model assigns **disproportionately high attention** to the audio modality across almost all samples, regardless of input characteristics.



(c) Attention Score

Average Key Distribution

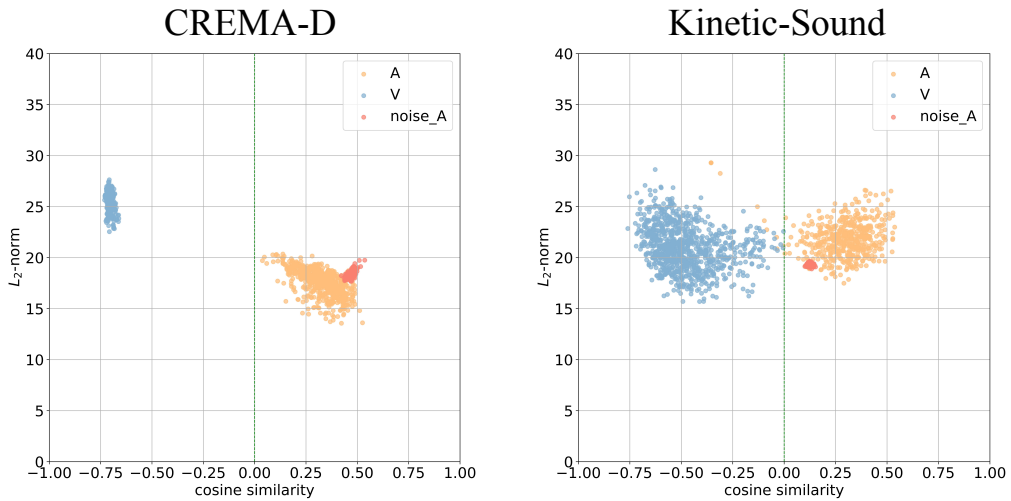
To investigate the underlying cause of this counterintuitive phenomenon, we analyzed the distribution of attention keys for each modality and their **cosine similarity with the query of the class token**, which directly determines the attention scores.

$$A_i = \frac{qK_i^T}{\sqrt{d}}, \qquad \text{expand to} \qquad \hat{k}_i^m = \frac{\sum_{j=1}^{L^m} k_{(i,j)}^m}{L^m}, m \in \{a, v\},$$

$$h_i = softmax(A_i)V_i, \qquad \sum_{j=1}^{L^m} \frac{qk_{(i,j)}^m}{\sqrt{d}} = \frac{L^m}{\sqrt{d}} ||q||_2 ||\hat{k}_i^m||_2 cos\theta_i^m.$$

A Modality is Biased

A modality is biased: the query of the class token, which determines the prediction of the model, remains significantly similar to the keys of the biased modality even when it contains no information. (In accordance with the "dominant modality" in *Imbalance Multimodal Learning*[2])



[2] Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation.In Proceedings of the IEEE/CVF conference on computervision and pattern recognition, pp. 8238–8247, 2022.

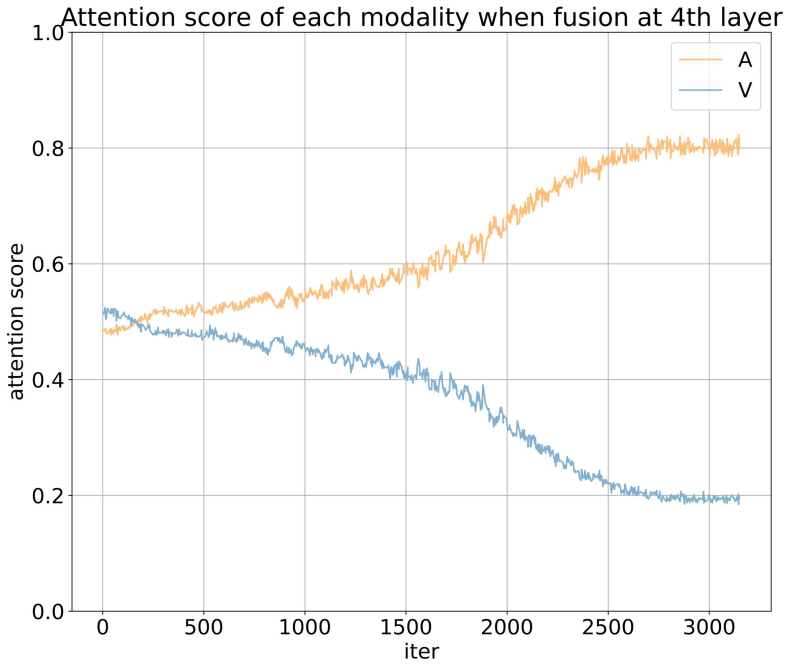
Effect of Biased Modality: Feed-forward Stage

Due to intrinsic differences between modalities, a modality may be favored by the model and provide higher quality features over time, becoming the biased modality.

Taking the **greedy nature** of multimodal deep neural networks [3] into account, the model tends to prioritize the modality a when it provides more informative features. As a result, the **biased modality accumulates higher attention scores and increasing cosine similarity with query.**

Consequently, this leads to a significant disparity in the average key distributions across modalities.

CREMA-D



[3] Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In International Conference on Machine Learning, pp. 24043–24055. PMLR,2022

Effect of Biased Modality: Back-forward Stage

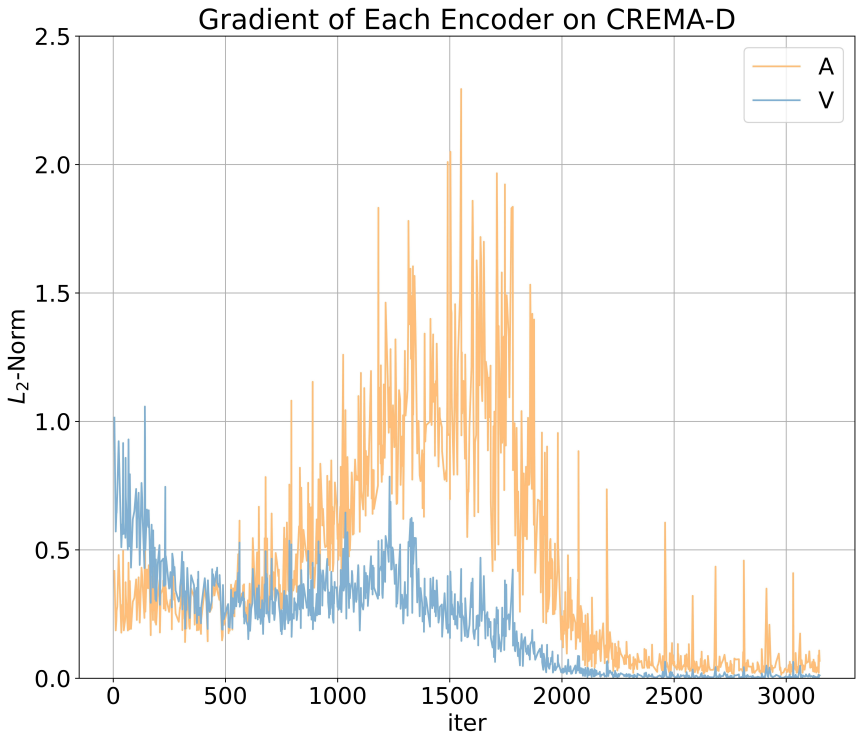
It, in turn, reinforces the optimization of its corresponding unimodal encoder parameters.

This dynamic further **exacerbates the inequality between feature qualities**, creating a self-reinforcing cycle.

$$\frac{\partial L}{\partial \theta^m} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial h_i}\frac{\partial h_i}{\partial z_i^m}\frac{\partial z_i^m}{\partial \theta^m}.$$

Using $s(\cdot)$ to denote $softmax(\cdot)$, we can expand the gradient $\frac{\partial h_i}{\partial z_i^m}$ to

$$\frac{\partial s(\frac{q[K_i^a, K_i^v]^T}{\sqrt{d}})}{\partial K_i^m}\frac{\partial K_i^m}{\partial z_i^m} + s(\frac{q[K_i^a, K_i^v]^T}{\sqrt{d}})\frac{\partial V_i^m}{\partial z_i^m}. \quad (8)$$

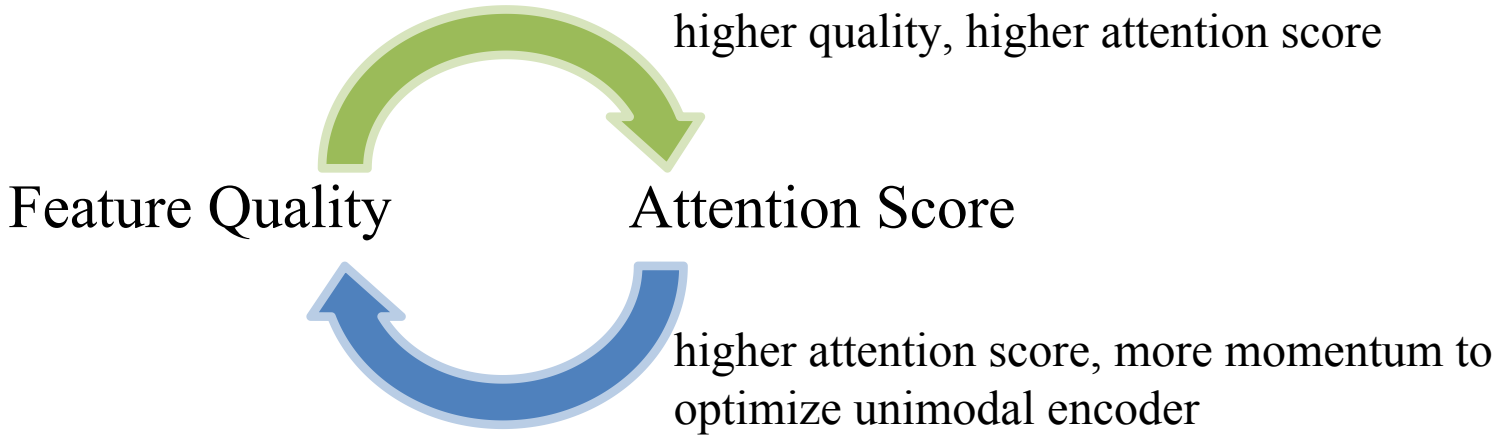


Modality Bias Triggers A Self-Reinforcing Cycle

Feed-forward stage: biased modality accumulates more attention score due to its more informative feature.

Backward propagation: the higher attention score provides more momentum to optimize the biased modality's encoder parameters.

Consequently, it not only creates a **significant distribution disparity**, but also amplifies the **inequality of feature quality**.



Rolling Query (RollingQ) Algorithm

Target: identifying and breaking the self-reinforcing cycle.

Identify: Monitoring distribution gap by **A**ttention **I**mbalance **R**ate (AIR) indicator

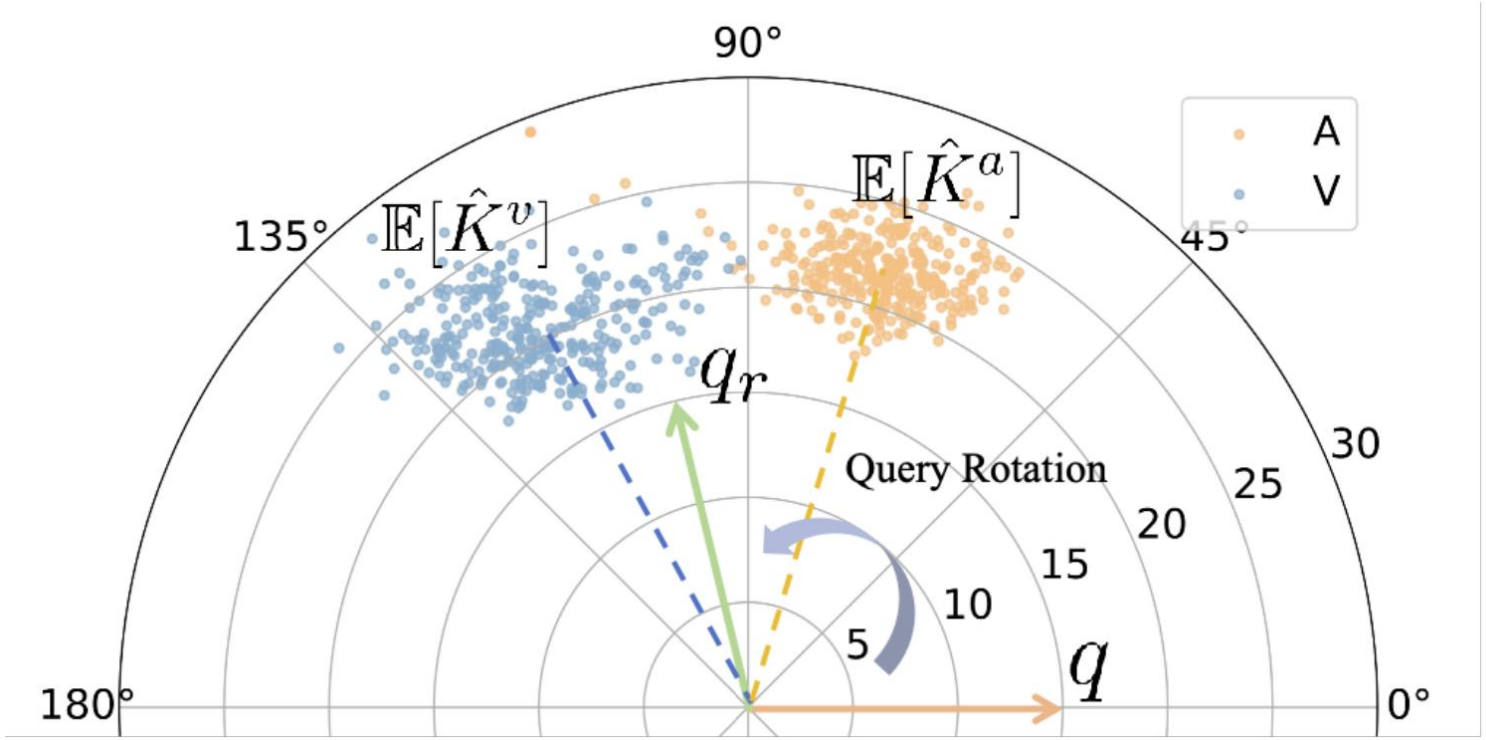
$$AIR = \mathbb{E}[cos\theta^a - cos\theta^v] \in [-2, 2].$$

Break the cycle: find a reasonable anchor that could assign higher attention scores to the unbiased modality rather than the biased one. (adjusting attention score)

$$q_b = (\alpha \frac{\mathbb{E}[\hat{K}^a]}{||\mathbb{E}[\hat{K}^a]||_2} + (1-\alpha)\frac{\mathbb{E}[\hat{K}^v]}{||\mathbb{E}[\hat{K}^v]||_2})||\mathbb{E}[Q]||_2, \qquad \alpha = \frac{1}{2}[1 + Tanh(-\rho AIR)],$$

■ Rolling Query (RollingQ) Algorithm

Rotate current query towards the rebalance anchor.



$$R_b = SVD(\mathbb{E}[Q], q_b),$$

$$q_r = qR_b.$$

Performance

Table 1. Validation on CREMAD (Audio+Visual), Kinetic-Sound (Audio+Visual) and MOSEI (Visual+Text) with Transformer backbone. The best results are presented in bold. We use ↑ to show the improvement in performance compared to not implementing the RollingQ.

	Dataset Metric	Fusion Layers	CREMA-D Acc	Kinetic-Sound Acc	CMU-MOSEI (V+T) Acc
Unimodal	Audio	0	47.6	53.9	-
	Visual	0	36.3	57.0	47.1
	Text	0	-	-	60.9
Static	Concat	1	49.3	68.0	62.8
	OGM (Peng et al., 2022)	1	51.2	68.2	62.7
	PMR (Fan et al., 2023)	1	50.1	68.2	<u>63.0</u>
Dynamic	Vanilla MT	1	48.8	67.0	62.7
	Vanilla MT*	2	51.5	69.1	62.2
	MulT (Tsai et al., 2019)	1	-	-	62.4
	MBT (Nagrani et al., 2021)	2	51.5	**72.2**	<u>63.0</u>
	JMT (Nagrani et al., 2021)	2	50.7	67.7	62.6
	MMML (Wu et al., 2024)	2	52.0	69.8	62.8
Ours	Vanilla MT+RollingQ	1	51.9 (↑ 3.1)	69.3 (↑ 2.3)	**63.2** (↑ 0.5)
	Vanilla MT*+RollingQ	2	<u>52.2</u> (↑ 0.7)	70.1 (↑ 1.0)	62.9 (↑ 0.7)
	MulT (Tsai et al., 2019)+RollingQ	1	-	-	62.5 (↑ 0.1)
	MMML (Wu et al., 2024)+RollingQ	2	**52.7** (↑ 0.7)	<u>70.7</u> (↑ 0.9)	**63.2** (↑ 0.4)

Further Verification

Pearson Coorelation Analysis

Table 2. The Pearson correlation between the attention score and whether it is a noise input. The coef closer to 1 or -1 indicates a stronger relation between attention score and input, while the $p < 0.01$ means that the coef is trustworthy.

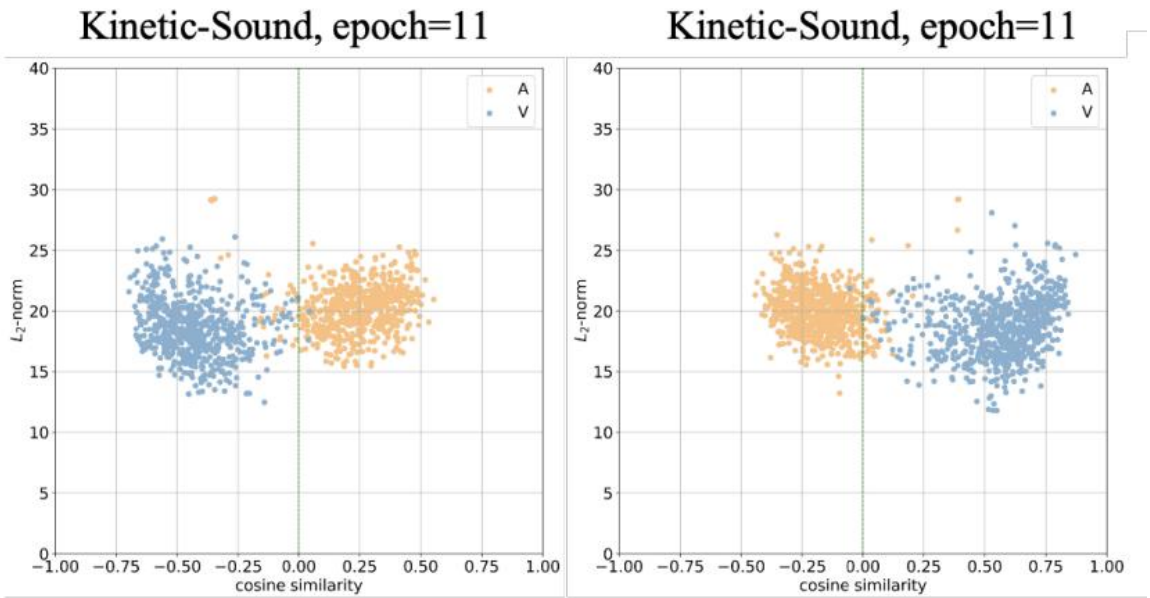
Dataset	CREMA-D		Kinetic-Sound	
Correlation	coef	p	coef	p
Vanilla MT	0.52	< 0.01	0.44	< 0.01
Vanilla MT+RollingQ	0.76	< 0.01	0.78	< 0.01

Test-time Adaptation for Noisy Biased Modality

Table 3. Experiment results when perturbing the audio modality (the biased modality) with Gaussian noise following Liang et al. (2021) on Kinetic-Sound Dataset.

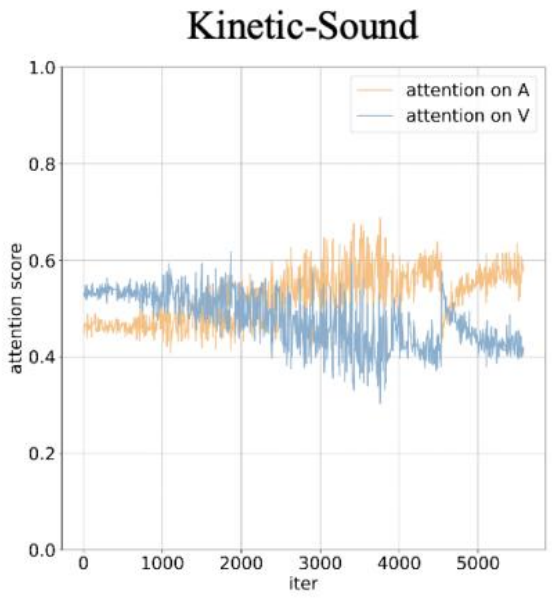
Noise level	Vanilla MT	Vanilla MT + RollingQ
0.00	67.0	69.3
0.25	62.7 (↓ 4.3)	67.2 (↓ 1.9)
0.50	52.9 (↓ 14.1)	58.2 (↓ 11.1)
0.75	43.2 (↓ 23.8)	47.5 (↓ 21.8)
1.00	34.7 (↓ 32.3)	40.6 (↓ 28.7)

What Has RollingQ Done to Revive Cooperation Dynamics?

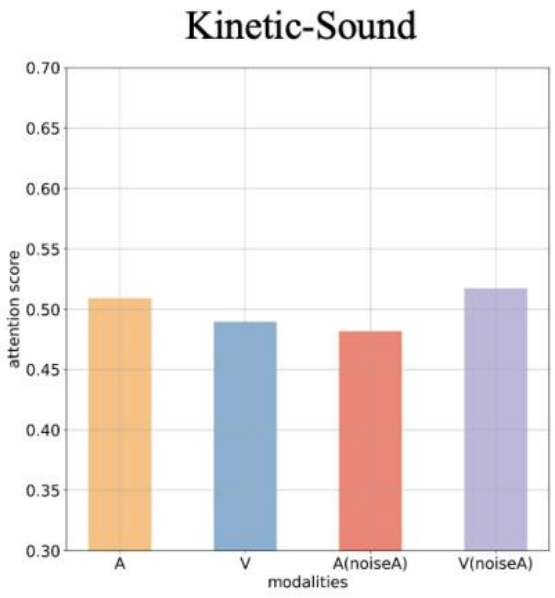


(a) Before Rotation

(b) After Rotation



(c) Training Process



(d) Attention Score

Simple yet Effective Method

Table 4. The accuracy, model parameters, and time complexity analysis on CREMA-D dataset. The GFLOPs are obtained from the thop library.

Method	Acc	Parameters	GFLOPs
Vanilla MT	48.8	59.87M	1489.13
MBT	51.5	114.21M	2746.90
MMML	**52.0**	77.88M	1828.29
JMT	50.7	62.23M	1494.87
Vanilla MT + RollingQ	51.9	**60.46M**	**1489.20**

中国人民大學
RENMIN UNIVERSITY OF CHINA

GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

Thank You for listening!

GeWu-lab

haotian_ni@buaa.edu.cn

hangliu@xmu.edu.cn

{yakewei; dihu}@ruc.edu.cn

2025-06-15