Supplementary Material: Robust Multimodal Emotion Recognition from Incomplete Modalities via Query-Based Unimodal and Cross-Modal Learning

Ryo Miyoshi* Mayu Otani Yuki Okafuji
CyberAgent

A. Datasets

In our study, we use two widely recognized datasets for evaluating our proposed method: the CMU-MOSEI [16] and MELD [8] datasets. We use pre-extracted features to ensure a fair comparison across methods. These features were provided by the MERBench project [5]. The features used in our experiments are publicly available in the MERBench repository (https://github.com/zeroQiaoba/MERTools).

B. Comparison of Computational Cost

This section compares the computational cost of our method with other representative state-of-the-art methods, including MulT [11], EMT-DLFR [9], UMDF [2], and CorrKD [3]. MulT is the de facto standard cross-modal Transformer-based method, EMT-DLFR is a reconstruction-based method, and UMDF and CorrKD are distillation-based methods. We evaluate the number of parameters, FLOPs, and inference time for each method. The number of parameters is calculated using the torchsummary library, and FLOPs are calculated using the fvcore library. Inference time is measured by averaging the time taken for 100 forward passes on a single NVIDIA RTX 3090 GPU with a batch size of 1.

Table 1 summarizes the number of parameters, FLOPs, and inference time for each method. We confirmed that the number of model parameters of the proposed method is comparable to or smaller than those of SOTA approaches. In terms of FLOPs and inference time, CorrKD achieves the lowest computational cost and fastest inference, while our method ranks second. Although the proposed method can be regarded as a query-based variant of MulT's cross-attention, it requires fewer FLOPs and shorter inference time than MulT, suggesting that its complexity has not increased.

*Corresponding author: miyoshi_ryo@cyberagent.co.jp

Table 1. Comparison of model size, FLOPs, and inference time. Inference time is reported as the mean and variance over 100 runs.

Model	Params (M)	FLOPs (G)	Inference Time (ms)
MulT (ACL 2019)	111.0	55.8	13.96 ± 0.159
EMT-DLFR (TAC 2023)	110.5	37.3	20.27 ± 0.253
UMDF (AAAI 2024)	114.1	41.2	21.54 ± 0.354
CorrKD (CVPR 2024)	139.3	27.2	6.44 ± 0.135
DQF (proposed)	112.6	30.0	13.28 ± 0.129

C. Comparative Experiments Under Various Imperfect Modality Conditions

This section presents additional experimental results of our proposed DQF method under various incomplete modality conditions.

C.1. Experimental Settings

We conduct experiments on two benchmark datasets, CMU-MOSEI and MELD, following the same settings as described in Section 4.1 of the main paper.

In this experiment, we evaluated the models under various missing conditions to assess the robustness of each method.

The training protocol follows that of the main paper, where each method was trained under the conditions specified for it. For methods designed to handle incomplete modalities, training was conducted under simulated missing conditions in which random frames were replaced with zero vectors.

For testing, we considered four settings formed by combining two missing patterns with two types of missing-value representations. The missing patterns were (i) randomly missing frames and (ii) consecutively missing frames, while the missing-value representations were (i) replacement with zero vectors and (ii) replacement with values sampled from a normal distribution. Among these, the case of randomly missing frames with zero-vector replacement is reported in the main paper. The missing ratios are consistent with those used in the main paper.

C.2. Experimental Results

C.2.1. Consecutive Missing Frames with Zero-Vector Replacement

In this experimental setting, we simulate scenarios where consecutive frames are missing, and the missing data are represented by zero vectors.

The results are presented in Table 2, 3. We observed that, on both CMU-MOSEI and MELD, our proposed method achieves the highest accuracy across all combinations of missing modalities.

For CMU-MOSEI, we observed that methods originally proposed under the complete-modality condition often achieved higher accuracy in this experimental setting, whereas methods designed for incomplete-modality conditions tended to show a slight decrease in performance.

For MELD, most methods exhibited improved accuracy under this setting. Our proposed method showed little variation in accuracy, suggesting that it is robust regardless of the missing pattern.

C.2.2. Random Missing Frames with Random-Noise Replacement

This experiment evaluates the setting in which randomly selected frames are replaced with values sampled from a normal distribution, thereby simulating randomly occurring noise.

The results are presented in Table 4, 5. We observed that, on both CMU-MOSEI and MELD, our proposed method consistently achieved the highest accuracy across all combinations of missing modalities.

We further compare these results with those in the main paper, where randomly selected frames were replaced with zero vectors. On CMU-MOSEI, most methods originally designed for complete-modality conditions exhibited a substantial drop in accuracy under the noise-based setting, whereas methods designed for incomplete-modality conditions did not show such degradation. In contrast, on MELD, this trend was not observed; the performance of all methods was comparable to that reported in the main paper.

C.2.3. Consecutive Missing Frames with Random-Noise Replacement

This experiment evaluates the setting in which consecutive frames are replaced with values sampled from a normal distribution, thereby simulating continuously occurring noise.

The results are presented in Table 6, 7. As shown, our proposed method outperformed all baselines across every missing-modality pattern.

We further compare these results with those in Table 2, 3, where the missing values were represented by zero vectors instead of noise vectors. On CMU-MOSEI, methods originally designed for complete-modality conditions exhibited a tendency to perform worse under the noise-vector replacement than under zero-vector replacement. In contrast, methods specifically proposed for incomplete-modality conditions did not show such degradation. On MELD, no such decline in accuracy was observed for either type of method, indicating that the impact of noise-based missing values is dataset-dependent.

C.3. Overall Findings

Across all experimental settings, the proposed method consistently achieved the highest accuracy on both CMU-MOSEI and MELD, demonstrating strong robustness to different missing patterns and missing-value representations.

A key observation is that methods originally developed for complete-modality conditions exhibit varying sensitivity to the type of missing-value representation. On CMU-MOSEI, these methods often suffered performance degradation when missing values were replaced with noise vectors, particularly under random or consecutive missing frames, whereas their performance was relatively higher when using zero-vector replacement. In contrast, methods explicitly designed for incomplete-modality conditions showed stable performance across both replacement strategies, indicating that their robustness is less affected by the nature of the missing values.

On MELD, the overall trend differed: most methods maintained comparable or even improved accuracy across all missing conditions, regardless of whether the missing values were represented by zeros or random noise. This suggests that the impact of missing-value representation is dataset-dependent.

Taken together, these findings highlight two important conclusions. First, robustness to incomplete modalities is strongly influenced by whether a method is designed under complete- or incomplete-modality assumptions. Second, the proposed method exhibits stable performance across both datasets and under diverse missing conditions, confirming its effectiveness as a robust solution for multimodal emotion recognition.

Table 2. Performance under incomplete modality conditions on CMU-MOSEI. This experimental setting consists of consecutive missing frames, which have been replaced with zero vectors. AUILC scores are reported for Acc.7 / Acc.2 / MAE. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete inputs. Higher is better for all metrics except MAE.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	49.8 / 84.6 / 0.574	48.5 / 83.4 / 0.594	47.1 / 81.3 / 0.632	49.8 / 84.6 / 0.577	48.5 / 82.3 / 0.611	47.1 / 80.8 / 0.631	48.3 / 81.8 / 0.618
	MFN [15]	49.4 / 83.6 / 0.600	50.0 / 83.8 / 0.593	48.0 / 81.2 / 0.633	49.8 / 83.8 / 0.595	47.9 / 81.1 / 0.635	48.3 / 81.4 / 0.629	48.1 / 81.2 / 0.632
	Graph-MFN [16]	49.2 / 84.5 / 0.590	49.4 / 84.2 / 0.588	48.3 / 81.4 / 0.616	49.3 / 84.2 / 0.590	48.2 / 81.2 / 0.627	48.4 / 81.1 / 0.627	48.3 / 80.8 / 0.628
	LMF [6]	51.7 / 86.9 / 0.545	51.8 / 86.7 / 0.546	50.1 / 83.7 / 0.596	51.7 / 86.6 / 0.549	49.9 / 83.7 / 0.600	50.0 / 83.6 / 0.599	49.7 / 83.5 / 0.600
	MCTN [7]	53.0 / 85.8 / 0.545	53.0 / 85.7 / 0.545	49.4 / 81.3 / 0.616	53.0 / 85.8 / 0.545	49.4 / 81.2 / 0.616	49.5 / 81.3 / 0.617	49.3 / 81.3 / 0.617
	MFM [10]	50.2 / 82.8 / 0.593	49.7 / 82.6 / 0.601	48.1 / 79.6 / 0.640	50.1 / 82.9 / 0.593	48.5 / 79.8 / 0.637	48.0 / 79.6 / 0.641	48.4 / 79.9 / 0.634
	MulT [11]	53.5 / 86.5 / 0.531	53.5 / 86.5 / 0.536	51.7 / 83.4 / 0.577	53.5 / 86.5 / 0.531	51.6 / 83.1 / 0.579	51.6 / 83.4 / 0.577	51.6 / 83.0 / 0.580
	MISA [1]	53.3 / 84.8 / 0.544	53.2 / 84.6 / 0.543	50.9 / 81.6 / 0.591	53.3 / 84.8 / 0.544	50.6 / 81.7 / 0.595	50.8 / 81.6 / 0.592	50.7 / 81.8 / 0.594
	DMD [4]	47.6 / 86.5 / 0.611	47.8 / 86.7 / 0.611	47.3 / 84.0 / 0.631	48.1 / 86.5 / 0.606	47.6 / 84.1 / 0.627	47.6 / 83.9 / 0.628	48.0 / 83.7 / 0.625
(2)	TFR-Net [13]	51.1 / 85.1 / 0.563	51.2 / 85.1 / 0.560	49.3 / 82.3 / 0.602	51.1 / 85.2 / 0.563	49.3 / 82.3 / 0.604	49.2 / 82.3 / 0.603	49.3 / 82.3 / 0.604
	DiCMoR [12]	47.1 / 79.3 / 0.653	47.2 / 79.3 / 0.652	45.6 / 76.5 / 0.692	47.1 / 79.3 / 0.653	45.6 / 76.4 / 0.693	45.6 / 76.6 / 0.692	45.7 / 76.5 / 0.692
	EMT-DLFR [9]	51.6 / 83.9 / 0.568	51.6 / 83.9 / 0.567	49.6 / 80.9 / 0.607	51.6 / 83.9 / 0.568	49.6 / 81.1 / 0.608	49.7 / 80.9 / 0.607	49.7 / 81.0 / 0.608
	UMDF [2]	51.4 / 83.7 / 0.569	51.5 / 83.8 / 0.566	50.0 / 81.5 / 0.602	51.1 / 83.7 / 0.572	49.4 / 80.8 / 0.611	49.8 / 81.2 / 0.606	49.3 / 80.7 / 0.614
	CorrKD [3]	53.1 / 86.8 / 0.541	51.6 / 86.6 / 0.556	49.9 / 82.6 / 0.594	52.3 / 86.7 / 0.549	50.6 / 82.7 / 0.592	49.6 / 82.8 / 0.597	50.2 / 83.0 / 0.594
	DQF(Ours)	**55.2 / 87.1 / 0.522**	**55.6 / 86.8 / 0.520**	**52.9 / 84.5 / 0.563**	**54.9 / 87.0 / 0.524**	**52.6 / 84.2 / 0.568**	**52.8 / 84.1 / 0.566**	**52.5 / 83.8 / 0.572**

Table 3. Performance evaluation under incomplete modality conditions on MELD. This experimental setting consists of consecutive missing frames, which have been replaced with zero vectors. AUILC scores are reported for WAR and UAR. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete modalities. Higher values indicate better performance.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	59.9 / 32.3	59.7 / 31.7	56.8 / 29.1	59.8 / 31.4	57.0 / 28.7	56.9 / 27.6	56.9 / 27.2
	MFN [15]	59.7 / 31.1	59.3 / 31.6	54.9 / 27.5	58.3 / 31.4	54.9 / 27.4	54.4 / 27.8	54.4 / 27.6
	Graph-MFN [16]	58.4 / 30.4	57.8 / 28.8	51.6 / 26.3	57.6 / 28.4	51.6 / 26.0	51.5 / 25.1	51.7 / 24.7
	LMF [6]	58.8 / 31.1	57.3 / 30.1	54.6 / 29.3	58.4 / 30.3	56.1 / 29.6	54.1 / 28.4	55.8 / 28.4
	MCTN [7]	51.5 / 20.0	51.5 / 20.0	50.4 / 17.8	51.5 / 20.0	50.3 / 17.8	50.3 / 17.7	50.4 / 17.7
	MFM [10]	51.6 / 20.8	52.2 / 21.6	50.4 / 19.3	52.2 / 21.7	50.3 / 19.3	50.4 / 19.5	50.4 / 19.6
	MulT [11]	61.9 / 39.2	61.5 / 39.7	57.5 / 33.7	61.8 / 39.4	57.7 / 33.2	57.6 / 33.4	57.9 / 33.0
	MISA [1]	52.7 / 22.7	52.7 / 22.7	51.8 / 20.8	52.7 / 22.7	51.6 / 20.7	51.7 / 20.8	51.8 / 20.9
	DMD [4]	61.9 / 33.8	61.8 / 33.9	57.7 / 28.3	61.7 / 33.7	57.7 / 28.3	57.7 / 28.3	57.7 / 28.3
(2)	TFR-Net [13]	60.6 / 39.5	60.0 / 39.2	58.3 / 34.6	59.9 / 39.2	58.2 / 34.6	57.7 / 33.6	57.4 / 33.7
	DiCMoR [12]	53.8 / 26.7	53.2 / 27.1	53.0 / 23.4	53.5 / 26.9	52.7 / 23.4	51.5 / 23.4	51.9 / 23.6
	EMT-DLR [9]	59.6 / 35.6	59.4 / 35.6	56.9 / 30.1	59.5 / 35.6	57.0 / 30.1	56.8 / 29.9	57.0 / 30.1
	UMDF [2]	59.8 / 34.7	59.8 / 34.5	56.9 / 29.3	59.8 / 34.5	56.8 / 29.3	56.9 / 29.3	56.9 / 29.4
	CorrKD [3]	61.2 / 33.0	61.1 / 32.7	56.8 / 29.2	61.2 / 32.8	56.9 / 29.1	56.9 / 28.6	57.1 / 28.7
	DQF(Ours)	**62.9 / 39.6**	**62.6 / 39.9**	**59.5 / 35.0**	**62.6 / 39.8**	**59.6 / 35.1**	**59.1 / 34.0**	**59.1 / 34.1**

Table 4. Performance under incomplete modality conditions on CMU-MOSEI. This experimental setting consists of random missing frames, which have been replaced with random noise vectors. The random noise vectors are sampled from a normal distribution. AUILC scores are reported for Acc.7 / Acc.2 / MAE. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete inputs. Higher is better for all metrics except MAE.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	49.0 / 84.6 / 0.588	48.6 / 83.4 / 0.594	44.9 / 77.5 / 0.677	49.2 / 84.6 / 0.583	45.7 / 76.9 / 0.680	45.0 / 76.3 / 0.686	45.6 / 76.2 / 0.639
	MFN [15]	47.9 / 83.2 / 0.625	50.0 / 84.2 / 0.592	45.9 / 75.5 / 0.690	47.8 / 83.4 / 0.622	46.1 / 76.3 / 0.685	46.1 / 75.9 / 0.692	46.2 / 76.6 / 0.691
	Graph-MFN [16]	47.6 / 83.6 / 0.614	48.8 / 83.6 / 0.600	46.1 / 76.1 / 0.678	46.3 / 82.4 / 0.635	46.4 / 77.1 / 0.675	46.1 / 76.5 / 0.673	46.3 / 76.7 / 0.677
	LMF [6]	51.2 / 86.7 / 0.554	51.7 / 86.6 / 0.547	48.2 / 79.5 / 0.632	51.0 / 86.5 / 0.557	48.5 / 78.4 / 0.636	48.3 / 79.6 / 0.635	48.4 / 78.9 / 0.639
	MCTN [7]	53.0 / 85.8 / 0.545	53.0 / 85.8 / 0.545	50.0 / 82.1 / 0.603	53.0 / 85.8 / 0.545	50.2 / 82.2 / 0.602	50.1 / 82.2 / 0.602	50.0 / 82.2 / 0.603
	MFM [10]	49.4 / 82.4 / 0.607	49.8 / 82.5 / 0.604	45.1 / 73.6 / 0.759	48.9 / 82.2 / 0.616	45.6 / 75.2 / 0.753	45.2 / 73.9 / 0.767	45.4 / 75.3 / 0.764
	MulT [11]	48.7 / 84.4 / 0.594	53.5 / 86.4 / 0.532	44.0 / 70.7 / 0.875	48.7 / 84.5 / 0.593	43.8 / 70.8 / 0.878	44.0 / 70.7 / 0.879	43.8 / 70.7 / 0.879
	MISA [1]	52.9 / 84.6 / 0.549	53.2 / 84.7 / 0.543	43.6 / 68.0 / 1.032	53.0 / 84.6 / 0.549	43.5 / 68.1 / 1.034	43.9 / 68.0 / 1.032	43.5 / 68.1 / 1.033
	DMD [4]	43.5 / 84.9 / 0.736	47.0 / 86.7 / 0.612	45.2 / 79.9 / 0.683	43.4 / 84.8 / 0.736	42.9 / 76.3 / 0.826	45.3 / 79.4 / 0.686	42.9 / 76.2 / 0.832
(2)	TFR-Net [13]	48.0 / 84.3 / 0.599	51.2 / 85.2 / 0.560	41.9 / 64.0 / 3.348	48.0 / 84.4 / 0.597	41.9 / 64.0 / 3.334	41.9 / 64.0 / 3.357	41.9 / 64.0 / 3.349
	DiCMoR [12]	46.6 / 78.9 / 0.660	47.0 / 79.4 / 0.653	45.7 / 79.9 / 0.683	46.6 / 79.0 / 0.661	45.3 / 76.0 / 0.697	45.6 / 76.9 / 0.684	45.2 / 75.9 / 0.697
	EMT-DLFR [9]	51.3 / 83.8 / 0.573	51.6 / 83.9 / 0.567	49.7 / 81.2 / 0.615	51.1 / 83.5 / 0.576	49.8 / 81.1 / 0.610	49.8 / 81.2 / 0.609	49.6 / 81.0 / 0.611
	UMDF [2]	52.0 / 82.9 / 0.562	51.3 / 83.7 / 0.568	49.7 / 81.2 / 0.607	51.6 / 82.9 / 0.568	49.9 / 79.2 / 0.612	49.5 / 80.6 / 0.613	49.5 / 79.3 / 0.618
	CorrKD [3]	52.8 / 86.9 / 0.545	51.9 / 86.6 / 0.553	49.9 / 82.9 / 0.595	52.0 / 86.8 / 0.552	50.2 / 83.2 / 0.594	49.6 / 83.1 / 0.597	49.9 / 83.3 / 0.596
	DQF(Ours)	**55.2 / 87.1 / 0.523**	**55.6 / 87.0 / 0.519**	**52.9 / 83.9 / 0.565**	**55.1 / 87.0 / 0.523**	**52.5 / 83.8 / 0.572**	**52.9 / 83.9 / 0.567**	**52.6 / 83.8 / 0.572**

Table 5. Performance evaluation under incomplete modality conditions on MELD. This experimental setting consists of random missing frames, which have been replaced with random noise vectors. AUILC scores are reported for WAR and UAR. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete modalities. Higher values indicate better performance.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	58.9 / 32.0	59.9 / 30.6	54.6 / 27.9	59.0 / 31.3	53.9 / 27.2	54.9 / 26.3	54.4 / 25.9
	MFN [15]	59.6 / 30.8	59.4 / 32.2	52.2 / 24.6	59.4 / 31.6	52.4 / 24.5	51.7 / 24.7	51.9 / 24.5
	Graph-MFN [16]	58.4 / 30.0	56.1 / 25.3	50.9 / 25.4	55.3 / 24.2	51.1 / 24.5	51.6 / 21.6	51.7 / 21.2
	LMF [6]	58.4 / 30.4	57.6 / 30.6	53.5 / 27.9	58.2 / 30.0	53.5 / 26.9	54.3 / 27.8	54.3 / 26.6
	MCTN [7]	51.5 / 20.0	51.5 / 20.0	49.6 / 16.6	51.5 / 20.0	49.6 / 16.5	49.6 / 16.5	49.6 / 16.5
	MFM [10]	51.7 / 21.0	49.1 / 21.8	49.5 / 17.0	49.0 / 21.7	49.5 / 17.0	48.7 / 18.0	48.7 / 18.7
	MulT [11]	62.1 / 38.0	60.8 / 40.0	55.2 / 28.1	61.3 / 38.7	55.3 / 27.0	54.7 / 28.4	55.0 / 27.3
	MISA [1]	52.7 / 22.7	52.7 / 22.7	50.9 / 18.8	52.7 / 22.7	51.0 / 18.8	50.9 / 18.8	51.0 / 18.9
	DMD [4]	61.6 / 33.5	61.9 / 34.0	53.1 / 27.2	61.5 / 33.7	52.8 / 26.6	52.8 / 27.0	52.8 / 26.3
(2)	TFR-Net [13]	59.7 / 37.9	60.7 / 39.6	57.4 / 34.9	59.8 / 38.1	56.1 / 33.3	57.2 / 33.5	56.5 / 33.1
	DiCMoR [12]	53.7 / 26.3	52.9 / 27.2	51.2 / 22.1	53.2 / 26.4	52.8 / 26.6	50.5 / 22.4	50.6 / 22.0
	EMT-DLR [9]	59.9 / 35.7	59.7 / 35.7	56.2 / 29.0	59.8 / 35.6	56.3 / 28.9	56.2 / 29.0	56.3 / 28.8
	UMDF [2]	59.8 / 35.7	60.1 / 34.5	56.3 / 28.9	60.1 / 34.5	56.2 / 28.9	56.3 / 28.5	56.3 / 28.6
	CorrKD [3]	61.1 / 32.9	61.3 / 32.8	56.2 / 28.7	61.4 / 32.8	56.2 / 28.6	56.5 / 28.3	56.6 / 28.3
	DQF(Ours)	**62.9 / 39.5**	**62.7 / 40.2**	**58.7 / 35.4**	**62.7 / 39.2**	**58.7 / 34.4**	**58.4 / 33.8**	**58.5 / 33.8**

Table 6. Performance under incomplete modality conditions on CMU-MOSEI. This experimental setting consists of consecutive missing frames, which have been replaced with random noise vectors. The random noise vectors are sampled from a normal distribution. AUILC scores are reported for Acc.7 / Acc.2 / MAE. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete inputs. Higher is better for all metrics except MAE.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	48.9 / 84.6 / 0.589	48.5 / 83.4 / 0.594	46.9 / 80.2 / 0.636	49.1 / 84.5 / 0.584	47.6 / 80.3 / 0.634	46.9 / 79.5 / 0.641	47.3 / 79.5 / 0.646
	MFN [15]	49.1 / 83.5 / 0.605	49.9 / 83.8 / 0.593	47.2 / 78.7 / 0.653	49.6 / 83.7 / 0.598	47.1 / 78.8 / 0.653	47.0 / 78.3 / 0.661	47.0 / 78.4 / 0.662
	Graph-MFN [16]	48.9 / 84.3 / 0.595	49.2 / 84.0 / 0.593	47.4 / 79.3 / 0.647	48.4 / 83.8 / 0.602	47.4 / 79.3 / 0.647	47.4 / 79.0 / 0.647	47.4 / 79.0 / 0.649
	LMF [6]	51.2 / 86.7 / 0.554	51.8 / 86.7 / 0.546	49.2 / 80.9 / 0.611	51.1 / 86.7 / 0.556	49.8 / 80.3 / 0.610	49.2 / 80.6 / 0.615	49.6 / 80.2 / 0.615
	MCTN [7]	53.0 / 85.8 / 0.545	53.0 / 85.8 / 0.545	49.4 / 81.4 / 0.616	53.0 / 85.8 / 0.545	49.4 / 81.3 / 0.615	49.5 / 81.5 / 0.615	49.4 / 81.4 / 0.616
	MFM [10]	49.8 / 82.6 / 0.599	49.5 / 82.3 / 0.607	45.3 / 73.6 / 0.734	49.4 / 82.4 / 0.601	45.3 / 74.0 / 0.735	45.3 / 74.3 / 0.729	45.5 / 74.7 / 0.731
	MulT [11]	52.5 / 86.1 / 0.545	53.5 / 86.5 / 0.531	49.2 / 80.8 / 0.655	52.6 / 86.1 / 0.543	48.9 / 80.7 / 0.656	49.2 / 80.8 / 0.656	48.9 / 80.6 / 0.659
	MISA [1]	53.2 / 84.7 / 0.546	53.2 / 84.7 / 0.543	44.1 / 70.6 / 0.836	53.2 / 84.7 / 0.546	44.1 / 70.6 / 0.864	44.0 / 70.5 / 0.866	44.1 / 70.4 / 0.864
	DMD [4]	44.1 / 85.6 / 0.683	47.6 / 86.6 / 0.614	47.1 / 83.4 / 0.639	43.9 / 85.4 / 0.690	44.0 / 81.9 / 0.713	47.0 / 82.9 / 0.616	43.8 / 81.6 / 0.728
(2)	TFR-Net [13]	44.1 / 82.9 / 0.642	51.2 / 85.1 / 0.560	41.9 / 64.0 / 3.344	44.2 / 82.9 / 0.641	41.9 / 64.0 / 3.312	41.9 / 64.0 / 3.349	41.9 / 64.0 / 3.317
	DiCMoR [12]	47.1 / 79.2 / 0.654	47.0 / 79.3 / 0.652	45.7 / 76.4 / 0.692	47.0 / 79.2 / 0.654	45.5 / 76.3 / 0.694	45.6 / 76.5 / 0.692	45.7 / 76.4 / 0.693
	EMT-DLFR [9]	51.4 / 83.8 / 0.571	51.6 / 83.9 / 0.567	49.6 / 80.8 / 0.623	51.3 / 83.7 / 0.573	49.6 / 81.9 / 0.713	49.6 / 80.7 / 0.616	49.6 / 80.8 / 0.612
	UMDF [2]	52.0 / 82.9 / 0.562	51.5 / 83.8 / 0.566	50.0 / 81.6 / 0.602	51.7 / 83.0 / 0.566	50.1 / 80.0 / 0.604	49.8 / 81.2 / 0.605	49.8 / 80.0 / 0.611
	CorrKD [3]	52.8 / 86.9 / 0.544	51.4 / 86.6 / 0.559	50.0 / 82.6 / 0.595	51.5 / 86.7 / 0.557	50.4 / 82.9 / 0.594	49.5 / 82.8 / 0.599	49.8 / 83.0 / 0.598
	DQF(Ours)	**55.2 / 87.1 / 0.523**	**55.6 / 86.8 / 0.520**	**52.7 / 84.1 / 0.565**	**54.9 / 86.9 / 0.525**	**52.5 / 83.9 / 0.571**	**52.7 / 83.9 / 0.568**	**52.6 / 83.7 / 0.574**

Table 7. Performance evaluation under incomplete modality conditions on MELD. This experimental setting consists of consecutive missing frames, which have been replaced with random noise vectors. AUILC scores are reported for WAR and UAR. The methods are grouped into two categories: (1) methods trained on complete inputs, and (2) methods explicitly designed to handle incomplete modalities. Higher values indicate better performance.

	Models	Incomplete modality						
		$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
(1)	TFN [14]	58.8 / 32.0	59.7 / 31.7	56.2 / 29.2	58.5 / 30.9	55.3 / 28.6	56.5 / 27.8	55.4 / 27.2
	MFN [15]	59.7 / 31.1	59.6 / 31.5	53.0 / 27.9	59.5 / 31.2	49.7 / 17.5	52.7 / 27.8	52.8 / 27.8
	Graph-MFN [16]	58.4 / 30.3	57.7 / 28.4	51.3 / 27.4	57.4 / 28.0	51.3 / 27.3	52.4 / 26.2	52.3 / 25.9
	LMF [6]	58.4 / 30.4	57.3 / 30.1	54.3 / 28.4	57.8 / 29.6	54.1 / 27.4	54.5 / 27.6	54.5 / 26.9
	MCTN [7]	51.5 / 20.0	51.5 / 20.0	50.4 / 17.8	51.5 / 20.0	50.4 / 17.9	50.3 / 17.8	50.4 / 17.8
	MFM [10]	51.7 / 20.8	50.9 / 22.1	49.7 / 17.5	50.8 / 22.0	49.7 / 17.6	49.5 / 18.0	49.5 / 18.1
	MulT [11]	62.1 / 39.1	61.7 / 39.5	57.8 / 32.8	62.0 / 39.1	58.1 / 32.5	58.0 / 32.7	58.2 / 32.5
	MISA [1]	52.7 / 22.7	52.7 / 22.7	51.6 / 20.0	52.7 / 22.7	51.6 / 20.0	51.6 / 20.0	51.8 / 20.1
	DMD [4]	61.8 / 33.8	61.9 / 34.0	55.8 / 29.2	61.7 / 33.9	57.6 / 29.0	57.4 / 29.1	55.3 / 28.9
(2)	TFR-Net [13]	60.4 / 38.7	60.0 / 39.2	58.4 / 35.2	60.0 / 38.5	58.1 / 34.3	57.3 / 33.8	57.6 / 33.7
	DiCMoR [12]	53.8 / 26.5	53.3 / 26.7	52.0 / 23.5	53.6 / 26.4	52.1 / 23.2	51.9 / 23.3	52.1 / 23.2
	EMT-DLR [9]	59.7 / 35.7	59.5 / 35.6	56.8 / 29.9	59.6 / 35.6	57.0 / 30.0	56.8 / 29.9	57.0 / 28.8
	UMDF [2]	59.8 / 34.7	59.8 / 34.5	56.9 / 29.3	59.8 / 34.6	56.7 / 29.3	56.9 / 29.3	56.9 / 29.4
	CorrKD [3]	61.1 / 32.9	61.1 / 32.7	56.8 / 29.2	61.1 / 32.7	56.9 / 29.1	56.9 / 28.6	57.0 / 28.6
	DQF(Ours)	**62.9 / 39.5**	**62.6 / 39.7**	**59.5 / 35.4**	**62.6 / 39.4**	**59.6 / 35.0**	**59.1 / 34.0**	**59.1 / 34.1**

D. Performance across Varying Missing Rates

To make our fine-grained missing-rate evaluation explicit, we further analyze how recognition performance changes as the missing rate increases when all modalities are affected simultaneously.

Figures 1a and 1b show the recognition performance at each missing rate under the most challenging configuration, where all three modalities ($\{v, a, t\}$) are simultaneously subject to random frame-level masking with zero-vector replacement. For CMU-MOSEI, we report Acc.7 / Acc.2 / MAE, and for MELD, WAR / UAR. The missing rate r is varied from 0.0 to 0.9 in increments of 0.1, and we exclude the degenerate case $r = 1.0$, where all modalities are fully masked and no input information remains. Across all missing rates on both datasets, DQF consistently outperforms the compared methods, indicating that its advantage is not limited to the aggregated AUILC score but holds at each level of modality incompleteness.

References

[1] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020. 3, 4, 5

[2] Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10074–10082, 2024. 1, 3, 4, 5

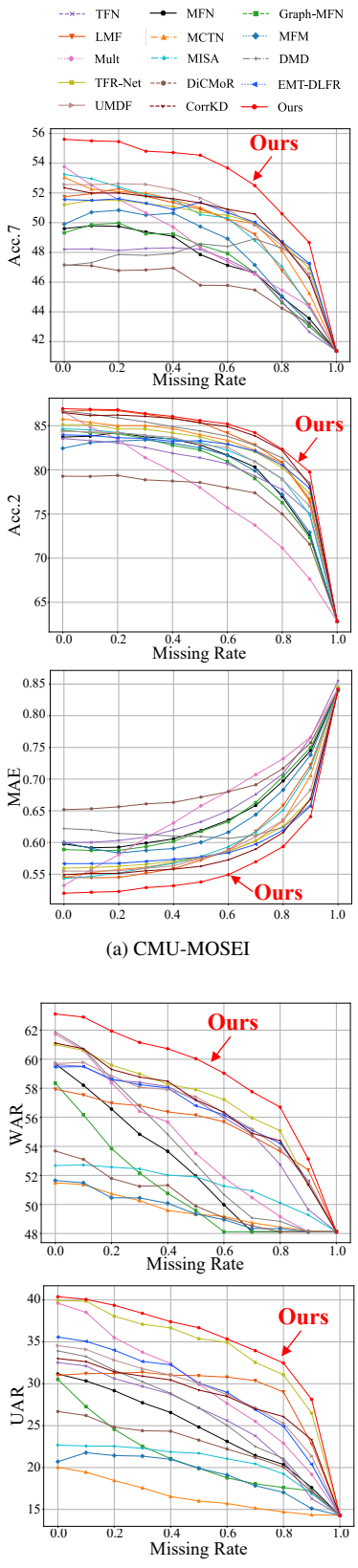
[3] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468, 2024. 1, 3, 4, 5

[4] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023. 3, 4, 5

[5] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*, 2024. 1

[6] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. 3, 4, 5

[7] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations be-



(a) CMU-MOSEI

(b) MELD

Figure 1. Performance comparison under various missing rates on CMU-MOSEI (top) and MELD (bottom).

tween modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6892–6899, 2019. 3, 4, 5

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, 2019. Association for Computational Linguistics. 1

[9] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2023. 1, 3, 4, 5

[10] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. 2019. 3, 4, 5

[11] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, page 6558. NIH Public Access, 2019. 1, 3, 4, 5

[12] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034, 2023. 3, 4, 5

[13] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407, 2021. 3, 4, 5

[14] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 3, 4, 5

[15] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3, 4, 5

[16] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246, 2018. 1, 3, 4, 5