

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

Ryo Miyoshi* Mayu Otani Yuki Okafuji
CyberAgent

Abstract

Multimodal emotion recognition (MER) aims to identify human emotions from inputs such as text, vision, and audio. However, existing methods often assume complete modality availability during training and inference, which is unrealistic in real-world scenarios due to sensor failures or privacy constraints. We propose Dual-Query Fusion (DQF), a framework that enables robust MER using only incomplete modality inputs, without relying on reconstruction or knowledge distillation. DQF introduces two types of learnable queries: Q-UA for extracting informative unimodal features, and Q-CA for adaptive cross-modal integration. These modules are designed to operate effectively even when some modalities are missing. Experiments on two public datasets demonstrate that DQF achieves superior performance and robustness compared to existing methods, even when trained exclusively on incomplete inputs. These results highlight the effectiveness and practicality of DQF for real-world MER tasks.

1. Introduction

Multimodal emotion recognition (MER) is a task that recognizes human emotional states by utilizing multiple modalities such as text, vision, and audio. This task plays a crucial role in the field of human-computer interaction, as it enables machines to understand and respond to human emotions more effectively.

Compared to unimodal emotion recognition approaches [1, 17, 35, 36, 49, 52], MER leverages complementary cues from different modalities [28, 32, 34]. However, real-world scenarios often involve missing modalities due to sensor failures, transmission errors, or privacy constraints. Despite this, many existing MER methods [12, 40, 48, 55, 56] assume complete modality inputs during both training and inference, which can significantly degrade performance when violated.

*Corresponding author: miyoshi.ryo@cyberagent.co.jp

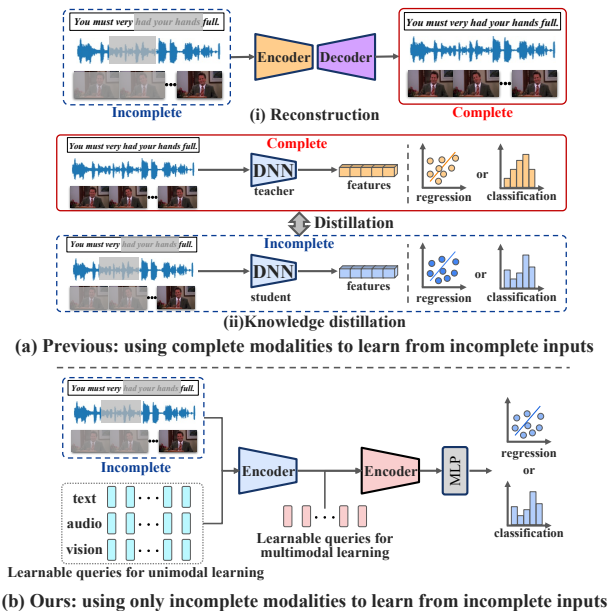


Figure 1. Comparison of approaches for robust MER under incomplete-modality conditions. (a) Previous methods introduce auxiliary tasks such as reconstruction or knowledge distillation and require access to complete multimodal sequences during training. (b) Our method employs learnable queries to extract and fuse features directly from incomplete inputs, enabling training without complete modalities while maintaining robustness and accuracy.

To improve robustness under incomplete modality settings, recent studies have introduced auxiliary tasks such as feature reconstruction [14, 23, 46] (Figure 1(a)-(i)) and knowledge distillation [7, 8, 15, 18] (Figure 1(a)-(ii)). While these approaches enhance robustness, they have notable limitations. Feature reconstruction may introduce irrelevant or unnecessary features that do not contribute to emotion recognition and become a bottleneck. Knowledge distillation, which encourages a student model trained on incomplete data to mimic a teacher trained on full data [20, 21], often results in trade-offs such as reduced performance under fully observed inputs. Importantly, both approaches require access to complete multimodal sequences during training, which is rarely feasible in practical settings.

Recently, large vision-language models (VLMs) have demonstrated general multimodal capabilities, including limited emotion recognition ability [25]. Furthermore, datasets, methods, and learning approaches have been proposed to enhance the MER capability of VLM [9, 26, 27]. However, these models are not tailored to MER tasks and exhibit lower accuracy and higher computational demands, limiting their use in real-world, resource-constrained environments.

To address these limitations, we propose a novel framework called **Dual-Query Fusion (DQF)**, which is designed to learn robust multimodal representations using only incomplete sequences during training (Figure 1(b)). Unlike previous methods that rely on auxiliary tasks, DQF introduces two types of learnable queries to enhance both feature extraction and cross-modal fusion. First, we employ a query-guided encoder (Q-UA), which incorporates learnable queries into the encoding process. These queries attend to task-relevant features in the observed inputs, acting as a semantic abstraction mechanism that mitigates the impact of incomplete data. Second, we introduce a query-based cross-modal attention module (Q-CA), which facilitates robust interaction between modalities via learnable queries. Those designs are inspired by query-driven architectures such as DETR [6] and Perceiver IO [16], but tailored specifically for the challenges of MER with incomplete inputs. Using queries to improve robustness against missing modalities while simultaneously promoting cross-modal interaction represents a distinct purpose from prior query-based approaches, and to the best of our knowledge, no existing work has employed queries in this manner.

We evaluate the proposed DQF framework on two public benchmark datasets. Extensive experiments demonstrate that DQF consistently outperforms prior state-of-the-art methods in both complete and incomplete modality scenarios. Our method achieves high performance even when trained exclusively on incomplete data, demonstrating strong robustness and real-world applicability.

2. Related Work

2.1. MER in the Complete Modality Condition

MER integrates text, audio, and visual modalities, whose heterogeneous cues make alignment challenging. Existing approaches can be broadly categorized into two paradigms.

Fusion-based methods directly combine multimodal features via attention or memory mechanisms. TFN [54] uses a Cartesian tensor product to capture higher-order interactions. MFM [47] employs memory for feature fusion, while MulT [48] applies directional cross-modal attention to adapt temporal features without explicit alignment.

Decoupling-based methods disentangle modality-invariant and modality-specific representations. MISA [13]

projects each modality into shared and specific subspaces, while DMD [22] separates exclusive and irrelevant components, leveraging dynamic graph distillation for cross-modal transfer.

2.2. MER in the Incomplete-modality Conditions

Since real-world data is often incomplete, recent studies address this challenge through reconstruction-based and distillation-based approaches.

Reconstruction-based methods recover missing modalities from the available ones. DiCMoR [50] and IMDer [51] use generative models to restore distributional and semantic consistency, while TFR-Net [53] and EMT-DLFR [43] employ Transformer-based architectures with local and global context. Although robust, these methods face inherent difficulties in recovering fine-grained cues, risk generating redundant or irrelevant features [33], may underuse discriminative unimodal signals [23, 31], and often assume fixed missing patterns [50, 51].

Distillation-based methods transfer knowledge from a teacher trained on complete data to a student trained on incomplete data. CorrKD [21] introduces contrastive and prototype-guided distillation, while UMDF [20] leverages multigranular attention and dynamic integration. While effective, these methods often underperform in fully observed conditions, as the student model prioritizes generalization on incomplete inputs, limiting exploitation of complete modality information.

Most prior methods still rely on complete data during training, restricting their real-world applicability. To overcome this, we propose a robust framework that directly integrates features from incomplete modality inputs.

2.3. Multimodal Learning with Learnable Queries

Recent multimodal models employ learnable queries to extract and align modality-specific features, particularly in vision–language settings. BLIP-2 [19] uses a Q-Former to distill image features into language-aligned embeddings for frozen LLMs, and InstructBLIP [10] extends this with instruction-aware queries for task-conditioned representation learning. These methods align inputs with LLMs but mainly support unidirectional mapping rather than joint multimodal fusion.

The paradigm also appears in other domains: DETR [6] employs fixed queries to extract object-level features, while Perceiver IO [16] uses query arrays to map latent features to structured outputs. These works show that query mechanisms can abstract task-relevant information independently of input completeness or structure—an important property for robust multimodal learning under missing inputs.

In contrast, our method leverages learnable queries not only for semantic abstraction but also to guide unimodal representation learning and cross-modal fusion. Unlike

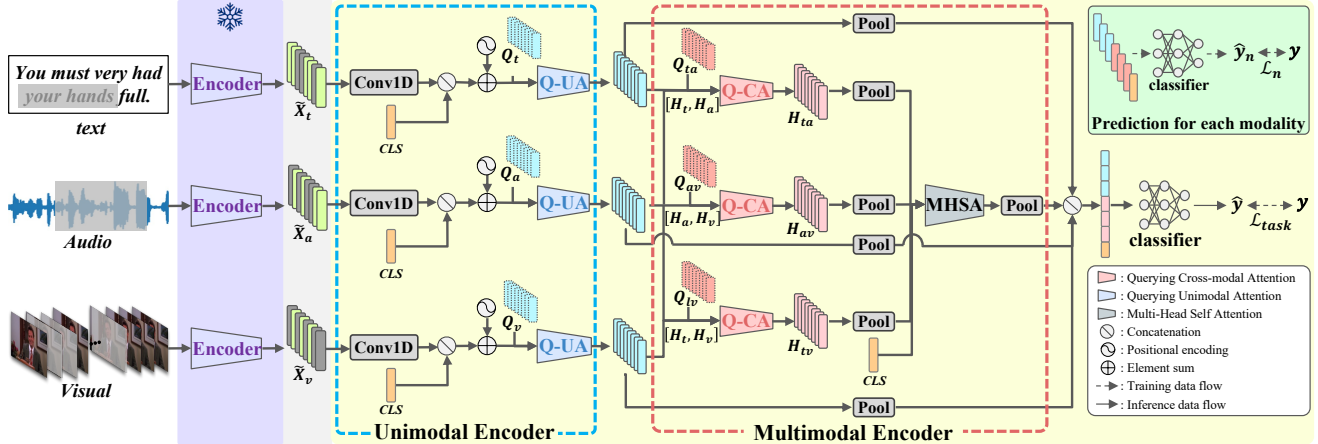


Figure 2. Overview of the DQF: Pretrained encoders first extract embeddings from each modality. Q-UA then uses learnable queries to derive informative unimodal features. Q-CA integrates these via cross-modal attention with symmetric query interaction. The fused representations are processed by MHSA, and predictions are made using multi-granularity features for robustness to missing modalities.

prior methods, which introduce queries mainly to bridge modalities or acquire class-specific representations, we use them to extract discriminative features from incomplete or redundant inputs and to mitigate modality overfitting. This query-driven design enables robust and flexible multimodal learning under incomplete modality conditions.

3. Proposed Method

In incomplete MER, it is essential to improve robustness to diverse missing modality patterns while effectively integrating heterogeneous modalities, which is a challenge unique to MER. However, many existing approaches rely on complete inputs during training, limiting their real-world applicability. To overcome these issues, we propose Dual-Query Fusion (DQF), which introduces two types of learnable queries: Q-UA for robust unimodal representation learning and Q-CA for cross-modal integration. Our model is trained entirely on incomplete inputs.

3.1. Problem Formulation

Multimodal emotion recognition (MER) typically takes three modalities as input: text, audio, and visual. We denote the input as $S = [X_t, X_a, X_v]$, where $X_t \in \mathbb{R}^{T_t \times d_t}$, $X_a \in \mathbb{R}^{T_a \times d_a}$, and $X_v \in \mathbb{R}^{T_v \times d_v}$ represent the feature sequences for text, audio, and visual modalities, respectively. Here, T_m and d_m denote the sequence length and embedding dimension of modality $m \in \{t, a, v\}$.

To simulate incomplete modality conditions commonly encountered in real-world scenarios, we apply a random masking strategy to each input modality. The masked input is denoted as $\tilde{X}_m = F(X_m, g_m) \in \mathbb{R}^{T_m \times d_m}$, where $F(\cdot)$ is a masking function and g_m is the masking ratio that determines the proportion of positions to be masked. Masked positions are replaced with zero vectors. The masking ratio

ratio g_m is randomly sampled in the range $0 \leq g_m \leq 1$ for each modality and mini-batch. We formalize the MER task under incomplete modality conditions, where models must learn from partially observed inputs.

3.2. Overall Architecture

An overview of the proposed method is illustrated in Figure 2. First, the embedded representations of each modality are extracted using pretrained encoders. Here, the parameters of these encoders are fixed. The extracted features are then input into the Querying Unimodal Attention (Q-UA), along with learnable queries, to extract modality-specific features. Next, the obtained features are input into the Querying Cross-Attention (Q-CA), which integrates features between two modalities. The integrated features are then passed to multi-head self-attention (MHSA) to capture interactions across all modalities. By integrating multimodal features with different granularities, the proposed method reduces the impact of modality heterogeneity. Finally, the features extracted at each granularity are concatenated and passed through an MLP for emotion recognition.

In the training phase, we introduce multiple emotion recognition objectives corresponding to different levels of feature integration. This strategy encourages each granularity (i.e., unimodal, bimodal, and trimodal combinations) to learn discriminative representations independently, enhancing robustness to various missing modality patterns while also promoting effective cross-modal interactions at each level. Specifically, features at each level are passed through their own MLP for prediction.

The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \sum_{n \in m_{\text{all}}} \mathcal{L}_n, \quad (1)$$

where $\mathcal{L}_{\text{task}}$ denotes the main loss computed from the final concatenated feature used for the primary emotion recognition output. \mathcal{L}_n denotes auxiliary losses computed from intermediate representations for each modality combination $m_{\text{all}} = \{t, a, v, ta, tv, av, tav\}$. These losses help guide the network to learn informative representations at each fusion level. We employ either cross-entropy loss or mean absolute error (MAE) depending on the dataset setting.

3.3. Unimodal Encoder Using Q-UA

Extracting meaningful features from incomplete inputs is challenging, as critical components may be missing. To address this, the proposed method introduces Querying Unimodal Attention (Q-UA), which learns robust modality-specific feature representations from incomplete inputs by leveraging learnable queries.

In the unimodal encoder, the embedded representation of each modality is first passed through a one-dimensional (1D) temporal convolution layer (Conv1D) to capture local temporal patterns:

$$\hat{X}_m = \text{Conv1D}(\tilde{X}_m). \quad (2)$$

The resulting features \hat{X}_m and the learnable queries $Q_m \in \mathbb{R}^{T_m \times d_m}$ are then fed into the Q-UA module.

An overview of Q-UA is shown in Figure 3-(a). Q-UA uses the incomplete features and modality-specific queries as inputs. First, the learnable queries are updated using MHSA followed by layer normalization:

$$\hat{Q}_m = \text{LayerNorm}(\text{MHSA}(Q_m) + Q_m). \quad (3)$$

These updated queries are then added element-wise to the input features and passed into a transformer encoder:

$$H_m = \text{Transformer}(\hat{X}_m + \hat{Q}_m). \quad (4)$$

The learnable queries in Q-UA function as task-adaptive extractors that guide the model toward informative patterns in incomplete inputs. Unlike conventional cross-attention mechanisms, in which queries directly attend to input features, Q-UA instead adopts an additive interaction where the updated query is element-wise added to the input before each transformer block.

Under incomplete-modality conditions, directly attending to inputs via self-attention can amplify noise or propagate uncertainty from missing regions. Instead, the learnable query serves as an adaptive signal that guides the encoder toward task-relevant features, while mitigating the influence of noisy or uncertain input regions.

The learnable query is dynamically updated at each layer through self-attention, allowing it to adapt to task objectives and context. This iterative update-and-injection process enables Q-UA to compensate for missing content, focus on

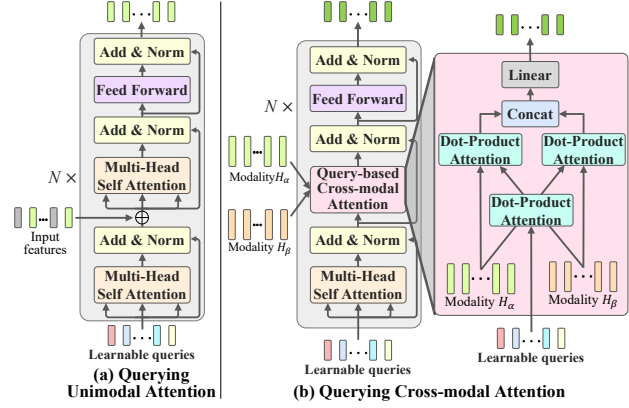


Figure 3. Overview of Q-UA and Q-CA.

task-relevant features, and suppress modality-specific noise. As the model is trained on diverse missing patterns, the queries are progressively optimized to highlight informative features and filter out irrelevant ones, resulting in robust unimodal representations even under severe modality incompleteness.

3.4. Multimodal Encoder Using Q-CA

While multimodal information can improve emotion recognition performance, modality heterogeneity poses a significant challenge. Differences in modality-specific characteristics can lead to inconsistent feature distributions across modalities, which complicate alignment and result in redundant or even conflicting representations during fusion. As a result, naive fusion strategies often cause the model to over-rely on dominant modalities, leading to task-irrelevant representations and overfitting to spurious correlations, which ultimately hinders generalization [38, 45]. This issue arises because features from a strong modality can dominate or interfere with those from weaker modalities during joint optimization. Consequently, interactions between modalities must be modeled in a way that both mitigates modality imbalance and encourages the extraction of complementary information.

To address modality imbalance and redundant interactions, we introduce Querying Cross-Attention (Q-CA), which enables cross-modal fusion via a shared learnable query. Instead of relying on direct feature-to-feature attention, Q-CA lets each modality interact with the shared query, encouraging mediated and balanced interactions. This design mitigates over-reliance on dominant modalities and allows complementary features from weaker ones to be preserved more effectively in the joint representation.

An overview of Q-CA is shown in Figure 3-(b). The input to Q-CA consists of features H_α and H_β from two modalities α and β , and a learnable query $Q_{\alpha\beta}$. First, the query is updated using self-attention as in Equation 3. The

updated query then sequentially attends to H_α and H_β , using them as keys while keeping the query as the value, so that the shared query representation is modulated by each modality and encodes their cross-modal dependencies.

$$\begin{aligned} H_{\alpha \rightarrow \beta} &= \text{Q-CA}_{\alpha \rightarrow \beta}(H_\alpha, H_\beta, Q_{\alpha\beta}), \\ &= \text{Attention}(\text{Attention}(Q_{\alpha\beta}, H_\alpha, Q_{\alpha\beta}), H_\beta, H_\beta), \end{aligned} \quad (5)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{(QW_q)(KW_k)^\top}{\sqrt{d_k}}\right)(VW_v), \quad (6)$$

where W_q , W_k , and W_v are projection matrices and d_k is the key dimension. The resulting features $H_{\alpha \rightarrow \beta}$ and $H_{\beta \rightarrow \alpha}$ are then concatenated and passed through a linear layer to produce the fused representation $H_{\alpha\beta}$:

$$H_{\alpha\beta} = \text{Linear}([H_{\alpha \rightarrow \beta}, H_{\beta \rightarrow \alpha}]). \quad (7)$$

After computing $H_{\alpha\beta}$ for each modality pair $\{ta, tv, av\}$, we apply average pooling to obtain compact fusion features:

$$\tilde{H}_{\alpha\beta} = \text{AvgPool}(H_{\alpha\beta}). \quad (8)$$

The pooled features from each pair, along with the [CLS] token, are then combined and processed by MHSA to capture higher-order cross-modal interactions:

$$\hat{H}_{ca} = \text{Pool}(\text{MHSA}([\tilde{H}_{ta}, \tilde{H}_{tv}, \tilde{H}_{av}, [\text{CLS}]))). \quad (9)$$

This hierarchical fusion strategy, which uses Q-CA for pairwise alignment and MHSA for global fusion, enables the model to better handle modality heterogeneity and extract coherent multimodal representations even in the presence of noisy or unbalanced modalities.

4. Experiments

4.1. Dataset and Evaluation Metrics

While several datasets exist for multimodal emotion recognition [4, 5, 29, 41, 57], we chose CMU-MOSEI [57] and MELD [41] for our experiments, as they are large-scale and serve as representative benchmarks for regression and classification tasks, respectively.

CMU-MOSEI consists of 22,856 video clips, with 16,326 samples for training, 1,871 for validation, and 4,659 for testing. Each sample is annotated with a sentiment score ranging from -3 to +3. We used three evaluation metrics: 7-class accuracy (Acc.7), binary accuracy (Acc.2), and mean absolute error (MAE).

MELD is a multiparty conversational emotion recognition dataset consisting of 13,708 utterances across 1,433

conversations. Each utterance is annotated with one of seven basic emotions. We evaluated performance using weighted average recall (WAR) and unweighted average recall (UAR), where WAR reflects general classification accuracy and UAR accounts for class imbalance.

In addition to conventional evaluation metrics, we report the Area Under Indicators Line Chart (AUILC) [57] to summarize overall model performance across different levels of modality incompleteness.

4.2. Implementation Details

Feature Extraction. To ensure fair comparison and reproducibility, we used the standardized MERBench [24] feature set, adopting the same pre-extracted features for all methods. This isolates the effect of the fusion mechanism from variability due to feature extractors. Visual features were obtained by cropping facial regions with OpenFace [3] and encoding them using the CLIP-pretrained ViT [11, 42]; audio features were extracted with Wav2Vec [2]; and textual features were generated using Baichuan-13B, a large-scale model on HuggingFace.

Experimental Setup. All models were implemented in PyTorch [37] and trained on an NVIDIA RTX 3090 GPU. To ensure fair comparison, we re-implemented both the proposed and baseline methods under complete and incomplete modality settings using official or public codebases. DQF was trained using Momentum SGD [44], with fixed hyperparameters across datasets: learning rate of 1e-5, batch size of 64, 100 epochs, six layers for Q-UA and Q-CA, and four attention heads in MHSA.

To ensure a comprehensive and fair evaluation, we compared our method against 14 reproducible SOTA methods. These include 9 methods designed for complete modality inputs and 5 recent methods tailored for incomplete modality scenarios. We followed the recommended training protocols for each method to maintain a fair comparison.

To evaluate the robustness of each method, we simulate modality incompleteness by randomly masking segments at the frame level within each modality’s sequence. The missing frames are replaced with zero vectors, a widely adopted strategy in prior MER studies [20, 21, 43, 50, 53], as it explicitly simulates the absence of observations (e.g., due to sensor failure or data loss). The missing rate is systematically varied as $\{0.0, 0.1, 0.2, \dots, 1.0\}$, enabling the evaluation of a broad spectrum of incompleteness conditions, including both partial and complete modality-level absence. We evaluate all seven combinations of modality incompleteness: one modality missing (3 cases), two modalities missing (3 cases), and all three modalities missing (1 case). For completeness, we include results up to the 1.0 missing rate, though the configuration where all modalities are entirely missing (i.e., 1.0 for all) is excluded from evaluation. This experimental design not only covers established robustness

testing settings but also extends them by offering a more comprehensive and fine-grained evaluation across varying missing patterns and severity levels. Final results are averaged over five runs with different random seeds.

4.3. Comparison with SOTA Methods under the Complete Modality Setting

To evaluate the performance on a standard MER task, we compared the methods when complete multimodal sequences were provided as input. Table 1 shows performance across the two benchmark datasets under the complete modality setting.

Results on CMU-MOSEI. On CMU-MOSEI, DQF achieves an Acc.7 of 55.6%, an Acc.2 of 86.9%, and a MAE of 0.520. Compared to the best-performing methods designed for incomplete modality conditions, DQF achieves improvements of +3.0% in Acc.7, +0.4% in Acc.2, and a reduction of 0.029 in MAE. It also outperforms the strongest complete-modality model by +1.8% in Acc.7, +0.2% in Acc.2, and a reduction of 0.012 in MAE.

Results on MELD. On MELD, DQF obtains a WAR of 63.1% and a UAR of 40.4%. Compared to the strongest methods designed for incomplete modalities, DQF improves by +2.0% in WAR and +0.5% in UAR. It also exceeds the performance of the best complete-modality model, with gains of +1.2% in WAR and +0.8% in UAR.

Overall findings. Models trained on complete inputs, such as MulT [48], MISA [13], and DMD [22], tend to perform well when evaluated under full modality conditions. In contrast, models designed for incomplete modality conditions generally perform worse than complete-modality models in this setting, likely due to their training being focused on handling degraded or missing inputs. Furthermore, VLMs [9, 26] exhibited lower accuracy than most other methods, underscoring the challenge of adapting them to MER tasks. As presented in the supplementary material, the proposed method achieves lower model parameters, computational cost, and inference time than SOTA methods in most cases, thereby confirming its efficiency. Despite being trained only on incomplete modality inputs, DQF achieves the best performance on both datasets. These results demonstrate its strong generalization ability to fully observed inputs.

4.4. Comparison with SOTA Methods under the Incomplete Modality Setting

To evaluate robustness, we compared models under varying incomplete modality configurations and missing rates. We considered seven cases, as detailed in Section 4.2. In this experiment, we comprehensively evaluated the model under simulated missing data. Specifically, we considered four settings formed by combining two types of missing patterns and two types of missing-value representations. The

Table 1. Performance comparison of all methods on CMU-MOSEI and MELD, evaluated under the complete modality setting. The results are grouped into two blocks according to training conditions: (1) methods trained on complete inputs and (2) methods explicitly trained to handle incomplete inputs. Additionally, methods marked with * cite results reported in each paper. Each result reports Acc.7 / Acc.2 / MAE for CMU-MOSEI, and WAR / UAR for MELD. Higher is better for all metrics except MAE.

	Models	CMU-MOSEI	MELD
(1)	TFN [54]	48.2 / 83.5 / 0.598	59.6 / 32.5
	MFN [55]	49.6 / 83.7 / 0.600	59.7 / 31.2
	Graph-MFN [57]	49.3 / 84.5 / 0.589	58.4 / 30.5
	LMF [30]	51.8 / 86.6 / 0.545	48.1 / 14.3
	MCTN [39]	53.0 / 85.8 / 0.544	51.5 / 20.0
	MFM [47]	49.9 / 82.4 / 0.600	51.7 / 20.7
	MulT [48]	53.8 / 86.6 / 0.532	61.7 / 39.6
	MISA [13]	53.3 / 84.7 / 0.543	52.7 / 22.7
	DMD [22]	47.1 / 86.7 / 0.621	61.9 / 33.9
	Emo-LLaMA* [9]	- / 67.7 / -	46.8 / -
AffectGPT* [26]	- / 80.9 / -	56.7 / -	
(2)	TFR-Net [53]	51.2 / 85.1 / 0.559	61.0 / 39.9
	DiCMoR [50]	47.2 / 79.3 / 0.652	53.7 / 26.7
	EMT-DLFR [43]	51.7 / 84.0 / 0.567	59.5 / 35.6
	UMDF [20]	52.6 / 84.4 / 0.555	59.7 / 34.5
	CorrKD [21]	52.4 / 86.5 / 0.549	61.1 / 33.0
	DQF(Ours)	55.6 / 86.9 / 0.520	63.1 / 40.4

missing patterns were (1) randomly missing frames and (2) consecutively missing frames. For the missing-value representation, we adopted (1) replacement with zero vectors and (2) replacement with values sampled from a normal distribution. In this paper, we report the results for randomly missing frames with zero-vector replacement, while the results for the other settings are provided in the supplementary material.

Result on CMU-MOSEI. Table 2 presents the AUILC for all seven combinations of incomplete modality configurations. Overall, methods trained on complete inputs tend to outperform those trained under incomplete conditions, reflecting the advantage of full-modality supervision. Nevertheless, our proposed method DQF, trained solely on incomplete inputs, achieves the highest performance across all configurations and all evaluation metrics. In particular, under the most challenging setting where all three modalities are simultaneously subject to missingness ($\{v, a, t\}$), DQF achieves an Acc.7 of 52.9%, exceeding the best baseline by +2.5%.

Result on MELD. Table 3 presents the AUILC under all combinations of incomplete modality configurations. Similar to CMU-MOSEI, methods trained on complete inputs generally outperform those trained under incomplete conditions, benefiting from access to full modality information during training. Despite being trained solely on incomplete inputs, DQF achieves the best performance across all con-

Table 2. Performance under randomly missing modality conditions on CMU-MOSEI. 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						
	$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
TFN [54]	49.8 / 84.5 / 0.574	48.6 / 83.4 / 0.594	45.3 / 79.4 / 0.668	49.8 / 84.6 / 0.577	46.9 / 80.3 / 0.650	45.7 / 78.4 / 0.668	46.8 / 78.5 / 0.659
MFN [55]	48.7 / 83.6 / 0.601	50.0 / 82.7 / 0.591	47.4 / 80.5 / 0.645	49.6 / 84.0 / 0.597	47.1 / 80.0 / 0.650	47.6 / 80.7 / 0.641	47.3 / 80.3 / 0.647
Graph-MFN [57]	49.1 / 84.7 / 0.593	49.4 / 83.9 / 0.591	47.9 / 80.7 / 0.636	48.9 / 83.9 / 0.595	47.7 / 80.3 / 0.641	47.7 / 80.4 / 0.639	47.4 / 79.9 / 0.645
LMF [30]	51.7 / 86.8 / 0.545	51.7 / 86.6 / 0.547	49.6 / 83.2 / 0.603	51.6 / 86.5 / 0.549	49.7 / 83.1 / 0.603	49.7 / 83.2 / 0.604	49.6 / 83.0 / 0.605
(1) MCTN [39]	53.0 / 85.6 / 0.545	53.0 / 85.6 / 0.544	49.9 / 82.0 / 0.604	53.0 / 85.8 / 0.545	50.1 / 82.1 / 0.604	50.1 / 82.2 / 0.603	50.0 / 82.1 / 0.604
MFM [47]	50.9 / 83.6 / 0.581	50.0 / 82.7 / 0.601	47.6 / 78.9 / 0.652	50.8 / 83.6 / 0.582	48.4 / 79.2 / 0.638	47.4 / 79.1 / 0.652	48.2 / 80.0 / 0.635
MuT [48]	53.6 / 86.5 / 0.536	53.7 / 86.6 / 0.533	48.4 / 78.1 / 0.657	53.5 / 86.5 / 0.537	48.4 / 77.2 / 0.659	48.4 / 78.0 / 0.658	48.4 / 77.1 / 0.662
MISA [13]	53.2 / 85.1 / 0.545	53.2 / 84.7 / 0.543	50.0 / 80.7 / 0.602	53.3 / 85.1 / 0.545	49.7 / 80.9 / 0.608	50.0 / 80.7 / 0.602	49.7 / 81.0 / 0.608
DMD [22]	46.4 / 83.1 / 0.630	46.5 / 83.0 / 0.629	45.1 / 78.7 / 0.678	46.2 / 83.3 / 0.630	45.1 / 78.5 / 0.682	45.1 / 78.9 / 0.679	45.1 / 78.3 / 0.684
TFR-Net [53]	51.4 / 85.2 / 0.560	51.2 / 85.2 / 0.560	49.6 / 81.8 / 0.604	51.2 / 85.2 / 0.560	49.8 / 81.8 / 0.603	49.6 / 81.8 / 0.605	49.8 / 81.9 / 0.602
DiCMoR [50]	47.0 / 79.3 / 0.657	47.0 / 79.3 / 0.653	45.7 / 77.1 / 0.683	46.8 / 79.3 / 0.658	45.7 / 76.9 / 0.689	45.6 / 77.0 / 0.684	45.6 / 76.8 / 0.690
EMT-DLFR [43]	51.6 / 84.2 / 0.567	51.6 / 83.9 / 0.567	49.8 / 81.3 / 0.600	51.5 / 84.0 / 0.568	50.0 / 81.5 / 0.600	49.9 / 81.2 / 0.600	50.0 / 81.5 / 0.602
(2) UMDf [20]	52.5 / 84.4 / 0.557	52.5 / 84.4 / 0.556	50.7 / 81.7 / 0.595	52.2 / 84.4 / 0.559	50.6 / 81.2 / 0.600	50.7 / 81.3 / 0.598	50.4 / 81.1 / 0.602
CorrKD [21]	53.1 / 86.8 / 0.541	51.9 / 86.6 / 0.553	49.9 / 82.8 / 0.595	52.7 / 86.8 / 0.546	50.4 / 82.5 / 0.591	49.8 / 83.2 / 0.596	50.2 / 83.3 / 0.592
DQF(Ours)	55.5 / 87.1 / 0.520	55.2 / 86.9 / 0.521	53.2 / 84.4 / 0.562	55.3 / 87.0 / 0.522	53.0 / 84.2 / 0.565	53.0 / 84.1 / 0.564	52.9 / 83.8 / 0.568

Table 3. Performance under randomly missing modality conditions on MELD. 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						
	$\{v\}$	$\{a\}$	$\{t\}$	$\{v, a\}$	$\{v, t\}$	$\{a, t\}$	$\{v, a, t\}$
TFN [54]	60.0 / 32.4	59.9 / 32.0	55.5 / 27.8	60.2 / 31.8	56.1 / 27.7	55.7 / 26.2	55.9 / 25.8
MFN [55]	59.6 / 32.9	58.9 / 33.8	53.1 / 32.8	58.9 / 33.2	53.2 / 32.3	49.6 / 28.1	48.9 / 27.5
Graph-MFN [57]	58.5 / 30.1	56.2 / 25.6	43.0 / 25.2	55.7 / 24.8	43.3 / 24.3	46.7 / 22.1	47.1 / 20.9
LMF [30]	48.1 / 14.3	48.1 / 14.3	48.1 / 14.3	48.1 / 14.3	48.1 / 14.3	48.1 / 14.3	48.1 / 14.3
(1) MCTN [39]	51.5 / 20.0	51.5 / 20.0	49.6 / 16.5	51.5 / 20.0	49.6 / 16.5	49.6 / 16.6	49.6 / 16.5
MFM [47]	51.8 / 21.0	52.1 / 21.6	49.7 / 18.8	52.2 / 21.9	49.6 / 18.9	49.4 / 19.0	49.5 / 19.3
MuT [48]	62.1 / 38.8	60.5 / 39.7	54.1 / 30.2	61.3 / 39.6	54.5 / 28.7	52.3 / 30.6	53.8 / 29.2
MISA [13]	52.7 / 22.7	52.7 / 22.7	51.3 / 20.7	52.7 / 22.7	51.3 / 20.7	51.2 / 20.6	51.3 / 20.7
DMD [22]	61.8 / 33.7	61.9 / 34.0	53.7 / 26.1	61.7 / 33.8	53.5 / 26.1	53.4 / 26.1	53.6 / 26.1
TFR-Net [53]	60.5 / 39.8	60.4 / 39.8	57.4 / 34.1	60.3 / 39.9	57.3 / 34.1	57.1 / 34.1	57.0 / 33.4
DiCMoR [50]	53.5 / 26.7	52.8 / 27.3	51.2 / 22.2	52.9 / 27.1	51.0 / 22.3	50.3 / 22.6	50.0 / 22.4
EMT-DLFR [43]	59.6 / 35.6	59.6 / 35.7	56.2 / 29.1	59.6 / 35.6	56.2 / 29.0	56.2 / 29.1	56.2 / 29.0
(2) UMDf [20]	59.8 / 34.7	60.1 / 34.5	56.3 / 28.9	60.0 / 34.6	56.2 / 28.9	56.3 / 28.6	56.3 / 28.7
CorrKD [21]	61.2 / 33.0	61.2 / 32.8	56.1 / 28.8	61.4 / 32.8	56.1 / 28.7	56.5 / 28.3	56.6 / 28.3
DQF(Ours)	62.9 / 40.3	62.5 / 40.0	59.1 / 36.2	63.0 / 40.0	59.2 / 35.8	58.6 / 35.0	59.0 / 34.9

figurations in both evaluation metrics. Notably, under the most challenging setting where all three modalities are simultaneously subject to missingness ($\{v, a, t\}$), DQF obtains a WAR of 59.0% and a UAR of 34.9%, outperforming the best baseline by +2.0% and +1.5%, respectively.

Overall findings. Despite being trained solely on partially observed inputs, DQF consistently outperforms existing methods, including those trained with full supervision, across all configurations, datasets, and evaluation metrics. These results underscore the strength of our query-based design in capturing task-relevant information under severe modality incompleteness.

4.5. Ablation Study

4.5.1. Effectiveness of Each Component

We conducted an ablation study on CMU-MOSEI to evaluate each component under incomplete modality conditions,

comparing DQF with three variants: (1) **w/o Q-UA**, where Q-UA is replaced by a Transformer encoder; (2) **w/o Q-CA**, where Q-CA is replaced by a Transformer that takes concatenated unimodal features as input; (3) **Q-UA w/ MuT**, where Q-CA is replaced by MuT [48], a widely used cross-modal attention model.

Table 4 shows the AUILC for all incomplete modality settings. Removing Q-CA (w/o Q-CA) leads to larger performance drops than removing Q-UA, indicating the importance of cross-modal query interaction. Replacing Q-CA with MuT improves performance over w/o Q-CA, confirming that explicitly modeling inter-modal attention is effective. However, DQF still performs best, showing the advantage of query-guided fusion over conventional attention strategies.

Q-UA contributes to robust unimodal encoding under missing inputs, while Q-CA enables efficient and selective

Table 4. Ablation study evaluating each component of DQF under incomplete multimodal conditions on CMU-MOSEI. Each item represents Acc.7 / Acc.2 / MAE.

Models	Incomplete modality						
	{v}	{a}	{t}	{v, a}	{v, t}	{a, t}	{v, a, t}
w/o Q-UA	52.7 / 86.3 / 0.551	52.5 / 86.3 / 0.552	50.8 / 83.6 / 0.590	52.7 / 86.3 / 0.552	50.8 / 83.2 / 0.592	50.7 / 83.3 / 0.591	50.8 / 83.2 / 0.593
w/o Q-CA	52.3 / 85.3 / 0.546	52.3 / 85.8 / 0.542	50.5 / 82.9 / 0.584	52.3 / 85.3 / 0.546	50.4 / 81.7 / 0.592	50.4 / 82.8 / 0.585	50.5 / 81.8 / 0.592
Q-UA w/ MulT	53.9 / 84.9 / 0.533	53.9 / 85.0 / 0.533	51.8 / 82.0 / 0.579	53.9 / 84.8 / 0.534	51.6 / 80.9 / 0.586	51.7 / 82.1 / 0.579	51.8 / 80.5 / 0.584
DQF	55.5 / 87.1 / 0.520	55.2 / 86.9 / 0.521	53.2 / 84.4 / 0.562	55.3 / 87.0 / 0.522	53.0 / 84.2 / 0.565	53.0 / 84.1 / 0.564	52.9 / 83.8 / 0.568

Table 5. Comparison of different number of layers and heads in the Q-UA and Q-CA modules under incomplete multimodal conditions on CMU-MOSEI. Each item represents Acc.7 / Acc.2 / MAE.

Layers (L) / Heads (H)	Incomplete modality						
	{v}	{a}	{t}	{v, a}	{v, t}	{a, t}	{v, a, t}
L=2 / H=2	54.6 / 86.5 / 0.524	54.7 / 86.6 / 0.523	52.6 / 84.2 / 0.564	54.4 / 86.5 / 0.526	52.4 / 84.0 / 0.568	52.4 / 83.7 / 0.568	52.3 / 83.5 / 0.572
L=2 / H=4	55.1 / 86.0 / 0.527	55.0 / 86.1 / 0.525	53.0 / 83.8 / 0.565	55.0 / 86.0 / 0.527	52.6 / 83.4 / 0.571	52.9 / 83.4 / 0.568	52.6 / 83.2 / 0.573
L=4 / H=4	54.6 / 86.6 / 0.524	54.5 / 86.4 / 0.526	52.6 / 84.2 / 0.564	54.5 / 86.4 / 0.525	52.4 / 83.9 / 0.567	52.4 / 83.8 / 0.569	52.4 / 83.4 / 0.570
L=6 / H=4	55.5 / 87.1 / 0.520	55.2 / 86.9 / 0.521	53.2 / 84.4 / 0.562	55.3 / 87.0 / 0.522	53.0 / 84.2 / 0.565	53.0 / 84.1 / 0.564	52.9 / 83.8 / 0.568
L=6 / H=6	54.8 / 86.4 / 0.529	55.1 / 86.7 / 0.524	53.1 / 84.3 / 0.565	54.7 / 86.7 / 0.529	52.6 / 83.8 / 0.573	52.9 / 84.0 / 0.567	52.5 / 83.3 / 0.574

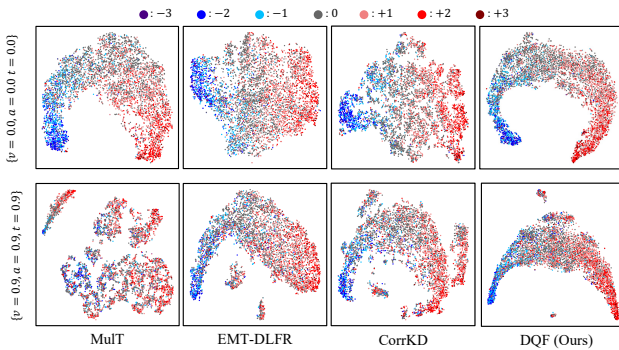


Figure 4. Visualization of the joint multimodal representation.

cross-modal fusion. Together, they support both modality compensation and alignment, which are crucial for robust multimodal learning. The consistent improvements over all ablated variants demonstrate the effectiveness of our query-based design under incomplete and uncertain inputs.

4.5.2. Effectiveness of the Number of Layers and Heads

We investigated the impact of varying the number of layers and attention heads in the Q-UA and Q-CA modules. Table 5 presents the AUIC for all incomplete modality settings. This result shows the effect of varying the number of layers and heads in the Q-UA and Q-CA modules. We observe that deeper architectures consistently improve performance, while excessively increasing the number of heads does not yield further gains. In particular, the configuration with six layers and four heads achieves the best balance, providing the highest accuracy and lowest MAE across most incomplete-modality conditions.

4.6. Qualitative Analysis

To evaluate feature robustness, we visualized t-SNE embeddings and compared DQF with three baselines: MulT [48]

(complete-modality), EMT-DLFR [43] (reconstruction-based), and CorrKD [21] (distillation-based).

Figure 4 shows the t-SNE embeddings under two conditions: (top) complete modality inputs, and (bottom) a severe missing scenario where all modalities are independently masked with a probability of 0.9. Under complete inputs, all models form reasonably well-separated clusters by emotion labels. However, under the missing modality condition, MulT’s feature space becomes notably fragmented, indicating reduced robustness. EMT-DLFR and CorrKD retain more structure, reflecting moderate resilience to missing inputs. DQF, in contrast, maintains a coherent and well-separated embedding space even under severe degradation, demonstrating its strong ability to extract task-relevant features despite high uncertainty. These qualitative trends align well with the quantitative performance, reinforcing that DQF produces more robust and semantically meaningful representations under severe input degradation.

5. Conclusion

We proposed Dual-Query Fusion (DQF), a query-based architecture for robust multimodal emotion recognition under incomplete modality conditions. DQF introduces two learnable queries: Q-UA for unimodal feature extraction and Q-CA for cross-modal integration, enabling strong performance without relying on full-modality supervision. Unlike prior approaches that still require complete data during training, DQF learns solely from incomplete inputs. Experiments on CMU-MOSEI and MELD demonstrate that DQF outperforms state-of-the-art methods and generalizes well to complete inputs, confirming its robustness to modality incompleteness. A current limitation is that our evaluation relies on simulated missing modalities on two benchmarks and focuses only on supervised MER. Validating DQF on real-world incomplete data and broader affective tasks is left for future work.

References

- [1] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, 2020. 1
- [2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 5
- [3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 5
- [4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 5
- [5] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016. 5
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1
- [8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1
- [9] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024. 2, 6
- [10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [12] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, page 2225. NIH Public Access, 2018. 1
- [13] 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. 2, 6, 7
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1
- [15] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1
- [16] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. 2
- [17] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dflew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020. 1
- [18] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 1
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [20] Mingcheng Li, Dingkan 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, 2, 5, 6, 7
- [21] Mingcheng Li, Dingkan 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, 2, 5, 6, 7, 8
- [22] 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. 2, 6, 7
- [23] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete mul-

- timodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023. 1, 2
- [24] 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. 5
- [25] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024. 2
- [26] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025. 2, 6
- [27] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Forty-second International Conference on Machine Learning*, 2025. 2
- [28] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8148–8156, 2021. 1
- [29] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Chsims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258, 2022. 5
- [30] 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. 6, 7
- [31] Wei Luo, Mengying Xu, and Hanjiang Lai. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *International conference on multimedia modeling*, pages 411–422. Springer, 2023. 2
- [32] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562, 2021. 1
- [33] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2302–2310, 2021. 2
- [34] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011. 1
- [35] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81, 2021. 1
- [36] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014. 1
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5
- [38] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. 4
- [39] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6892–6899, 2019. 6, 7
- [40] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017. 1
- [41] 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. 5
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [43] 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. 2, 5, 6, 7, 8
- [44] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 5
- [45] Antonio Tejero-de Pablos. Complementary-contradictory feature regularization against multimodal overfitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5679–5688, 2024. 4
- [46] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017. 1

- [47] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. 2019. [2](#), [6](#), [7](#)
- [48] 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](#), [2](#), [6](#), [7](#), [8](#)
- [49] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022. [1](#)
- [50] 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. [2](#), [5](#), [6](#), [7](#)
- [51] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [52] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814, 2021. [1](#)
- [53] 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. [2](#), [5](#), [6](#), [7](#)
- [54] 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. [2](#), [6](#), [7](#)
- [55] 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. [1](#), [6](#), [7](#)
- [56] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. [1](#)
- [57] 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. [5](#), [6](#), [7](#)