Label-Guided Enhancement: A Distillation Framework for Uncertainty-Aware Multimodal Emotion Recognition

Abstract

Multimodal Emotion Recognition (MER) is a vital technology for capturing nuanced human emotions by integrating complementary textual, acoustic, and visual cues. However, real-world MER systems frequently encounter issues with missing or conflicting modalities, arising from sensor failures, privacy constraints, or contradictory emotional signals. These issues compromise the efficacy of existing attention-based fusion models. In this paper, we propose CUMDF, a ***C**ounterfactual-based **U**ncertain **M**issing Modality **D**istillation **F**ramework* that addresses these challenges through three core innovations. Specially, we introduce a Label-Guided Multimodal Masked Transformer (LG-MMT) to align features with target sentiment semantics and improving robustness under incomplete or conflicting data. Furthermore, we design the Adaptive and Generalized Knowledge Extractors to disentangle modality-specific information from shared cross-modal patterns, enhancing representational diversity and coherence. Finally, we design a Modality Attribution-based Counterfactual Inference (MACI) mechanism that quantifies each modality's causal contribution via counterfactual predictions and dynamically adjusts distillation weights to focus the student model on under-optimized modalities. Experimental results on three benchmark datasets demonstrate that CUMDF outperforms state-of-the-art approaches, highlighting the importance of uncertainty modeling in MER.

1 Introduction

With the explosive growth of user-generated content on social media platforms, multimodal signals (i.e. text, audio and visual) have become indispensable for the accurate inference of sentiment and emotion. Multimodal Emotion Recognition (MER) employs a combination of cues from these streams to enhance the accuracy of inference in comparison to unimodal systems, with extensive applications in human-computer interaction, virtual assistants and affective computing in the fields of healthcare and education.

Despite the noteworthy advancements, two under-explored challenges impede the robustness and generalizability of existing MER systems. Firstly, the absence of modalities is a common occurrence in real-world implementations due to various factors, including sensor failures, bandwidth limitations, privacy filters, and occlusion. This results in irregular availability of modalities. Models that have been trained on fully observed data encounter difficulties when attempting to generalize under these incomplete conditions. Second, modality conflicts—scenarios where different modalities convey contradictory sentiments (e.g., cheerful speech paired with a frustrated facial expression)—are prevalent in spontaneous human communication. It is evident that standard fusion strategies are often insensitive to such conflicts, resulting in semantically incoherent feature representations and compromised sentiment prediction.

To address these challenges, researchers have explored various approaches to enhance MER robustness. Early efforts focused on simple fusion techniques using handcrafted features with classical classifiers like SVMs (Rozgic et.al, 2012; Wang et.al, 2014) and ensemble trees (Cummins et.al, 2018). However, these methods proved incapable of capturing cross-modal dependencies, which are critical for emotion recognition (Smith et al., 2017). The advent of deep learning gave rise to a more sophisticated array of fusion paradigms. These include early fusion (Morency et al., 2011;

Perez et al., 2013; Yu et al., 2021) which aggregates raw embeddings, and late fusion (Shutova et al., 2016; Morvant et al., 2014; Evangelopoulos et al., 2013) which combines modality-specific predictions. Nevertheless, these methodologies continue to encounter challenges in effectively handling complex interactions between modalities.

More recently, attention-based mechanisms have become increasingly prominent. These include global attention modules (Zhang et al., 2020) and bi-GRU multi-attention (Lian et al., 2021), which attempt to focus on salient multimodal features. Similarly, modality-weighted fusion techniques, including gated networks (Arevalo et al., 2017) and self-adaptive path selection (Yang et al., 2023), have been proposed to dynamically adjust each modality's contribution. Mai et al. (Mai et.al, 2023) introduced intra-modal and inter-modal comparative and semi-comparative learning. This instructional strategy is employed to facilitate a comprehensive exploration of cross-modal interactions, to ensure the maintenance of inter-class relationships, and to address any existing modal gaps. Despite their sophistication, these approaches typically assume complete modality availability and lack mechanisms to handle missing or conflicting signals.

With regard to the specific issue of modality missingness, knowledge distillation (KD) has been identified as a promising approach. MCTN (Pham et al., 2019) employs cycle-consistent translation between modalities with a view to enhancing representation consistency across different input combinations. In alternative approaches, the reconstruction of absent modalities is undertaken at the feature level (Sun et al., 2023) or through the estimation of these modalities by means of completion-based KD (Sun et al., 2024). Further refinements have been proposed, including margin-aware distillation (Wei et al., 2023) to regularize modality importance, graph-based KD (Deng et al., 2025) to aggregate multi-source information, and hierarchical cross-modal distillation (Wei et al., 2024) to improve teacher-student alignment. Zhang et al. (2024) proposed a novel approach to address the challenges posed by absent modality and optimization imbalance, namely sample-weighted distillation and prototype regularization network. While these approaches have advanced the field, they still suffer from three fundamental limitations when dealing with real-world MER scenarios:

Lack of semantic guidance: Existing methods generally perform cross-modal alignment without fine-grained label-aware supervision, rendering them susceptible to spurious correlations in noisy modalities. In the absence of explicit semantic anchoring to emotional content, models are unable to distinguish relevant features from noise, particularly when modalities provide conflicting signals.

Inadequate knowledge disentanglement: Current approaches fail to effectively separate modality-specific knowledge from shared information. This limits their ability to preserve unique modality features while exploiting cross-modality commonalities, which is critical for robust performance when certain modalities are missing or unreliable.

Consistent distillation strategies: Most methods apply identical distillation weights across all modalities and samples, ignoring the varying power and reliability of each modality. This becomes problematic when dealing with conflicting modalities or different missing patterns across samples.

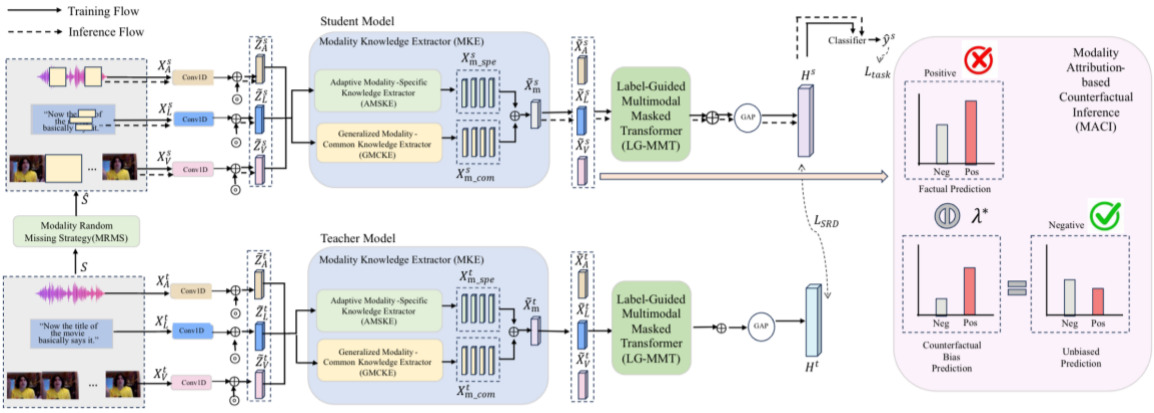


Figure 1: The proposed CUMDF consists of three core components: (1) the Modality Domain Knowledge Extractor (2) the Label-Guided Multimodal Masked Transformer, and (3) the Modality Contribution-based Counterfactual Inference.

To address these limitations, the Counterfactual-based Uncertain Missing Modality Distillation Framework (CUMDF) is proposed, enhancing MER performance under incomplete or conflicting modalities. As demonstrated in Figure 1, CUMDF introduces three key innovations:

First, we propose a Label-Guided Multimodal Masked Transformer (LG-MMT), which employs learnable label embeddings as query inputs within the attention mechanism. In contrast to previous fusion methodologies, the LG-MMT model employs audio-text-label interactions with label embeddings as anchors to guide emotion feature fusion. This approach focuses attention on emotion-relevant features to stabilize cross-modal fusion and address issues related to missing or inconsistent signals.

Secondly, we propose the Adaptive Modality-Specific (AMSKE) and the Generalized Modality-Common (GMCKE) Knowledge Extractors, which decompose modality features into discriminative, modality-specific traits and cross-modal commonalities. The explicit modelling of private and shared knowledge has been shown to enhance feature expressiveness and compatibility during fusion, particularly in the presence of noisy or absent streams.

Finally, the Modality Attribution-based Counterfactual Inference (MACI) module adapts distillation by estimating each modality's causal impact via full-model and counterfactual prediction comparisons, dynamically adjusting distillation weights to promote fairness and enhance under-optimized modal streams.

2 Methods

2.1 Preliminary

Multimodal Emotion Recognition (MER) is commonly framed as a regression task. Given a full-modality video sample set $S = [X_L, X_A, X_V]$, where $X_L^t \in \mathbb{R}^{T_L * d_L}$, $X_V^t \in \mathbb{R}^{T_V * d_V}$,and $X_A^t \in \mathbb{R}^{T_A * d_A}$, respectively. And t represents the teacher model, T_m is the sequence length (temporal dimension) and d_m is the embedding dimension of modality $m \in \{L, A, V\}$. Meanwhile, we further extend the traditional MER task to encompass modality-missing scenario, the modality missing version features are denoted as X_m^s, where s represents the student model, $m \in \{L, A, V\}$. Our goal is to find the utterance-level sentiments by employing the multimodal data with missing modalities.

To handle intra-modality missingness, we develop the Modality Random Missing Strategy (MRMS), which generates datasets by dropping frame-level features with a drop ratio p (0 to 0.7, 0.1 increments). Unlike prior work (Yuan et.al, 2021), this ensures at least one modality remains per sample for balanced evaluation, replacing missing segments with zero vectors. Incomplete features X_m^s (for $m \in \{L, A, V\}$) are then processed by the student model similarly to the teacher model (as shown in Eq. 1):

$$X'^t_m = Conv1D(X_m^t, k_m^t) + PE(T, d) \in \mathbb{R}^{T*d} \quad (1)$$

where PE denotes to the position in the sequence and d represents the feature dimension.

2.2 Modality Knowledge Extractor (MKE)

To comprehensively represent multimodal inputs, the framework uses Adaptive Modality-Specific (AMSKE) and Generalized Modality-Common (GMCKE) Knowledge Extractors to decompose unimodal features into modality-specific traits ($X_{m\_spe}^t$) and transferable cross-modal knowledge ($X_{m\_com}^t$). Formally, the teacher model's extractors are defined as:

$$X_{m\_spe}^t = encoder_{m\_spe}(X'^t_m; \theta_{m\_spe}) \quad (2)$$

$$X_{m\_com}^t = encoder_{m\_com}(X'^t_m; \theta_{m\_com}) \quad (3)$$

where $encoder_{m\_spe}$ and $encoder_{m\_com}$ denote the AMSKE and GMCKE, respectively. In our framework, AMSKE employs a 1-layer Transformer to learn sequential information of modalities, and GMCKE is a fine-grained mamba block. We use the language modality as an example to present the structure of GMCKE, which is calculated as follows:

$$\begin{aligned} M_{LA}^t &= X'^t_L X'^{t\,T}_A \\ S_{LA}^t &= Softmax(tan(M_{LA}^t)) \\ R_{LA}^t &= S_{LA}^t X'^t_A \\ \tilde{X}_{LA}^t &= R_{LA}^t \oplus (R_{LA}^t \odot X'^t_L) \\ X_{L\_com}^t &= \tilde{X}_{LA}^t \oplus \tilde{X}_{LV}^t \end{aligned} \quad (4)$$

where $\odot$ represents element-wise multiplication, $\oplus$ denotes feature concatenation. The final output

of the modality domain knowledge extractor $\tilde{X}_m^t$ is a combination of $X_{m\_spe}^t$ and $X_{m\_com}^t$.

2.3 Label-guided Multimodal Masked Transformer (LG-MMT)

In traditional MER methods, Transformers are often used for cross-modal relationship modeling. However, when modalities are absent or in semantic conflict, Transformers may encounter challenges due to an absence of global semantic guidance. To solve this, we propose the Label-guided Multimodal Masked Transformer (LG-MMT), as shown in Figure 2, which uses label embeddings for each sample as semantic anchors to guide attention during fusion.

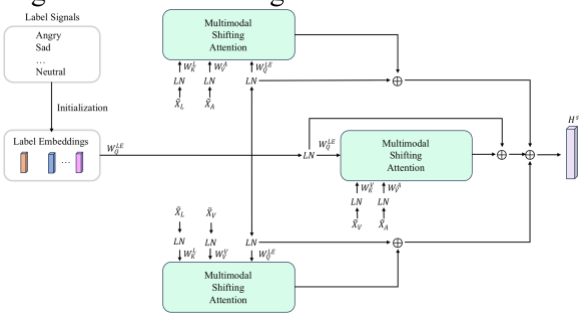


Figure 2: Structure of the proposed LG-MMT

2.3.1 Label Embedding Generation

For each sample i with a label $y_i \in \{1, \dots, C\}$, we map it into a dense vector $e_i \in \mathbb{R}^d$ using a learnable label embedding matrix $E_y \in \mathbb{R}^{C*d}$:

$$e_i = E_y[y_i] \tag{4}$$

here, y_i denotes the sample label (discrete class like "happy" or "sad"), with label embedding $e_i \in \mathbb{R}^d$ encoding its semantic information. This embedding is crucial for guiding the attention mechanism in subsequent layers.

2.3.2 Fusion of Label Embedding with Modality Features

Let $\tilde{Z}_{m1}^i \in \mathbb{R}^d$ and $\tilde{Z}_{m2}^i \in \mathbb{R}^d$ be the feature sequences of modalities $m1$ and $m2$ with $m1, m2 \in \{L, A, V\}$, respectively. The label embedding e_i is fused with the modality features to obtain the label-enhanced features. For example, to obtain the label-enhanced language feature $\tilde{Z}_L^i$, e_i is repeated along the time dimension T_L to align with the temporal dimension of the language feature:

$$\tilde{Z}_L^i = \tilde{X}_L^t \odot (W_e * e_i + b_e) \tag{5}$$

where W_e and b_e are learnable matrix and bias vector. In this step, the label embedding is broadcast across the entire sequence, enabling the model to use the label's semantic information throughout the sequence.

2.3.3 Label-guided Attention Computation

Label-enhanced features $\tilde{Z}_{m1}^i$ and $\tilde{Z}_{m2}^i$ are fed into the Label-guided attention, where the label embedding e_i acts as query Q, $\tilde{Z}_{m1}^i$ and $\tilde{Z}_{m2}^i$ are served as key K and value V. For instance, label-guided cross-attention between language and acoustic modalities focuses on features aligning with label sentiment, enhancing multimodal fusion:

$$\begin{aligned} Q &= W_Q^{LE} * e_i \\ K_L &= W_K^L * \tilde{Z}_L^i \\ V_A &= W_V^A * \tilde{Z}_A^i \end{aligned} \tag{6}$$

where W_Q^{LE}, W_K^L and W_V^A are learnable weight matrices for the query, key, and value transformations, respectively. Next, we compute the attention score and apply it to the value vectors to obtain the weighted features:

$$head_{LA} = Softmax\left(\frac{QK_L^T}{\sqrt{d_k}} + Mask\right)V_A \tag{7}$$

$$Mask_m = \begin{cases} 0, & if\ token\ on\ the\ padding\ position \\ 1, & otherwise \end{cases} \tag{8}$$

where $Mask$ represents the mask matrix to mask the attention tokens. By improving semantic relevance between modalities and strengthening intra-modality cohesion, the attention matrix's weights more accurately reflect cross-modal feature correlations.

After computing the attention heads, we combine them by concatenating the outputs from each attention head and passing them through a feed-forward neural network (FFN) for further processing. This results in the final feature representation for each modality:

$$\begin{aligned} H_{LA} &= LayerNorm \\ &\left(\tilde{Z}_L + FFN(Concat(head_{LA}^1, \dots, head_{LA}^h)W_O)\right) \\ H^s &= GAP\left(Concat(H_{LA}^s, H_{LV}^s, H_{AV}^s)\right) \\ \hat{y}^s &= Softmax\left(W_{s\_pre}H^s + b_s\right) \end{aligned} \tag{9}$$

Here, H_{LA}^s is the output of language and acoustic features after attention and FFN processing; similarly, LG-MMT captures H_{LV}^s and H_{AV}^s. Global Average Pooling (GAP) reduces fused feature dimensionality to obtain H^s, which averages features to retain critical information. Finally, H^s feeds into a Softmax-connected fully connected layer to compute prediction scores $\hat{y}^s$.

2.4 Similarity-based Representation Distillation

We introduce the Similarity-based Representation Distillation (SRD) to align the student and teacher model representations, especially under missing modalities, via cosine similarity loss on fused features. The similarity loss is defined as follows:

$$L_{SRD} = \alpha_{st}\left(1 - \frac{W_s H^t \odot W_t H^s}{\| W_s H^t \| \| W_t H^s \|}\right) \quad (10)$$

where α_{st} is the weighting vector to balance hierarchical consistency and final representation similarity, H^t and H^s represent the joint fused representations from the teacher and student models, respectively. The L_{SRD} penalizes any divergence between H^t and H^s, encouraging the student model to approximate the teacher's multimodal representation as closely as possible.

2.5 Modality Attribution-based Counterfactual Inference

To quantify individual modality impacts on predictions, we introduce the modality attribution-based counterfactual inference (MACI) module. By excluding specific modalities in counterfactual predictions, we assess their necessity for decision-making. Formally, the removal of modality *m* from the student model was shown to result in the recompilation of output distributions. This process defined uni-modal and counterfactual predictions as follows:

$$\hat{y}_m^s = Softmax\left(W_s \tilde{X}_m^s + b_s\right) \quad (11)$$

$$\hat{y}_{con_m}^s = Softmax\left(W_{s_{con}} \tilde{X}_{w/o\_m}^s + b_{s_{con}}\right) \quad (12)$$

here, $\hat{y}_{con_m}^s$ is the counterfactual prediction excluding modality m, with $\tilde{X}_{w/o\_m}^s = \frac{1}{N}\sum_i^N \tilde{X}_m^s$ as the student model's hidden representation replacing m with its training-set average. The comparison of $\hat{y}_{con_m}^s$ with the full-modality prediction $\hat{y}_m^s$, enables modality *m*'s impact. Finally, the attribution score, designated as δ_i^m, is thus defined as the absolute difference between predictions $\hat{y}_{m(i)}^s$ and $\hat{y}_{con_m(i)}^s$:

$$\delta_i^m = \hat{y}_{m(i)}^s - \hat{y}_{con_m(i)}^s \quad (13)$$

where δ_i^m captures the predictive difference for the $i\text{-}th$ sample when modality m is omitted. Here, $\hat{y}_{con_m(i)}^s$ represents the counterfactual prediction of the student model, while $\hat{y}_{m(i)}^s$ corresponds to the prediction of the student model under the same condition. To further quantify the relative importance of each modality across the dataset, we define the normalized modality contribution as follows:

$$X_{con_m} = \frac{\sum_{i=1}^N |\delta_i^m|}{\sum_{m' \in M} \sum_{i=1}^N |\delta_i^{m'}|} \quad (14)$$

where X_{con_m} denotes the attribution proportion of modality m, these insights inform modality re-weighting, feature selection, and model compression in resource-constrained settings. To ensure the model learns modal relative importance while maintaining prediction consistency, we define the counterfactual loss as:

$$\lambda_{con_m} = \frac{X_{con_m}}{\sum_{m' \in M} X_{con_{m'}}} \quad (15)$$

$$L_{cou} = \lambda_{con_m} \sum_{m \in M} X_{con_m} \quad (16)$$

where λ_{con_m} is a normalization coefficient that ensures the modality attributions sum to 1.

To avoid overfitting and promote robust learning, we use a regularization loss L_{reg} that penalizes the magnitudes of the attention mechanism's weight matrices (query, key, value projection). Mathematically, it's defined as:

$$L_{reg} = \lambda_Q^{le} \left\| W_Q^{le} \right\|^2 + \lambda_K^m \| W_K^m \|^2 + \lambda_V^m \| W_V^m \|^2 \quad (17)$$

where W_Q^{le}, W_K^m, and W_V^m represent the weight matrices for the query, key, and value transformations in the attention layers of the model, while λ_Q^{le}, λ_K^m, and λ_V^m denote their corresponding regularization coefficients. Finally, we defined the task-specific loss L_{task} to directly optimize the model's performance on the sentiment prediction task We utilize the conventional cross-entropy loss, which is characterized as:

$$L_{task} = -\frac{1}{N}\sum_{n=1}^{N} y_n \, log \, \hat{y}_n \quad (18)$$

where y_n denotes the true label for the n-th sample, while $\hat{y}_n$ represents the student model's predicted probability of the correct class. Finally, we define the final loss function of CUMDF is defined as Eq. (19):

$$L = \lambda_1 L_{reg} + \lambda_2 L_{cou} + \lambda_3 L_{SRD} + L_{task} \quad (19)$$

where λ_1, λ_2, and λ_3 are weights that control the relative importance of each loss term.

3 Experiments

3.1 Datasets and Evaluation Metrics

We conduct extensive experiments on three MER datasets with word-aligned data, including MOSI (Zadeh et.al,2016), MOSEI (Zadeh et.al,2018), and

IEMOCAP (Busso et.al, 2008). The statistics of the dataset are shown in Appendix A.1.

3.2 Implementation Details

All models are constructed using the Pytorch framework, utilizing NVIDIA Tesla V100 GPUs and torch version 1.8.2. The details of parameter implementations are listed in Appendix A.2.

3.3 Comparison with the state-of-the-art

We compare our CUMDF with eight exemplary and replicable state-of-the-art (SOTA) approaches, including complete-modality methods: Self-MM (Yu et.al, 2021), CubeMLP (Sun et.al, 2022), and DMD (Li et.al, 2023) and missing-modality methods: MCTN (Pham et.al, 2019), EMT (Sun et.al, 2022) IF-MMIN (Zuo et.al, 2023) and TransM (Wang et.al, 2020). To simulate missing inter-modalities, we remove modalities from samples in Table 1 and Table 2. "{l}" means only language is present.

Table 1: Performance comparison results on MOSI and MOSEI

Dataset	Models	Testing Conditions													
		$\{l\}$		$\{a\}$		$\{v\}$		$\{l,a\}$		$\{l,v\}$		$\{a,v\}$		$\{l,a,v\}$	
		MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1
MOSI	Self-MM	0.810	66.25	0.764	41.37	0.752	39.15	0.753	68.94	0.742	73.89	0.852	47.90	0.814	83.15
	CubeMLP	0.804	65.78	0.782	43.10	0.748	41.20	0.801	64.76	0.695	68.12	0.823	49.04	0.803	79.82
	DMD	0.795	67.45	0.801	43.65	0.761	42.90	0.698	69.42	0.735	67.93	0.831	51.25	0.892	82.14
	MCTN	0.821	75.56	**0.715**	58.96	0.797	58.12	0.710	**76.43**	0.703	73.62	0.864	62.47	0.802	83.75
	EMT	0.845	63.87	0.807	39.25	0.739	43.47	0.721	63.15	0.734	65.85	0.845	48.36	0.789	83.67
	IF-MMIN	0.830	56.14	0.841	46.89	0.815	45.13	0.745	61.27	0.785	64.94	0.802	66.92	0.762	81.24
	TransM	0.857	58.21	0.851	**66.83**	0.823	51.09	0.749	58.45	0.739	68.92	0.804	65.37	0.785	82.67
	CUMDF	**0.766**	**81.93**	**0.745**	62.29	**0.726**	**60.48**	**0.689**	76.38	**0.675**	**80.41**	**0.731**	**75.29**	**0.718**	**84.37**
MOSEI	Self-MM	0.763	67,85	0.731	43.83	0.725	41.74	0.775	69.43	0.688	72.52	0.803	48.95	0.752	82.74
	CubeMLP	0.772	71.65	0.749	43.52	0.736	37.67	0.722	75.94	0.725	74.67	0.824	49.58	0.739	83.07
	DMD	0.796	70.37	0.784	42.37	0.755	38.42	0.684	74.55	0.713	72.68	0.817	50.26	0.807	81.75
	MCTN	0.809	72.93	0.695	44.26	0.788	39.18	0.692	74.21	0.691	73.63	0.848	62.91	0.728	82.78
	EMT	0.791	67.46	**0.716**	39.55	0.719	32.57	0.700	71.63	0.708	70.05	0.822	48.52	0.780	83.13
	IF-MMIN	0.823	68.59	0.758	41.47	0.781	33.96	0.719	71.26	0.752	70.73	0.796	48.95	0.743	82.76
	TransM	0.801	69.75	0.782	42.56	0.803	34.98	0.727	73.28	0.729	71.19	0.773	50.34	0.775	82.91
	CUMDF	**0.742**	**72.48**	0.735	**57.53**	**0.717**	**58.45**	**0.678**	**79.72**	**0.676**	**81.03**	**0.735**	**71.57**	**0.704**	**83.25**

As illustrated in Table 1, CUMDF demonstrates efficacy in modality missing scenarios in comparison to baselines for both MOSI and MOSEI. In the context of MOSI in the language scenario, CUMDF exhibits the lowest MAE of 0.766 and the highest F1 of 81.93%. This performance surpasses the second-best by 6.37% in terms of F1. This phenomenon can be attributed to the alignment of features with label sentiment priors by LG-MMT, a process that ensures the extraction of robust information.

In the {a,v} scenario on MOSI, CUMDF's F1 is 75.29%, which is 9.92% higher than the next best. This demonstrates the efficacy of MACI, which adjusts modal contributions based on causal impact, a feature that is particularly advantageous when language is missing.

As demonstrated in Table 2, the F1 score performance of the CUMDF and baseline models is shown for various modality circumstances on IEMOCAP. In the "Happy" emotion category, the CUMDF achieves the highest F1 scores in three out of seven testing conditions, including single-modality scenarios {l} (82.4%), {a} (69.3%), and {v} (68.1%). Similarly, the "Sad" category exhibited consistent superiority of our method over baselines across five distinct testing conditions. The adaptability of the model's knowledge extraction and fusion mechanisms is emphasized

by its consistent superiority across diverse emotion categories.

Table 2: F1 score performance comparison results on IEMOCAP

Models	Metrics	Testing Conditions						
		$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	$\{l,a,v\}$
Self-MM	Happy	68.1	53.4	51.3	72.5	69.7	62.0	88.6
	Sad	69.2	53.0	53.8	68.9	68.1	61.8	87.9
	Angry	67.5	54.2	53.3	67.1	68.9	57.2	85.7
	Neutral	56.2	48.5	50.7	58.3	56.2	52.6	69.8
CubeMLP	Happy	66.2	52.3	50.1	69.4	68.5	56.2	**88.9**
	Sad	68.5	51.8	54.7	71.8	69.5	57.5	86.7
	Angry	65.2	53.7	51.6	69.5	67.8	56.6	85.4
	Neutral	55.7	48.6	50.3	58.5	56.9	52.8	70.7
DMD	Happy	69.8	55.7	52.2	79.3	77.5	65.8	83.1
	Sad	65.5	55.7	53.6	78.3	73.4	68.7	82.8
	Angry	65.3	54.1	51.4	81.7	80.4	59.5	84.6
	Neutral	54.6	51.3	49.5	65.3	62.7	54.9	67.1
MTCN	Happy	77.5	63.2	61.7	81.3	80.4	66.5	85.5
	Sad	76.2	64.3	60.5	82.9	81.5	64.3	84.0
	Angry	77.1	61.5	58.4	83.7	81.5	**68.4**	85.1
	Neutral	60.5	51.2	50.6	65.8	62.7	56.5	67.1
EMT	Happy	68.3	54.2	51.5	72.1	69.4	60.3	89.0
	Sad	65.7	54.6	53.4	70.5	68.73	58.1	88.5
	Angry	65.2	53.6	50.9	69.7	69.5	54.8	86.1
	Neutral	53.8	50.4	48.5	57.2	54.3	51.8	**71.8**
IF-MMIN	Happy	80.3	66.8	64.5	83.4	81.7	67.2	90.3
	Sad	79.2	65.9	62.3	82.6	79.5	70.4	85.2
	Angry	80.1	**67.6**	61.2	83.4	82.3	59.7	84.9
	Neutral	61.2	50.8	49.5	62.4	52.9	55.7	67.2
TransM	Happy	82.3	67.7	66.9	83.5	**82.6**	69.8	87.3
	Sad	81.7	69.5	66.3	**84.1**	81.8	70.3	86.9
	Angry	81.5	67.6	65.5	82.5	81.6	68.1	85.2
	Neutral	61.2	52.3	43.1	64.9	62.7	57.2	71.5
CUMDF	Happy	**82.4**	**69.3**	**68.1**	**84.2**	82.1	**70.1**	87.6
	Sad	**82.5**	**71.7**	**67.5**	83.3	**82.4**	**72.4**	**88.7**
	Angry	**82.7**	67.4	**66.3**	**83.6**	**82.9**	67.3	**86.2**
	Neutral	**63.3**	**54.1**	**52.4**	**68.3**	**64.5**	**57.8**	71.2

To illustrate performance variation at different drop rates, Figure 3 plots the performance curves of models across intra-modality drop ratios (0.1–0.7). CUMDF maintains significantly higher F1 scores for the "happy" and "sad" categories. This is due to its AMSKE and GMCKE modules, which extract modality-specific and modality-common knowledge to compensate for missing information.

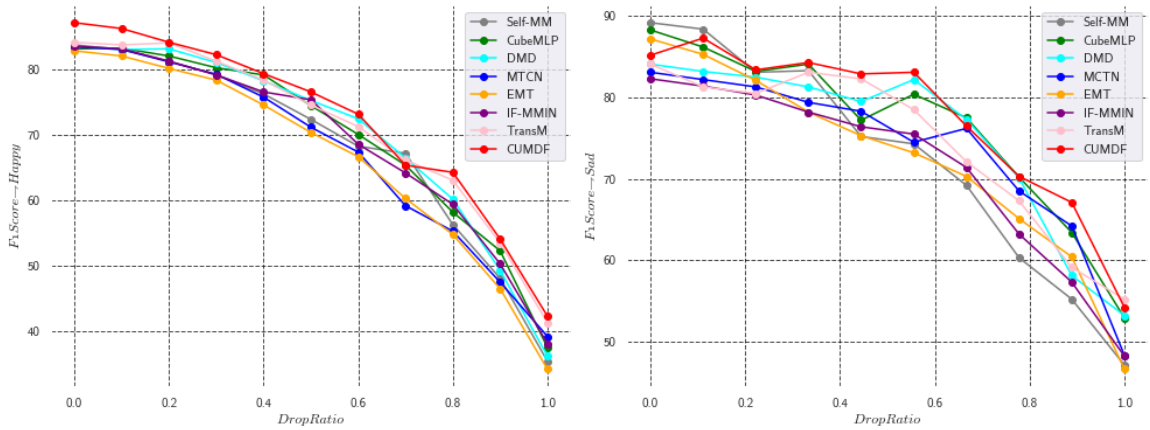


Figure 3: F1-score results of different intra-modality drop ratios on IMEOCAP.

3.4 Ablation Studies of Modules

To validate the contribution of each component in our framework, we conducted comprehensive ablation studies on the MOSI dataset with a drop ratio of 0.2, as shown in Table 3.

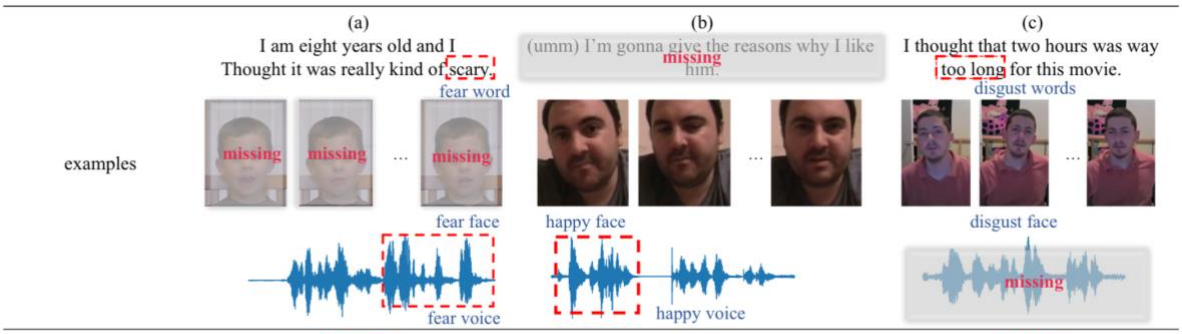
Ablation studies show removing LG-MMT causes significant performance drops, confirming the label-guided attention mechanism focuses on emotion-relevant features and improves cross-modal alignment. Removing MACI also degrades performance, validating counterfactual reasoning to measure modal causal impacts and adaptively weight them during fusion. Removing MKE leads to severe performance deterioration, indicating disentangling modality-specific/shared knowledge is crucial for capturing unique traits and cross-modal commonalities.

Table 3: Module ablation performance results on CMU-MOSI with a drop ratio of $p = 0.2$.

MKE	LG-MMT	MACI	MAE	F1
			0.820	82.36
	√	√	0.782	84.15
√		√	0.798	83.82
√	√		0.764	84.09
√	√	√	**0.758**	**84.57**

3.5 Case Study

Figure 4 shows CUMDF versus baseline IF-MMIN on three MOSI test samples, with missing modalities marked by grey rectangles. CUMDF outperforms the baseline across missing modality scenarios by using counterfactual reasoning to prioritize reliable modal signals, yielding predictions closer to ground truth labels.



Ground Truth	Negative	Positive	Negative
IF-MMIN(baseline)	Positive	Negative	Positive
CUMDF(ours)	Negative	Positive	Negative

Figure 4: The case of emotion recognition base on our proposed CUMDF.

Conclusion

In this paper, we introduce CUMDF, a novel distillation framework for multimodal sentiment analysis that effectively addresses modality missingness and conflicts. We design a Label-Guided Multimodal Masked Transformer (LG-MMT) that incorporates label embeddings as semantic anchors to guide cross-modal feature fusion. The proposed Adaptive Modality-Specific and Generalized Modality-Common Knowledge Extractors disentangle unique modality traits from shared patterns, enhancing representational coherence. Additionally, our Modality Attribution-based Counterfactual Inference (MACI) mechanism dynamically adjusts modality contributions based on their causal impact on prediction. Extensive experiments on MOSI, MOSEI, and IEMOCAP datasets demonstrate that CUMDF consistently outperforms state-of-the-art approaches across various modality missing scenarios.

Limitation

Our CUMDF has the following limitations: (1) The Label-Guided Multimodal Masked Transformer and counterfactual inference mechanisms increase computational complexity compared to simpler fusion approaches. (2) The effectiveness of CUMDF depends on the quality of initial modality representations and could benefit from more advanced feature encoders, though this

is not the focus of our work. (3) While we evaluated our framework on standard EMR benchmarks, it lacks validation on more diverse scenarios such as real-time streaming data with dynamic modality availability or datasets from specialized domains like healthcare or education. Testing across additional multimodal tasks beyond emotion recognition could further validate the framework's generalizability and effectiveness in broader contexts.

Acknowledgments

This work was supported by the National Social Science Fund of China (22BTQ048), Scientific Research Project of Education Department of Jilin Province (JJKH20250758KJ), and the Projects of Jilin University of Finance and Economics (2023YB021 and 2024PY010).

References

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6892–6899.

Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal emotion recognition. *IEEE Transactions on Affective Computing*.

Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Rohit Prasad. 2012. Ensemble of SVM trees for multimodal emotion recognition. In *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4.

Nicholas Cummins, Shahin Amiriparian, Sandra Ottl, Maurice Gerczuk, Maximilian Schmitt, Björn Schuller. 2018. Multimodal bag-of-words for cross domains emotion recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4954–4958.

Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. 2014. Microblog Emotion recognition Based on Cross-Media Bag-of-Words Model. In *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS)*, pages 76–80.

Louis-Philippe Morency, Rada Mihalcea, and Paresh Doshi. 2011. Towards multimodal emotion recognition: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)*, pages 169–176.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal emotion recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 973–982.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal emotion recognition. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 10790–10797.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In*, The 2016 Conference of the North Ameri can Chapter of the Association for Computational Linguistics (NAACL) HLT 2016: Human Language Technologies*, pages 160–170.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. 2014. Majority vote of diverse classifiers for late fusion. Structural, Syntactic, and Statistical Pattern Recognition - In *Joint IAPR International Workshop, S+SSPR* 2014, pages 153–162. Springer.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Konstantinos Rapantzikos, and Georgios Skoumas. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7): 1553–1568.

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. In *Proceedings of the International Conference on Learning Representations*, pages 1–17.

Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Self-adaptive Context and Modal-interaction Modeling for Multimodal Emotion Recognition. In *Proceedings of the 61th annual meeting of the Association for Computational Linguistics*, pages 6267–6281.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, et al. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 3183–3192.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29: 985–1000.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2023. Hybrid contrastive learning of tri-modal representation for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 14(3): 2276–2289.

Yuhang Sun, Zhizhong Liu, Quan Z. Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. 2024. Similar modality completion-based multimodal emotion recognition under uncertain missing modalities. *Information Fusion*, 110: 114.

Shicai Wei, Yang Luo, and Chunbo Luo. 2023. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049.

Yuanyue Deng, Jintang Bian, Shisong Wu, Jianhuang Lai, and Xiaohua Xie. 2025. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. *Information Fusion*, 114: 102711.

Puling Wei, Juan Yang, and Yali Xiao. 2024. Hierarchical Cross-Modal Interaction and Fusion Network Enhanced with Self-Distillation for Emotion Recognition in Conversations. *Electronics*, 13: 2645.

Yujuan Zhang, Fang'ai Liu, Xuqiang Zhuang, Ying Hou, and Yuling Zhang. 2024. Prototype-based sample-weighted distillation unified framework adapted to missing modality emotion recognition. *Neural Networks*, 177: 106397.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal emotion recognition. In *Proceedings of the 2021 ACM Multimedia Conference (MM '21)*, pages 4400–4407.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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*, pages 2236–2246.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335-359.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. CubeMLP: An MLP-based Model for Multimodal Emotion recognition and Depression Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, pages 3722–3729.

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

Sun Licai, Lian Zheng, Liu Bin, and Tao Jianhua. 2022. Efficient Multimodal Transformer with Dual-Level Feature Restoration for Robust Multimodal Emotion recognition. *arXiv preprint arXiv:2208.07589.*

Hamlin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal emotion recognition. In *Proceedings of The Web Conference*, pages 2514–2520.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7216–7223.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, and Zhe Li. 2023. Transformer-based multimodal feature fusion for emotion recognition with missing modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8004–8012.

Jeffery Pennnington, Richard Socher and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.* Pages 1532–1543.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio and Stefan Scherer. 2014. COVAREP: A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1532-1543.

Zengqun Zhao, Qingshan Liu and Shanmin Wang (2021) Learning deep global multi-scale and local attention features for facial expression recognition

in the wild. *IEEE Trans Image Process* 30:6544–6556.

Penn Phonetics Laboratory (2013) p2fa-vislab: A script for audio/transcript alignment. GitHub. Available from: *https://github.com/ucbvislab/p2favislab/*

Jiandian Zeng, Tianyi Liu and Jiantao Zhou. 2022. Tag-assisted multimodal emotion recognition under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1545–1554.

A Appendices

A.1 Datasets and Evaluation Metrics

We conduct extensive experiments on three MER datasets with word-aligned data, including MOSI, MOSEI, and IEMOCAP. The MOSI dataset consists of 2,199 video clips containing opinion comments from 89 independent speakers debating YouTube movie reviews. The sample comprises 41 female and 48 male speakers. Every video clip is labelled with sentiment strength, from -3 (very negative) to $+3$ (extremely positive). The MOSEI is a dataset consisting of 22,856 video clips, which has 16,326, 1,871, and 4,659 samples in train, valid, and test data. Each sample of MOSI and MOSEI is labeled by human annotators with a sentiment score of -3 (strongly negative) to +3 (strongly positive).

On the MOSI and MOSEI datasets, we utilize weighted F1 score computed for positive/negative classification results as evaluation metrics. The IEMOCAP dataset is intended for multi-label emotion recognition. It comprises 302 videos, including 151 recorded chat videos. Every sentence in the dialogue snippets is categorized by a specific emotion: happiness, sadness, anger, surprise, fear, and ten additional emotions. In our research, we follow the idea from (Wang et.al, 2019) and concentrate on identifying four fundamental emotions: happiness, sadness, anger, and neutrality.

According to a previous experimental study (Wang et.al, 2023), our model's performance leverages two metrics, adopting a dual approach for classification and regression predictions. For regression, we provide MAE as a measurement. For the classification tasks, we provide accuracy F1-score (F1) as measurement.

A.2 Implementation Details

A.2.1 Feature Extraction

For the language modality, we transform the video transcripts into pre-trained GloVe (Pennnington et.al, 2014) word embeddings to acquire a 300-dimensional vector. In the audio modality, we employ the COVAREP (Degottex et.al, 2014) to extract 74-dimensional low-level audio features, which can process Mel-cepstral coefficients, fundamental frequency, voiced/unvoiced segments, normalized amplitude quotient, quasi-open quotient, glottal source parameters, harmonic model, phase distortions, and formants. MA-Net (Zhang et.al, 2021) is utilized as the video feature extractor for the video modality. Renowned for its considerable success in facial expression recognition, we initially employed the MTCNN face detection algorithm to identify faces. Subsequently, we use the pre-trained MA-Net model to extract 1024-dimensional video features.

To attain word-level alignment among the three modalities, we first process the video and audio streams via P2FA (Penn, 2013) to provide aligned timestamps. Thereafter, the video and audio attributes are averaged throughout these synchronized intervals. In the CMU-MOSI dataset, the sequence lengths for all three modalities are established at 50. Conversely, the other two dataset maintain a sequence length of 20 throughout all three modalities.

Additionally, we partition these two datasets into training sets, validation sets, and test sets according to established ratios, with the dataset sizes specified in Table 1.

Table 1: Statistics of Datasets

Dataset	CMU-MOSI	CMU-MOSEI	IEMOCAP
Training Set	1284	16326	2717
Validation Set	229	1871	798
Test set	686	4659	938

A.2.2 Experimental Setup

All models are constructed using the Pytorch framework, utilizing NVIDIA Tesla V100 GPUs and torch version 1.8.2. To facilitate a fair comparison, we re-implement the state-of-the-art (SOTA) approaches utilizing publicly accessible codebases and integrate them with our experimental frameworks. All experimental findings are averaged across numerous trials with five distinct random seeds.

The hyperparameters of our proposed model utilize the configurations outlined in (Zeng et.al,2022) as detailed in Table 2. The learning rate is established as 0.001, and the concealed size is designated as 300. We employed the Adam optimizer to reduce the overall loss function and enhance the proposed AUMDF. The duration is established at 20, and the loss weight is designated at 0.1. The batch size for the CMU-MOSI dataset is set at 35, while the other two are designated as 32.

Table 2: Hyperparameter settings of CUMDF

Description	Symbol	Value
Epoch	b	20
Dropout rate	d	0.8
Hidden size	h	300
Learning rate	lr	0.001
Drop ratio	p	[0.1-0.7]
Maximum language length	m_l	25
Maximum audio length	m_a	150
Maximum video length	m_v	100
Loss weights	$\lambda_1, \lambda_2, \lambda_3$	0.1
Early stop	es	20