BALM: A Model-Agnostic Framework for Balanced Multimodal Learning under Imbalanced Missing Rates

Supplementary Material

A. Training procedure of BALM

The overall training procedure of BALM, including both the feature calibration and gradient rebalancing stages, is summarized in Algorithm 1.

Algorithm 1 Training procedure of the proposed plug-in framework BALM

Require: Multimodal dataset $\mathbb{D}$; model $(\phi^m, \mathcal{F}, \mathbf{F}_{pred})$; missing rates $\mathbf{r} = (r_a, r_v, r_l)$; total epochs E; hyper-parameters τ, ρ, α.

Initialize modulation loss $\mathcal{L}_{mod} \leftarrow 0$ and iteration $t \leftarrow 0$
Generate missing mask $e \sim \mathcal{M}(x; \mathbf{r})$
Generate incomplete dataset $\tilde{\mathbb{D}} \leftarrow \{(\tilde{x}_i, y_i)\}$
for *epoch* $= 1$ **to** E **do**
  for each $(\tilde{x}_i, y_i) \in \tilde{\mathbb{D}}$ **do**
    Calibrate unimodal features: $\hat{x}_i^m \leftarrow \texttt{FCM}(\tilde{x}_i)$
    Encode modalities: $z_i^m \leftarrow$ Eq. 11
    Fuse embeddings: $h_i \leftarrow$ Eq. 12
    Compute multimodal prediction: $\hat{y}_i \leftarrow$ Eq. 13
    Compute task loss: $\mathcal{L}_{task} \leftarrow$ Eq. 14
    Aggregate total loss: $\mathcal{L} \leftarrow \mathcal{L}_{task} + \tau \mathcal{L}_{mod}$
    Update parameters: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$
    Compute unimodal predictions: $\hat{y}_i^m \leftarrow$ Eq. 17
    Compute unimodal loss: $\mathcal{L}_{task}^m \leftarrow$ Eq. 18
    Estimate coefficients: $\mu^m \leftarrow$ Eq. 21
    Modulate encoder gradients: $\theta^{\phi^m} \leftarrow$ Eq. 22
    Update spatial modulation loss: $\mathcal{L}_{mod} \leftarrow$ Eq. 26
    $t \leftarrow t + 1$
  end for
end for

B. Theoretical Analysis

B.1. Gradient Imbalance under IMR

This section provides a formal justification for Eq. 6 in the main paper, showing how the imbalance in gradient magnitudes arises under the Imbalanced Missing Rate (IMR) condition.

Assumption 1 (Independent Missingness). Each modality m is independently missing according to a Bernoulli variable $e_i^m \in \{0, 1\}$ with $P(e_i^m = 1) = (1 - r_m)$ and $P(e_i^m = 0) = r_m$,, where $r_m \in [0, 1)$ denotes the missing probability of modality m. The missingness indicators $\{e_i^m\}$ are assumed independent of the target label Y (i.e., Missing Completely At Random-MCAR).

Lemma 1 (Expected Gradient Scaling under IMR). Under above Assumption, for a differentiable per-sample loss $\ell_m = \ell(f_m(X_m), Y)$ with finite variance, the expected gradient of modality m with respect to its parameters θ_m satisfies:

$$\mathbb{E}[\nabla_{\theta_m} \mathcal{L}_{\text{IMR}}] = (1 - r_m)\, \mathbb{E}[\nabla_{\theta_m} \ell(f_m(X_m), Y)]\,. \quad (28)$$

Proof. Let $\mathcal{L}_{\text{IMR}}$ denote the expected loss under the IMR distribution:

$$\mathcal{L}_{\text{IMR}} = \mathbb{E}_{X,Y,e}\left[\sum_{m=1}^{M} e^m\, \ell(f_m(X_m; \theta_m), Y)\right], \quad (29)$$

where $f_m(\cdot; \theta_m)$ represents the modality-specific encoder or prediction branch parameterized by θ_m, and e^m indicates its availability. Taking the gradient with respect to θ_m yields:

$$\nabla_{\theta_m} \mathcal{L}_{\text{IMR}} = \mathbb{E}_{X,Y,e}[e^m\, \nabla_{\theta_m} \ell(f_m(X_m; \theta_m), Y)]\,. \quad (30)$$

By the linearity of expectation, we can separate the stochastic variable e^m:

$$\mathbb{E}[\nabla_{\theta_m} \mathcal{L}_{\text{IMR}}] = \mathbb{E}_e[e^m]\ \mathbb{E}_{X,Y}[\nabla_{\theta_m} \ell(f_m(X_m; \theta_m), Y)]\,. \quad (31)$$

Since $e^m \sim \text{Bernoulli}(1 - r_m)$, we have $\mathbb{E}[e^m] = 1 - r_m$. Substituting this into Eq. 31 gives:

$$\mathbb{E}[\nabla_{\theta_m} \mathcal{L}_{\text{IMR}}] = (1 - r_m)\, \mathbb{E}[\nabla_{\theta_m} \ell(f_m(X_m; \theta_m), Y)]\,. \quad (32)$$

Eq. 32 shows that, in expectation, the gradient propagated to the parameters θ_m of each modality-specific encoder is linearly scaled by its availability ratio $(1 - r_m)$.

Corollary 1 (Gradient Imbalance Effect). Taking the L_2 norm of Eq. 32 yields

$$\mathbb{E}[\|\nabla_{\theta_m} \mathcal{L}_{\text{IMR}}\|] = (1 - r_m)\, \mathbb{E}[\|\nabla_{\theta_m} \ell(f_m(X_m; \theta_m), Y)\|]\,. \quad (33)$$

Therefore, modalities with higher missing rates (r_m large) receive proportionally weaker gradient updates on their encoder parameters, causing slower convergence and optimization dominance by low-missing-rate modalities.

The above analysis reveals that, under IMR, each modality receives a gradient update scaled by its availability ratio $(1 - r_m)$. This induces biased optimization dynamics where dominant modalities converge faster while rare ones lag behind. To counteract this imbalance, the Gradient

Rebalancing Module (GRM) introduced later (Eq. 21 and Eq. 22) adaptively rescales the gradient flow through modulation coefficients μ_m that act as an inverse correction to the implicit scaling factor $(1 - r_m)$. In essence, GRM restores equilibrium among modalities by dynamically adjusting both the magnitude and the direction of their gradients, thereby stabilizing multimodal optimization under heterogeneous missing conditions.

B.2. Gradient Rebalancing Rationale

Proposition 1. Building on **Lemma 1**, which shows that the expected gradient magnitude of each modality is scaled by its availability ratio $(1 - r_m)$, we explain the rationale behind the proposed Gradient Rebalancing Module (GRM).
Rationale. GRM compensates for the imbalance in gradient magnitudes through adaptive modulation of each modality's update:

$$\begin{aligned} \mu^m &= \rho \frac{\sum_{m' \in \{a,v,l\},\, m' \neq m} \Delta_{\mathrm{KL}}^{m'}}{\sum_{m' \in \{a,v,l\}} \Delta_{\mathrm{KL}}^{m'}}, \\ \theta_{(t+1)}^{\phi^m} &= \theta_{(t)}^{\phi^m} - \alpha\, \mu^m \frac{\partial \mathcal{L}_{\mathrm{task}}}{\partial \theta_{(t)}^{\phi^m}}. \end{aligned} \tag{34}$$

Here, Δ_{KL}^m quantifies the learning discrepancy of modality m between its unimodal and fused distributions, and ρ is a hyperparameter controlling the modulation intensity. A smaller μ^m indicates that modality m is learning faster (*i.e.*, larger Δ_{KL}^m) and thus its gradient update is attenuated, whereas slower modalities are amplified. This mechanism counteracts the imbalance identified in Lemma 1 and encourages all modality-specific gradients to approach an equilibrium:

$$\left\| \mu^m \nabla_{\theta^{\phi^m}} \mathcal{L}_{\mathrm{task}} \right\| \approx \left\| \mu^{m'} \nabla_{\theta^{\phi^{m'}}} \mathcal{L}_{\mathrm{task}} \right\|, \quad \forall m, m'. \tag{35}$$

Here, $\nabla_{\theta^{\phi^m}} \mathcal{L}_{\mathrm{task}}$ denotes the gradient of the task loss with respect to the parameters of the m-th encoder, and $\| \cdot \|$ denotes the Euclidean norm measuring its magnitude.

Such modulation is consistent with previous analyses [4, 14] showing that gradient reweighting based on learning discrepancy stabilizes multimodal training. GRM extends this idea to imbalanced missing-rate conditions, using KL-divergence as a continuous signal of learning disparity rather than a fixed prior ratio $(1 - r_m)$.

C. Benchmark Datasets

C.1. Dataset Description

To evaluate the effectiveness of BALM, we conduct experiments on two widely used multimodal emotion and sentiment benchmarks: IEMOCAP [1] and CMU-MOSEI [18]. Their key statistics are summarized in Table 5.

Table 5. Statistical overview of the IEMOCAP and CMU-MOSEI datasets.

Dataset	Dialogues			Utterances		
	train	valid	test	train	valid	test
IEMOCAP	120		31	5810		1623
CMU-MOSEI	2249	300	676	16326	1871	4659

IEMOCAP contains dyadic interactions between actors performing both scripted and improvised dialogues designed to elicit diverse emotions. The corpus comprises five sessions, each segmented into multiple utterances annotated with categorical emotion labels. Following the label processing in [11], we adopt the common six-class setting. Since the original dataset only provides training and test splits, we further divide the training set into training and validation subsets using a ratio of r (default 0.1).

CMU-MOSEI consists of 22,856 video-based utterances from over 1,000 YouTube speakers, each annotated with a sentiment score in the range $[-3, 3]$. Following [9], this dataset is trained as regression task, and evaluated as negative/positive classification task. Positive and negative classes are assigned for < 0 and > 0 scores, respectively. The official partitioning protocol is adopted to ensure consistency with previous studies.

Similar to prior studies [6, 7, 12], we use Accuracy (Acc) and Weighted F1 Score (W-F1) as our main evaluation metrics.

C.2. Multimodal Feature Extraction

For each utterance, multimodal features are extracted from acoustic, lexical, and visual modalities. The details of the extraction process for the two datasets are described as follows.

For the **IEMOCAP** dataset, we follow the feature extraction procedures outlined in [11] to obtain feature vectors for each modality. Specifically, we employ the RoBERTa-Large [10] model to extract 1024-dimensional textual features. RoBERTa is fine-tuned for emotion recognition on conversation transcripts, and the embeddings of the [CLS] tokens from the last layer are used as textual representations. Acoustic features are extracted using openSMILE [2] and then reduced to 1,582 dimensions via a fully connected layer, while visual features are obtained from a pre-trained DenseNet [8], resulting in 342-dimensional representations for each utterance.

Similarly, we adopt the feature extraction methods described in [9] for **CMU-MOSEI**. Pre-trained wav2vec[1] [15] is leveraged to extract 512-dimensional acoustic features for each utterance. For the textual modality, the pre-trained

[1] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

DeBERTa-Large model[2] [5] is exploited to encode word sequences into 1024-dimensional representations. Visual features are obtained through a two-step process: faces are first detected and aligned using the MTCNN [19] face detection algorithm, and the aligned frames are subsequently processed with MA-Net[3] [22] to produce frame-level features. Finally, we aggregate these frame-level facial features into 1024-dimensional utterance-level representations using average pooling.

D. Baseline Models

To evaluate the performance of BALM, we compare it with state-of-the-art methods for incomplete or imbalanced multimodal learning, as well as typical MER backbones.

D.1. Incomplete or Imbalanced Multimodal Models

The following baselines focus on addressing the challenges of missing modalities and uneven multimodal contribution.

MMIN [21] learns robust joint representations by imagining the features of absent modalities from the available ones via cycle-consistent autoencoders, thereby handling uncertain missing conditions effectively.

SDR-GNN [3] integrates spectral analysis into a hypergraph framework to impute missing data and explicitly retains high-frequency signals typically lost in conventional GNNs.

Mi-CGA [13] utilizes a reconstruction module to approximate missing inputs and leverages cross-modal graph attention to capture comprehensive inter-modal dependencies.

MoMKE [17] adopts a dual-phase learning scheme in which a learnable router dynamically fuses outputs from pretrained unimodal encoders to derive a more comprehensive representation for incomplete data.

GCNet [9] captures speaker and temporal dependencies via graph neural networks to handle incomplete conversations and employs a dual-task framework to simultaneously predict target labels and restore missing features.

Ada2I [12] rectifies learning imbalances by dynamically re-weighting feature and modality contributions under the supervision of a learning discrepancy metric.

RedCore [16] employs variational encoders to construct robust cross-modal representations and dynamically regulates auxiliary supervision based on reconstruction difficulty.

MCE [20] optimizes training dynamics via game-theoretic evaluations and promotes semantic robustness through subset prediction to facilitate balanced feature capability despite imbalanced missing rates.

[2]https://huggingface.co/microsoft/deberta-large

[3]https://github.com/zengqunzhao/MA-Net

Table 6. Hyper-parameters Setting

Parameters	IEMOCAP	CMU-MOSEI
Audio dim d_a	1582	512
Lexical dim d_l	1024	1024
Visual dim d_v	342	1024
d_{global}	737	640
ρ	(1.1,1.6]	
τ	(0.2,0.8]	
batch size	16	32
epoch	80	25

D.2. MER Backbones

To assess overall effectiveness, we further compare against mainstream backbones renowned for their multimodal context modeling and fusion mechanisms.

MMGCN [7] leverages a deep spectral graph network to fuse multimodal features. The framework defines utterances as interconnected nodes and captures long-range dependencies along with speaker context for emotion classification.

MMDFN [6] employs a gated graph architecture to selectively regulate cross-modal information flow. By dynamically filtering feature propagation across layers, it mitigates the accumulation of redundant data while strengthening inter-modal synergy.

E. Implementation Details

We conduct experiments under varying missing configurations to evaluate the performance of different baselines on multimodal emotion recognition datasets. For each dataset $\mathbb{D}$ and missing setting $\mathbf{r} = (r_A, r_L, r_V)$, we generate modality-missing masks e for the train/validation/test sets independently using the masking operator $\mathcal{M}(\cdot, \mathbf{r})$. Masking is applied prior to any processing or training, ensuring that complete data remain hidden from all models during both training and inference, while the missing masks remain fixed. The best model selected on the incomplete validation set is then evaluated on the incomplete test set.

For all baselines, we adopt their official implementations and model-specific hyperparameter settings (including learning rates) provided in their documentation. For **GCNet**, **SDR-GNN**, and **Mi-CGA**, specifically, we employ variants that omit reconstruction losses to ensure no baseline has access to any complete data.

For integrating **BALM** to **GCNet**, **MMGCN**, and **MMDFN**, `FCM` is added as a additional module to their architect while `GCM` is used as a separated module monitoring the training phase, the unimodal prediction heads is trained in parallel with the backbone using the same optimizer and learning rate, main architect and hyper-parameters of the backbone remain unchanged. Since **CMU-MOSEI** is

Table 7. Weighted-F1 sensitivity to hyperparameters under two contrastive missing-rate configurations of audio, language and visual.

Config A		ρ					**Config B**		ρ				
$(0.5, 0.3, 0.7)$		1.0	1.2	1.4	1.6	1.8	$(0.5, 0.7, 0.3)$		1.0	1.2	1.4	1.6	1.8
τ	0.1	64.38	64.92	63.71	63.90	65.93	τ	0.1	59.49	61.37	**62.41**	59.39	61.26
	0.4	65.50	64.71	65.53	63.80	64.31		0.4	59.98	60.62	61.80	60.22	60.45
	0.7	64.88	64.14	63.59	64.14	63.57		0.7	60.21	59.05	60.86	61.34	59.11
	1.0	65.10	63.71	64.26	**66.30**	65.67		1.0	60.35	60.58	60.73	59.48	59.87
	1.3	63.64	64.50	64.03	64.16	64.37		1.3	61.34	60.21	59.02	58.39	59.09

treated as a regression task, we omit the softmax function ($\delta(\cdot)$ in Eq. 13, Eq. 17) and use sigmoid function to map predicted scalars to probabilities of negative/positive class for Distribution-driven Modulation in GRM.

All experiments are implemented in PyTorch[4], and tracked with Comet.ml[5]. Additional training configurations and hyper-parameters of **BALM** are summarized in Table 6.

Code Availability and Reproducibility. We release our full implementation and configurations at: https://github.com/np4s/BALM_CVPR2026.git. The repository includes source code, configuration files, and brief guidelines with scripts to run experiments or adapt the framework to other datasets.

Masking Operator. Given a dataset of N samples with M modalities and a missing-ratio vector $\mathbf{r} = [r_1, \ldots, r_M]$, our masking operator $\mathcal{M}(\cdot; \mathbf{r})$ generates missing masks based on the hypothetical ratios of missing patterns. For a missing pattern $\hat{e} = [\hat{e}^1, \ldots, \hat{e}^M]$ among the 2^M possible patterns, its ratio is computed as:

$$\hat{r} = \prod_{m=1}^{M} \hat{e}^m (1 - r_m) \times (1 - \hat{e}^m) r_m. \tag{36}$$

Thus, $n = \lfloor N\hat{r} \rfloor$ samples are randomly assigned to the missing pattern $\hat{e}$, generating n missing-mask vectors with $e_i = \hat{e}$. Since each sample must retain at least one modality, the $N \prod_{m=1}^{M} r_m$ masks corresponding to the pattern where all modalities are missing are redistributed uniformly among the M patterns where exactly $M - 1$ modalities are missing. Consequently, the error ε_m of $\mathcal{M}(\cdot; \mathbf{r})$, defined as the discrepancy between the intended missing ratio r_m and the realized missing ratio $\frac{1}{N} \sum_{i=1}^{N} (1 - e_i^m)$ for modality m, satisfies $\varepsilon_m \leq \frac{1}{M} \prod_{m'=1}^{M} r_{m'}$.

F. Additional Experiment Results

F.1. Hyperparameter Sensitivity under Contrasting Missing Rates (Detailed Results)

Table 7 provides the complete numerical results of MMGCN$_{+\text{BALM}}$ on IEMOCAP for the two contrasting IMR configurations.

Specifically, under *Config–A* $(0.5, 0.3, 0.7)$, w-F1 remains stable across the (τ, ρ) grid, mostly lying within 63–64%, with limited sensitivity to either hyperparameter. The best score 66.30% occurs at (τ=1.0, ρ=1.6), while all other settings fluctuate within a narrow range of about 2%.

In contrast, *Config–B* $(0.5, 0.7, 0.3)$ shows noticeably larger spread, ranging from 58.39% to 62.41%. Performance improves around small τ with moderate ρ (best at (τ=0.1, ρ=1.4)=62.41%), whereas higher ρ consistently leads to degradation across all τ. This pattern indicates stronger τ–ρ interaction when the lexical modality has the highest missing rate.

F.2. Average performance on different modality combinations (Detailed Results)

Table 8 and Table 9 report the complete results for Accuracy (Acc) and weighted-F1 (w-F1), respectively, across all modality combinations-*unimodal* (A, V, L), *bimodal* (AV, AL, LV), and *trimodal* (ALV)-under different missing-rate settings on IEMOCAP. For each combination, unspecified modalities are ablated from training and evaluating. We compare GCNet and MMGCN with their variants enhanced by BALM. Across all unimodal, bimodal, and trimodal configurations, the BALM-enhanced models consistently achieve higher Acc and W-F1, demonstrating improved robustness under varying missing-modality conditions. The averaged performance across all configurations is presented in Fig. 6 of the main paper.

Table 10 reports the full MMGCN results across all modality combinations before and after integrating the FCM module. In this experiment, MMGCN and MMGCN$_{+\text{BALM}}$ are trained under the specified condition and tested on the complete test-set of modality combinations. The averaged performance is summarized in Table 3 of the main paper. Overall, FCM provides consistent improvements across settings, highlighting the effectiveness of BALM in addressing the inconsistent representations arising from IMR.

F.3. BALM under SMR settings (Numerical Results)

Table 11 and Table 12 report the detail results from Fig. 4 in the main paper. By addressing challenges of missing modalities with BALM, MMGCN$_{+\text{BALM}}$ and MMDFN$_{+\text{BALM}}$ are

[4]https://pytorch.org/
[5]https://comet.ml

Table 8. Performance across unimodal, bimodal, and trimodal configurations under varying missing rates (Accuracy)

MR Setting	Model	A	V	L	AV	AL	LV	ALV
(0.5, 0.5, 0.7)	GCNet	46.33	37.28	57.42	48.18	59.09	54.16	58.47
	GCNet$_{\text{+BALM}}$	46.95	31.44	59.09	50.89	59.64	54.41	59.33
	MMGCN$_{\text{+BALM}}$	51.45	31.98	60.63	46.77	61.18	62.60	60.44
(0.5, 0.7, 0.5)	GCNet	49.04	31.85	56.99	48.49	56.69	53.73	57.86
	GCNet$_{\text{+BALM}}$	47.13	27.60	54.59	46.21	58.10	55.76	55.82
	MMGCN$_{\text{+BALM}}$	51.02	29.94	59.89	51.26	62.23	52.50	61.49
(0.7, 0.5, 0.5)	GCNet	44.18	30.19	56.44	38.32	56.93	56.81	58.23
	GCNet$_{\text{+BALM}}$	48.37	30.75	56.90	47.57	60.01	58.60	57.86
	MMGCN$_{\text{+BALM}}$	48.92	29.94	62.97	46.03	62.54	60.07	62.85
(0.3, 0.5, 0.7)	GCNet	47.87	32.72	53.54	48.55	60.20	55.51	59.46
	GCNet$_{\text{+BALM}}$	51.57	32.10	54.53	53.05	60.38	54.53	61.12
	MMGCN$_{\text{+BALM}}$	52.93	28.34	60.57	52.43	62.78	60.38	62.78
(0.5, 0.3, 0.7)	GCNet	45.41	32.35	59.64	45.96	59.64	57.92	58.60
	GCNet$_{\text{+BALM}}$	51.26	35.00	59.64	51.76	61.80	60.20	60.69
	MMGCN$_{\text{+BALM}}$	50.89	28.34	63.71	50.40	64.08	61.31	64.39
(0.7, 0.5, 0.3)	GCNet	46.89	35.49	54.47	48.86	56.01	54.78	58.41
	GCNet$_{\text{+BALM}}$	44.55	35.43	55.51	52.00	54.10	55.88	59.33
	MMGCN$_{\text{+BALM}}$	45.96	28.90	61.61	46.95	61.86	59.09	61.74
Average	GCNet	46.62	**33.31**	56.42	46.39	58.09	55.49	58.51
	GCNet$_{\text{+BALM}}$	48.31	32.05	56.71	**50.25**	59.01	56.56	59.02
	MMGCN$_{\text{+BALM}}$	**50.20**	29.57	**61.56**	48.97	**62.45**	**59.33**	**62.28**

Table 9. Performance across unimodal, bimodal, and trimodal configurations under varying missing rates (w-F1)

MR Setting	Model	A	V	L	AV	AL	LV	ALV
(0.5, 0.5, 0.7)	GCNet	45.91	34.50	57.76	48.44	59.06	54.02	58.52
	GCNet$_{\text{+BALM}}$	46.56	27.57	59.47	50.56	59.77	53.93	59.51
	MMGCN$_{\text{+BALM}}$	50.66	25.81	61.05	48.66	60.44	62.12	60.39
(0.5, 0.7, 0.5)	GCNet	44.14	29.65	56.58	49.40	56.71	53.81	58.15
	GCNet$_{\text{+BALM}}$	48.36	26.36	54.82	46.89	58.36	56.05	58.75
	MMGCN$_{\text{+BALM}}$	50.07	28.88	59.50	51.01	61.85	51.77	61.37
(0.7, 0.5, 0.5)	GCNet	42.68	25.13	56.54	35.41	57.30	57.26	58.35
	GCNet$_{\text{+BALM}}$	45.31	26.62	54.77	46.54	59.99	58.63	58.01
	MMGCN$_{\text{+BALM}}$	48.85	28.88	62.18	46.00	62.08	60.29	62.78
(0.3, 0.5, 0.7)	GCNet	43.77	32.29	52.87	48.67	60.68	54.84	59.40
	GCNet$_{\text{+BALM}}$	50.54	28.40	54.75	51.26	60.57	54.52	61.40
	MMGCN$_{\text{+BALM}}$	52.51	26.86	60.95	51.91	62.73	60.61	62.17
(0.5, 0.3, 0.7)	GCNet	43.47	32.25	59.85	44.77	60.09	57.73	58.88
	GCNet$_{\text{+BALM}}$	49.61	34.11	59.75	50.92	62.21	60.43	60.78
	MMGCN$_{\text{+BALM}}$	50.87	26.86	63.64	50.33	63.28	61.33	63.73
(0.7, 0.5, 0.3)	GCNet	46.08	33.43	54.02	49.48	56.14	54.54	58.61
	GCNet$_{\text{+BALM}}$	42.91	32.98	55.51	51.46	54.09	56.45	58.98
	MMGCN$_{\text{+BALM}}$	45.19	25.48	61.54	50.63	61.15	59.21	61.92
Average	GCNet	44.34	**31.21**	56.27	46.03	58.33	55.37	58.65
	GCNet$_{\text{+BALM}}$	47.22	29.34	56.51	49.61	59.17	56.67	59.57
	MMGCN$_{\text{+BALM}}$	**49.69**	27.13	**61.48**	**49.76**	**61.92**	**59.22**	**62.06**

able to achieve a more robust performance as the shared missing rate increases. Notably, for IEMOCAP, other methods addressing missing modalities can suffer up to over 3% performance reduction, e.g. SDR-GNN and GCNet when SMR increase from 0.5 to 0.6, Mi-CGA when SMR increase from 0.6 to 0.7; while both MMGCN$_{\text{+BALM}}$ and MMDFN$_{\text{+BALM}}$ only see a $1\% - 2\%$ drop across missing rates. Although more subtle, such trend in the performance's consistency can also be seen for CMU-MOSEI.

F.4. Gradient Rebalancing Module Analysis

To further investigate the two sub-modules of Gradient Rebalancing Module (GRM), we introduce two variants: **BALM-D** consisting `FCM` and *distribution modulation* only, and its counterpart **BALM-S** consisting `FCM` and *spatial modulation* only.

Fig. 1 shows the learning progress of the modalities during training under IMR setting $(r_A, r_L, r_V) = (0.5, 0.7, 0.3)$, while Table 13 displays quantitative performance of the two variants. The more robust and over-

Table 10. Performance of MMGCN on different modality combinations after training under IMR setting $(r_A, r_L, r_V) = (0.5, 0.7, 0.3)$, before and after plugged with FCM module.

Modality	IEMOCAP				CMU-MOSEI			
	MMGCN		+FCM		MMGCN		+FCM	
	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
A	31.05	25.97	39.99	34.75	62.85	48.51	62.85	48.51
L	61.92	61.01	62.48	61.56	83.46	83.68	85.75	85.76
V	19.78	11.67	20.70	12.07	62.91	48.66	62.85	50.06
AL	64.45	63.11	63.77	62.15	83.57	83.78	85.72	85.71
AV	32.29	26.92	42.88	38.07	62.91	48.64	62.85	48.86
LV	61.98	61.03	62.91	61.92	85.39	85.40	85.91	85.87
ALV	65.19	63.94	64.20	62.80	85.55	85.55	85.88	85.83
Avg.	48.09	44.81	**50.99**	**47.62**	75.23	69.17	**75.97**	**70.09**

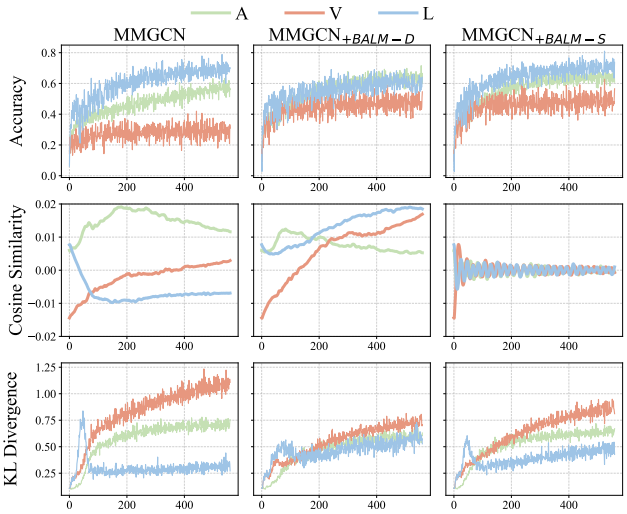


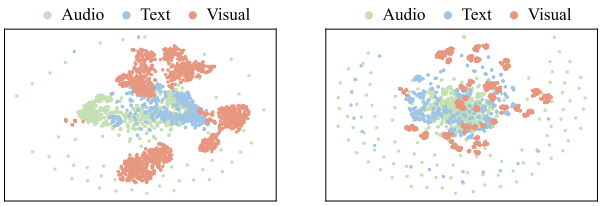
Figure 1. Modality discrepancies on IEMOCAP with different variants of BALM.

all better performance of MMGCN$_{\textbf{+BALM}}$ (from Table 1,2 in main paper) when compared to MMGCN$_{\textbf{+BALM-D}}$ and MMGCN$_{\textbf{+BALM-S}}$ further highlight the complementary nature of *distribution-* and *spatial-driven* Modulation. Fig. 1 suggests that these two-sided modulations act as each other's regulator, keeping the backbone from being over-balanced toward either perspective (e.g., the cosine similarity of MMGCN$_{\textbf{+BALM-S}}$), thus, giving a more general model.

F.5. Distribution-driven Modulation Analysis

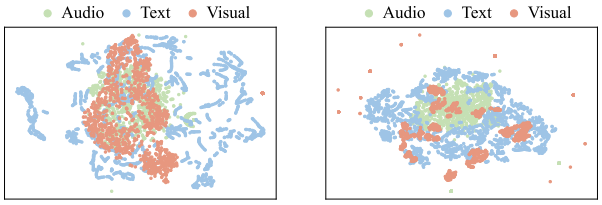
We introduce the **BALM-M** variant, in which the KL divergence (i.e., $\mathbf{KL}(\cdot)$ in Eq. 19) is replaced with an Mean Squared Error (MSE) loss, while all other computations in GRM remain unchanged. Detailed results for GCNet$_{\textbf{+BALM-M}}$, MMGCN$_{\textbf{+BALM-M}}$, and MMDFN$_{\textbf{+BALM-M}}$, along with their performance gaps relative to the corresponding baseline enhanced by BALM, are reported in Table 14 and Table 15.

The improvements observed in MMDFN$_{\textbf{+BALM-M}}$ over MMDFN$_{\textbf{+BALM}}$ under most settings on IEMOCAP, up to 1.67% under $(0.3, 0.5, 0.7)$, suggest that MSE can also serve as an effective measure of the distribution gap between multimodal and unimodal predictions for certain backbones. Nonetheless, KL divergence consistently yields better performance across different backbones and across both datasets, likely due to its softer characterization of distribution differences, verifies its advantage as a more suitable choice for distribution quantification as we aim to generalize BALM across diverse models.



(a) Original (b) Calibrated

Figure 8. Original input of MMGCN (i.e., $\tilde{x}$) and calibrated input of MMGCN$_{\text{+BALM}}$ (i.e., $\hat{x}$), under IMR setting $(r_A, r_L, r_V) = (0.5, 0.7, 0.3)$ on IEMOCAP.



(a) Embeddings of MMGCN (b) Embeddings of MMGCN$_{\text{+BALM}}$

Figure 9. The embeddings (i.e., z in Eq. 11) corresponding to the inputs from Fig. 8.

F.6. Multimodal Representation

Fig. 8, 9, 10 visualize the features comparison at different stages between MMGCN and MMGCN$_{\text{+BALM}}$ using t-SNE. As Fig. 8 depicts, after being calibrated with FCM, the outliers of audio and lexical modalities become more visible, which is a better reflection of the dataset's properties in this specific scenario where these two modalities suffer loss at high rates, while visual features are also preprocessed into more defined clusters. Consequently, the embeddings in Fig.9 also show a better learning of modal-specific encoders, resulting in multimodal fused feature with more distinct border between *hap-exc* (happy - excited) and *ang-fru* (angry - frustrated) clusters in Fig. 10.

Table 11. Average performance of five runs under different SMR settings on IEMOCAP.

SMR	Mi-CGA		SDR-GNN		GCNet		MMGCN$_{+BALM}$		MMDFN$_{+BALM}$	
	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
0.0	$64.45_{\pm 0.76}$	$64.49_{\pm 0.94}$	$64.86_{\pm 0.78}$	$64.65_{\pm 0.83}$	$64.90_{\pm 0.59}$	$65.01_{\pm 0.68}$	$65.63_{\pm 1.92}$	$65.49_{\pm 1.90}$	$70.02_{\pm 0.95}$	$69.89_{\pm 0.81}$
0.1	$63.43_{\pm 1.27}$	$63.42_{\pm 1.24}$	$63.43_{\pm 1.07}$	$63.35_{\pm 0.99}$	$64.10_{\pm 0.73}$	$63.99_{\pm 0.63}$	$64.28_{\pm 0.71}$	$64.05_{\pm 0.74}$	$68.90_{\pm 0.99}$	$68.78_{\pm 1.06}$
0.2	$63.71_{\pm 1.68}$	$63.76_{\pm 1.68}$	$62.28_{\pm 1.48}$	$62.00_{\pm 1.41}$	$63.18_{\pm 1.57}$	$62.83_{\pm 1.67}$	$63.73_{\pm 1.84}$	$63.54_{\pm 1.89}$	$68.56_{\pm 0.73}$	$68.45_{\pm 0.72}$
0.3	$60.46_{\pm 1.26}$	$59.93_{\pm 1.46}$	$60.54_{\pm 3.23}$	$59.92_{\pm 2.92}$	$62.50_{\pm 0.61}$	$62.14_{\pm 0.80}$	$62.76_{\pm 1.16}$	$62.67_{\pm 1.21}$	$67.43_{\pm 1.38}$	$67.44_{\pm 1.37}$
0.4	$59.99_{\pm 1.96}$	$60.18_{\pm 2.02}$	$59.16_{\pm 1.25}$	$58.85_{\pm 1.16}$	$60.31_{\pm 1.01}$	$60.01_{\pm 1.38}$	$62.12_{\pm 0.97}$	$61.90_{\pm 0.94}$	$66.21_{\pm 0.98}$	$66.15_{\pm 0.95}$
0.5	$58.07_{\pm 1.49}$	$58.09_{\pm 1.64}$	$58.79_{\pm 2.18}$	$58.20_{\pm 1.71}$	$58.68_{\pm 0.85}$	$58.54_{\pm 0.80}$	$61.77_{\pm 1.53}$	$61.82_{\pm 1.72}$	$64.93_{\pm 1.47}$	$64.78_{\pm 1.48}$
0.6	$58.43_{\pm 1.09}$	$58.42_{\pm 1.26}$	$55.71_{\pm 2.32}$	$55.27_{\pm 2.15}$	$55.00_{\pm 2.37}$	$54.83_{\pm 2.47}$	$60.75_{\pm 1.93}$	$60.71_{\pm 1.94}$	$63.64_{\pm 2.33}$	$63.27_{\pm 2.29}$
0.7	$55.79_{\pm 0.85}$	$55.40_{\pm 1.67}$	$54.41_{\pm 2.47}$	$53.65_{\pm 2.46}$	$53.91_{\pm 0.83}$	$53.52_{\pm 1.21}$	$59.68_{\pm 2.09}$	$59.59_{\pm 2.20}$	$61.63_{\pm 2.36}$	$61.80_{\pm 2.15}$

Table 12. Average performance of five runs under different SMR settings on CMU-MOSEI.

SMR	Mi-CGA		SDR-GNN		GCNet		MMGCN$_{+BALM}$		MMDFN$_{+BALM}$	
	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
0.0	$86.23_{\pm 0.44}$	$86.21_{\pm 0.35}$	$86.60_{\pm 0.56}$	$86.58_{\pm 0.51}$	$87.00_{\pm 0.32}$	$86.94_{\pm 0.39}$	$87.33_{\pm 0.16}$	$87.29_{\pm 0.19}$	$86.98_{\pm 0.38}$	$86.96_{\pm 0.36}$
0.1	$85.06_{\pm 0.43}$	$85.04_{\pm 0.50}$	$85.82_{\pm 0.30}$	$85.80_{\pm 0.31}$	$85.92_{\pm 0.29}$	$85.87_{\pm 0.23}$	$85.81_{\pm 0.22}$	$85.73_{\pm 0.16}$	$86.13_{\pm 0.53}$	$86.06_{\pm 0.50}$
0.2	$83.60_{\pm 0.68}$	$83.47_{\pm 0.56}$	$84.85_{\pm 0.36}$	$84.77_{\pm 0.34}$	$84.57_{\pm 0.83}$	$84.56_{\pm 0.71}$	$85.55_{\pm 0.19}$	$85.45_{\pm 0.18}$	$84.78_{\pm 0.36}$	$84.65_{\pm 0.34}$
0.3	$82.26_{\pm 0.50}$	$82.23_{\pm 0.47}$	$83.34_{\pm 0.39}$	$83.18_{\pm 0.26}$	$83.83_{\pm 0.30}$	$83.73_{\pm 0.29}$	$83.90_{\pm 0.22}$	$83.67_{\pm 0.34}$	$83.87_{\pm 0.13}$	$83.80_{\pm 0.18}$
0.4	$81.33_{\pm 0.38}$	$81.19_{\pm 0.42}$	$82.35_{\pm 0.56}$	$81.99_{\pm 0.72}$	$82.58_{\pm 0.53}$	$82.36_{\pm 0.50}$	$82.71_{\pm 0.50}$	$82.44_{\pm 0.48}$	$82.86_{\pm 0.55}$	$82.65_{\pm 0.50}$
0.5	$80.06_{\pm 0.74}$	$80.05_{\pm 0.71}$	$80.88_{\pm 0.72}$	$80.59_{\pm 0.83}$	$81.44_{\pm 0.64}$	$81.26_{\pm 0.63}$	$81.92_{\pm 0.22}$	$81.63_{\pm 0.21}$	$81.83_{\pm 0.32}$	$81.59_{\pm 0.41}$
0.6	$79.01_{\pm 0.61}$	$78.86_{\pm 0.53}$	$79.75_{\pm 0.60}$	$79.49_{\pm 0.56}$	$79.92_{\pm 0.79}$	$79.86_{\pm 0.68}$	$80.28_{\pm 0.19}$	$80.07_{\pm 0.21}$	$80.46_{\pm 0.28}$	$80.16_{\pm 0.18}$
0.7	$78.44_{\pm 0.31}$	$78.16_{\pm 0.35}$	$78.64_{\pm 0.75}$	$78.46_{\pm 0.64}$	$78.55_{\pm 1.16}$	$78.45_{\pm 1.01}$	$79.05_{\pm 0.23}$	$78.71_{\pm 0.23}$	$79.31_{\pm 0.35}$	$78.95_{\pm 0.24}$

Table 13. Performance of MMGCN on IEMOCAP under different IMR settings when integrated with two variants of BALM.

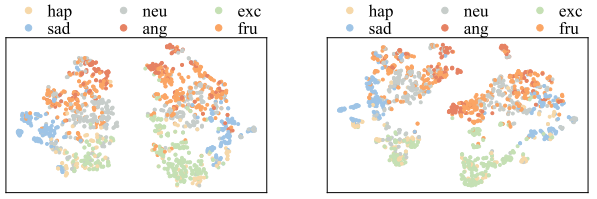
IMR Settings	IEMOCAP				CMU-MOSEI			
	+BALM-D		+BALM-S		+BALM-D		+BALM-S	
	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
(0.3, 0.5, 0.7)	64.51	64.02	62.29	62.47	80.79	80.69	81.04	80.39
(0.3, 0.7, 0.5)	61.49	60.80	59.15	59.51	78.01	77.82	78.32	77.77
(0.5, 0.3, 0.7)	63.83	63.50	65.13	65.18	84.04	83.76	84.07	83.86
(0.5, 0.7, 0.3)	60.63	60.53	60.87	60.90	77.99	77.66	79.25	78.94
(0.7, 0.3, 0.5)	64.76	64.65	64.26	64.49	84.56	84.31	84.31	84.22
(0.7, 0.5, 0.3)	60.57	60.60	60.44	60.63	82.55	82.43	81.65	81.53

Table 14. Performance of **BALM-M** variant on IEMOCAP under different IMR settings.

MR Setting	GCNet$_{+BALM-M}$		MMGCN$_{+BALM-M}$		MMDFN$_{+BALM-M}$	
	Acc	w-F1	Acc	w-F1	Acc	w-F1
(0.3, 0.5, 0.7)	$59.46_{\downarrow 1.85}$	59.93	$65.50_{\uparrow 1.11}$	65.26	$69.01_{\uparrow 1.67}$	68.79
(0.3, 0.7, 0.5)	$58.23_{\uparrow 0.87}$	58.47	$60.57_{\uparrow 0.50}$	60.69	$64.08_{\uparrow 0.68}$	63.54
(0.5, 0.3, 0.7)	$61.43_{\uparrow 0.12}$	61.57	$65.13_{\downarrow 0.24}$	65.19	$69.32_{\uparrow 0.56}$	69.06
(0.5, 0.7, 0.3)	$56.87_{\downarrow 0.77}$	57.12	$60.63_{\downarrow 1.66}$	60.68	$64.39_{\uparrow 0.37}$	64.10
(0.7, 0.3, 0.5)	$58.16_{\downarrow 1.17}$	58.31	$63.89_{\downarrow 1.11}$	63.90	$68.45_{\uparrow 1.23}$	68.29
(0.7, 0.5, 0.3)	$54.71_{\downarrow 5.55}$	54.97	$61.18_{\downarrow 0.56}$	61.25	$65.13_{\downarrow 1.37}$	64.70

Table 15. Performance of **BALM-M** variant on CMU-MOSEI under different IMR settings.

MR Setting	GCNet$_{+BALM-M}$		MMGCN$_{+BALM-M}$		MMDFN$_{+BALM-M}$	
	Acc	w-F1	Acc	w-F1	Acc	w-F1
(0.3, 0.5, 0.7)	$81.70_{\uparrow 0.16}$	81.44	$81.29_{\downarrow 0.74}$	80.69	$81.40_{\downarrow 0.66}$	80.96
(0.3, 0.7, 0.5)	$78.59_{\uparrow 0.08}$	78.10	$78.40_{\downarrow 0.33}$	78.10	$77.44_{\downarrow 1.37}$	77.34
(0.5, 0.3, 0.7)	$82.83_{\downarrow 0.44}$	82.66	$82.97_{\downarrow 1.57}$	82.97	$83.65_{\downarrow 0.69}$	83.55
(0.5, 0.7, 0.3)	$77.55_{\downarrow 1.95}$	77.54	$77.96_{\downarrow 1.87}$	78.00	$78.70_{\downarrow 0.66}$	78.50
(0.7, 0.3, 0.5)	$82.97_{\downarrow 0.66}$	82.96	$84.26_{\downarrow 0.58}$	84.07	$84.48_{\downarrow 0.30}$	84.36
(0.7, 0.5, 0.3)	$81.15_{\downarrow 0.94}$	81.11	$82.72_{\uparrow 0.22}$	82.29	$82.31_{\downarrow 0.22}$	82.16



(a) MMGCN (b) MMGCN$_{+BALM}$

Figure 10. The multimodal fusion (i.e., h in Eq. 12) corresponding to the inputs from Fig. 8, colored by ground truth.

F.7. Shared- and Imbalanced Missing Rates Relation

Given a dataset of N samples with M modalities, and the binary observation indicator e_i^m of modality m for sample i with $e_i^m = 0$ denotes missing. Following [3, 9], the shared missing rate is defined as:

$$r_{\text{shared}} = \frac{\sum_{m=1}^{M}\sum_{i=1}^{N}(1-e_i^m)}{MN}, \tag{37}$$

while the modality-specific missing rate under IMR is given by:

$$r_m = \frac{\sum_{i=1}^{N}(1 - e_i^m)}{N}. \tag{38}$$

Thus, we can assume the IMR settings equivalent to a given SMR are those that satisfy

$$\sum_{m=1}^{M} r_m = M r_{\text{shared}}. \tag{39}$$

Accordingly, in Fig. 2 of the main paper, we compare the performance of models addressing missing modalities under $SMR = 0.5$ with IMR configurations satisfying $r_A + r_L + r_V = 1.5$. The detailed results are provided in Table 1 for the sampled IMR settings and in Table 11 for $SMR = 0.5$. These experiments show that a shared missing rate alone cannot fully capture a model's behavior, particularly under highly imbalanced missing-rate conditions, underscoring the necessity of investigating IMR scenarios.

F.8. Modality Discrepancies

In this section, we provide extended visual comparisons of modality discrepancies for MMGCN vs. MMGCN$_{\text{+BALM}}$ and MMDFN vs. MMDFN$_{\text{+BALM}}$ on IEMOCAP under different IMR configurations (from Fig.11 to Fig. 22).

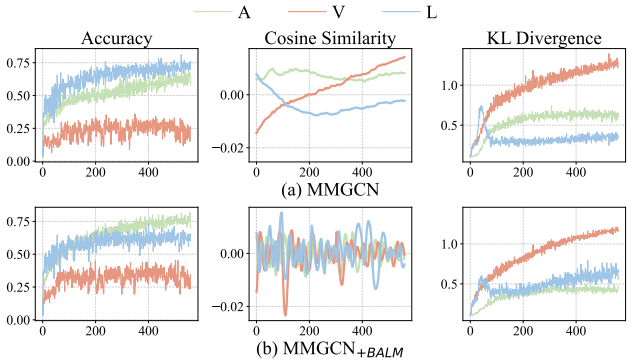


Figure 11. $(r_A, r_L, r_V) = (0.3, 0.5, 0.7)$

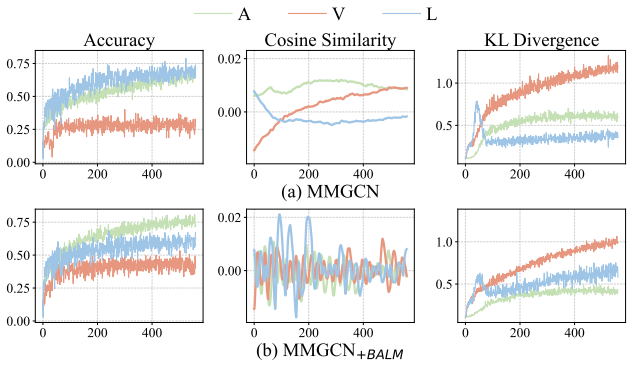


Figure 12. $(r_A, r_L, r_V) = (0.3, 0.7, 0.5)$

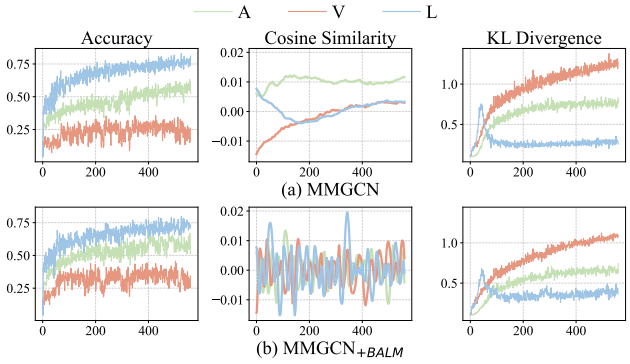


Figure 13. $(r_A, r_L, r_V) = (0.5, 0.3, 0.7)$

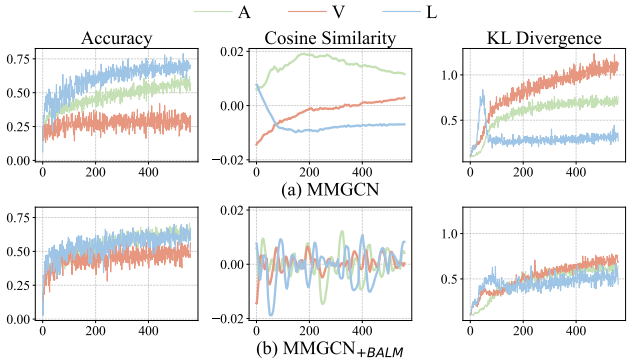


Figure 14. $(r_A, r_L, r_V) = (0.5, 0.7, 0.3)$

References

[1] 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. 2

[2] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, 2013. 2

[3] Fangze Fu, Wei Ai, Fan Yang, Yuntao Shou, Tao Meng, and Keqin Li. Sdr-gnn: Spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowledge-Based Systems*, page 112825, 2024. 3, 7

[4] Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. *Advances in Neural Information Processing Systems*, 37:133328–133344, 2024. 2

[5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and

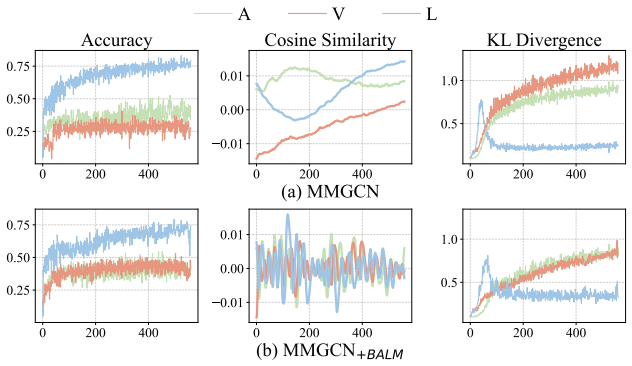


Figure 15. $(r_A, r_L, r_V) = (0.7, 0.3, 0.5)$

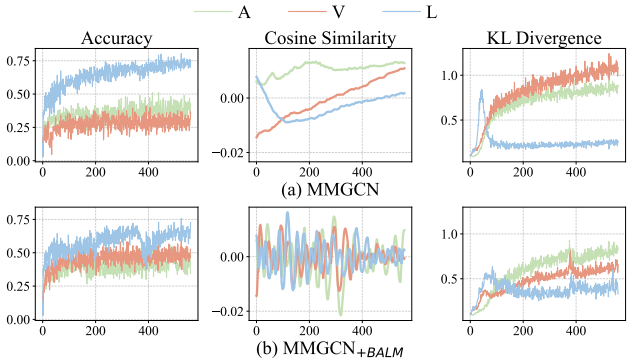


Figure 16. $(r_A, r_L, r_V) = (0.7, 0.5, 0.3)$

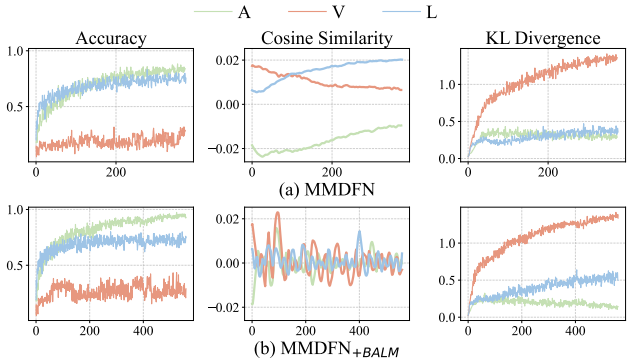


Figure 17. $(r_A, r_L, r_V) = (0.3, 0.5, 0.7)$

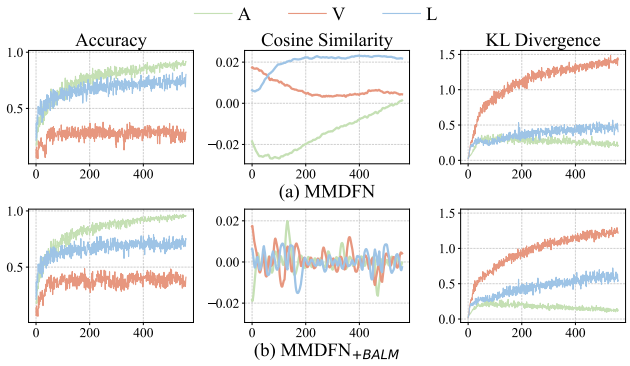


Figure 18. $(r_A, r_L, r_V) = (0.3, 0.7, 0.5)$

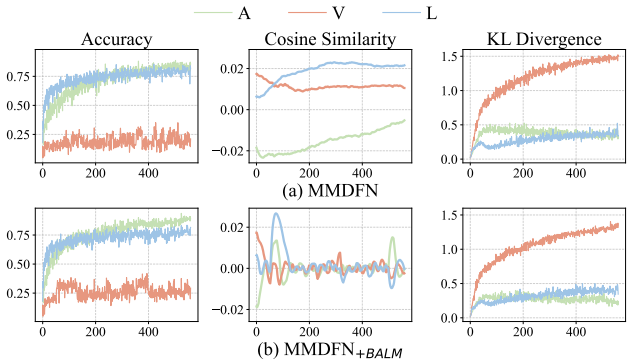


Figure 19. $(r_A, r_L, r_V) = (0.5, 0.3, 0.7)$

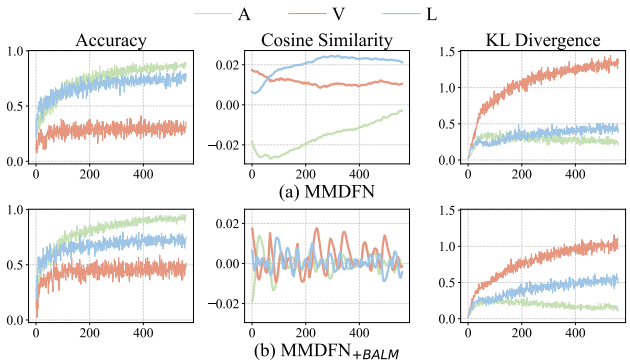


Figure 20. $(r_A, r_L, r_V) = (0.5, 0.7, 0.3)$

Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 3

[6] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE, 2022. 2, 3

[7] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online, 2021. Association for Computational Linguistics. 2, 3

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convo-

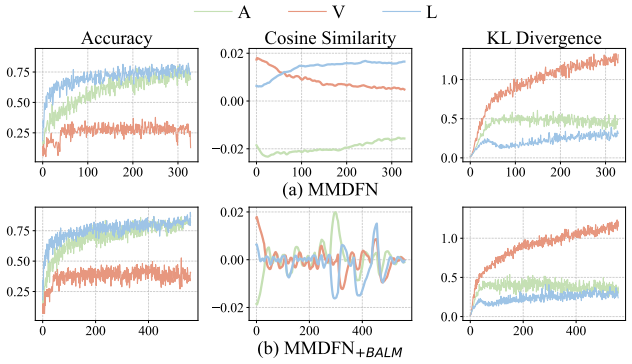


Figure 21. $(r_A, r_L, r_V) = (0.7, 0.3, 0.5)$

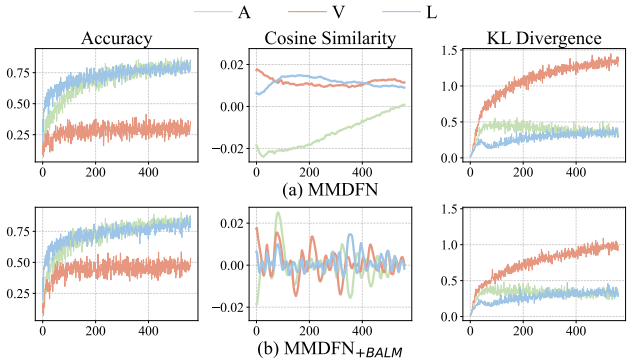


Figure 22. $(r_A, r_L, r_V) = (0.7, 0.5, 0.3)$

lutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[9] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 7

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[11] Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26: 776–788, 2024. 2

[12] Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9330–9339, 2024. 2, 3

[13] Cam-Van Thi Nguyen, Hai-Dang Kieu, Quang-Thuy Ha, Xuan-Hieu Phan, and Duc-Trong Le. Mi-cga: Cross-modal graph attention network for robust emotion recognition in the presence of incomplete modalities. *Neurocomputing*, 623:129342, 2025. 3

[14] 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. 2

[15] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 2

[16] Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15173–15182, 2024. 3

[17] Wenxin Xu, Hexin Jiang, et al. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM Multimedia 2024*, 2024. 3

[18] 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. 2

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 3

[20] Binyu Zhao, Wei Zhang, and Zhaonian Zou. Mce: Towards a general framework for handling missing modalities under imbalanced missing rates. *Pattern Recognition*, page 112591, 2025. 3

[21] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2608–2618, 2021. 3

[22] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 3