CyIN: Cyclic Informative Latent Space for Bridging Complete and Incomplete Multimodal Learning

Ronghao Lin[1,2], **Qiaolin He**[1], **Sijie Mai**[3], **Ying Zeng**[1], **Aolin Xiong**[1],
Li Huang[1,4], **Yap-peng Tan**[2], **Haifeng Hu**[1,5*]

[1] School of Electronics and Information Technology, Sun Yat-Sen University, China
[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3] School of Computer Science, South China Normal University, China
[4] Desay SV Automotive Co., Ltd, China
[5] Pazhou Laboratory, China

{linrh7,heqlin5,zengy268,xiongaolin}@mail2.sysu.edu.cn,
sijiemai@m.scnu.edu.cn,Li.Huang@desaysv.com
eyptan@ntu.edu.sg, huhaif@mail.sysu.edu.cn

Abstract

Multimodal machine learning, mimicking the human brain's ability to integrate various modalities has seen rapid growth. Most previous multimodal models are trained on perfectly paired multimodal input to reach optimal performance. In real-world deployments, however, the presence of modality is highly variable and unpredictable, causing the pre-trained models in suffering significant performance drops and fail to remain robust with dynamic missing modalities circumstances. In this paper, we present a novel Cyclic INformative Learning framework (CyIN) to bridge the gap between complete and incomplete multimodal learning. Specifically, we firstly build an informative latent space by adopting token- and label-level Information Bottleneck (IB) cyclically among various modalities. Capturing task-related features with variational approximation, the informative bottleneck latents are purified for more efficient cross-modal interaction and multimodal fusion. Moreover, to supplement the missing information caused by incomplete multimodal input, we propose cross-modal cyclic translation by reconstruct the missing modalities with the remained ones through forward and reverse propagation process. With the help of the extracted and reconstructed informative latents, CyIN succeeds in jointly optimizing complete and incomplete multimodal learning in one unified model. Extensive experiments on 4 multimodal datasets demonstrate the superior performance of our method in both complete and diverse incomplete scenarios. [1]

1 Introduction

To obtain the optimal multimodal performance, previous methods implicitly assume that every modality present at training will also be available at inference time. However, in real-world scenarios, multimodal data may be missing due to numerous factors such as fail sensors, causing uncertainty of the presence of input modalities [1]. The multimodal model exhibit pronounced sensitivity to such incomplete multimodal input, resulting in severe performance degradation when deploying the pre-trained multimodal model in downstream inference, especially for Transformer-based models [2]. This issue is summarized as missing modality issue [3–5], and the methods devoted to addressing it are called incomplete multimodal learning methods [6–8].

* Corresponding author.
[1] Code is released at https://github.com/RH-Lin/CyIN.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-0594

Current methods focus on enhancing the robustness of multimodal models by designing various delicate modules in multimodal learning to deal with diverse missing circumstances, mainly divided into alignment and generation methods. The former leverage technologies such as contrastive learning [4, 9, 10], canonical correlation analysis [11–13], and data augmentation like mixup or noisy input [5, 7, 14] to align the representations with complete and incomplete multimodal inputs, while the latter introduce generative models such as autoencoders [15–17], variational autoencoders [18–21], graph-based networks [22, 23], and diffusion models [24] to reconstruct the missing information.

Although these methods have alleviated the missing modality issue by bridging the information gap in varying degrees, they suffer from insufficient exploitation in missing information and task-unrelated noise interruption no matter in alignment or generation [6, 24]. Moreover, previous methods typically require training separate models tailored to each possible combination of missing modalities [16, 22]. Consequently, their capacity in generalization and robustness maybe largely limited due to unknown and dynamic missing circumstances in real-world scenarios. Besides, most of previous methods inevitably sacrifice the complete multimodal performance when addressing incomplete input, failing in jointly combining complete and incomplete multimodal learning in a single model [7].

In this paper, we propose CyIN, a novel Cyclic INformative Learning framework that unifies complete and incomplete multimodal learning within a unified model. Firstly, we constructs an effective informative latent space via token- and label-level Information Bottlenecks (IB) to encourage the information flow in multimodal interaction. The former builds information bridge on token embeddings at low-level perception and adopt cyclic interaction among various unimodal representations, while the latter utilize the ground truth labels as guidance to introduce high-level semantics to the informative space. Combining these two IB objectives, we efficiently capture task-relevant features and filter out the redundant noise by sampling the bottleneck latents with variational approximation. Besides, to address missing modality issue, we introduce cross-modal cyclic translation to perform forward and reverse propagation between remained and missing modalities, thereby reconstructing the missing information in the informative space. By jointly optimizing both cyclic IB and translation process, CyIN seamlessly bridges complete and incomplete multimodal learning in a unified informative latent space. The key contributions of our paper can be summarized as:

- **Informative Latent Bottleneck Space.** Built by token- and label-level IB, the proposed informative latent space efficiently bridge complete and incomplete learning in one unified framework, where multimodal fusion and missing information reconstruction can both benefit from the purified bottleneck latents.
- **Cyclic Interaction and Translation.** The presented cyclic information processing progress greatly boost the performance of cross-modal interaction in complete multimodal learning and enhance the translation quality in incomplete multimodal learning.
- Extensive experiments on 4 datasets validate that CyIN achieves state-of-the-art performance in multimodal learning and maintain robustness across various missing modality scenarios.

2 Preliminary

2.1 Complete and Incomplete Multimodal Representation Learning

Considering multimodal inputs with multiple unimodal data source $X_u \in \{X_0, X_1, ..., X_U\}$, multimodal learning aims at integrating complementary information from paired modalities $u \in \{u_0, u_1, ..., u_U\}$ to learn multimodal representations that can drive inference on a variety of downstream tasks [25, 26], where $|u| = U$ denotes the total number of modalities. The performance hinges on the strategy of multimodal fusion to effectively lverages informaiotn from each modality [27, 28], which in turn requires the completeness of modalities.

As shown in Figure 1, each modality raw input is firstly processed by modality-specific encoders $E_u : X_u \mapsto F_u$ to produce unimodal representations F_u, and then merged by multimodal fusion decoders $D_M : \{F_u\} \mapsto F_M$ to generate the multimodal representations F_M. With multimodal input denoted as $X_u \equiv X_{complete}$, the process of complete multimodal learning can be formulated as:

$$F_M = D_M(F_0, F_1, ..., F_U),\ F_u = E_u(X_u),\ \text{where } u \in \{u_0, u_1, ..., u_U\} \tag{1}$$

When there are missing modalities in the multimodal inputs, denoted as $X_u \equiv X_{incomplete} = \{X^{remain}, X^{miss}\}$, the learning process of multimodal models is turned into incomplete multimodal learning, where X^{remain} denotes the original unimodal data input while X^{miss} denotes the zero vector without corresponding unimodal information. Thus, the incomplete multimodal representations F_M in Equ. 1 should be denoted as:

$$F_M = D_M(F^{remain}, F^{miss}),\ F^{remain} = E_u(X_u), F^{miss} = \mathbf{0} \text{ , where } u \in \{u_0, u_1, ..., u_U\} \quad (2)$$

The final prediction $\hat{y}_m$ is obtained by *MLP* head on the multimodal representations $\hat{y}_M = MLP(F_M)$. Then, the optimization objective can be computed according to the specific regression (Mean Absolute Error) or classification (Cross Entropy) tasks, denoted as:

$$\mathcal{L}_{task} = \begin{cases} \mathbb{E}\left|y_{gt} - \hat{y}_M\right| \text{ , for regression} \\ -\mathbb{E}\left[y_{gt} \log \hat{y}_M\right] \text{ , for classification} \end{cases} \quad (3)$$

Due to the fact that the missing information will interrupt the original multimodal fusion space [4, 22] and semantic ambiguity issue may raised by overfitting on the remained modalities [2, 14], the performance of incomplete multimodal learning decreases significantly compared with complete multimodal learning. Thus, the demand of jointly improving complete multimodal fusion and enhancing the reconstruction performance for incomplete input with missing modalities in one unified model remains challenging for the general multimodal systems [7].

2.2 Information Bottleneck

The approach of Information Bottleneck (IB) is proposed to compress the source state S into a compact bottleneck latent B, which contains the least necessary features from S while preserves the most relevant information about the target state T [29–31]. Using mutual information to provide the the relevance between states, such trade-off can be written as the following minimization problem:

$$\min_{p(B|S)} I(S; B) - \beta\, I(B; T) \quad (4)$$

where $I(\cdot|\cdot)$ denotes the mutual information between two states and the hyper-parameter β controls the trade-off degree. By introducing parameterized networks, the former mutual information term $I(B; S)$ can be modeled by IB encoder $E_S : S \mapsto B$ to extract information from source state, while the second term $I(T; B)$ can be conducted by IB decoder $D_T : B \mapsto T$ referring to the target state.

For the purpose of optimizing, VIB [32] utilize variational approximation to replace the intractable mutual information terms, yielding the following upper bound:

$$I(S; B) - \beta\, I(B; T) \leq \mathbb{E}_{S \sim p(S)}\left[KL(p_\theta(B|S) \parallel q(B))\right] - \beta\, \mathbb{E}_{B \sim p(B|S)} \mathbb{E}_{S \sim p(S)}[\log q_\phi(T|B)] \quad (5)$$

where $p_\theta(B|S) \sim \mathcal{N}(\mu, \sigma^2)$ and $q(B) \sim \mathcal{N}(0, \mathbf{I})$ satisfy the Gaussian Distribution. Denoting loss objective $\mathcal{L}_{ib} = I(B; S) - \beta\, I(T; B)$, optimizing $\mathcal{L}_{ib}$ is equivalent to minimizing the upper bound in Equ. 5. Details derivations are presented in Appendix B.

Utilizing reparameterization trick [32, 33] to encourage gradient propagation, the bottleneck latents B can be sampled from $p_\theta(B|S)$, denoted as:

$$B = \mu + \sigma \odot \mathbf{z} \text{ , where } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \quad (6)$$

Serving as a bridge carrying compactly relevant information between S and T, the bottleneck latent B can effectively construct an informative space to control the information flow from the source state S to the target state T. More related works and background can be found in Appendix A

3 Methodology

3.1 Multimodal Informative Latent Space

As shown in Figure 1 , both complete and incomplete multimodal learning require task-relevant informative representations, regardless of modeling intra- or inter-modal relationships. Therefore, constructing a unified informative latent space can not only be beneficial in bridging the modality gap in the fusion of multimodal representations, but strengthening the reconstruction of the most meaningful features among various modalities. To enhance the efficiency of informative space, we introduce two complementary IB mechanisms to encourage the compactness of the bottleneck latents at both perceptual low-level and semantics high-level, referring to token- and label-level IB.

 https://doi.org/10.52202/085713-0594

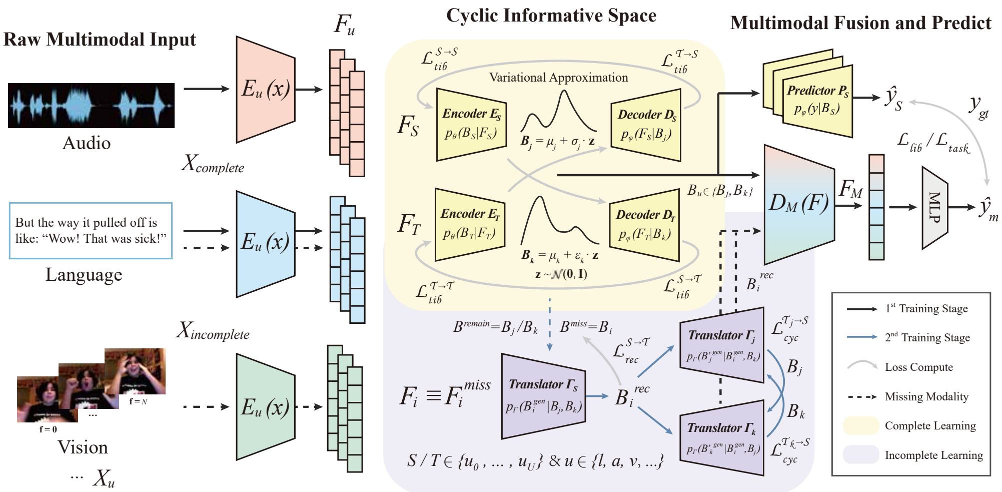


Figure 1: Framework overview. The proposed CyIN build a cyclic informative space to jointly train the complete and incomplete multimodal ealrning.

Token-level Information Bottleneck. Without losing generality, for arbitrary two modalities in multimodal inputs $X_u \in \{X_0, ..., X_U\}$, we denote one modality input X_S as the source state while another one X_T as the target state in Equ. 4. Let $X_S/X_T = \{x_u^0, ..., x_u^L\}$ denote the sequence of token-level inputs for modality u where $S/T \in \{u_0, u_1, .., u_U\}$, where L denotes the sequence length. Then, producing by the corresponding modality-specific encoder $E_u : X_u \mapsto F_u$, the unimodal representations can be obtained as $F_S/F_T = \{f_u^i\}|_{i=1}^L \in \mathbb{R}^{L\times C}$, where f_u denotes token embedding and C denotes the feature dimension of specific modality.

With IB encoder $E_S : F_S \mapsto B_S$, the token-level bottleneck latents $B_S = \{b_S^i\}|_{i=1}^L$ can be attained by applying information bottleneck on all token embeddings, denoted as:

$$\mathcal{L}_{tib}^{S\to T} \approx \frac{1}{L}\sum_i^L \{\mathbb{E}_{f_S^i\sim p(f_S^i)}\left[KL(p_\theta(b_S^i|f_S^i) \parallel q(b_S^i))\right] - \beta\, \mathbb{E}_{b_S^i\sim p(b_S^i|f_S^i)}\mathbb{E}_{f_S^i\sim p(f_S^i)}[\log q_\phi(f_T^i|b_S^i)]\} \tag{7}$$

Since unimodal representations F_S/F_T are concatenated by token embeddings with continuous value, we can formulate posterior probability $q_\phi(f_T|b_S)$ with IB decoder $D_T : B_S \mapsto F_T$ to project bottleneck of the source representation into the target one. Besides, given $p_\theta(b_S^i|f_S^i) \sim \mathcal{N}(\mu_B^i, (\sigma_B^i)^2)$ and $q(b_S^i) \sim \mathcal{N}(0, \mathbf{I})$, then Equ. 7 can be derived as

$$\mathcal{L}_{tib}^{S\to T} \approx \frac{1}{L}\sum_i^L \{KL(\mathcal{N}(\mu_B^i, (\sigma_B^i)^2) \parallel \mathcal{N}(0,\mathbf{I})) + \beta\, \mathbb{E}_{b_S}[\parallel f_T - D_T(b_S) \parallel^2]\} \tag{8}$$

Cyclic Interaction. With unimodal representations $F_u \in \{F_0, ..., F_U\}$ from various modalities, the source and target modalities in token-level IB can be cyclically chosen in an iterative way to filter the redundant noise contained in each unimodal representation and enhance cross-modal interaction.

Considering the source state and target state as the same modality when $F_S = F_T$, the optimization of token-level IB focuses on learning modality-specific features and capturing intra-modal dynamics, denoted as $\mathcal{L}_{tib}^{S\to S}$. The two terms in Equ. 8 focuses on compressing sufficient information from each token embedding and re-projecting to the original sequence, respectively.

On the other hand, when the source state and target state comes from diverse modalities with $F_S \neq F_T$, the token-level IB aims at modeling invariant information across different modalities and seizing the inter-modal dynamics. Then Equ. 8 aims at integrating modality-shared features for the source modality embeddings to match the target one. Besides, interchanging the role of modalities as source and target states in Equ. 8, referring as $\mathcal{L}_{tib}^{S\to T}$ and $\mathcal{L}_{tib}^{T\to S}$, the modality-shared features can be further explored by such directional information flow.

Combing the intra- and inter-modal settings, the final loss of token-level IB can be denoted as:

$$\mathcal{L}_{tib} = \mathbb{E}_{S\cup T}[\mathcal{L}_{tib}^{S\to S} + \frac{1}{2}(\mathcal{L}_{tib}^{S\to T} + \mathcal{L}_{tib}^{T\to S})] \text{ , where } S/T \in \{u_0, ..., u_U\} \tag{9}$$

Label-level Information Bottleneck. Except for token-level IB focusing on low-level perception, we introduce label-level IB to inject the supervision of high-level semantics in the information flow. For unimodal representation F_S from each modality X_S as the source state, we utilize the ground truth label y_{gt} (regression scores or recognition classes) as the supervised target. Given N batched samples $\{X_S^i\}|_{i=1}^N$ for each modality, with IB encoder $E_S : F_S \mapsto B_S$ and predictor $P_S : B_S \mapsto \hat{y}_S$, the label-level bottleneck latents can be constrained with the following loss:

$$\mathcal{L}_{lib}^S \approx \frac{1}{N}\sum_i^N \{\mathbb{E}_{F_S^i\sim p(F_S^i)}\left[KL(p_\theta(B_S^i|F_S^i) \parallel q(B_S))\right] - \beta\, \mathbb{E}_{B_S^i\sim p(B_S^i|F_S^i)}\mathbb{E}_{F_S^i\sim p(F_S^i)}[\log q_\phi(y_{gt}|B_S^i)]\} \tag{10}$$

Iterating computing $\mathcal{L}_{lib}^S$ through all input modalities $S \in \{u_0, u_1, ..., u_U\}$, we can derive the overall label-level IB loss. In practice, for regression task, the posterior probability $q_\phi(y_{gt}|B_S)$ is formulated as $q_\phi(y_{gt}|B_S) = e^{-\|y_{gt}-\hat{y}_S\|} = e^{-\|y_{gt}-P_S(B_S)\|}$, then we have:

$$\mathcal{L}_{lib} \approx \mathbb{E}_S\{\frac{1}{N}\sum_i^N \left[KL(\mathcal{N}(\mu_B^i, (\sigma_B^i)^2) \parallel \mathcal{N}(0, \mathbf{I}))\right] + \beta\, \mathbb{E}_{B_S}[\| y_{gt} - P_S(B_S) \|]\} \tag{11}$$

While for classification task with V classes, the posterior probability $q_\phi(y_{gt}|B_S)$ is computed as $q_\phi(y_{gt}|B_S) = \prod^V \hat{y}_S^{y_{gt}} = \prod^V [P_S(B_S)]^{y_{gt}}$, then we have:

$$\mathcal{L}_{lib} \approx \mathbb{E}_S\{\frac{1}{N}\sum_i^N \left[KL(\mathcal{N}(\mu_B^i, (\sigma_B^i)^2) \parallel \mathcal{N}(0, \mathbf{I}))\right] - \beta\, \mathbb{E}_{B_S}[\sum^V y_{gt} \log P_S(B_S)]\} \tag{12}$$

Optimized by token- and label-level IB loss, we can build an informative space which controls the information flow inside and across modalities in the guidance of task-related semantics.

3.2 Cross-modal Cyclic Translation

Since the bottleneck latents are learned in the informative space, task-irrelevant features and unimodal inherent noise are sufficiently filtered out. Hence, with incomplete multimodal inputs, reconstructing missing information in the built informative latent space becomes significantly easier than attempting reconstruction in the original space. Here we present the cross-modal cyclic translation with forward and reverse propagation to enhance model's robustness under incomplete multimodal scenarios.

Forward Propagation. In order to reconstruct the missing information, we leverage Cascaded Residual Autoencoder (CRA) [34] as the translator $\Gamma_{S\to T} : B_S \mapsto B_T$ across various modalities as the machine translation task [35]. CRA has been demonstrated sufficiently in mitigating the modality gap and translating information from the source modality S to the target one T, denoted as:

$$B_{S\to T}^{rec} = \Gamma_{S\to T}(B_S) = p_\Phi(B_T|B_S) \tag{13}$$

where translator Γ consists of a series of stacked Residual Autoencoders $RA_i^S(\cdot)$ denoted as:

$$r_T = \begin{cases} RA_1^S(B_S),\ i = 1 \\ RA_i^S(B_S + \sum_{i=1}^{n-1} RA_i^S(B_S)),\ i > 1 \end{cases} \tag{14}$$

where r_T denotes the output of each RA block and the last output of RA_i is the overall translated information denoted as $B_{S\to T}^{rec}$. For the translation process $S \to T$, we align the translated information with the original one by a reconstruction loss computed as:

$$\mathcal{L}_{rec}^{S\to T} = \| B_T - B_{S\to T}^{rec} \|^2 \tag{15}$$

Reverse Propagation. To improve translation performance and encourage sufficient exploration of inter-modal dynamics, we further apply cyclic consistent learning [17, 36] to reverse the translated

https://doi.org/10.52202/085713-0594

direction of information flow as back-translation trick [37]. Since forward propagation denotes the reconstruction process of information flow $B_S \rightarrow B^{rec}_{S\rightarrow T}$ from source to target modality, we adopt reverse propagation to translate the reconstructed information back as $B^{rec}_{S\rightarrow T} \rightarrow B^{cyc}_S$, denoted as:

$$B^{cyc}_S = \Gamma_{T\rightarrow S}(B^{rec}_{S\rightarrow T}) = p_\Phi(B_S|B^{rec}_{S\rightarrow T}) \tag{16}$$

where translator $\Gamma_{T\rightarrow S}$ shares the model weights with the one used in forward propagation. Thus, the reconstruction loss for the reverse propagation can be denoted as:

$$\mathcal{L}^{T\rightarrow S}_{cyc} = \| B_S - B^{cyc}_S \|^2 \tag{17}$$

Generalize to Multiple Remained Modalities. For multimodal learning with more than two modalities as input, the missing circumstance of modalities is highly uncertain when encountering incomplete multimodal input. When the number of remained modalities are more than one, how to effectively integrating features from these modalities to reconstruct the missing one is yet to address due to the diverse incomplete input circumstances. Thus, we present a generalizable solution to employ the paired cross-modal translator to generate missing information regardless of the number of remained modalities.

Considering data $X^{incomplete} \equiv \{X^{remain}, X^{miss}\}$ with remained $\{u_j, ..., u_k\}$ and missing u_i modality, we aims at translating the informative bottleneck latents $B^{remain} = \{B_j, ..., B_k\}$ of the remained modalities X^{remain} into the latents B^{miss} of missing modality X^{miss}, denoted as:

$$B^{miss} \coloneqq B^{rec}_i = \Gamma_u(B^{remain}) = p_\Phi(B_i|B_j, ..., B_k) \tag{18}$$

Instead of training more translators to fit in various circumstances of the remained modalities input, we leverage the additive characteristic of Gaussian Distribution to integrate the translated missing information. Since the original bottleneck latents $\{B^{remain}, B^{miss}\}$ are all sampled from the informative space, we directly combine the outputs of each cross-modal translator according to the source modality and denote the translated latents of missing modalities as:

$$B^{rec}_i = \sum^{|u|}_{j\neq i} B^{rec}_{j\rightarrow i} = \sum^{|u|}_{j\neq i} \Gamma_{j\rightarrow i}(B_j) = \prod^{|u|}_{j\neq i} p^j_\Phi(B_i|B_j) \tag{19}$$

where B^{rec}_i can be denoted as the translation from a special form of Gaussian Mixture Model $\sum \Gamma_{j\rightarrow i}(\mathcal{N}(\mu_j, \sigma^2_j))$. Regardless of the specific settings of remained modalities, we utilize B^{rec}_i as the final translated informative latents to supplement the information of missing modality B^{miss}.

Combining forward and reverse propagation, the objective for cross-modal cyclic translation is:

$$\mathcal{L}_{tran} = \mathcal{L}^{S\rightarrow T}_{rec} + \mathcal{L}^{T\rightarrow S}_{cyc} \text{ , where } S/T \in \{u_i, u_j, ..., u_k\} \text{ and } S \neq T \tag{20}$$

3.3 Complete and Incomplete Multimodal Fusion

We jointly conduct complete and incomplete multimodal learning in one unified multimodal fusion network, boosting the robustness with incomplete input and maintaining the performance with complete ones. For simplicity, we employ Multimodal Transformer [38] to present multimodal fusion decoder $D_M : \{B_0, B_1, ..., B_U\} \mapsto F_M$ for bottleneck latents output by the cyclic information space. Note that the fusion module can be replaced by any networks.

Considering bottleneck latents of arbitrary two modalities $\{B_j, B_k\}$, the multimodal fusion ecoder D_M consists of a series of multi-head cross-modal attention layers, composed of:

$$\begin{aligned}
&y^h_r = CMAttention_r(B_j, B_k) = Softmax\left[B_j W^h_q \cdot (W^h_k B_k)^T/\sqrt{C}\right] B_k W^h_v \\
&Y^H_r = Concat(y^1_r, .., y^h_r)W^o \\
&Z_r = Y^H_r + LayerNorm(Y^H_{r-1}) \text{ , where } r \in \{1, ..., R\} \\
&M_r = FeedForward(LayerNorm(Z_r)) + LayerNorm(Z_r)
\end{aligned} \tag{21}$$

where H denotes the number of heads, C denotes the dimension of bottleneck latents, W_q, W_k, W_v, W_o denotes the weight matrices and R denotes the number of cross-modal attention layers. The output of last layer $M^{j,k}_R$ is used as the bimodal fusion latent of modalities $\{u_j, u_k\}$.

Iteratively extracting $M_R^{j,k}$ with paired modalities from $u \in \{u_0, ..., u_U\}$ for both complete and incomplete multimodal learning, the final multimodal representation can be denoted as:

$$F_M = \begin{Bmatrix} D_M(B_0, B_1, ..., B_U) \\ D_M(B^{remain}, B^{miss}) \end{Bmatrix} = Concat([M_R^{0,1}, M_R^{1,2}, ..., M_R^{U-1,U}]) \quad (22)$$

Optimization Objective To sum up, our framework involves four learning objectives, including task prediction loss $\mathcal{L}_{task}$ in Equ. 3, token-level information bottleneck loss $\mathcal{L}_{tib}$ in Equ. 9, label-level information bottleneck loss $\mathcal{L}_{lib}$ in Equ. 11 or Equ. 12, and cross-modal cyclic translation loss $\mathcal{L}_{tran}$ in Equ. 20. The total loss can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \frac{1}{\beta}(\mathcal{L}_{tib} + \mathcal{L}_{lib}) + \gamma\mathcal{L}_{tran} \quad (23)$$

where β denotes the balancing trade-off weight of mutual information among bottleneck latents and representations and γ denotes the weight of translation training for incomplete multimodal learning.

Multi-stage Training Since both complete and incomplete multimodal learning requires an effective informative bottleneck space, we divide the training process into two stages. The first stage set $\gamma = 0$ to make the multimodal learning focus on the construction and stabilization of informative space under complete multimodal learning, while the second stage set $\gamma > 0$ to gradually introduce the training of cross-modal translator to enhance the incomplete multimodal learning.

4 Experiments

In this section, we conduct comprehensive experiments for complete and incomplete multimodal learning on 4 datasets to evaluate the performance of the proposed framework.

Tasks and Datasets. *Multimodal Regression task*: **MOSI** and **MOSEI** are multimodal sentiment analysis datasets contain 2,199 and 22,856 YouTube opinion video clips, respectively, where each clip is annotated with a continuous sentiment score ranging from –3 (strongly negative) to +3 (strongly positive) as Likert scale. *Multimodal Classification task*: **IEMOCAP** dataset provides 7,369 video-recoded conversation, annotated with six emotion categories: {happy, sad, neutral, angry, excited, and frustrated}. **MELD** dataset includes 13,391 utterances drawn from multi-party conversations of television series Friend, labeled with seven emotions: {neutral, surprise, fear, sadness, joy, disgust, and anger}. These two datasets are provided for multimodal emotion recognition.

Implementation Details. All experiments are performed on H800 GPU with Pytorch 2.4.1 on CUDA 12.4. For audio and vision modality, we leverage ImageBind [39] as a feature extractor for better alignment performance. For language modality, we use pre-trained BERT [40] on MOSI and MOSEI while sBERT [41] on IEMOCAP and MELD for fair comparison with baselines. Detailed about hyper-parameters in each dataset is presented in Appendix D.

Evaluation Protocols. Following [8, 17, 22, 24], we evaluate the performance of different methods by (1) complete multimodal input ($u \in \{l, a, v\}$) and under (2) fixed missing protocol ($u \in \{l\}/\{v\}/\{a\}/\{l, a\}/\{l, v\}/\{a, v\}$) and (3) random missing protocol with missing rates $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$. Here $u \in \{l, a, v\}$ denote language, audio and vision modalities. Note that diverse with previous methods, we leverage one unified model for evaluation on various input circumstances with 10 runs. The practical details of missing protocols can be found in Appendix E. Details about baselines and evaluation metrics can be found in Appendix F-G.

Quantitative Comparisons. The average results under both complete input or various incomplete multimodal input circumstances are reported in Table 1. Compared with previous baselines, CyIN reaches superior performance on most metrics in scenarios with both complete and incomplete multimodal input. Specifically, with the purify capability of informative space, CyIN sufficiently explore the modality-specific and -shared features and conduct efficient multimodal fusion based on the bottleneck latents when all multimodal input are present.

Under fixed missing protocols, CyIN consistently outperforms baselines across fixed one or two missing modality settings, demonstrating strong generalization even without specialized tuning for any specific modalities. The ability to integrate informatio from both dominant and inferior modalities highlights the flexibility of the informative bottleneck space. Besides under random missing protocols, CyIN remarkably enhances the model's robustness to various random modality missing scenarios,

https://doi.org/10.52202/085713-0594

Table 1: Performance comparison between the proposed CyIN and baselines with the average results in complete modality setting $u \in \{l, a, v\}$ and incomplete modality settings, with fixed missing protocols including modality settings $u \in \{l\}/\{v\}/\{a\}/\{l,a\}/\{l,v\}/\{a,v\}$ and random missing protocols including missing rates $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$

Setting	Models	MOSI				MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
Complete $u \in \{l, a, v\}$	CCA	27.7	74.9	1.106	0.541	46.1	82.9	0.654	0.666	61.7	61.5	51.2	46.6
	DCCA	25.1	73.6	1.220	0.422	39.3	73.3	0.787	0.425	56.8	55.5	47.7	37.0
	DCCAE	19.7	69.5	1.642	0.357	38.9	73.5	0.782	0.437	57.4	56.4	48.0	36.9
	CPM-Net	16.4	65.5	1.337	0.348	35.9	75.4	0.873	0.375	55.7	56.2	42.3	38.0
	CRA	34.8	83.2	0.916	0.741	51.4	85.5	0.553	0.765	63.4	62.2	57.6	54.8
	MCTN	43.0	84.6	0.752	0.783	47.9	84.2	0.592	0.721	58.4	57.8	56.3	52.4
	MMIN	43.2	85.0	0.744	0.782	52.9	84.9	0.537	0.769	62.5	62.7	60.6	56.2
	GCNet	43.6	85.8	0.732	0.792	52.6	85.9	0.531	0.778	63.0	63.0	60.4	58.5
	IMDer	43.8	85.7	0.724	0.796	**53.8**	85.1	0.532	0.756	64.4	64.8	61.1	59.7
	LNLN	44.0	84.3	0.762	0.766	52.6	85.1	0.542	0.772	62.9	62.5	58.2	57.1
	CyIN	**48.0**	**86.3**	**0.712**	**0.801**	53.2	**86.1**	**0.530**	**0.774**	**66.1**	**66.0**	**61.6**	**59.8**
Fixed Missing	CCA	21.8	57.7	1.264	0.339	43.1	69.7	0.744	0.446	45.9	41.0	49.1	39.4
	DCCA	19.7	58.8	1.418	0.261	41.3	70.0	0.799	0.392	41.3	38.9	47.0	34.4
	DCCAE	21.6	62.8	1.444	0.306	39.5	67.6	0.806	0.370	42.2	40.2	46.9	33.9
	CPM-Net	17.1	60.1	1.353	0.332	38.6	73.8	1.095	0.139	41.3	39.8	33.8	32.8
	CRA	27.3	67.8	1.158	0.404	45.4	78.5	0.672	0.593	47.9	45.9	50.7	45.4
	MCTN	28.5	63.0	1.104	0.392	44.9	73.2	0.717	0.409	40.6	34.3	52.4	42.0
	MMIN	31.3	68.4	1.093	0.433	46.2	77.7	0.661	0.588	50.8	50.2	53.9	43.4
	GCNet	29.5	69.5	1.065	0.538	45.5	73.6	0.697	0.551	52.8	51.9	50.4	46.7
	IMDer	31.4	70.6	1.043	0.533	47.1	76.6	0.680	0.583	54.7	54.4	53.1	**49.8**
	LNLN	29.7	64.8	1.102	0.428	46.9	77.6	0.663	0.581	53.6	52.1	49.6	44.7
	CyIN	**32.8**	**72.2**	**1.037**	**0.599**	**47.6**	**78.6**	**0.656**	**0.594**	**57.4**	**56.6**	**54.4**	49.4
Random Missing	CCA	23.1	66.3	1.220	0.420	44.1	74.8	0.725	0.526	49.9	49.3	48.8	39.8
	DCCA	21.9	64.8	1.279	0.323	40.9	68.6	0.797	0.380	43.3	42.2	47.3	33.8
	DCCAE	19.7	62.7	1.566	0.276	39.0	65.4	0.810	0.345	43.2	41.7	47.1	34.3
	CPM-Net	16.9	63.9	1.340	0.325	34.3	74.0	1.497	0.078	53.1	53.7	41.6	33.6
	CRA	28.6	68.4	1.145	0.558	46.6	80.1	0.647	0.635	49.9	49.4	52.0	47.6
	MCTN	30.4	67.2	1.052	0.573	45.4	73.2	0.692	0.550	35.3	37.5	52.4	44.1
	MMIN	33.3	70.9	1.014	0.584	47.5	79.3	0.644	0.635	49.7	49.7	53.9	45.8
	GCNet	33.8	73.8	0.989	0.623	46.6	79.2	0.680	0.630	55.3	55.3	47.8	47.7
	IMDer	34.6	74.9	0.950	0.644	47.3	78.9	0.660	0.611	55.8	56.1	52.1	49.5
	LNLN	34.2	72.8	0.978	0.627	47.9	79.2	0.639	0.642	55.5	55.8	51.7	49.0
	CyIN	**35.0**	**75.7**	**0.943**	**0.650**	**48.3**	**79.9**	**0.633**	**0.650**	**57.5**	**57.5**	**54.8**	**50.5**

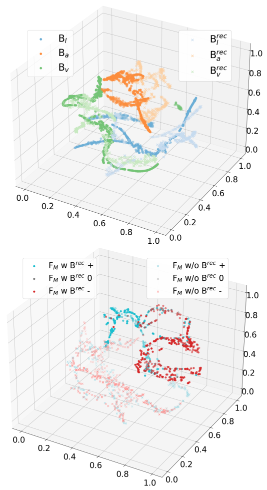
with minimal performance loss even at severe missing rate. The result illustrates broader application of CyIN for downstream multimodal tasks as in real-world where the presence of modalities is highly dynamic and unpredictable. More results and analysis are reported in Appendix H

Qualitative Comparisons. We project the features from test set of MOSI dataset into a 3D t-SNE space [42]. As shown in Figure 2(a), we firstly visualize the distribution of translated B_u^{rec} and original unimodal latents B_u to show the reconstruction quality. The cross-modal translated latents closely clustered to the original one, illustrating the effectiveness of CyIN in transferring information across various modalities despite the huge modality gap.

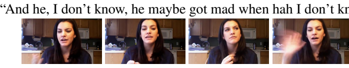
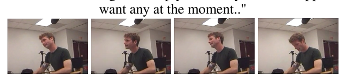
Besides, we present multimodal latents F_M with and without reconstructed bottleneck B^{rec} with random missing $MR = 0.7$. We can observe that inferring with missing modalities leads to serious interference to the multimodal representations, while the reconstructed latents can competently supplement the missing information during multimodal fusion, yielding better clustering result.

Case Study. As shown in Figure 2(b), we compare model predictions with and without CyIN on some examples from test set of MOSI and IEMOCAP datasets under fixed missing modality protocol. In samples$\#1$ and $\#3$, removing the semantic utterance causes the pretrained model to make incorrect, even opposite, predictions, which implicitly showing the dominant role of language modality in multimodal affective computing tasks [8, 16, 43, 44]. On the other hand, the presence of other modalities also have delicate contribution in the accuracy of prediction, referring to the subjective biased prediction when missing audio or vision modality in sample $\#2$ and $\#4$. In contrast, by reconstructing the missing information with the proposed CyIN, predictions become more stable and precise, demonstrating the superiority and robustness of the informative bottleneck space.

Ablation Study. We conduct ablation study with the proposed modules on MOSI and IEMOCAP datasets with complete and the most severe random missing protocol $MR = 0.7$, as shown in Table 2. Token- and label-level IB are both crucial for constructing the informative latent space, as removing either results in performance degradation. The cross-modal cyclic interaction and translation mainly effect in capturing modality-shared features in multimodal fusion and enhancing



(a) Feature Distribution

#	Multimodal Input ($u \in \{language, vision, audio\}$)	Modal Status	Ground Truth	Predict w CyIN	Predict w/o CyIN
1	"And he, I don't know, he maybe got mad when hah I don't know"  Speculative and Wondering Tone	✗ ✓ ✓	-0.250	-0.364	0.263
2	"So I think it was cool to actually see Ray Park in action, great action star"  Satisfied and Relief Tone	✓ ✗ ✗	2.400	2.609	1.110
3	"You needn't be so grand simply because you don't happen to want any at the moment.."  Accusatory and Agitated Tone	✗ ✓ ✓	Angry	Angry	Frustrated
4	"You won't have trouble, You won't have trouble."  Calm and Fast Tone	✓ ✗ ✓	Happy	Happy	Neutral

(b) Case Study on MOSI (#1-2) and IEMOCAP (#3-4)

Figure 2: (a) Feature distribution of translated unimodal latents and multimodal latents and (b) Examples on the test set of MOSI and IEMOCAP datasets when inferring with and without reconstructed information from CyIN. The ✓ and ✗ in modal status denotes the remained and missing modalities.

incomplete reconstruction, respectively. Moreover, we demonstrate the efficiency of informative bottleneck space in both complete and incomplete multimodal learning. Without the constrain of bottleneck, the task-irrelevant redundancy and heterogeneous noise in the original feature space raises difficulty in integrating information from various modalities and reconstructing missing modalities. Lastly, reconstructed from the remained modalities, the translated latents can productively increase the robustness to missing modality issue especially in severe incomplete input circumstance.

Table 2: Ablation study of the proposed CyIN on MOSI and IEMOCAP dataset with complete multimodal settings $u \in \{l, a, v\}$ and random missing protocols with missing rates $MR = 0.7$.

Setting	Model Variants	MOSI				IEMOCAP	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑
Complete $u \in \{l, a, v\}$	**CyIN**	**48.0**	**86.3**	0.712	**0.801**	**66.1**	**66.0**
	w/o $\mathcal{L}_{tib}$	43.9	84.3	0.737	0.789	65.2	64.8
	w/o $\mathcal{L}_{lib}$	47.3	85.4	**0.693**	0.800	63.6	63.9
	w/o Cyclic Interaction	43.4	85.9	0.742	0.795	64.3	63.6
	w/o Cyclic Translation	43.3	85.7	0.743	0.797	65.2	65.0
	w/o Informative Space	42.0	83.1	0.747	0.782	62.3	62.1
Random Missing $MR = 0.7$	**CyIN**	**28.0**	65.9	**1.117**	**0.530**	**48.6**	**49.0**
	w/o $\mathcal{L}_{tib}$	27.3	63.5	1.223	0.509	46.3	46.2
	w/o $\mathcal{L}_{lib}$	25.1	63.7	1.220	0.516	43.2	42.9
	w/o Cyclic Interaction	26.5	**67.8**	1.134	0.473	47.8	48.6
	w/o Cyclic Translation	24.5	63.9	1.171	0.473	47.1	46.6
	w/o Informative Space	23.7	62.9	1.240	0.430	44.1	43.6
	w/o Translated Latents	23.4	56.7	1.299	0.441	41.2	39.2

Hyper-parameter Sensitivity. We empirically evaluate the effect of various hyper-parameter settings of β and γ in loss function of Equ. 23 under $MR = 0.7$. The results in Table 3 shows that setting $\beta = 8 - 32$ gives better performance as a balanced trade-off for mututal information between $I(S; B)$ and $I(B; T)$, where improper bottleneck strength harms performance. Besides, setting $\gamma = 10$ yields the best overall performance across metrics, indicating that moderate reconstruction enforces better consistency across modalities. We observe that a small γ results in weak regularization, making it hard to align modalities under severe missing conditions. In contrast, a large γ hurts performance by focusing too much on consistency and ignoring the need to effectively build informative space.

Moreover, to ensure generality and stability under diverse missing modality scenarios, CyIN assumes each modality have equal contribution in multimodal learning. However, in practice, different modalities may contribute unequally depending on the task [45]. For instance, language plays a dominant role in multimodal sentiment analysis, as highlighted by prior works [8, 46, 47]. To

https://doi.org/10.52202/085713-0594

Table 3: Hyper-parameter sensitivity on β and γ of CyIN on random missing protocols with missing rate $MR = 0.7$.

Model Variants		MOSI			
		Acc7↑	F1↑	MAE↓	Corr↑
γ=10	β=2	24.6	59.4	1.248	0.461
	β=4	27.5	61.3	1.226	0.478
	β=8	**30.0**	63.6	1.199	0.519
	β=16	28.0	65.9	**1.117**	**0.530**
	β=32	26.4	**66.0**	1.118	0.506
	β=64	23.3	64.1	1.206	0.463
β=16	γ=1	25.8	64.1	1.197	0.437
	γ=5	27.5	63.1	1.135	0.488
	γ=10	**28.0**	**65.9**	**1.117**	**0.530**
	γ=15	26.8	62.9	1.272	0.519
	γ=20	25.5	60.4	1.311	0.490

Table 4: Hyper-parameter sensitivity on different ratios of CRA layer for translation among language, vision, audio modalities of CyIN on MOSI dataset on random missing protocols with missing rate $MR = 0.7$.

# Layer (la:lv:av)	MOSI			
	Acc7↑	F1↑	MAE↓	Corr↑
1:1:1	28.0	65.9	1.117	0.530
2:2:1	28.5	67.2	1.122	0.521
4:4:1	**28.7**	**67.5**	**1.114**	0.532
1:1:2	27.5	66.3	1.119	0.531
1:1:4	27.4	65.6	1.116	**0.534**

partially address this, we conduct experiments by varying the number of CRA layers in the cross-modal translation modules in Table 4. Specifically, we allocated more layers to reconstruct the language modality comparing with audio and vision modalities, reflecting its higher importance and allowing the model to learn representation with more semantics. This design show the flexibility of CyIN architecture without hardcoding any modality-specific priors in it.

Computational Efficiency. Compared with the state-of-the-art methods, the proposed CyIN achieves the lowest total parameter count and significantly lower FLOPs as shown in Table 5. The inference time of CyIN is $\sim 3\times$ faster than GCNet and $\sim 5\times$ faster than IMDer. Due to the cyclic translation process during training, the training time of CyIN is slightly slower than GCNet which constructs graph neural networks for modality reconstruction, but still faster then IMDer which utilizes score-based diffusion model for cross-modal generation.

Table 5: Comparion of training and inference computation efficiency on MOSI dataset.

Model	Total Param (M)	Total Training Time	Inference Time (/iteration)	FLOPs (T)
GCNet	144.34 M	**1.47h**	70.62s	3.747
IMDer	168.31 M	1.89h	103.75s	5.466
CyIN (ours)	**123.49 M**	1.61h	**22.41s**	**1.594**

5 Conclusion

In this paper, we present a novel framework named CyIN, which constructs an informative latent space to jointly conduct complete and incomplete multimodal learning. Guided by token- and label-level Information Bottlenecks, CyIN succeeds in learning a compact yet semantically rich bottleneck latent which purifies task-related features and improves robust multimodal fusion. The proposed cyclic interaction and translation mechanism further encourage sufficient exploration in inter-modal dynamics and enhances the reconstruction quality in missing modalities. Comprehensive experiments on 4 multimodal datasets demonstrate that CyIN not only surpass previous methods in complete multi-modal learning but also retains superior performance and stability under various incomplete scenarios, highlighting its effectiveness and generalization capacity for real-world multimodal applications.

Limitations

Our framework still has several limitations. 1) It treats all modalities equally to improve the generalization, while ignoring possible imbalance contribution from various modalities. We will try to adaptively allocate weights according to the information abundance of each modality. 2) The performance of current cross-modal translators can be replaced by more recent generative approaches such as diffusion models or flow-based models to attain optimal reconstruction performance. 3) We suppose broader training on diverse multimodal understanding datasets or integration with Large Language Models could strengthen the generalization and scalability of the proposed framework.

https://doi.org/10.52202/085713-0594

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62076262, 61673402, 61273270, 60802069) and by the International Program Fund for Young Talent Scientific Research People, Sun Yat-Sen University.

References

[1] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024. 1, 18

[2] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18156–18165, 2022. doi: 10.1109/CVPR52688.2022.01764. 1, 3, 18

[3] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2302–2310, May 2021. doi: 10.1609/aaai.v35i3.16330. URL https://ojs.aaai.org/index.php/AAAI/article/view/16330. 1

[4] Ronghao Lin and Haifeng Hu. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 12 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00628. URL https://doi.org/10.1162/tacl_a_00628. 2, 3, 18

[5] Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539, 2024. doi: 10.1109/TMM.2023.3267882. 1, 2

[6] Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=XgwTH95kCl. 1, 2

[7] Ronghao Lin and Haifeng Hu. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105:102216, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102216. URL https://www.sciencedirect.com/science/article/pii/S1566253523005328. 2, 3

[8] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mYEjc7qGRA. 1, 7, 8, 9, 24, 25, 29

[9] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17782–17800. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/poklukar22a.html. 2, 18

[10] Rui Liu, Haolin Zuo, Zheng Lian, Björn W. Schuller, and Haizhou Li. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873, 2024. doi: 10.1109/TAFFC.2024.3378570. 2

[11] Harold Hotelling. *Relations Between Two Sets of Variates*, pages 162–190. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_14. URL https://doi.org/10.1007/978-1-4612-4380-9_14. 2, 25

 https://doi.org/10.52202/085713-0594

[12] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1247–III–1255. JMLR.org, 2013. 25

[13] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1083–1092. JMLR.org, 2015. 2, 25

[14] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.50. URL https://aclanthology.org/2022.naacl-main.50. 2, 3, 18

[15] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/5da713a690c067105aeb2fae32403405-Paper.pdf. 2, 25

[16] 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*, volume 33, pages 6892–6899, 2019. 2, 8, 25, 29

[17] 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.203. URL https://aclanthology.org/2021.acl-long.203. 2, 5, 7, 18, 24, 25

[18] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5580–5590, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[19] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. *Variational mixture-of-experts autoencoders for multi-modal deep generative models*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[20] Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2022. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.11.019. URL https://www.sciencedirect.com/science/article/pii/S0893608021004470.

[21] Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1301–1309, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136. 2

[22] 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*, 45(7):8419–8432, 2023. doi: 10.1109/TPAMI.2023.3234553. 2, 3, 7, 18, 24, 25

[23] Yuntao Shou, Tao Meng, Wei Ai, Fuchen Zhang, Nan Yin, and Keqin Li. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. *Information Fusion*, 112:102590, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102590. URL https://www.sciencedirect.com/science/article/pii/S1566253524003683. 2

[24] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BuGFwUS9B3. 2, 7, 18, 24, 25

[25] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023. ISSN 2731-538X. doi: 10.1007/s11633-022-1410-8. URL https://www.mi-research.net/en/article/doi/10.1007/s11633-022-1410-8. 2, 18

[26] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.738. URL https://aclanthology.org/2024.findings-acl.738/. 2

[27] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.09.005. URL https://www.sciencedirect.com/science/article/pii/S1566253520303675. 2, 18, 25

[28] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(3), October 2023. ISSN 1551-6857. doi: 10.1145/3617833. URL https://doi.org/10.1145/3617833. 2

[29] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL https://arxiv.org/abs/physics/0004057. 3, 18

[30] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. doi: 10.1109/ITW.2015.7133169. 18

[31] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ry_WPG-A-. 3, 18

[32] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg. 3, 18

[33] Max Welling Diederik P. Kingma. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 3

[34] 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 (CVPR)*, July 2017. 5, 25

[35] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. January 2015. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. 5

[36] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, WWW '20, page 2514–2520, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380000. URL https://doi.org/10.1145/3366423.3380000. 5

 https://doi.org/10.52202/085713-0594

[37] Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics, July 2018. URL https://sites.google.com/site/wnmt18/home,https://sites.google.com/site/wnmt18/. 2nd Workshop on Neural Machine Translation and Generation, WNMT 2018 ; Conference date: 15-07-2018 Through 20-07-2018. 6

[38] 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 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656. 6

[39] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023. 7

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. 7, 30

[41] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410. 7

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html. 8

[43] 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. URL https://doi.org/10.1145/3394171.3413678. 8, 18, 29

[44] Ronghao Lin and Haifeng Hu. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26: 2740–2755, 2023. doi: 10.1109/TMM.2023.3303711. 8, 29

[45] 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*, 2022. 9

[46] Ronghao Lin and Haifeng Hu. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–18, 2023. doi: 10.1109/TAFFC.2023.3282410. 9, 18

[47] 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. 9, 18

[48] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.723. URL https://aclanthology.org/2021.emnlp-main.723. 18

[49] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore, December 2023. Association for Computational Linguistics.

[50] Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Z9AZsU1Tju. 18

[51] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In Yoav Goldberg, Zornitsa ozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.534. URL https://aclanthology.org/2022.emnlp-main.534. 18

[52] Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.996. URL https://aclanthology.org/2023.emnlp-main.996/.

[53] Ronghao Lin, Ying Zeng, Sijie Mai, and Haifeng Hu. End-to-end semantic-centric video-based multimodal affective computing. *arXiv preprint arXiv:2408.07694*, 2024. 18

[54] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qXZVSy9LFR. 18

[55] Jin Li, Shoujin Wang, Qi Zhang, Shui Yu, and Fang Chen. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2787–2798, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714606. URL https://doi.org/10.1145/3696410.3714606. 18, 30

[56] Vittorio Pipoli, Alessia Saporita, Kevin Marchesini, Costantino Grana, Elisa Ficarra, and Federico Bolelli. Im-fuse: A mamba-based fusion block for brain tumor segmentation with incomplete modalities. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 225–235, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-032-04984-1. 18, 30, 31

[57] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, apr 2024. ISSN 0360-0300. doi: 10.1145/3656580. URL https://doi.org/10.1145/3656580. Just Accepted. 18

[58] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9), April 2024. ISSN 0360-0300. doi: 10.1145/3649447. URL https://doi.org/10.1145/3649447. 18

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 18

 https://doi.org/10.52202/085713-0594

[60] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. doi: 10.1109/TPAMI.2023.3275156. 18

[61] Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia*, 26:5753–5768, 2024. doi: 10.1109/TMM.2023.3338769. 18

[62] Jingjing Tang, Qingqing Yi, Saiji Fu, and Yingjie Tian. Incomplete multi-view learning: Review, analysis, and prospects. *Applied Soft Computing*, 153:111278, 2024. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2024.111278. URL https://www.sciencedirect.com/science/article/pii/S1568494624000528. 18

[63] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415, 2022. doi: 10.1109/TPAMI.2020.3037734. 18, 25

[64] Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. 3 2017. 18

[65] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning (ICML)*, 2023. 18

[66] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. *Deep Multi-view Information Bottleneck*, pages 37–45. doi: 10.1137/1.9781611975673.5. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611975673.5. 18

[67] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

[68] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. Multi-view information-bottleneck representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):10085–10092, May 2021. doi: 10.1609/aaai.v35i11.17210. URL https://ojs.aaai.org/index.php/AAAI/article/view/17210. 18

[69] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 18

[70] Changhee Lee and Mihaela van der Schaar. A variational information bottleneck approach to multi-omics data integration. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1513–1521. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lee21a.html. 18

[71] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 32971–32998. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6818dcc65fdf3cbd4b05770fb957803e-Paper-Conference.pdf. 19

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. 24

[73] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2017.02.003. URL https://www.sciencedirect.com/science/article/pii/S1566253517300738. 25

[74] Geetha A.V., Mala T., Priyanka D., and Uma E. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102218. URL https://www.sciencedirect.com/science/article/pii/S1566253523005341. 26

[75] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS. 30

[76] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA. 30

[77] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 38–54, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72946-1. 30, 31

 https://doi.org/10.52202/085713-0594

A Related Works

A.1 Multimodal Representation Learning

Multimodal representation learning aims at constructing multimodal system to conduct multimodal understanding tasks including multimodal sentiment analysis [43, 46, 48–50] or emotion recognition [51–54], multimodal recommendation [55] and multimodal segmentation [56] and so on. The common framework of multimoda learning can be divided into two main modules: modality-specific representation learning and multimodal fusion, according to the purpose of reflecting heterogeneity and interconnections between modality elements [57].

On the one hand, effective extraction of task-relevant features within each individual modality forms the foundation of multimodal representation learning [47, 53]. Across various modalities, these modality-specific encoders emphasize critical features and suppress unrelated noise, thereby shortening the path from raw data to semantically meaningful unimodal representations. On the other hand, multimodal fusion aims to bridge modality gaps and exploit inter-modal complementarity [27, 57]. Early fusion schemes simply concatenate unimodal representations while late-fusion approaches combine unimodal predictions via weighted voting or gating. Hybrid fusion combine early and late fusion to conduct hierarchical and more delicated cross-modal interaction [58].

With the emerge of attention mechanism [59], multimodal model with Transformer-based architecture efficiently capture intra- and inter-modal dynamics by interleaving unimodal representation with self- and cross-modal attention layers [25, 60]. However, most of previous multimodal methods assume paired, fully observed modalities at both training and inference stages to achieve the state-of-the-art performance of multimodal understanding.

A.2 Missing Modality Issue

The missing modality issue occurs when pre-trained multimodal models suffer from severe performance degradation at downstream inference due to the absent input of modality data [4, 61]. In real world, the presence of multimodal input could not be guaranteed due to numerous reasons, such as sensor failure, hardware malfunctions, privacy limitations, environmental disruptions, and data transmission problems [1]. Empirical analysis have shown that the efficacy of multimodal methods critically depends on complete presence of modalities [14], and that Transformer-based encoders are especially sensitive to missing inputs [2]. Thus, addressing missing modality issue is essential for robust performance in real-world applications, which is named as incomplete multimodal learning.

Taking alternative perspective of incomplete multimodal learning, the missing modality issue can be regarded as missing specific-view information issue in the multi-view learning [62]. Previous methods mainly focus on aligning complete and incomplete multimodal representations [9, 63] or generating the missing information with incomplete multimodal input [17, 22, 24]. While alignment-based models avoid the complexity of explicit generation, they often yield suboptimal inference with uncertain and heterogeneous information asymmetry. Diversely, generative-based models remain susceptible to the task-unrelated noise and massive redundancy in the unimodal features, which can easily destabilize the information reconstruction or generation process.

Moreover, previous methods require adjusting according to specific missing scenarios to achieve optimal performance, restricting their applications in the uncertain circumstance of the real world. Beside, most of them sacrifice performance of complete multimodal learning to increase the robustness to the missing modality issue. Therefore, achieving a single unified framework that jointly optimize both complete and incomplete multimodal learning remains an open challenge.

A.3 Deep Learning with Information Bottleneck

Original introduced by Tishby et al. [29], Information Bottleneck (IB) formulates representation learning as an information-theoretic approach that defines task-relevant representation as trade-off between feature compression and task prediction [31]. Due to the excellent interpretability of representation learning and generalization estimation of the deep neural networks [30, 64, 65], IB has gained continuous interests in deep learning, especially in multi-view problem [66–68]. VIB [32] introduce variational approximation to generate the bottleneck latents in the information flow, sharing similar objective form with VAE [69] in generation. DeepIMV [70] utilize product-

https://doi.org/10.52202/085713-0594

of-experts to integrate the marginal specific-view representation into a joint latent representation. FactorCL [71] factorizes task-relevant information into shared and unique information with multi-view redundancy defined by mutual information. Nevertheless, they can neither be employed to varying missing circumstance nor succeed in achieving optimal intra- and inter-modal information extraction performance.

In this paper, we adopt the principles of IB to construct an effective informative latent space by designing information flow to features among modalities and guidance of semantics, paving the way for enhanced performance in both complete and incomplete multimodal learning.

B Derivations of the Variational Information Bottleneck

In this section, we derive the formula of the original variational information bottleneck and the proposed token- and label-level information bottleneck in detail, as shown in Equ. 5, Equ. 8 and Equ. 11-12. Besides, we provide the differentiable sampling process of the informative latents presented in Equ. 6.

Equ. 5 represents the information bottleneck loss $\mathcal{L}_{ib}$ for the information flow as $(S \to B \to T)$, where S denotes the source state, B denotes the bottleneck latents and T denotes the target state. Recall that the loss is firstly given by the constrained of mutual information, formulated as:

$$\min \mathcal{L}_{ib}(S,T) = \min_{p(B|S)} I(S;B) - \beta\, I(B;T) \tag{24}$$

where $\beta > 0$ is a trade-off parameters to balance the two mutual information terms. Directly obtaining the mutual information among $S/B/T$ is unachievable. Therefore, we tend to obtain the upper bound of $\mathcal{L}_{ib}(S,T)$ to transfer the objective minimization problem to an evidence upper bound optimization problem.

Starting with the first term $I(S;B)$, writing is out in full form as:

$$\begin{aligned} I(S;B) &= \int dB\, dS\, p(S,B) \log \frac{p(S,B)}{p(S)p(B)} = \int dB\, dS\, p(S,B) \log \frac{p(B|S)}{p(B)} \\ &= \int dB\, dS\, p(S,B) \log p(B|S) - \int dB\, dS\, p(B) \log p(B) \end{aligned} \tag{25}$$

Considering that the computation of marginal distribution of the bottleneck latents $p(B) = \int dSp(B|S)p(S)$ might be difficult, let $q(B)$ be a variational approximation to this marginal distribution. With the non-negative Kullback Leibler (KL) divergence $KL(p(B) \parallel q(B)) \geq 0$, we have $\int dB\, p(B) \log p(B) \geq \int dB\, p(B) \log q(B)$, the following upper bound can be derived:

$$\begin{aligned} I(S;B) &\leq \int dBdS\, p(S,B) \log p(B|S) - \int dB\;\, p(B) \log q(B) \\ &= \int dBdS\, p(S)p(B|S) \log \frac{p_\theta(B|S)}{q(B)} \end{aligned} \tag{26}$$

where $p_\theta(B|S)$ can be learned by IB encoder $E_S : S \mapsto B$.

Then for the second term $I(B;T)$, consider it in the same full form as:

$$I(B;T) = \int dT\, dB\, p(T,B) \log \frac{p(T,B)}{p(T)p(B)} = \int dT\, dB\, p(T,B) \log \frac{p(T|B)}{p(T)} \tag{27}$$

Since $p(T|B)$ is intractable, introducing IB decoder $D_T : B \mapsto T$, let $q_\phi(T|B)$ be a variational approximation to $p(T|B)$. Similarly, with KL divergence $KL(p(T|B) \parallel q_\phi(T|B)) \geq 0$, we have $\int dT\, p(T|B) \log p(T|B) \geq \int dT\, p(T|B) \log q_\phi(T|B)$. Hence $I(B;T)$ has a lower bound as follows:

$$\begin{aligned} I(B;T) &\geq \int dT\, dB\, p(T,B) \log \frac{q_\phi(T|B)}{p(T)} \\ &= \int dT\, dB\, p(T,B) \log q_\phi(T|B) - \int dT\, p(T) \log p(T) \\ &= \int dT\, dB\, p(T,B) \log q_\phi(T|B) + H(T) \end{aligned} \tag{28}$$

https://doi.org/10.52202/085713-0594

where $H(T)$ is the entropy of target state T, determined only by the distribution of T itself, no matter as the unimodal token embeddings or ground truth labels. Thus, $H(T)$ can be dropped in the loss function. When the source state S is diverse from the target state T, we introduce the relationship that B is independent of T given S:

$$p(T,B) = \int dS\, p(S,T,B) = \int dS\, p(S)p(T|S)p(B|S) \tag{29}$$

Note that when $S = T$, we have $p(T,B) = p(S,B) = \int p(S)p(B|S) = \int p(S)p(T|S)p(B|S)$ where above expression still stand. Besides, $p(T|S)$ denotes the known distribution for the joint and paired data samples. Utilizing data samples within each batch as the empirical data distribution, we have $p(S,T) = p(S)p(T|S) = \frac{1}{N}\sum_{n=1}^{N} \delta_{S_n}(S)\delta_{T_n}(T)$. Hence, $I(B;T)$ can be generally denoted as:

$$\begin{aligned} I(B;T) &\geq \int dS\, dT\, dB\, p(S)p(T|S)p(B|S)\log q_\phi(T|B) \\ &\approx \mathbb{E}_{S\sim p(S)}\left[\int dB\, p(B|S)\log q_\phi(T|B)\right] \end{aligned} \tag{30}$$

Combining the bound of $I(S;B)$ and $I(B;T)$, we can derive the following upper bound for $\mathcal{L}_{ib}$ as Equ. 5:

$$\begin{aligned} &I(S;B) - \beta\, I(B;T) \\ &\leq \int dBdS\, p(S)p(B|S)\log\frac{p_\theta(B|S)}{q(B)} - \beta\, \mathbb{E}_{S\sim p(S)}\left[\int dB\, p(B|S)\log q_\phi(T|B)\right] \\ &= \mathbb{E}_{S\sim p(S)}\left[KL(p_\theta(B|S)\parallel q(B))\right] - \beta\, \mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)] \end{aligned} \tag{31}$$

Finally, the optimization objective of variational information bottleneck can be rewritten as:

$$\min \mathcal{L}_{ib} = \min \mathbb{E}_{S\sim p(S)}\left[KL(p_\theta(B|S)\parallel q(B))\right] - \beta \max\ \mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)] \tag{32}$$

Now, we derive the practical expression of two minimization and maximization terms of the above upper bound.

First for minimizing $KL(p_\theta(B|S)\parallel q(B))$, in practice, we utilize standard Gaussian Distribution $\mathcal{N}(0,1)$ as the prior distribution of $q(B)$ and model the approximate posterior distribution $p_\theta(B|S)$ as a multivariate Gaussian distribution $\mathcal{N}(\mu,\varepsilon)$. Here μ and ε denotes the mean and variance vectors of the latent Gaussian Distribution. Since each dimension of $q(B)$ and $p_\theta(B|S)$ are independent, we can derive the situation with one-dimensional Gaussian Distribution of $KL(p_\theta(B|S)\parallel q(B))$ as:

$$\begin{aligned} &KL(p_\theta(B|S)\parallel q(B)) = KL(\mathcal{N}(\mu_B,\sigma_B^2)\parallel \mathcal{N}(0,\mathbf{I})) \\ &= \int \mathcal{N}(\mu_B,\sigma_B^2)\log\frac{\mathcal{N}(\mu_B,\sigma_B^2)}{\mathcal{N}(0,\mathbf{I})}\, db \\ &= \int \mathcal{N}(\mu_B,\sigma_B^2)\log\frac{\frac{1}{\sqrt{2\pi\sigma_B^2}}\exp\left(-\frac{(b-\mu_B)^2}{2\sigma_B^2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{b^2}{2}\right)}\, db \\ &= \int \mathcal{N}(\mu_B,\sigma_B^2)\left(-\log\sqrt{\sigma_B^2} - \frac{(b-\mu_B)^2}{2\sigma_B^2} + \frac{b^2}{2}\right) db \\ &= -\frac{1}{2}\int \mathcal{N}(\mu_B,\sigma_B^2)\left(\log\sigma_B^2 + \frac{(b-\mu_B)^2}{\sigma_B^2} - b^2\right) db \\ &= -\frac{1}{2}\left[\log\sigma_B^2\int \mathcal{N}(\mu_B,\sigma_B^2)\, db + \frac{1}{\sigma_B^2}\int \mathcal{N}(\mu_B,\sigma_B^2)(b-\mu_B)^2\, db - \int \mathcal{N}(\mu_B,\sigma_B^2)b^2\, db\right] \end{aligned} \tag{33}$$

Since any probability density function integrates to 1 over its domain, we have $\int \mathcal{N}(\mu_B,\sigma_B^2)\, db = 1$. According to the definition of the variance and second raw moment of Gaussian Distribution, we have $\int \mathcal{N}(\mu_B,\sigma_B^2)(b-\mu_B)^2\, db = \sigma_B^2$ and $\int \mathcal{N}(\mu_B,\sigma_B^2)b^2\, db = \mu_B^2 + \sigma_B^2$. Thus, the analytical solution

https://doi.org/10.52202/085713-0594

of $KL(\mathcal{N}(\mu_B, \sigma_B^2) \parallel \mathcal{N}(0, \mathbf{I}))$ is computed as:

$$\begin{aligned} KL(\mathcal{N}(\mu_B, \sigma_B^2) \parallel \mathcal{N}(0, \mathbf{I})) &= -\frac{1}{2}\left[\log \sigma_B^2 \cdot 1 + \frac{1}{\sigma_B^2} \cdot \sigma_B^2 - (\mu_B^2 + \sigma_B^2)\right] \\ &= -\frac{1}{2}\left(\log \sigma_B^2 + 1 - \mu_B^2 - \sigma_B^2\right) \end{aligned} \tag{34}$$

When each dimension $d \in \{0, ..., C\}$ is independent in multivariate Gaussian distribution $\mathcal{N}(\mu, \varepsilon)$, we have:

$$\begin{aligned} KL(p_\theta(B|S) \parallel q(B)) &= \int p_\theta(b|S) \log \frac{p_\theta(b_1|S) \cdots p_\theta(b_d|S)}{q(b_1) \cdots q(b_d)}\, db \\ &= \int p_\theta(b|S)\left[\sum_{d=1}^{C} \log p_\theta(b_d|S) - \sum_{d=1}^{C} \log q(b_d)\right] db \\ &= \sum_{d=1}^{C} \int p_\theta(b|S)\left[\log p_\theta(b_d|S) - \log q(b_d)\right] db \\ &= \sum_{d=1}^{C} \int p_\theta(b_d|S)\left[\log p_\theta(b_d|S) - \log q(b_d)\right] db_d \\ &= \sum_{d=1}^{C} KL\left(p_\theta(b_d|S) \parallel q(b_d)\right) \end{aligned} \tag{35}$$

Then we have:

$$\min KL(\mathcal{N}(\mu_B, \sigma_B^2) \parallel \mathcal{N}(0, \mathbf{I})) = \min -\frac{1}{2}\mathbb{E}_C\left(\log \sigma_B^2 + 1 - \mu_B^2 - \sigma_B^2\right) \tag{36}$$

where C is the feature dimension of the bottleneck latents B.

While for maximizing $\mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)]$, we leverage diverse expression to conduct computation based on the property of specific target state T, divided into token- and label-level IB in this paper.

B.1 Derivation of Token-level Information Bottleneck

For unimodal representation $F_u = \{f_u\}|_{i=1}^{L} \in \mathbb{R}^{L\times C}$ as target state T, we utilize IB decoder $D_T : B \mapsto T$ to enable information flow across the source and target representations F_S/F_T with a bottleneck bridge B. Since each dimension is independent for token embeddings of F_S/F_T, we have:

$$q_\phi(T|B) = \mathcal{N}(\mu_T, \sigma_T^2) = \prod_{i=1}^{L}\prod_{d=1}^{C} \mathcal{N}(\mu_T^d, (\sigma_T^d)^2) = \prod_{i=1}^{L}\left(\prod_{d=1}^{C} \frac{1}{\sqrt{2\pi}\sigma_T^d}\right) \exp\left[-\sum_{i=1}^{C} \frac{(f_T^d - \mu_T^d)^2}{2(\sigma_T^d)^2}\right] \tag{37}$$

Then we have:

$$\log q_\phi(T|B) = -\sum_{i=1}^{L}\left[\frac{C}{2}\log 2\pi + \frac{1}{2}\sum_{d=1}^{C} \log(\sigma_T^d)^2 + \frac{1}{2}\sum_{d=1}^{C} \frac{(f_T^d - \mu_T^d)^2}{(\sigma_T^d)^2}\right] \tag{38}$$

For simplify, assuming that variance $(\sigma_T^d)^2$ of each dimension $d \in \{0, .., C\}$ is consistent and fixed as a constant, IB decoder only needs to output the mean $\mu^i = D_T(b_S)$ of the projected token embedding. Then the following equation holds:

$$\begin{aligned} \max \mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)] &\equiv \min \sum_{i=1}^{L}\left[\frac{1}{2}\sum_{d=1}^{C}(f_T^d - \mu_T^d)^2\right] \\ &\equiv \min \sum_{i=1}^{L} \mathbb{E}_{b_S} \parallel f_T^d - D_T(b_S) \parallel^2 \end{aligned} \tag{39}$$

https://doi.org/10.52202/085713-0594

Combining Equ. 36 and Equ. 39, we can derive Equ. 8 as the objective of $S \rightarrow T$ token-level information bottleneck:

$$
\begin{aligned}
\mathcal{L}_{tib}^{S\rightarrow T} &\approx \mathbb{E}_{F_S\sim p(F_S)}\left[KL(p_\theta(B_S|F_S) \parallel q(B_S))\right] - \beta\, \mathbb{E}_{B_S\sim p(B_S|F_S)}\mathbb{E}_{F_S\sim p(F_S)}[\log q_\phi(F_T|B_S)] \\
&= \frac{1}{L}\sum_{i}^{L}\{KL(\mathcal{N}(\mu_B^i,(\sigma_B^i)^2) \parallel \mathcal{N}(0,\mathbf{I})) + \beta\, \mathbb{E}_{b_S}[\parallel f_T - D_T(b_S) \parallel^2]\} \\
&= \frac{1}{L}\sum_{i}^{L}\{\left[-\frac{1}{2}\mathbb{E}_C\left(\log\sigma_B^2 + 1 - \mu_B^2 - \sigma_B^2\right)\right] + \beta\, \mathbb{E}_{b_S}[\parallel f_T - D_T(b_S) \parallel^2]\}
\end{aligned}
\tag{40}
$$

B.2 Derivation of Label-level Information Bottleneck

For ground truth label $y_{gt} \in \mathbb{R}^1/\mathbb{R}^V$ as target state T, we utilize two forms of the posterior distribution $q_\phi(T|B)$ according to the continuous (1 as the regression value) or discrete (V as the one-hot recognition classes) forms of the supervision. We leverage unimodal predictor $P_S : B_S \mapsto \hat{y}_S$ for each source modality S to model the posterior distribution $q_\phi(T|B)$ and output the prediction $\hat{y}_S$ based on information of single modality.

With continuous labels $y_{gt} \in \mathbb{R}^1$ in regression task, we consider $q_\phi(T|B)$ as one-dimensional Gaussian Distribution, computed as:

$$
\begin{aligned}
q_\phi(T|B) &= e^{-\|y_{gt}-\hat{y}_S\|} = e^{-\|y_{gt}-P_S(B_S)\|} \\
\log q_\phi(T|B) &= \parallel y_{gt} - P_S(B_S) \parallel
\end{aligned}
\tag{41}
$$

Then we turn the maximization of posterior distribution $q_\phi(T|B)$ into:

$$
\max \mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)] \equiv \min \sum_{i=1}^{L} \mathbb{E}_{B_S} \parallel y_{gt} - P_S(B_S) \parallel \tag{42}
$$

Combining Equ. 36 and Equ. 42, we can derive Equ. 11 as the objective of label-level information bottleneck:

$$
\mathcal{L}_{lib} \approx \frac{1}{N}\sum_{i}^{N}\left[-\frac{1}{2}\mathbb{E}_C\left(\log\sigma_B^2 + 1 - \mu_B^2 - \sigma_B^2\right)\right] + \beta\, \mathbb{E}_{B_S}[\parallel y_{gt} - P_S(B_S) \parallel] \tag{43}
$$

With one-hot discrete labels $y_{gt} \in \mathbb{R}^V$ with V classes in classification task, we consider $q_\phi(T|B)$ as multi-dimensional Bernoulli Distribution, computed as:

$$
\begin{aligned}
q_\phi(T|B) &= \prod^{V} \hat{y}_S^{y_{gt}} = \prod^{V}[P_S(B_S)]^{y_{gt}} \\
\log q_\phi(T|B) &= \sum^{V} y_{gt}\log P_S(B_S)
\end{aligned}
\tag{44}
$$

Then we turn the maximization of posterior distribution $q_\phi(T|B)$ into:

$$
\max \mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log q_\phi(T|B)] \equiv \min \mathbb{E}_{B_S}[\sum^{V} y_{gt}\log P_S(B_S)] \tag{45}
$$

Combining Equ. 36 and Equ. 45, we can derive Equ. 12 as the objective of label-level information bottleneck:

$$
\mathcal{L}_{lib} \approx \frac{1}{N}\sum_{i}^{N}\left[-\frac{1}{2}\mathbb{E}_C\left(\log\sigma_B^2 + 1 - \mu_B^2 - \sigma_B^2\right)\right] + \beta\, \mathbb{E}_{B_S}[\sum^{V} y_{gt}\log P_S(B_S)] \tag{46}
$$

https://doi.org/10.52202/085713-0594

B.3 Reparameterization Trick for Informative Bottleneck Latent

Since the informative bottleneck latents B is sampled from $p_\theta(B|S) \sim \mathcal{N}(\mu, \sigma^2)$ where the sampling process is not differentiable. For the purpose of optimizing the IB encoder $E_S : S \mapsto B$ through gradient decent algorithm, we need to identify the gradient of B to the weight parameters θ of the encoder, denoted as follows according to the Chain Role:

$$\frac{\partial B}{\partial \theta} = \frac{\partial B}{\partial \mu}\frac{\partial \mu}{\partial \theta} + \frac{\partial B}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial \theta} \tag{47}$$

where μ and σ are directly output by the encoder so that $\frac{\partial \mu}{\partial \theta}$ and $\frac{\partial \sigma^2}{\partial \theta}$ is easily tractable.

To compute $\frac{\partial B}{\partial \mu}$ and $\frac{\partial B}{\partial \sigma^2}$ from the sampled bottleneck latent B, we firstly sample a random vector $\mathbf{z}$ from nonparametric distribution $\mathcal{N}(0, \mathbf{I})$ and utilize a transformation function $g(\mathbf{z}, \mu, \sigma^2)$ to attain the bottleneck B, shown as:

$$B = \mu + \sigma \odot \mathbf{z}\text{ , where } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \tag{48}$$

Since the process of sampling $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is independent to parameters μ and σ, the derivation of $\frac{\partial B}{\partial \mu}$ and $\frac{\partial B}{\partial \sigma^2}$ are transformed into the derivation of $\frac{\partial g(\cdot)}{\partial \mu}$ and $\frac{\partial g(\cdot)}{\partial \sigma^2}$, which is differentiable.

C Theoretical Grounding of combining IB with Cyclic Translation

We present further explanation for the theoretical analysis of CyIN in combining Information Bottleneck (IB) with cross-modal cyclic translation. Except for task prediction term, the optimization objective for IB loss and cyclic translation in Equ. 23 is deduced as follows:

$$\begin{aligned} \mathcal{L} &= \frac{1}{\beta}(\mathcal{L}_{tib} + \mathcal{L}_{lib}) + \gamma \mathcal{L}_{tran} \\ &= \frac{1}{\beta}\mathcal{L}_{IB} + \gamma(\mathcal{L}_{rec} + \mathcal{L}_{cyc}) \end{aligned} \tag{49}$$

Considering multimodal learning with two modalities without loss of generality, the theoretical modeling of CyIN can be formulated as follows:

$$\begin{array}{ccccc} S_1 \rightarrow & B_1 & \rightleftharpoons & B_2 & \leftarrow S_2 \\ & \downarrow & & \downarrow & \\ & T_1 & & T_2 & \end{array}$$

where S denotes the source state and T denotes the target state.

- The left and right side with $S_1 \to B_1 \to T_1$ and $S_2 \to B_2 \to T_2$ denote the chain of information bottleneck. Note that the source and target states have $S/T \in \{F_{S_1}, F_{S_2}\}\{F_{T_1}, F_{T_2}\}$ in the token-level IB or $S \in \{F_{S_1}, F_{S_2}\}, T \in y$ in the label-level IB, both of which have been theoretically deduced above, formulated as:

$$\mathcal{L}_{tib} = \underbrace{[I(S_1; B_1) - \beta I(B_1; T_1)]}_{IB\ for\ modality\ 1} + \underbrace{[I(S_2; B_2) - \beta I(B_2; T_2)]}_{IB\ for\ modality\ 2} \tag{50}$$

- The middle part $B_1 \rightleftharpoons B_2$ denote the cyclic translation with $B_1 \rightharpoonup B^{rec}_{2\to 1} \rightharpoonup B^{cyc}_1 \sim B_1$ and $B_2 \rightharpoonup B^{rec}_{1\to 2} \rightharpoonup B^{cyc}_2 \sim B_1$, where $\rightharpoonup$ denotes the translation process while $\sim$ denotes alignment with the original bottleneck. Considering Γ as the translator network, the translation process can be divided into cross-modal reconstruction across diverse unimodal latents B_1/B_2 and cyclic translation from the reconstructed latent $B^{rec}_{1\to 2}/B^{rec}_{2\to 1} = \Gamma_{1\to 2}(B_1)/\Gamma_{2\to 1}(B_2)$ back to the origin $B^{cyc}_2/B^{cyc}_1 = \Gamma_{1\to 2}(B^{rec}_{2\to 1})/\Gamma_{2\to 1}(B^{rec}_{1\to 2})$, formulated as:

$$\begin{aligned} \mathcal{L}_{rec} &= \underbrace{D_{KL}(B_1; B^{rec}_{2\to 1})]}_{Reconstruction\ from\ modality\ 1 \rightharpoonup 2} + \underbrace{D_{KL}(B_2; B^{rec}_{1\to 2})}_{Reconstruction\ from\ modality\ 2 \rightharpoonup 1} \\ \mathcal{L}_{cyc} &= \underbrace{D_{KL}(B_2; B^{cyc}_2)}_{Translating\ Back\ from\ modality\ 2 \rightharpoonup 1} + \underbrace{D_{KL}(B_1; B^{cyc}_1)}_{Translating\ Back\ from\ modality\ 1 \rightharpoonup 2} \end{aligned} \tag{51}$$

https://doi.org/10.52202/085713-0594

Minimizing above losses, we can obtain informative unimodal bottleneck latents B_u for each modality u, which contains inter-modal features to conduct productive multimodal fusion. Comparing with directly fusing unimodal representations F_u, fusion process implemented on the bottleneck latents B_u, denoted as $F_M = D_M(B_1, B_2)$, can benefit from the compression ability of information bottleneck and lead to less reconstruction difficulty in cross-modal translation.

When both modalities u_1 and u_2 are presented, known as complete multimodal learning, we can regard the cyclic translation $B_1 \rightleftharpoons B_2$ as sort of cross-modal interaction enhancing the extraction of modality-shared dynamics. Such cross-modal synergies can benefit the multimodal fusion process $F_M = D_M(B_1, B_2)$ to attain the final discriminative representation.

When one of the modalities is missing, the framework is required to conduct incomplete multimodal learning. Assuming u_1 is missing, we can obtain bottleneck latents B_2 from modality u_2 to reconstruct the bottleneck latents B_1^{rec}, denoted as $B_2 \rightharpoonup B_1^{rec}$, which can be further decoded as the supplementary information for the multimodal fusion process $F_M = D_M(B_1^{rec}, B_2)$.

The aforementioned two-modality scenarios can be generalized to arbitrary modality pairs without constraint, thereby facilitating efficient scaling of CyIN for tasks involving multiple modalities.

D Hyper-parameter Setting

Since missing modalities lead to highly performance fluctuations under different missing scenarios, to guarantee fair and consistent comparison among various methods, we conduct evaluation of each experiment ten times using fixed 10 random seeds. We utilize AdamW [72] as the training optimizer.

The split of datasets and hyper-parameters of CyIN are reported in Table 6.

Table 6: Splits of datasets and the corrsponding hyper-parameters settings.

Multimodal Task	Regression		Classification	
Dataset	MOSI	MOSEI	IEMOCAP	MELD
Train	1,284	16,326	5,228	9,765
Valid	229	1,871	519	1,102
Test	686	4,659	1,622	2,524
Training Epochs	50	30	50	50
Batch Size	128	128	256	128
Learning Rate of Language Model	4e-5	1e-5	5e-6	3e-5
Learning Rate of Other Parameters	1e-3	5e-4	1e-4	5e-4
Weight Decay	1e-2	1e-5	1e-3	1e-4
Dimension C_U for Unimodal Representation	256	64	64	32
Dimension C_{ib} in IB Encoder or Decoder	256	64	256	64
Dimension C_B of Bottleneck Latent B	128	256	128	32
# Layers of $RA(\cdot)$ in CRA	8	16	8	4
Dimension of $RA(\cdot)$ in CRA	$[64, 32, 16]$	$[128, 64, 32]$	$[128, 64, 32]$	$[128, 64, 32]$
# Layers of Cross-modal Attention	2	8	2	4
# Heads of Attention H	8	2	4	8
β	16	32	4	4
γ	10	5	10	5
1^{st}:2^{nd} Training Stage	1:9	3:7	3:7	1:4

E Details about Evaluation Protocols

The evaluation under the circumstance of (1) complete multimodal input is consistent with other state-of-the-art baselines, referring to $u \in \{l, a, v\}$, where l, a, and v denote language, acoustic, and visual modalities, respectively.

Following [8, 17, 22, 24], to evaluate model robustness under missing modality scenarios, we adopt two missing data protocols: (2) fixed and (3) random missing protocols.

(2) In the fixed missing protocol, a consistent subset of modalities is removed across all iterations. Specifically, we discard either one modality (e.g., $u \in \{l, a\}/\{l, v\}/\{a, v\}$) or two modalities (e.g., $u \in \{l\}/\{a\}/\{v\}$).

(3) In contrast, the random missing protocol simulate the real-world scenarios by randomly selecting missing modality combinations for each sample in every batch, where either one or two modalities may be absent. Following [22, 24], the random missing rate (MR) is defined as:

$$MR = 1 - \frac{\sum_{i=1}^{N} |u_i|}{N \times U} \tag{52}$$

where $|u_i|$ denotes the number of available modalities for the i-th sample, U is the total number of modalities, and N is the total number of samples. Given a realistic application scenario, each sample is guaranteed to have at least one available modality, ensuring $|u_i| \in \{1, ..., U\}$, and consequently, $MR \leq \frac{U-1}{U}$. In our multimodal regression and classification tasks with $U = 3$ modalities, we consider missing rates $MR \in \{0.0, 0.1, \ldots, 0.7\}$, where $MR = 0.7$ approximates the maximum tolerable missingness. Notably, $MR = 0.0$ represents the complete multimodal scenario where all modalities ($u \in l, a, v$) are present.

F Baselines

Baselines are reproduced by the open-source codes. Following [27], we conduct fifty-times of random grid search for the best hyper-parameters of each model. The descriptions of baselines are presented as follows.

CCA [11] projects paired modalities into a shared low-dimensional space by maximizing their canonical correlations, providing a classical linear imputation for incomplete data.

DCCA [12] replaces the linear projections with deep neural encoders, enabling nonlinear correspondences between modalities.

DCCAE [13] augments DCCA with autoencoder-based reconstruction objectives [15], jointly preserving each modality's internal structure while learning maximally canonical correlations.

CRA [34] employs a cascade of residual autoencoders that iteratively refine the reconstruction of representations with complete inputs from the ones with partial inputs.

MCTN [16] leverages encoder–decoder recurrent neural networks to translate between source and target modalities in cyclic translation way.

MMIN [17] extends cycle consistency learning with a cascaded residual architecture by imagining missing information from the observed views.

CPM-Net [63] utilize partial multi-view clustering to tackle incomplete multimodal learning issue, embedding all views into a structured latent space where missing features can be inferred via data transmission and cluster-aware classification.

GCNet [22] introduces two complementary graph neural networks to model temporal and speaker relationships in conversational data, and then reconstructs missing features from the learned graph representations.

IMDer [24] adopts score-based diffusion models to learn distribution of missing modalities in iterative diffusion and denoising process, regularized by the remained modalities.

LNLN [8] employs a dominant modality correction module to ensure the quality of dominant modality representations and conduct dominant modality based multimodal learning to resist input noise.

G Evaluation Metrics

The formulas and computation for evaluation metrics on multimodal regression and classification tasks are presented as follows.

For multimodal regression tasks on MOSI and MOSEI, we adopt well-representative metrics including **Acc7, F1, MAE, and Corr**, to evaluate models' performance [73]:

https://doi.org/10.52202/085713-0594

Acc7 (Seven-class Accuracy) measures the proportion of correct predictions across the seven integer labels in $[-3, +3]$, formulated as:

$$\text{Acc7} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(\lfloor \hat{y}_n \rceil = \lfloor y_n \rceil\right) = \frac{1}{7}\sum_{k=-3}^{3} \frac{\left|\{\, n : y_n = k \wedge \hat{y}_n = k \}\right|}{\left|\{\, n : y_n = k \}\right|} \tag{53}$$

where N is the total number of samples, $\hat{y}_n$ is the predicted label, y_n is the ground-truth label, $\lfloor \cdot \rceil$ denotes the round to nearest integer and $\mathbb{I}(\cdot)$ is the indicator function. With $k \in \{-3, -2, -1, 0, 1, 2, 3\}$, $\{n : y_n = k\}$ denotes the set of samples with true label k, and $\{n : y_n = k \wedge \hat{y}_n = k\}$ is the subset correctly predicted as k.

F1 (Binary F1-score) assesses positive–negative discrimination by collapsing regression labels into two classes (negative/positive), formulated as:

$$F1 = \frac{1}{N}\sum_{n=1}^{N} w_c \cdot 2 \cdot \frac{(Precision_c * Recall_c)}{(Precision_c + Recall_c)} \tag{54}$$

$$Precision = \frac{TP}{TP + FP}\,,\, Recall = \frac{TP}{TP + FN} \tag{55}$$

where C is the number of classes, w_c is the weight of class c (typically $w_c = N_c/N$), and TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively.

MAE (Mean Absolute Error) quantifies the average absolute deviation between predicted and ground-truth scores, formulated as:

$$MAE(\hat{y}, y) = \frac{\sum_{n=1}^{N} |\hat{y}_n - y_n|}{N} \tag{56}$$

Corr (Pearson Correlation) captures the linear agreement between predictions and labels, indicating any systematic bias or skew in model outputs, formulated as:

$$\text{Corr}(\hat{y}, y) = \frac{\sum_{n=1}^{N}\left(\hat{y}_n - \bar{\hat{y}}\right)\left(y_n - \bar{y}\right)}{\sqrt{\sum_{n=1}^{N}\left(\hat{y}_n - \bar{\hat{y}}\right)^2 \sum_{n=1}^{N}\left(y_n - \bar{y}\right)^2}} \tag{57}$$

where $\bar{\hat{y}}$ and $\bar{y}$ are the means of the predicted scores and ground truth labels.

For multimodal classification tasks on IEMOCAP and MELD, we report **Acc** and **wF1** to balance the weight of scores from each class [74].

Acc (Binary Accuracy) describes the unweighted proportion of correctly predicted samples out of the total number of samples, formulated as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{58}$$

wF1 (Weighted F1-score) balances weighted precision and recall based on the true instances for each class, and the formulation of wF1 in classification is the same as Equ. 54.

H More Experiment Results

H.1 Detailed Results Under Various Missing Scenarios

The detail experiment results under various missing modality scenarios with fixed and random missing protocols are presented in Table 7 and Table 8, respectively. Note that the reported results are the average values under 10 fixed random seeds.

With fixed missing protocol including incomplete settings $u \in \{l\}/\{v\}/\{a\}/\{l, a\}/\{l, v\}/\{a, v\}$, CyIN demonstrates superior or competitive performance across nearly all settings, , significantly outperforming other models in both unimodal and bimodal missing scenarios. From Table 7, we can observe that language modality plays the most essential role in these multimodal datasets due

Table 7: Performance comparison between the proposed CyIN and baselines with fixed missing protocols including incomplete settings $u \in \{l\}/\{v\}/\{a\}/\{l,a\}/\{l,v\}/\{a,v\}$.

Fix u	Models	MOSI				MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
$\{l\}$	CCA	26.4	74.7	1.098	0.567	44.6	80.2	0.677	0.621	56.1	53.9	49.7	43.1
	DCCA	23.3	73.5	1.196	0.520	41.5	68.0	0.810	0.377	51.7	50.7	48.9	35.3
	DCCAE	25.8	66.4	1.278	0.508	41.6	67.9	0.806	0.336	50.1	49.2	47.7	32.0
	CPM-Net	17.2	63.9	1.335	0.345	40.3	74.7	0.832	0.230	50.4	50.5	42.0	34.8
	CRA	36.3	82.8	0.900	0.749	49.3	84.9	0.576	0.741	53.8	51.3	55.3	51.6
	MCTN	41.5	83.7	0.777	0.765	49.6	81.8	0.594	0.715	58.4	57.8	56.1	52.3
	MMIN	42.3	85.0	0.743	0.791	**52.8**	84.6	0.543	**0.761**	57.0	57.5	59.2	54.9
	GCNet	42.6	85.0	0.740	0.795	51.0	84.8	0.562	0.750	58.7	58.3	59.9	57.9
	IMDer	43.0	85.1	**0.717**	0.794	51.8	85.0	0.552	0.756	60.4	60.1	59.9	58.2
	LNLN	43.9	83.5	0.761	0.760	52.5	84.3	0.547	0.765	62.8	62.4	58.3	57.1
	CyIN	**44.0**	**85.2**	0.742	**0.795**	52.6	**85.4**	**0.541**	0.751	**63.4**	**63.0**	**60.0**	**58.6**
$\{a\}$	CCA	17.5	46.8	1.422	0.112	41.5	63.3	0.808	0.272	39.1	34.8	47.7	35.7
	DCCA	14.1	43.3	1.821	0.052	41.3	59.4	0.810	0.356	30.0	24.8	46.1	33.4
	DCCAE	19.7	56.5	1.406	0.123	37.4	50.8	0.841	0.301	35.6	34.0	47.3	34.4
	CPM-Net	17.3	56.6	1.371	0.313	39.4	**74.2**	0.916	0.139	38.1	37.2	26.3	29.7
	CRA	19.7	52.0	1.438	0.102	41.8	71.1	**0.781**	**0.420**	40.7	40.1	46.6	37.5
	MCTN	15.5	42.2	1.431	0.018	41.4	62.9	0.842	0.096	22.3	11.3	48.2	31.6
	MMIN	20.1	52.6	1.429	0.080	39.3	70.9	0.790	0.404	42.6	40.0	48.2	31.5
	GCNet	19.8	57.5	1.383	0.387	40.8	56.6	0.817	0.389	44.6	42.2	40.9	34.0
	IMDer	19.6	58.3	1.363	0.392	42.1	69.9	0.795	0.372	47.0	46.8	45.6	**42.5**
	LNLN	15.6	47.7	1.442	0.075	41.0	68.9	0.795	0.338	39.2	33.4	39.3	33.7
	CyIN	**20.7**	**60.2**	**1.355**	**0.412**	**42.3**	70.8	0.782	0.363	**50.3**	**48.9**	**48.5**	37.7
$\{v\}$	CCA	15.5	26.1	1.445	0.125	41.4	48.6	0.837	0.239	23.6	9.1	**48.4**	31.6
	DCCA	17.4	51.7	1.437	0.081	41.3	72.9	0.800	0.387	26.6	22.4	45.8	32.9
	DCCAE	**20.9**	**61.8**	1.659	0.407	41.4	73.1	0.798	0.389	25.2	21.8	45.5	32.9
	CPM-Net	16.5	56.8	1.368	0.323	38.8	**74.2**	0.862	0.205	28.0	20.9	31.8	32.1
	CRA	16.6	52.1	1.396	0.013	39.0	72.0	0.781	0.426	29.9	26.3	44.2	36.8
	MCTN	15.5	42.2	1.431	0.018	41.4	62.9	0.842	0.096	23.5	8.9	**48.4**	31.6
	MMIN	19.7	50.5	1.471	0.075	39.5	69.8	0.789	0.395	35.3	34.1	**48.4**	31.6
	GCNet	16.0	47.9	1.390	0.066	41.4	58.5	0.833	0.268	44.8	43.4	39.1	33.5
	IMDer	17.0	52.0	1.426	0.032	41.4	62.7	0.826	0.386	46.8	45.8	46.3	38.6
	LNLN	15.5	42.2	1.447	0.125	41.8	70.9	0.771	0.405	47.9	47.3	48.1	31.3
	CyIN	19.8	57.7	**1.362**	**0.408**	**42.1**	71.7	**0.767**	**0.430**	**49.0**	**47.7**	48.2	**39.3**
$\{l,a\}$	CCA	27.8	74.9	1.106	0.542	44.5	81.8	0.663	0.639	60.1	58.5	51.1	46.5
	DCCA	21.9	59.7	1.377	0.422	41.3	72.8	0.795	0.401	53.9	52.5	47.9	35.9
	DCCAE	22.2	69.4	1.425	0.414	38.7	68.2	0.802	0.394	54.0	52.7	47.3	34.5
	CPM-Net	17.2	64.0	1.335	0.345	39.9	74.5	0.980	0.145	43.9	44.0	35.1	35.8
	CRA	36.4	83.4	0.885	0.750	50.5	84.6	0.565	0.756	60.7	60.0	56.4	53.3
	MCTN	41.5	83.7	0.777	0.765	47.9	84.3	0.592	0.721	58.4	57.8	56.3	52.4
	MMIN	43.0	85.0	0.744	0.782	52.8	84.6	0.538	0.768	62.2	62.5	59.4	55.4
	GCNet	41.5	84.8	0.742	0.796	49.2	84.4	0.581	0.753	62.9	63.0	57.4	55.8
	IMDer	44.3	84.9	0.731	**0.797**	52.8	85.1	0.560	0.775	63.7	63.9	60.0	58.3
	LNLN	44.0	83.3	0.761	0.766	52.7	84.6	0.546	0.768	57.9	57.3	58.1	57.1
	CyIN	**45.0**	**85.1**	**0.728**	0.792	**53.1**	**85.5**	**0.541**	**0.786**	**65.2**	**64.7**	**60.5**	**58.3**
$\{l,v\}$	CCA	26.1	75.6	1.093	0.570	44.7	80.7	0.674	0.625	56.6	54.5	49.7	43.1
	DCCA	24.1	71.2	1.234	0.398	41.2	73.3	0.791	0.420	50.0	49.9	47.2	35.3
	DCCAE	21.7	67.3	1.327	0.340	39.7	73.1	0.791	0.394	52.9	50.4	46.8	34.1
	CPM-Net	17.5	62.9	1.339	0.333	38.1	73.4	1.582	0.072	42.8	43.5	41.7	35.2
	CRA	36.3	81.5	0.917	0.741	51.3	85.1	0.555	0.761	58.3	55.7	56.4	53.3
	MCTN	41.5	83.7	0.777	0.765	47.9	84.3	0.592	0.721	58.3	58.2	56.3	52.4
	MMIN	42.3	85.0	0.742	0.791	52.8	85.1	0.534	0.768	59.3	59.4	60.1	55.5
	GCNet	39.0	84.4	0.771	0.796	49.1	84.8	0.582	0.750	56.6	56.2	60.2	58.3
	IMDer	45.8	85.1	**0.701**	**0.799**	52.9	85.2	0.556	0.761	60.2	60.0	59.8	58.0
	LNLN	43.7	83.5	0.760	0.766	52.7	85.0	0.544	0.770	62.1	61.5	58.2	57.1
	CyIN	**46.2**	**85.3**	0.727	0.794	**53.5**	**85.2**	**0.531**	**0.778**	**63.2**	**62.9**	**60.9**	**58.8**
$\{a,v\}$	CCA	17.4	48.1	1.420	0.117	41.6	63.6	0.806	0.278	39.6	35.1	48.0	36.1
	DCCA	17.5	53.6	1.445	0.093	41.2	73.6	0.790	0.408	35.4	33.3	46.1	33.7
	DCCAE	19.4	55.2	1.569	0.046	37.9	72.7	0.796	0.404	35.5	33.2	46.5	35.7
	CPM-Net	16.8	56.2	1.367	0.330	35.2	72.0	1.395	0.041	44.5	42.4	25.6	28.9
	CRA	18.5	55.0	1.410	0.066	40.4	**73.2**	0.773	0.453	44.1	42.0	45.3	40.1
	MCTN	15.5	42.2	1.431	0.018	41.4	62.9	0.842	0.102	22.7	11.5	48.8	31.6
	MMIN	20.3	52.3	1.427	0.081	39.8	71.0	0.774	0.431	48.2	47.4	48.3	31.5
	GCNet	17.9	57.2	1.363	0.385	41.2	72.2	0.805	0.393	49.0	48.1	44.7	40.9
	IMDer	18.4	58.1	1.319	0.386	41.5	71.6	0.788	0.446	50.0	49.9	46.7	43.0
	LNLN	15.6	48.7	1.441	0.076	40.8	71.7	0.775	0.441	51.5	50.4	35.5	31.8
	CyIN	**21.0**	**59.4**	**1.305**	**0.391**	**41.7**	72.9	**0.773**	**0.458**	**53.0**	**52.1**	**48.4**	**43.5**

https://doi.org/10.52202/085713-0594

Table 8: Performance comparison between the proposed CyIN and baselines with random missing protocols including missing rates $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$.

MR	Models	MOSI				MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
	CCA	26.9	72.7	1.133	0.512	45.6	81.2	0.671	0.635	58.9	58.7	49.3	41.9
	DCCA	24.1	70.1	1.219	0.418	40.1	73.1	0.789	0.415	52.4	51.4	47.5	35.3
	DCCAE	20.2	68.4	1.620	0.336	38.7	72.5	0.787	0.418	51.0	50.0	46.4	35.5
	CPM-Net	16.8	63.9	1.336	0.339	34.5	74.3	1.780	0.060	55.1	55.1	42.5	33.2
	CRA	33.6	79.3	0.972	0.699	50.1	84.2	0.576	0.736	60.0	59.0	56.0	53.0
0.1	MCTN	38.9	79.9	0.846	0.720	47.1	81.4	0.620	0.680	42.0	43.3	53.6	46.8
	MMIN	41.0	82.1	0.808	0.741	51.5	83.5	0.562	0.738	53.9	53.9	55.8	49.0
	GCNet	41.2	84.5	0.806	0.751	50.3	83.1	0.653	0.696	60.7	60.8	56.3	54.4
	IMDer	42.3	84.1	0.796	0.753	51.4	83.9	0.600	0.721	60.9	61.3	58.4	**56.7**
	LNLN	40.8	83.3	0.790	0.756	51.4	83.7	0.566	0.745	61.2	62.0	56.2	55.0
	CyIN	**42.5**	**84.6**	**0.783**	**0.759**	**51.9**	**84.5**	**0.562**	**0.748**	**64.0**	**63.9**	**58.6**	56.6
	CCA	25.6	70.4	1.161	0.489	45.0	79.3	0.688	0.605	56.0	55.8	49.1	41.3
	DCCA	22.8	68.7	1.242	0.394	40.4	72.3	0.792	0.400	49.4	48.5	47.2	34.6
	DCCAE	20.9	67.6	1.584	0.325	38.9	70.8	0.795	0.389	48.6	47.7	46.7	35.2
	CPM-Net	16.9	63.9	1.336	0.338	34.3	74.2	1.457	0.064	53.5	54.2	41.9	33.1
	CRA	31.9	76.2	1.025	0.660	48.9	82.7	0.598	0.691	56.6	55.7	54.2	50.9
0.2	MCTN	36.0	75.5	0.918	0.676	46.8	78.7	0.642	0.645	39.6	41.7	53.3	46.0
	MMIN	38.3	78.6	0.872	0.701	50.2	82.0	0.597	0.693	52.6	52.7	55.2	48.1
	GCNet	39.0	81.0	0.832	0.725	48.1	81.9	0.622	0.682	59.2	59.3	53.2	51.8
	IMDer	39.3	82.8	0.878	0.729	49.7	82.2	0.610	0.694	59.9	60.3	55.9	53.9
	LNLN	39.9	79.6	0.837	0.717	50.1	82.3	0.591	0.713	60.5	61.1	54.2	52.9
	CyIN	**40.1**	**83.0**	**0.824**	**0.732**	**51.3**	**82.9**	**0.590**	**0.721**	**61.9**	**61.8**	**57.2**	**54.8**
	CCA	24.2	68.6	1.193	0.454	44.6	77.1	0.707	0.569	53.1	52.9	48.9	40.6
	DCCA	23.1	67.2	1.251	0.375	40.8	71.4	0.794	0.390	46.2	45.2	47.3	34.2
	DCCAE	20.4	66.2	1.545	0.310	39.1	69.2	0.801	0.371	46.3	45.3	46.9	34.8
	CPM-Net	16.9	64.0	1.336	0.336	34.1	74.0	1.180	0.111	53.0	53.8	40.4	34.6
	CRA	31.3	72.3	1.077	0.624	48.0	81.3	0.632	0.661	53.5	52.7	52.9	49.1
0.3	MCTN	33.5	72.3	0.977	0.634	46.1	76.1	0.668	0.599	37.7	40.0	53.0	45.4
	MMIN	35.7	74.6	0.947	0.648	48.7	80.6	0.631	0.659	51.2	51.3	54.8	47.3
	GCNet	36.3	75.4	0.968	0.652	48.5	80.7	0.673	0.665	57.5	57.6	50.9	49.7
	IMDer	37.7	76.6	0.923	0.677	48.6	80.9	0.643	0.660	58.6	58.9	54.2	51.8
	LNLN	37.5	76.5	0.890	0.683	**49.3**	80.8	0.613	0.683	59.3	60.2	52.6	50.9
	CyIN	**38.1**	**78.4**	**0.886**	**0.694**	49.2	**81.5**	**0.606**	**0.685**	**60.3**	**60.4**	**55.8**	**52.7**
	CCA	22.8	66.7	1.222	0.423	44.3	75.3	0.723	0.535	50.1	49.7	48.7	39.9
	DCCA	21.2	64.6	1.283	0.304	41.1	69.9	0.796	0.378	42.4	41.2	47.1	33.7
	DCCAE	19.2	62.9	1.574	0.273	39.3	66.6	0.808	0.347	43.4	42.2	46.9	34.1
	CPM-Net	16.9	63.8	1.336	0.335	34.2	73.9	1.439	0.061	52.8	53.6	41.9	33.2
	CRA	28.3	68.2	1.160	0.552	46.7	79.9	0.646	0.626	50.0	49.5	52.1	47.9
0.4	MCTN	30.8	67.9	1.050	0.583	45.3	73.2	0.695	0.551	35.4	37.8	52.6	44.5
	MMIN	33.3	71.2	1.021	0.523	47.4	79.4	0.655	0.629	49.9	50.0	54.3	46.3
	GCNet	33.9	74.6	0.986	0.628	47.1	79.2	0.672	0.623	56.1	56.3	47.7	47.0
	IMDer	34.1	75.9	0.944	0.655	47.2	78.7	0.664	0.618	56.8	57.2	52.0	49.2
	LNLN	34.5	73.3	0.952	0.644	47.2	79.2	0.640	0.646	58.7	59.3	51.5	49.3
	CyIN	**34.7**	**76.5**	**0.935**	**0.666**	**47.5**	**79.3**	**0.635**	**0.649**	**59.3**	**59.4**	**54.4**	**50.3**
	CCA	21.7	63.8	1.255	0.384	43.6	73.0	0.744	0.489	46.9	46.2	48.6	39.1
	DCCA	21.6	63.9	1.289	0.304	41.3	67.4	0.800	0.362	39.5	37.8	47.2	33.3
	DCCAE	19.5	60.9	1.558	0.246	39.3	63.4	0.818	0.317	40.1	38.7	47.4	34.0
	CPM-Net	16.9	63.9	1.335	0.334	34.2	74.0	1.626	0.056	52.7	53.3	41.1	33.1
	CRA	26.7	64.1	1.209	0.514	45.4	78.6	0.671	0.603	46.4	46.1	50.5	45.7
0.5	MCTN	25.7	59.7	1.161	0.503	44.6	70.7	0.717	0.510	33.3	35.7	52.0	43.3
	MMIN	30.7	68.0	1.081	0.550	46.1	78.0	0.669	0.605	48.3	48.2	53.4	45.0
	GCNet	30.7	71.0	1.038	0.598	45.0	78.0	0.699	**0.618**	54.3	54.4	45.6	45.5
	IMDer	31.8	72.2	0.995	**0.605**	46.2	76.4	0.687	0.570	54.8	55.3	50.0	46.8
	LNLN	31.9	69.3	1.024	0.593	46.7	77.8	0.668	0.606	53.6	54.1	49.9	46.7
	CyIN	**32.5**	**72.6**	**0.990**	0.599	**47.1**	**78.7**	**0.658**	0.611	**55.3**	**55.5**	**53.6**	**48.2**
	CCA	20.6	62.0	1.279	0.351	43.0	69.7	0.764	0.437	43.3	42.2	48.4	38.2
	DCCA	20.6	60.3	1.324	0.252	41.3	64.4	0.803	0.355	37.3	36.3	47.3	32.8
	DCCAE	19.0	57.4	1.545	0.231	39.1	59.2	0.827	0.293	37.5	35.2	47.7	33.4
	CPM-Net	17.0	64.0	1.341	0.299	34.3	73.9	1.304	0.093	52.6	**53.2**	41.5	33.7
	CRA	25.1	61.4	1.264	0.454	43.9	77.2	0.696	0.577	42.7	42.6	49.6	44.1
0.6	MCTN	24.1	58.8	1.190	0.464	44.1	67.4	0.745	0.452	30.6	32.9	51.4	41.8
	MMIN	28.3	62.9	1.158	0.486	44.8	76.4	0.688	0.577	46.3	46.2	52.1	43.3
	GCNet	29.3	65.9	1.127	0.513	43.8	77.2	0.705	**0.584**	51.8	51.8	41.2	43.9
	IMDer	**30.2**	67.8	1.090	0.564	44.6	75.3	0.704	0.544	52.2	51.7	47.6	44.8
	LNLN	28.9	64.9	1.150	0.522	45.6	76.1	0.689	0.565	48.1	47.7	49.3	45.0
	CyIN	29.2	**68.6**	**1.069**	**0.571**	**45.8**	**77.5**	**0.683**	0.581	**52.8**	52.4	**52.3**	**46.5**
	CCA	19.9	59.7	1.298	0.326	42.7	68.3	0.775	0.409	41.2	39.8	48.5	37.7
	DCCA	19.6	58.6	1.343	0.214	41.3	61.9	0.806	0.360	36.2	35.0	47.2	32.5
	DCCAE	18.6	55.5	1.539	0.209	38.8	55.9	0.833	0.277	35.7	32.7	47.8	33.2
	CPM-Net	17.1	63.8	1.361	0.291	34.2	73.9	1.692	0.104	**52.2**	**52.5**	41.6	34.1
	CRA	23.6	57.3	1.311	0.401	43.0	**76.5**	0.713	0.548	40.1	40.0	48.7	42.7
0.7	MCTN	23.5	56.1	1.223	0.432	43.7	64.9	0.760	0.413	28.8	30.8	51.2	41.2
	MMIN	25.6	59.0	1.212	0.440	43.8	75.3	0.707	0.545	45.9	45.8	51.7	41.9
	GCNet	26.3	63.9	1.166	0.494	43.3	74.5	0.733	0.542	47.3	47.2	39.9	41.8
	IMDer	27.0	65.1	1.126	0.526	43.1	75.1	0.714	0.470	47.7	48.2	46.3	43.0
	LNLN	26.1	62.7	1.200	0.472	45.0	74.8	0.705	0.536	46.9	45.9	48.5	43.1
	CyIN	**28.0**	**65.9**	**1.117**	**0.530**	**45.1**	74.9	**0.700**	**0.553**	48.6	49.0	**51.8**	**44.4**

https://doi.org/10.52202/085713-0594

to the rich semantics and high recognition-related information in utterance [16, 43, 44]. Despite better performance in this dominant modality, CyIN pays more attention on the inferior modalities including audio and vision modalities. This result illustrates the generalization ability of the proposed framework, since CyIN has not spectacularly designed extra delicate modules to concentrate training on dominant modalities as LNLN [8]. Besides, with one modality missing and two modalities as input, CyIN sufficiently integrate the information from bimodal interaction and reach higher performance, validating the flexibility of informative bottleneck latents in various combination of modalities.

With random missing protocol including missing rates $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$, CyIN suffers from less performance degradation as the missing rate increases. The random missing case is mostly simulating diverse missing situation in the real-world, when the presence of modalities is not fixed or known. As shown in Table 8, compared with baselines, the robustness of CyIN under various input data sparsity highlights the strength of informative bottlenecks extracted by token- and label-level IB, ensuring stable performance even when the presence of multimodal input are severely missing at $MR = 0.7$. Moreover, CyIN reaches state-of-the-art performance on 4 datasets regardless of multimodal regression or classification tasks. While most baselines reach suboptimal performance on either one of the tasks due to the limitation of generalization.

The experiment results in various missing scenarios clearly demonstrates that CyIN is highly effective in both fixed or random missing scenarios. The general framework enables the informative space to maintain powerful performance despite the presence of input modalities, making it a robust and practical solution for real-world multimodal applications.

H.2 Comparison of Performance Stability

Table 9: Performance stability comparison between the proposed CyIN and baselines with the average results on MOSI and IEMOCAP dataset with fixed missing protocols including modality settings $u \in \{l\}/\{a, v\}$ and random missing protocols including missing rates $MR = 0.7$.

Setting	Models	MOSI				IEMOCAP	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑
Fixed $u \in \{l\}$	MCTN	41.5 ± 0.71	83.7 ± 0.37	0.777 ± 0.005	0.765 ± 0.003	58.4 ± 0.40	57.8 ± 0.41
	MMIN	42.3 ± 0.82	85.0 ± 0.45	0.743 ± 0.009	0.791 ± 0.005	57.0 ± 0.52	57.5 ± 0.49
	GCNet	42.6 ± 0.53	85.0 ± 0.31	0.740 ± 0.003	0.795 ± 0.002	58.7 ± 0.33	58.3 ± 0.42
	IMDer	43.0 ± 0.59	85.1 ± 0.32	**0.717** ± 0.005	0.794 ± 0.002	60.4 ± 0.47	60.1 ± 0.44
	LNLN	43.9 ± 0.33	83.5 ± 0.24	0.761 ± 0.008	0.760 ± 0.006	62.8 ± 0.51	62.4 ± 0.48
	CyIN	**44.0** ± 0.25	**85.2** ± 0.29	0.742 ± 0.004	**0.795** ± 0.002	**63.4** ± 0.10	**63.0** ± 0.10
Fixed $u \in \{a, v\}$	MCTN	15.5 ± 0.71	42.2 ± 0.37	1.431 ± 0.005	0.018 ± 0.004	22.7 ± 0.40	11.5 ± 0.24
	MMIN	20.3 ± 0.95	52.3 ± 1.32	1.427 ± 0.009	0.081 ± 0.005	48.2 ± 0.51	47.4 ± 0.35
	GCNet	17.9 ± 0.83	57.2 ± 1.25	1.363 ± 0.008	0.385 ± 0.002	49.0 ± 0.48	48.1 ± 0.31
	IMDer	18.4 ± 0.85	58.1 ± 1.58	1.319 ± 0.016	0.386 ± 0.026	50.0 ± 0.30	49.9 ± 0.38
	LNLN	15.6 ± 0.75	48.7 ± 0.45	1.441 ± 0.008	0.076 ± 0.006	51.5 ± 0.56	50.4 ± 0.48
	CyIN	**21.0** ± 0.35	**59.4** ± 0.27	**1.305** ± 0.010	**0.391** ± 0.019	**53.0** ± 0.40	**52.1** ± 0.46
Random Missing $MR = 0.7$	MCTN	23.5 ± 1.15	56.1 ± 1.23	1.223 ± 0.018	0.432 ± 0.026	28.8 ± 0.64	30.8 ± 0.69
	MMIN	25.6 ± 1.04	59.0 ± 1.60	1.212 ± 0.017	0.440 ± 0.019	45.9 ± 0.61	45.8 ± 0.65
	GCNet	26.3 ± 2.25	63.9 ± 2.86	1.166 ± 0.038	0.494 ± 0.032	47.3 ± 0.79	47.2 ± 0.81
	IMDer	27.0 ± 1.30	65.1 ± 1.50	1.126 ± 0.022	0.526 ± 0.023	47.7 ± 1.20	48.2 ± 1.12
	LNLN	26.1 ± 1.07	62.7 ± 1.63	1.200 ± 0.019	0.472 ± 0.021	46.9 ± 1.66	45.9 ± 1.79
	CyIN	**28.0** ± 0.90	**65.9** ± 1.71	**1.117** ± 0.018	**0.530** ± 0.021	**48.6** ± 0.71	**49.0** ± 0.70

We evaluate the performance stability of the models on MOSI dataset under three representative circumstances, including fixed missing protocols $u \in \{l\}/\{a, v\}$ and random missing protocols $MR = 0.7$. Based on the overall average performance reported in Table 7 and Table 8, we further computed the standard deviations of the corresponding results on each missing scenarios with 10 runs, to demonstrate the performance variation degree in real-world circumstances. The final stability measure are reported in Table 9.

Under fixed missing modality protocols such as $u \in \{l\}/\{a, v\}$, the standard deviations of all methods show relatively small, reflecting that when the missing modalities are predetermined, models can learn to compensate in a more determine way. In contrast, random missing protocol introduce much higher variance, especially in severe missing scenario with $MR = 0.7$. We summarize this into two reasons: First reason is imbalance contribution of modalities (with dominant modality like

https://doi.org/10.52202/085713-0594

language containing more semantics compared with the inferior ones like audio or vision in specific tasks), and the second one is random missing setting is more closed to simulate the unpredictable inference circumstances in real-world, exposing each model's real robustness.

Across both fixed and random missing scenarios, CyIN not only achieves the superior or competitive performance on most metrics but also maintains relatively low fluctuations, highlighting the remarkable balance between performance and stability for the constructed informative space.

H.3 Generalization Performance on Different Language Models

To further validate the generalization performance, We train and evaluate CyIN with different sizes of Pretrained language models including BERT [40], RoBERTa [75], and DeBERTa-V3 [76] as shown in 10. The experiment results shows a clear benefit from scaling up the PLM backbone, as larger models like DeBERTa-V3 yield better textual representations and thus higher multimodal performance when all modalities are available. The similar trend occurs when evaluating with the most severe missing circumstance $MR = 0.7$, which indicates the effective generalization ability of CyIN.

Table 10: Comparison of the proposed CyIN using different sizes of language models on MOSI dataset, under both complete and randomly incomplete multimodal learning settings.

Setting	Model Variants (PLM)	MOSI			
		Acc7↑	F1↑	MAE↓	Corr↑
Complete $u \in \{l, a, v\}$	BERT	48.0	86.3	0.712	0.801
	RoBERTa	49.5	88.2	0.692	0.823
	DeBERTa-V3	**50.3**	**90.1**	**0.671**	**0.841**
Random Missing $MR = 0.7$	BERT	28.0	65.9	1.117	0.530
	RoBERTa	29.3	67.2	0.992	0.556
	DeBERTa-V3	**32.2**	**69.5**	**0.965**	**0.578**

H.4 Various Scales of Other Multimodal Tasks

We have extended experiments of CyIN on 6 non-affective datasets with 2–4 modalities and random missing rates in Table 11 - 14, including the following tasks: **Multimodal Recommendation (3 modalities: visual, audio, textual)** [55] (Amazon Baby, Tkitok, Allrecipes datasets), **Multimodal Face Anti-spoofing (3 modalities: RGB, Depth, IR)** [77] (CASIA-SURF dataset), **Multimodal Dense Prediction (2 modalities: RGB, Depth)** [77] (NYUv2 dataset), and **Multimodal Medical Segmentation (4 modalities: FLAIR, T1, T1c, T2)** [56] (BraTS2023 dataset).

Table 11: Experiment comparison for **Multimodal Recommendation** task of accuracy and fairness performance (%) on three datasets, including Amazon Baby, Tiktok and Allrecipes dataset. The modality setting is set in random missing protocol with 0.4 missing rate as the SOTA [55] paper.

Dataset	Model	Recall ↑		Precision ↑		NDCG ↑		F ↑		F_{fuse} ↑	
		K=10	K=20	K=10	K=20	K=10	K=20	K=10	K=20	K=10	K=20
Amazon Baby	SOTA [55]	5.26	8.54	0.56	0.45	2.76	3.60	87.24	90.12	1.11	0.90
	CyIN(ours)	**5.43**	**8.56**	**0.57**	0.45	**2.89**	**3.68**	**90.10**	**92.70**	**1.14**	0.90
Tiktok	SOTA [55]	4.81	7.39	0.48	0.37	2.95	3.60	88.15	92.27	0.96	0.74
	CyIN(ours)	**4.91**	**7.57**	**0.49**	**0.38**	**3.15**	**3.81**	**89.02**	**92.31**	**0.98**	**0.75**
Allrecipes	SOTA [55]	2.49	3.36	0.24	0.16	1.33	1.55	96.82	89.52	0.49	0.33
	CyIN(ours)	**2.61**	**3.43**	**0.26**	**0.17**	**1.45**	**1.66**	**99.57**	**91.94**	**0.52**	**0.34**

The experimental results illustrate that the proposed CyIN reaches comparable performance with the state-of-the-art (SOTA) methods, which further demonstrates the effectiveness of CyIN in generalizing to diverse numbers of modalities or multimodal fusion tasks. Moreover, the scale of datasets varies from 1,449 samples (NYU v2) to 65,671 samples (Tiktok), 87k samples (CASIA-SURF), 139,110 samples (Amazon Baby), which partially show the scaling performance of the proposed CyIN. Due to the computation resource limitation, we leave exploration on larger scale datasets in the future work.

Table 12: Experiment comparison for **Multimodal Face Anti-spoofing** task with Average Classification Error Rate (ACER) (↓) metric on CASIA-SURF dataset. The modalities setting is set in fix missing protocol as the SOTA [77] paper.

Modality Setting RGB	Depth	IR	SOTA [77]	**CyIN(ours)**
✓			7.33	**4.48**
	✓		**2.13**	2.83
		✓	10.41	**7.75**
✓	✓		**1.02**	2.26
✓		✓	3.88	**3.06**
	✓	✓	1.38	**1.20**
✓	✓	✓	0.69	**0.66**
Average			3.84	**3.18**

Table 13: Experiment comparison for **Multimodal Dense Prediction** task with mIoU (↑) metric on NYUv2 dataset. The modalities setting is set in fix missing protocol as the SOTA [77] paper.

Modality Setting RGB	Depth	SOTA [77]	**CyIN(ours)**
✓		**44.06**	43.93
	✓	41.82	**43.84**
✓	✓	49.89	**50.46**
Average		45.26	**46.07**

Table 14: Experiment comparison for **Multimodal Medical Segmentation** task with Dice similarity coefficient (DSC) % (↑) metric on BraTS2023 dataset. The average results are reported conducted in fixed modality setting as the original SOTA [56] paper.

Class	SOTA[3]	**CyIN(ours)**
Enhancing Tumor	74.6	**75.1**
Tumor Core	85.0	**85.8**
Whole Tumor	90.6	**91.2**

I Social Impacts

Devoted in bridging complete and incomplete multimodal learning, the proposed CyIN shows strong potential in real-world field, such as healthcare, education, social media analysis, marketing, advertising, and human-computer interaction. By integrating data from multiple sources such as text, audio, image and video, CyIN enables multimodal system for a more detailed understanding of human emotions and intentions.

However, the use of rich and multi-source data introduces risks related to privacy violations, unauthorized surveillance, and potential misuse for manipulation. In scenarios with missing modalities, efforts to reconstruct data may introduce biases, fabricate faking information, particularly affecting underrepresented groups and raising fairness concerns. These concerns highlight the need of careful and responsible deployment for the proposed method in real-world applications.

https://doi.org/10.52202/085713-0594

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction clearly reflect the paper's contributions and scope. They accurately state the design and capabilities of the proposed CyIN framework in jointly handling complete and incomplete multimodal learning.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations have been discussed in Section Limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-0594

Answer: [Yes]

Justification: The paper have provided a complete and correct derivation proof for the mentioned equation in Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed information on the hyperparameter settings necessary for reproducibility in Appendix D. Moreover, as stated in the abstract, the source code is publicly released.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/085713-0594

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in this paper are publicly accessible and available upon request. Moreover, as stated in the abstract, the source code is now publicly released, ensuring transparency and reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details including dataset splits, hyper-parameters, and type of optimizer and so on are specified in Appendix D. Besides, following previous methods, the best hyper-parameters are chosen by fifty-times of random grid search for each model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: As shown in Appendix H.2, we report the performance variation to demonstrate the superiority of the proposed method across various missing scenarios.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the required computer resources in Implementation Details of Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This research conforms all the requirements of moral and ethical norms in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed both positive and negative social impacts in Appendix I.

https://doi.org/10.52202/085713-0594

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper utilizes public used datasets and pre-trained models, which poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The used data have been correctly cited in the paper and the license and terms of use are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: As shown in Abstract, the source code is publicly released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/085713-0594

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.