arXiv:2602.04920v1 [cs.LG] 4 Feb 2026

CyIN：用于衔接完整与不完整多模态学习的循环信息潜在空间

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] 中山大学电子与信息技术学院，中国[2] 南洋理工大学电气与电子工程学院，新加坡 [3] 华南师范大学计算机科学学院，中国[4] 德赛西威汽车有限公司，中国[5] 琶洲实验室，中国

{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

摘要

模仿人脑整合多种模态能力的多模态机器学习近年来发展迅速。大多数以往的多模态模型都在完美配对的多模态输入上训练，以达到最佳性能。然而，在真实世界部署中，模态的存在情况高度多变且难以预测，导致预训练模型性能显著下降，无法在模态动态缺失的情况下保持稳健。本文提出一种新颖的循环信息学习框架（CyIN），以弥合完整与不完整多模态学习之间的差距。具体而言，我们首先在不同模态之间循环采用令牌级和标签级信息瓶颈（IB），构建信息潜在空间。通过变分近似捕获任务相关特征后，信息瓶颈潜变量得到提炼，从而实现更高效的跨模态交互与多模态融合。此外，为补充不完整多模态输入造成的信息缺失，我们提出跨模态循环翻译，通过正向与反向传播过程，利用保留的模态重构缺失模态。借助提取和重构的信息潜变量，CyIN成功在一个统一模型中联合优化完整与不完整多模态学习。在4个多模态数据集上进行的大量实验表明，我们的方法在完整场景和多种不完整场景下均表现出优越性能。 [1]

1 引言

为获得最佳多模态性能，先前的方法隐含地假设，训练期间存在的每种模态在推理时也都可用。然而，在现实场景中，多模态数据可能会因多种因素而缺失，例如传感器故障，从而导致输入模态是否存在变得不确定 [1]。多模态模型对此类不完整多模态输入表现出明显的敏感性，因此，在下游推理中部署预训练的多模态模型时，性能会严重下降，基于 Transformer 的模型尤其如此 [2]。这一问题被概括为模态缺失问题 [3–5], 而致力于解决该问题的方法称为不完整多模态学习方法 [6–8]。

* 通讯作者。[1] 代码发布于 https://github.com/RH-Lin/CyIN。

第39届神经信息处理系统大会（NeurIPS 2025）。

当前方法主要通过在多模态学习中设计各种精巧模块来增强多模态模型的鲁棒性，以应对各种缺失情形，大致可分为对齐方法和生成方法。前者利用对比学习[4, 9, 10], 、典型相关分析 [11–13], 等技术，以及混合增强或噪声输入等数据增强方法[5, 7, 14] ，使完整和不完整多模态输入的表示对齐；后者则引入自编码器 [15–17], 变分自编码器 [18–21],基于图的网络 [22, 23], 和扩散模型 [24] 来重建缺失信息。

尽管这些方法通过不同程度地弥合信息差距，缓解了模态缺失问题，但无论是在对齐还是生成方面，它们都未能充分利用缺失信息，且容易受到任务无关噪声的干扰 [6, 24]。此外，以往的方法通常需要针对每种可能的缺失模态组合分别训练模型 [16, 22]。因此，由于现实场景中的缺失情形未知且动态变化，这些方法的泛化能力和鲁棒性可能会受到很大限制。此外，大多数以往方法在处理不完整输入时不可避免地会牺牲完整多模态性能，未能在单一模型中联合实现完整与不完整多模态学习 [7]。

本文提出 CyIN，这是一种新颖的循环信息学习框架，可在统一模型中整合完整与不完整多模态学习。首先，我们通过令牌级和标签级信息瓶颈（IB）构建有效的信息潜在空间，以促进多模态交互中的信息流动。前者在低层感知阶段基于词元嵌入搭建信息桥梁，并在各种单模态表示之间进行循环交互；后者则利用真实标签作为指导，将高层语义引入信息空间。结合这两个信息瓶颈目标，我们通过变分近似对瓶颈潜变量进行采样，从而高效捕获任务相关特征并过滤冗余噪声。此外，为解决模态缺失问题，我们引入跨模态循环翻译，在保留模态与缺失模态之间进行正向与反向传播，从而在信息空间中重建缺失信息。通过联合优化循环信息瓶颈和翻译过程，CyIN 在统一信息潜在空间中无缝衔接完整与不完整多模态学习。本文的主要贡献总结如下：

- **信息潜在瓶颈空间**。 通过 token 级和标签级信息瓶颈构建，所提出的信息潜在空间在一个统一框架中高效地衔接完整学习与不完整学习，使多模态融合和缺失信息重建都能受益于经过提纯的瓶颈潜变量。
- **循环交互与转换**。 所提出的循环信息处理过程显著提升了完整多模态学习中的跨模态交互性能，并增强了不完整多模态学习中的转换质量。
- 在 4 个数据集上进行的大量实验验证了 CyIN 在多模态学习中取得了最先进的性能，并且在各种模态缺失场景下保持了鲁棒性。

2 预备知识

2.1 完整与不完整多模态表示学习

考虑具有多个单模态数据源 $X_u \in \{X_0, X_1, ..., X_U\}$ 的多模态输入，多模态学习旨在整合配对模态 $u \in \{u_0, u_1, ..., u_U\}$ 中的互补信息，以学习能够支持各种下游任务推理的多模态表示 [25, 26], ，其中 $|u| = U$ 表示模态总数。模型性能取决于多模态融合策略能否有效利用各模态的信息 [27, 28],，而这反过来又要求模态完整。

如图 1所示，每种模态的原始输入首先由模态特定编码器处理E_u： $X_u \mapsto F_u$ 以生成单模态表征 F_u，然后由多模态融合解码器合并 D_M： $\{F_u\} \mapsto F_M$ 以生成多模态表征 F_M。将多模态输入记为 $X_u \equiv X_{complete}$，完整多模态学习的过程可表示为：

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

当多模态输入中存在缺失模态（记为 $X_u \equiv X_{incomplete} = \{X^{remain}, X^{miss}\}$）时，多模态模型的学习过程就转变为不完整多模态

学习，其中 X^{remain} 表示原始单模态数据输入，而 X^{miss} 表示没有对应单模态信息的零向量。因此，式F_M 中的不完整多模态表示应记为 1：

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

最终预测值 $\hat{}y_m$ 由多层感知机（MLP）头基于多模态表示 $\hat{}y_M$ $=MLP(F_M)$ 得出。然后，可根据具体的回归（平均绝对误差）或分类（交叉熵）任务计算优化目标，表示为：

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

由于缺失信息会中断原有的多模态融合空间[4, 22]，且对剩余模态过拟合可能引发语义歧义问题 [2, 14],因此，与完整多模态学习相比，不完整多模态学习的性能会显著下降。因此，如何在一个统一模型中同时改进完整多模态融合，并提升含有缺失模态的不完整输入的重建性能，仍是通用多模态系统面临的挑战 [7]。

2.2 信息瓶颈

信息瓶颈（IB）方法旨在将源状态 S 压缩为紧凑的瓶颈潜变量 B，其中包含来自 S 的最少必要特征，同时保留有关目标状态 T [29–31]的最相关信息。利用互信息表示状态之间的相关性，这种权衡可写为以下最小化问题：

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

其中，$I(\cdot|\cdot)$表示两个状态之间的互信息，超参数 β 控制权衡程度。通过引入参数化网络，前一个互信息项$I(B;S)$可由信息瓶颈编码器 E_S： $S \mapsto B$ 建模，以从源状态中提取信息；而第二项 $I(T;\ B)$可由信息瓶颈解码器 D_T： $B \mapsto T$实现，该解码器参照目标状态。

为进行优化，VIB [32] 利用变分近似替代难以处理的互信息项，从而得到以下上界：

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

其中，$p_\theta(B|S)\ \sim \mathcal{N}(\mu, \sigma^2)$和 $q(B)\ \sim \mathcal{N}(0, \mathrm{I})$均满足高斯分布。令损失目标 $\mathcal{L}_{ib} = I(B;\ S)$ $- \beta\, I(T;\ B)$，优化 $\mathcal{L}_{ib}$ 等价于最小化公式 5 中的上界。详细推导见附录 B。

Util化重参数化技巧 [32, 33] 以促进梯度传播，瓶颈潜变量 s
B c可从 $p_\theta(B|S)$中采样，记作：

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

作为在 S 和 T之间传递紧凑相关信息的桥梁，瓶颈潜变量B 能够有效构建信息空间，以控制从源状态S 到目标状态 T的信息流。更多相关工作和背景请参阅附录 A

3 方法论

3.1 多模态信息潜在空间

如图 1 所示，无论是完整还是不完整多模态学习，都需要与任务相关的信息表征，无论建模的是模态内还是模态间关系。因此，构建统一信息潜在空间不仅有助于弥合多模态表征融合中的模态差距，还能加强对各种模态中最具意义特征的重构。为提高信息潜在空间的效率，我们引入两种互补的信息瓶颈机制，以促进感知低层和语义高层瓶颈潜变量的紧凑性，分别称为令牌级和标签级信息瓶颈。

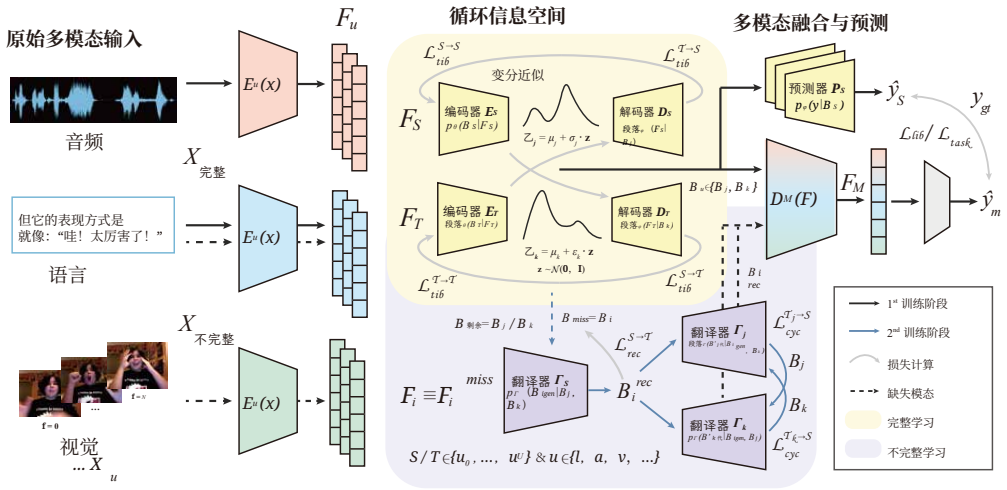


图1：框架概览。所提出的 CyIN 构建了一个循环信息空间，以联合训练完整和不完整的多模态学习。

令牌级信息瓶颈。 一般性不失地，对于多模态输入 $X_u \in \{X_0, ..., X_U\}$中的任意两种模态，我们将一种模态输入 X_S称为公式 X_T4中的源状态，将另一种模态输入 $X_S/X_T = \{x_u^0, ..., x_u^L\}$ 称为目标状态。令 u 表示模态 $S/T \in \{u_0, u_1, .., u_U\}$的令牌级输入序列，其中 L 表示序列长度。随后，通过相应的模态特定编码器 E_u： $X_u \mapsto F_u$，可得到单模态表示 $F_S/F_T = \{f_u^i\}|_{i=1}^L \in \mathbb{R}^{L\times C}$，其中 f_u 表示令牌嵌入， C 表示特定模态的特征维度。

借助信息瓶颈编码器 E_S： $F_S \mapsto B_S$，通过对所有词元嵌入应用信息瓶颈，可以获得令牌级瓶颈潜变量 $B_S = \{b_S^i\}|_{i=1}^L$，表示为：

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

由于单模态表征 F_S/F_T 由具有连续值的词元嵌入拼接而成，我们可以将后验概率表示为 $q_\phi(f_T|b_S)$， 借助信息瓶颈解码器 D_T： $B_S \mapsto F_T$ 将源表示的瓶颈投影到目标表示中。此外，给定 $p_\theta(b_S^i|f_S^i) \sim \mathcal{N}(\mu_B^i、(\sigma_B^i)^2)$和 $q(b_S^i)$， $\sim \mathcal{N}(0,\ \mathbf{I})$，则可以推导出公式 7 如下

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

循环交互。 对于来自各种模态的单模态表示 $F_u \in \{F_0, ..., F_U\}$，可以通过迭代方式循环选择词元级信息瓶颈中的源模态和目标模态，以滤除每个单模态表示中包含的冗余噪声，并增强跨模态交互。

当 $F_S = F_T$ 时，将源状态和目标状态视为同一模态，此时词元级信息瓶颈的优化侧重于学习模态特定特征并捕捉模态内动态，记作 $\mathcal{L}_{tib}^{S\to S}$。公式 8 中的两项分别侧重于从每个词元嵌入中压缩充分信息，并将其重新投影至原始序列。

另一方面，当源状态和目标状态来自具有$F_S \neq F_T$的不同模态时，词元级信息瓶颈旨在对不同模态之间的不变信息进行建模，并捕捉模态间动态。随后，公式 8 旨在整合模态共享特征，使源模态嵌入与目标模态嵌入相匹配。此外，在公式 8 中互换模态作为源状态和目标状态的角色，分别记为 $\mathcal{L}_{tib}^{S\to T}$和 $\mathcal{L}_{tib}^{T\to S}$，通过这种方向性信息流可以进一步探索模态共享特征。

结合模态内设置与模态间设置，词元级信息瓶颈的最终损失可表示为 as:

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

标签级信息瓶颈。与侧重低层感知的词元级信息瓶颈不同，我们引入标签级信息瓶颈，将高层语义的监督注入信息流中。对于来自每个模态 X_S、作为源状态的单模态表示 F_S，我们使用真实标签 y_{gt}（回归分数或识别类别）作为监督目标。给定每个模态的 N 个批量样本 $\{X_S^i\}|_{i=1}^N$，以及信息瓶颈编码器 E_S： $F_S \mapsto B_S$ 和预测器 P_S： $B_S \mapsto \hat{}y_S$，可通过以下损失约束标签级瓶颈潜变量：

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

遍历所有输入模态 $S \in \{u_0, u_1, ..., u_U\}$ 迭代计算 $\mathcal{L}_{lib}^S$，即可得到整体标签级信息瓶颈损失。实际应用中，对于回归任务，后验概率 $q_\phi(y_{gt}|B_S)$ 表示为 $q_\phi(y_{gt}|B_S) = e^{-\|y_{gt}-\hat{y}_S\|} = e^{-\|y_{gt}-P_S(B_S)\|}$，于是有：

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

对于包含 V 个类别的分类任务，后验概率 $q_\phi(y_{gt}|B_S)$计算为$q_\phi(y_{gt}|B_S) = \prod^V \hat{}y_S^{y_{gt}} = \prod^V [P_S(B_S)]^{y_{gt}}$，因此我们有：

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

通过令牌级和标签级信息瓶颈损失进行优化，我们可以构建一个信息空间，在任务相关语义的引导下控制模态内部及模态之间的信息流。

3.2 跨模态循环翻译

由于瓶颈潜变量是在信息空间中学习得到的，任务无关特征和单模态固有噪声已得到充分过滤。因此，对于不完整多模态输入，在构建的信息潜在空间中重构缺失信息，要比尝试在原始空间中重构容易得多。本文提出采用正向与反向传播的跨模态循环翻译，以增强模型在不完整多模态场景下的鲁棒性。

前向传播。为了重构缺失信息，我们采用级联残差自编码器（CRA） [34] 作为翻译器 $\Gamma_{S\to T}$： $B_S \mapsto B_T$ 将不同模态之间的转换视为机器翻译任务 [35]。已有充分研究证明，CRA 能够有效缓解模态差距，并将信息从源模态 S 转换到目标模态 T，表示如下：

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

其中，翻译器 Γ 由一系列堆叠的残差自编码器 $RA_i^S(\cdot)$ 组成，表示如下：

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

其中，r_T 表示每个 RA 块的输出，而 RA_i 的最后一个输出即为整体翻译信息，记作 $B_{S\to T}^{rec}$。对于翻译过程 $S \to T$，我们通过如下重建损失使翻译信息与原始信息对齐：

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

反向传播。为了提升翻译性能并充分探索模态间动态，我们进一步采用循环一致性学习 [17, 36] 来反向转换已翻译的

信息流的方向，作为回译技巧 [37]。由于前向传播表示信息流的重构过程 $B_S \rightarrow B_{S\rightarrow T}^{rec}$，即从源模态到目标模态，我们采用反向传播将重构后的信息转换回 $B_{S\rightarrow T}^{rec} \rightarrow B_S^{cyc}$，记作：

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

其中，翻译器 $\Gamma_{T\rightarrow S}$ 与用于前向传播的翻译器共享模型权重。因此，反向传播的重建损失可表示为：

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

泛化至多个保留模态。对于以两个以上模态作为输入的多模态学习，遇到不完整多模态输入时，模态的缺失情形具有很高的不确定性。当保留模态的数量多于一个时，如何有效整合这些模态的特征以重构缺失模态，仍有待解决，因为不完整输入情形多种多样。因此，我们提出一种可泛化的解决方案，利用成对跨模态翻译器生成缺失信息，而不受保留模态数量的限制。

考虑具有保留模态 $\{u_j, ..., u_k\}$ 和缺失模态 u_i 的数据 $X^{incomplete} \equiv \{X^{remain}, X^{miss}\}$，我们的目标是将保留模态 X^{remain} 的信息瓶颈潜变量 $B^{remain} = \{B_j, ..., B_k\}$ 翻译为缺失模态 X^{miss} 的潜变量 B^{miss}，表示为：

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

我们不再训练更多翻译器以适应保留模态输入的各种情形，而是利用高斯分布的可加性来整合翻译后的缺失信息。由于原始瓶颈潜变量 $\{B^{remain}, B^{miss}\}$ 均从信息空间中采样，我们直接根据源模态组合各个跨模态翻译器的输出，并将缺失模态的翻译后潜变量表示为：

$$B_i^{rec} = \sum_{j\neq i}^{|u|} B_{j\rightarrow i}^{rec} = \sum_{j\neq i}^{|u|} \Gamma_{j\rightarrow i}(B_j) = \prod_{j\neq i}^{|u|} p_\Phi^j(B_i|B_j) \tag{19}$$

其中，B_i^{rec} 可表示为从一种特殊形式的高斯混合模型 $\sum \Gamma_{j\rightarrow i}(\mathcal{N}(\mu_j,\ \sigma_j^2))$ 得到的翻译结果。无论其余模态的具体设置如何，我们都将 B_i^{rec} 用作最终翻译得到的信息潜变量，以补充缺失模态 B^{miss} 的信息。

结合正向与反向传播，跨模态循环翻译的目标为 ：

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

3.3 完整与不完整多模态融合

我们在一个统一的多模态融合网络中联合开展完整与不完整多模态学习，从而提升模型对不完整输入的鲁棒性，并保持其在完整输入下的性能。为简洁起见，我们采用 MultimodalTransformer [38] 来表示多模态融合解码器 D_M：$\{B_0, B_1, ..., B_U\} \mapsto F_M$ 用于输出循环信息空间中的瓶颈潜变量。请注意，融合模块可以替换为任意网络。

对于任意两个模态 $\{B_j, B_k\}$ 的瓶颈潜变量，多模态融合解码器 D_M 由一系列多头跨模态注意力层组成，其构成如下：

$$\begin{aligned}
y_r^h &= CMAttention_r(B_j, B_k) = Softmax\left[B_j W_q^h \cdot (W_k^h B_k)^T / \sqrt{C}\right] B_k W_v^h \\
Y_r^H &= Concat(y_r^1, .., y_r^h) W^o \\
Z_r &= Y_r^H + LayerNorm(Y_{r-1}^H) \text{ , where } r \in \{1, ..., R\} \\
M_r &= FeedForward(LayerNorm(Z_r)) + LayerNorm(Z_r)
\end{aligned} \tag{21}$$

其中，H 表示注意力头的数量，C 表示瓶颈潜变量的维度，W_q，W_k，W_v，W_o 表示权重矩阵，R 表示跨模态注意力层的数量。最后一层的输出 $M^{j,\ k}_{\ R}$ 用作模态 $\{u_j, u_k\}$ 的双模态融合潜变量。

针对完整与不完整多模态学习，通过配对模态从 $u \in \{u_0, ..., u_U\}$ 中迭代提取 $M^{j}_{R}{}^{k}$，最终的多模态表示可表示为：

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

优化目标总而言之，我们的框架包含四个学习目标，包括任务预测损失 $\mathcal{L}_{task}$，见公式3；令牌级信息瓶颈损失 $\mathcal{L}_{tib}$，见公式9；标签级信息瓶颈损失 $\mathcal{L}_{lib}$，见公式11 或公式12；以及跨模态循环翻译损失 $\mathcal{L}_{tran}$，见公式20。总损失可写为：

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

其中，β 表示瓶颈潜变量与表示之间互信息的权衡系数，γ 表示用于不完整多模态学习的翻译训练权重。

多阶段训练由于完整与不完整多模态学习都需要有效的信息瓶颈空间，我们将训练过程分为两个阶段。第一阶段设置 $\gamma = 0$，使多模态学习专注于在完整多模态学习场景下构建并稳定信息空间；第二阶段设置 $\gamma > 0$，逐步引入跨模态翻译器的训练，以增强不完整多模态学习。

4 实验

In t本节中，我们针对完整与不完整多模态学习开展了全面实验 1
lear在 4 个数据集上进行训练，以评估所提出框架的性能。

任务与数据集。 多模态回归任务：**MOSI** 和 **MOSEI** 是多模态情感分析数据集，分别包含 2,199 个和 22,856 个 YouTube 意见视频片段，每个片段都按照李克特量表标注了一个连续情感分数，范围从 –3（极度负面）到 +3（极度正面）。多模态分类任务：**IEMOCAP** 数据集提供了 7,369 段视频录制的对话，并标注了六种情绪类别：{高兴、悲伤、中性、愤怒、兴奋和沮丧}。**MELD** 数据集包含 13,391 个话语，取自电视剧《老友记》的多方对话，并标注了七种情绪：{中性、惊讶、恐惧、悲伤、喜悦、厌恶和愤怒}。这两个数据集用于多模态情感识别。

实现细节。 所有实验均在配备 Pytorch 2.4.1 和 CUDA 12.4 的 H800 GPU 上进行。对于音频和视觉模态，我们采用 ImageBind [39] 作为特征提取器，以获得更好的对齐性能。对于语言模态，在 MOSI 和 MOSEI 上使用预训练的 BERT [40]，在 IEMOCAP 和 MELD 上使用 sBERT [41]，以便与基线方法进行公平比较。各数据集的超参数详情见附录 D。

评估协议。 基于 [8, 17, 22, 24], ，我们通过以下方式评估不同方法的性能：(1) 完整多模态输入（$u \in \{l, a, v\}$）；(2) 在固定缺失协议（$u \in \{l\}/\{v\}/\{a\}/\{l, a\}/\{l, v\}/\{a, v\}$）下；以及 (3) 在缺失率为$MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$的随机缺失协议下。此处 $u \in \{l, a, v\}$ 分别表示语言、音频和视觉模态。请注意，与以往方法不同，我们采用一个统一模型，在各种输入情形下运行 10 次进行评估。缺失协议的具体细节见附录E。有关基线方法和评估指标的详细信息见附录 F-G。

定量比较。 完整输入或各种不完整多模态输入情形下的平均结果报告于表 1中。与以往基线方法相比，CyIN 在完整和不完整多模态输入场景下的大多数指标上均取得了更优性能。具体而言，借助信息空间的净化能力，当所有多模态输入均存在时，CyIN 能够充分探索模态特定特征和模态共享特征，并基于瓶颈潜变量执行高效的多模态融合。

在固定缺失协议下，CyIN 在固定缺失一种或两种模态的设置中始终优于基线方法，即使未针对任何特定模态进行专门调优，也展现出强大的泛化能力。整合主导模态与次要模态信息的能力凸显了信息瓶颈空间的灵活性。此外，在随机缺失协议下，CyIN 显著增强了模型面对各种随机模态缺失场景的鲁棒性，

表1：所提出的 CyIN 与基线方法的性能比较，展示了完整模态设置 $u \in \{l, a, v\}$ 以及不完整模态设置下的平均结果；其中固定缺失协议包括模态设置 $u \in \{l\}/\{v\}/\{a\}/\{l, a\}/\{l, v\}/\{a, v\}$，随机缺失协议包括缺失率 $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$

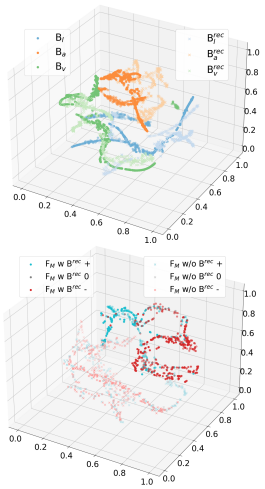
设置	模型	MOSI				MOSEI				IEMOCAP		MELD	
		准确率7↑	F1↑	MAE↓	相关系数↑	准确率7↑	F1↑	MAE↓	相关系数↑	Acc↑	wF1↑	Acc↑	wF1↑
完整 $u \in \{l, a, v\}$	CCA	27.7	74.9	1.106	0.541	46.1	82.9	0.654	0.666	61.7	61.5	51.2	46.6
	DCCA	25.1	73.6	1.220	0.422	39.3	73.3	0.787	0.425	56.8	55.5	47.7	37.0
	DCCAE	19.7	69.5	1.642	0.357	38.9	73.5	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**
固定缺失	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
随机缺失	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**

即使在严重缺失率下，性能损失也极小。该结果说明，CyIN 可更广泛地应用于真实世界中的下游多模态任务，其中模态的存在高度动态且不可预测。更多结果与分析见附录 H

定性比较。 我们将 MOSI 数据集测试集中的特征投影到 3D t-SNE 空间中 [42]。如图 2(a) 所示，我们首先将转换后的 B_u^{rec}和原始单模态潜变量的分布进行可视化 B_u，以展示重构质量。跨模态转换潜变量与原始潜变量紧密聚类，说明尽管存在巨大的模态差距，CyIN 仍能有效地在不同模态之间传递信息。
此外，我们展示了 F_M 包含和不包含重构瓶颈 B^{rec} 且存在随机缺失 $MR = 0.7$时的多模态潜变量。我们可以观察到，使用缺失模态进行推理会对多模态表示造成严重干扰，而重构潜变量能够在多模态融合过程中有效补充缺失信息，从而获得更好的聚类结果。

案例研究。 如图 2(b)所示，我们在固定模态缺失协议下，对比了 MOSI 和 IEMOCAP 数据集测试集中的一些样本在使用和不使用 CyIN 时的模型预测结果。在样本#1 和#3中，移除语义话语会导致预训练模型作出错误甚至相反的预测，这也间接表明了语言模态在多模态情感计算任务中的主导作用 [8, 16, 43, 44]。另一方面，其他模态的存在也对预测准确率产生了细微影响，这可从样本#2 和#4中缺失音频或视觉模态时出现的主观偏置预测看出。相比之下，通过使用所提出的 CyIN 重构缺失信息，预测结果变得更加稳定和精确，证明了信息瓶颈空间的优越性和鲁棒性。

消融研究。 我们在 MOSI 和 IEMOCAP 数据集上，针对完整输入和最严重的随机缺失协议 $MR = 0.7$开展了所提出模块的消融研究，如表 2所示。令牌级和标签级信息瓶颈对于构建信息潜在空间都至关重要，移除其中任一项都会导致性能下降。跨模态循环交互与转换主要用于在多模态融合中捕获模态共享特征，并增强



(a) 特征分布

#	多模态输入（$u \in \{language, vision, audio\}$）	模态状态	真实值	预测 使用 CyIN	预测 不使用 CyIN
1	“而他，我不知道，他可能生气了，当时，哈，我也不知道”  推测和疑惑语气	✗ ✓ ✓	-0.250	-0.364	0.263
2	“所以我觉得能亲眼看到 Ray Park 的表现很酷，他真是一位出色的动作明星。”  满足和释然的语气	✓ ✗ ✗	2.400	2.609	1.110
3	“你不必因为碰巧不想要任何东西 就表现得如此了不起……”  指责且激动的语气	✗ ✓ ✓	愤怒	愤怒	沮丧
4	“你不会有麻烦的，你不会有麻烦的。”  平静且快速的语气	✓ ✗ ✓	高兴	高兴	中性

(b) MOSI（#1-2）和 IEMOCAP（#3-4）上的案例研究

图 2：(a) 转换后的单模态潜变量和多模态潜变量的特征分布；(b) 在使用和不使用 CyIN 重构信息进行推理时，MOSI 和 IEMOCAP 数据集测试集上的示例。模态状态中的 ✓和 ✗分别表示保留的模态和缺失的模态。

不完整重建。此外，我们证明了信息瓶颈空间在完整与不完整多模态学习中的有效性。若不受瓶颈空间的约束，原始特征空间中与任务无关的冗余和异质噪声会增加整合不同模态信息及重构缺失模态的难度。最后，利用剩余模态进行重构后，翻译后的潜变量能够有效提升对模态缺失问题的鲁棒性，尤其是在输入严重不完整的情况下。

表 2：在完整多模态设置 $u \in \{l, a, v\}$ 和缺失率为 $MR = 0.7$的随机缺失协议下，所提出的 CyIN 在 MOSI 和 IEMOCAP 数据集上的消融研究。

设置	模型变体	MOSI				IEMOCAP	
		准确率7↑	F1↑	MAE↓	相关系数↑	Acc↑	wF1↑
完整 $u \in \{l, a, v\}$	**CyIN**	**48.0**	**86.3**	0.712	**0.801**	**66.1**	**66.0**
	不含 $\mathcal{L}_{tib}$	43.9	84.3	0.737	0.789	65.2	64.8
	不含 $\mathcal{L}_{lib}$	47.3	85.4	**0.693**	0.800	63.6	63.9
	无循环交互	43.4	85.9	0.742	0.795	64.3	63.6
	无循环翻译	43.3	85.7	0.743	0.797	65.2	65.0
	无信息空间	42.0	83.1	0.747	0.782	62.3	62.1
随机缺失 $MR = 0.7$	**CyIN**	**28.0**	65.9	**1.117**	**0.530**	**48.6**	**49.0**
	无 $\mathcal{L}_{tib}$	27.3	63.5	1.223	0.509	46.3	46.2
	无 $\mathcal{L}_{lib}$	25.1	63.7	1.220	0.516	43.2	42.9
	无循环交互	26.5	**67.8**	1.134	0.473	47.8	48.6
	无循环翻译	24.5	63.9	1.171	0.473	47.1	46.6
	无信息空间	23.7	62.9	1.240	0.430	44.1	43.6
	无翻译后的潜变量	23.4	56.7	1.299	0.441	41.2	39.2

超参数敏感性。 我们通过实验评估了不同超参数设置对 β 和 γ 在公式 23 的损失函数中，在 $MR = 0.7$条件下的影响。表 3 中的结果表明，设置$\beta = 8 - 32$能够在以下各项之间实现互信息的平衡折中，从而获得更好的性能：$I(S;\ B)$和 $I(B;\ T)$；其中，不恰当的瓶颈强度会损害性能。此外，设置 $\gamma = 10$ 在各项指标上都取得了最佳的总体性能，表明适度的重构能够促进模态间更好的一致性。我们观察到，较小的 γ 会导致正则化较弱，使得在严重缺失的条件下难以对齐模态。相反，较大的 γ 会因过度关注一致性而忽视有效构建信息空间的需求，从而损害性能。

此外，为确保在各种模态缺失场景下的通用性和稳定性，CyIN 假设各模态在多模态学习中的贡献相等。然而，实际中，不同模态的贡献可能因任务而异 [45]。例如，正如先前研究 [8, 46, 47]所指出的，语言在多模态情感分析中起主导作用。为

表 3：在缺失率为 $MR = 0.7$ 的随机缺失协议下，CyIN 的 β 和 γ 的超参数敏感性。

模型变体		MOSI			
		准确率7↑	F1↑	MAE↓	相关系数↑
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

表 4：在缺失率为 $MR = 0.7$ 的随机缺失协议下，CyIN 在 MOSI 数据集上对语言、视觉和音频模态进行转换时，CRA 层不同比例的超参数敏感性。

# 层数 (la:lv:av)	MOSI			
	准确率7↑	F1↑	MAE↓	相关系数↑
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**

对此进行部分解决，我们通过改变表中跨模态转换模块的 CRA 层数量开展实验4。具体而言，我们分配更多层来重构语言模态，而分配给音频和视觉模态的层数较少，以体现语言模态更高的重要性，并使模型能够学习包含更多语义的表征。这种设计展现了 CyIN 架构的灵活性，无需在其中硬编码任何模态特定先验。

计算效率。与最先进的方法相比，如表 5 所示，所提出的 CyIN 总参数量最低，浮点运算次数也显著更少。CyIN 的推理时间比 GCNet 快 $\sim 3\times$，比 IMDer 快 $\sim 5\times$。由于训练期间采用循环翻译过程，CyIN 的训练时间略长于用于模态重建的图神经网络模型 GCNet，但仍快于使用基于分数的扩散模型进行跨模态生成的 IMDer。

表 5：MOSI 数据集上训练和推理计算效率的比较。

模型	总参数量（M）	总训练时间	推理时间（/次迭代）	浮点运算次数（T）
GCNet	144.34 M	**1.47 小时**	70.62秒	3.747
IMDer	168.31 M	1.89小时	103.75秒	5.466
CyIN（我们的方法）	**123.49 M**	1.61h	**22.41s**	**1.594**

5 结论

在本论文中，我们提出了一个名为CyIN的新颖框架，该框架构建信息潜在空间，以联合开展完整与不完整多模态学习。在token级与标签级信息瓶颈的引导下，CyIN成功学习到一个紧凑而语义丰富的瓶颈潜变量，从而纯化任务相关特征并提升稳健的多模态融合。所提出的循环交互与转换机制进一步促进对模态间动态的充分探索，并提升缺失模态的重构质量。在4个多模态数据集上的全面实验表明，CyIN不仅在完整多模态学习中超越了以往方法，而且在各种不完整场景下仍保持更优的性能与稳定性，凸显了其在真实世界多模态应用中的有效性与泛化能力。

局限性

我们的框架仍存在若干局限性。1）该框架对所有模态一视同仁，以提升泛化能力，却忽略了不同模态可能存在的贡献不平衡。我们将尝试根据各模态的信息丰富程度自适应地分配权重。2）当前跨模态翻译器的性能可通过采用更新的生成式方法（如扩散模型或基于流的模型）进一步提升，以获得最佳重建性能。3）我们认为，在多样化的多模态理解数据集上进行更广泛的训练，或与大语言模型集成，有望增强所提框架的泛化能力和可扩展性。

致谢

本研究得到国家自然科学基金委员会（62076262、61673402、61273270、60802069）和中山大学青年人才科研人员国际项目基金的资助。

参考文献

[1] Renjie Wu、Hu Wang、Hsiang-Ting Chen 和 Gustavo Carneiro。《缺失模态深度多模态学习：综述》。arXiv 预印本 arXiv:2409.07825，2024。1, 18[2] Mengmeng Ma、Jian Ren、Long Zhao、Davide Testuggine 和 Xi Peng。《多模态变换器能否适应缺失模态？》载于《2022 IEEE/CVF计算机视觉与模式识别会议（CVPR）》，第18156–18165 页，2022。doi: 10.1109/CVPR52688.2022. 01764。1, 3, 18[3] Mengmeng Ma、Jian Ren、Long Zhao、Sergey Tulyakov、Cathy Wu 和 Xi Peng。《Smil：严重缺失模态下的多模态学习》。《AAAI人工智能会议论文集》，35(3):2302–2310，2021年5月。doi: 10.1609/aaai.v35i3.16330。网址 https://ojs.aaai.org/index.php/AAAI/article/view/16330。1[4] Ronghao Lin 和 Haifeng Hu。《MissModal：提升多模态情感分析应对缺失模态的鲁棒性》。《计算语言学协会汇刊》，11:1686–1702，2023年12月。ISSN 2307-387X。doi: 10.1162/tacl_a_00628。网址 https://doi. org/10.1162/tacl_a_00628。2, 3, 18[5] Ziqi Yuan、Yihe Liu、Hua Xu 和 Kai Gao。《用于鲁棒多模态情感分析的基于噪声模仿的对抗训练》。《IEEE多媒体汇刊》，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 和 Lihua Zhang。《通过分层表示学习实现鲁棒的不完整多模态情感分析》。《第三十八届神经信息处理系统年会论文集》，2024。网址 https://openreview.net/forum?id=XgwTH95kCl。1, 2[7] Ronghao Lin 和 Haifeng Hu。《适应与探索：用于表示学习的多模态混合增强》。《信息融合》，105:102216，2024。ISSN 1566-2535。doi: https://doi.org/10. 1016/j.inffus.2023.102216。网址 https://www.sciencedirect.com/science/article/ pii/S1566253523005328。2, 3[8] Haoyu Zhang、Wenbin Wang 和 Tianshu Yu。《利用不完整数据实现鲁棒的多模态情感分析》。《第三十八届神经信息处理系统年会论文集》，2024。网址 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 和 Danica Kragic。《几何多模态对比表示学习》。载于 Kamalika Chaudhuri、Stefanie J egelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），《第39届国际机器学习会议论文集》，第162卷《机器学习研究论文集》，第17782–17800页。PMLR，2022年7月17–23日。网址 https://proceedings.mlr.press/v162/poklukar22a.html。2, 18

[10] Rui Liu、Haolin Zuo、Zheng Lian、Björn W. Schuller 和李海洲。基于对比学习的模态不变特征获取，用于在缺失模态情况下实现鲁棒的多模态情感识别。*IEEE*情感计算汇刊，15(4):1856–1873，2024。doi: 10.1109/TAFFC.2024.3378570。2

[11] 哈罗德·霍特林。 两组变量之间的关系，第162–190页。施普林格纽约，纽约，纽约州，1992。ISBN 978-1-4612-4380-9。doi: 10.1007/978-1-4612-4380-9_14。网址 https://doi.org/10.1007/978-1-4612-4380-9_14。 2，25

[12] Galen Andrew、Raman Arora、Jeff Bilmes和Karen Livescu。深度典型相关分析。见第30届国际机器学习会议论文集——第28卷，ICML'13，第III–1247–III–1255页。JMLR.org，2013年。25[13] Weiran Wang、Raman Arora、Karen Livescu和Jeff Bilmes。论深度多视图表示学习。见第32届国际机器学习会议论文集——第37卷，ICML'15，第1083–1092页。JMLR.org，2015年。2, 25[14] Devamanyu Hazarika、Yingting Li、Bo Cheng、Shuai Zhao、Roger Zimmermann和Soujanya Poria。分析多模态情感分析中的模态鲁棒性。见2022年计算语言学协会北美分会会议：人类语言技术论文集，第685–696页，西雅图，美国，2022年7月。计算语言学协会。doi: 10.18653/v1/2022.naacl-main.50。网址 https://aclanthology.org/2022.naacl-main.50。2, 3, 18[15]Yoshua Bengio、Pascal Lamblin、Dan Popovici和Hugo Larochelle。深度网络的贪心逐层训练。载于B. Schölkopf、J. Platt和T. Hoffman编，《神经信息处理系统进展》，第19卷。麻省理工学院出版社，2006年。网址 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和Barnabás Póczos。见于翻译：通过模态间循环翻译学习鲁棒联合表示。见AAAI人工智能会议论文集，第33卷，第6892–6899页，2019年。2, 8, 25, 29

[17] Jinming Zhao、Ruichen Li 和 Qin Jin。面向不确定缺失模态情感识别的缺失模态想象网络。载于 计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第2608–2618页，线上，2021年8月。计算语言学协会。doi: 10.18653/v1/2021.acl-long.203。网址https://aclanthology.org/2021.acl-long.203。 2、5、7、18、24、25

[18] Mike Wu 和 Noah Goodman。用于可扩展弱监督学习的多模态生成模型。载于《第32届神经信息处理系统国际会议论文集》，NIPS'18，第5580–5590页，美国纽约州雷德胡克，2018年。Curran Associates Inc. 2[19] Yuge Shi、N. Siddharth、Brooks Paige 和 Philip H. S. Torr。用于多模态深度生成模型的变分混合专家自动编码器。Curran Associates Inc.，美国纽约州雷德胡克，2019年。[20] Miguel Vasco、Hang Yin、Francisco S. Melo 和 Ana Paiva。利用多模态生成模型中的层级结构，实现有效的跨模态推理。《神经网络》，146:238–255，2022年。ISSN 0893-6080。doi: https://doi.org/10.1016/j.neunet.2021.11.019。网址：https: // www.sciencedirect.com/science/article/pii/S0893608021004470。[21] Miguel Vasco、Hang Yin、Francisco S. Melo 和 Ana Paiva。如何感知世界：利用多模态感知中的层级结构，构建稳健的强化学习智能体。载于《第21届自主智能体与多智能体系统国际会议论文集》，AAMAS '22，第1301–1309页，里奇兰，南卡罗来纳州，2022年。国际自主智能体与多智能体系统基金会。ISBN 9781450392136。2[22] Zheng Lian、Lan Chen、Licai Sun、Bin Liu 和 Jianhua Tao。GCNet：用于对话中不完整多模态学习的图补全网络。《IEEE模式分析与机器智能汇刊》，45(7):8419–8432，2023年。doi: 10.1109/TPAMI.2023.3234553。2, 3, 7, 18, 24, 25[23] Yuntao Shou、Tao Meng、Wei Ai、Fuchen Zhang、Nan Yin 和 Keqin Li。通过信息瓶颈进行对抗对齐与图融合，用于对话中的多模态情感识别。《信息融合》，112:102590，2024年。ISSN 1566-2535。doi: https://doi.org/10.1016/j.inffus.2024.102590。网址：https://www.sciencedirect.com/ science/article/pii/S1566253524003683。2

[24] Yuanzhi Wang、Yong Li 和 Zhen Cui。不完整多模态扩散情感识别。载于《第三十七届神经信息处理系统大会》，2023年。网址：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 和 Wen Gao。大规模多模态预训练模型：综合综述。《机器智能研究》，20(4):447–482，2023年。ISSN 2731-538X。doi: 10.1007/ s11633-022-1410-8。网址：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 和 Dong Yu。MM-LLMs：多模态大语言模型的最新进展。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，《计算语言学协会年会 ACL 2024 论文集》，第12401–12430页，泰国曼谷，2024年8月。计算语言学协会。doi: 10.18653/v1/2024.findings-acl.738。网址：https://aclanthology.org/2024.findings-acl.738/。2[27] Dimitris Gkoumas、Qiuchi Li、Christina Lioma、Yijun Yu 和 Dawei Song。是什么造成了差异？多模态语言分析融合策略的实证比较。《信息融合》，66:184–197，2021年。ISSN 1566-2535。doi: https://doi.org/10.1016/j.inffus.2020.09.005。网址：https://www.sciencedirect.com/science/article/pii/S1566253520303675。2, 18, 25[28] Muhammad Arslan Manzoor、Sarah Albarri、Ziting Xian、Zaiqiao Meng、Preslav Nakov 和 Shangsong Liang。多模态表征学习：关于其演进、预训练及应用的综述。《ACM多媒体计算、通信与应用汇刊》，20(3)，2023年10月。ISSN 1551-6857。doi: 10.1145/3617833。网址：https://doi.org/10.1145/3617833。2[29] Naftali Tishby、Fernando C. Pereira 和 William Bialek。信息瓶颈方法。载于《第37届 Allerton 通信、控制与计算年会论文集》，第368–377页，1999年。网址：https://arxiv.org/abs/physics/0004057。3, 18[30] Naftali Tishby 和 Noga Zaslavsky。深度学习与信息瓶颈原理。载于《2015年 IEEE 信息论研讨会 (ITW)》，第1–5页，2015年。doi: 10.1109/ITW.2015. 7133169。18[31] Andrew Michael Saxe、Yamini Bansal、Joel Dapello、Madhu Advani、Artemy Kolchinsky、Brendan Daniel Tracey 和 David Daniel Cox。论深度学习的信息瓶颈理论。载于《国际学习表征会议》，2018年。网址：https: //openreview.net/forum?id=ry_WPG-A-。3, 18[32] Alexander A. Alemi、Ian Fischer、Joshua V. Dillon 和 Kevin Murphy。深度变分信息瓶颈。载于《国际学习表征会议》，2017年。网址：https://openreview.net/forum?id=HyxQzBceg。3, 18[33] Max Welling Diederik P. Kingma。自动编码变分贝叶斯。载于《第二届国际学习表征会议》，ICLR 2014，2014年。3[34] Luan Tran、Xiaoming Liu、Jiayu Zhou 和 Rong Jin。通过级联残差自编码器填补缺失模态。载于《IEEE计算机视觉与模式识别会议 (CVPR) 论文集》，2017年7月。5, 25[35] Dzmitry Bahdanau、Kyung Hyun Cho 和 Yoshua Bengio。通过联合学习对齐与翻译实现神经机器翻译。2015年1月。《第三届国际学习表征会议》，ICLR 2015；会议日期：2015年5月7日至2015年5月9日。5[36] Zilong Wang、Zhaohong Wan 和 Xiaojun Wan。Transmodality：一种基于 Transformer 的端到端多模态情感分析融合方法。载于《2020年万维网大会论文集》，WWW '20，第2514–2520页，美国纽约州纽约，2020年。美国计算机协会。ISBN 9781450370233。doi: 10.1145/3366423.3380000。网址：https://doi.org/ 10.1145/3366423.3380000。5

[37] Cong Duy Vu Hoang、Philipp Koehn、Gholamreza Haffari 和 Trevor Cohn。用于神经机器翻译的迭代回译。载于《第2届神经机器翻译与生成研讨会论文集》，第18–24页。计算语言学协会，2018年7月。网址 https://sites.google.com/site/wnmt18/home,https: //sites.google.com/site/wnmt18/。第2届神经机器翻译与生成研讨会（WNMT 2018）；会议日期：2018年7月15日至2018年7月20日。6[38] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J. Zico Kolter、Louis-Philippe Morency 和 Ruslan Salakhutdinov。用于未对齐多模态语言序列的多模态变换器。载于《第57届计算语言学协会年会论文集》，第6558–6569页，意大利佛罗伦萨，2019年7月。计算语言学协会。doi：10.18653/v1/P19-1656。网址 https://aclanthology.org/P19-1656。6[39] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：一个嵌入空间，绑定万物。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第15180–15190页，2023年6月。7[40] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》第1卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。doi：10.18653/v1/N19-1423。网址 https://aclanthology.org/N19-1423。7, 30[41] Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生BERT网络生成句子嵌入。载于 Kentaro Inui、Jing Jiang、Vincent Ng 和 Xiaojun Wan 编，《2019年自然语言处理实证方法会议暨第9届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集》，第3982–3992页，中国香港，2019年11月。计算语言学协会。doi：10.18653/v1/D19-1410。网址 https://aclanthology.org/D19-1410。7[42] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究期刊》，9(86):2579–2605，2008年。网址 http://jmlr.org/papers/v9/ vandermaaten08a.html。8[43] Devamanyu Hazarika、Roger Zimmermann 和 Soujanya Poria。MISA：用于多模态情感分析的模态不变与模态特定表示。载于《第28届ACM国际多媒体会议论文集》，第1122–1131页，2020年。网址 https://doi.org/ 10.1145/3394171.3413678。8, 18, 29[44] Ronghao Lin 和 Haifeng Hu。通过混合模态注意力动态转换多模态表示，用于多模态情感分析。《 IEEE多媒体汇刊》，26: 2740–2755，2023年。doi：10.1109/TMM.2023.3303711。8, 29[45] Xiaokang Peng、Yake Wei、Andong Deng、Dong Wang 和 Di Hu。通过即时梯度调制实现平衡多模态学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年。9[46] Ronghao Lin 和 Haifeng Hu。用于多模态情感分析的多任务动量蒸馏。《IEEE情感计算汇刊》，第1–18页，2023年。doi：10.1109/TAFFC. 2023.3282410。9, 18[47] Zheng Lian、Licai Sun、Yong Ren、Hao Gu、Haiyang Sun、Lan Chen、Bin Liu 和 Jianhua Tao。MERBench：多模态情感识别统一评测基准。arXiv预印本 arXiv:2401.03429，2024年。9, 18[48] Wei Han、Hui Chen 和 Soujanya Poria。通过层次互信息最大化改进多模态融合，用于多模态情感分析。载于《2021年自然语言处理实证方法会议论文集》，第9180–9192页，线上及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。doi：10.18653/v1/2021.emnlp-main.723。网址 https://aclanthology.org/ 2021.emnlp-main.723。18

[49] Haoyu Zhang、Yu Wang、Guanghao Yin、Kejun Liu、Yuanyuan Liu 和 Tianshu Yu。学习语言引导的自适应超模态表示，用于多模态情感分析。载于*2023*年自然语言处理实证方法会议论文集，第756–767页，新加坡，2023年12月。计算语言学协会。[50] Xiongye Xiao、Gengshuo Liu、Gaurav Gupta、Defu Cao、Shixuan Li、Yaxing Li、Tianqing Fang、Mingxi Cheng 和 Paul Bogdan。用于多模态学习的神经启发信息论层次感知。载于第十二届国际学习表征会议，2024。网址https://openreview.net/forum?id=Z9AZsU1Tju。18

[51] Guimin Hu、Ting-En Lin、Yi Zhao、Guangming Lu、Yuchuan Wu 和 Yongbin Li。UniMSE：迈向统一的多模态情感分析与情感识别。载于 Yoav Goldberg、Zornitsa ozareva 和 Yue Zhang（编），*2022*年自然语言处理实证方法会议论文集，第 7837–7851 页，阿布扎比，阿联酋，2022年12月。计算语言学协会。doi：10.18653/v1/2022.emnlp-main.534。网址https://aclanthology.org/2022.emnlp-main.534。18

[52] Dongyuan Li、Yusong Wang、Kotaro Funakoshi 和 Manabu Okumura。Joyful：面向多模态情感识别的联合模态融合与图对比学习。载于 Houda Bouamor、Juan Pino 和 Kalika Bali（编），*2023*年自然语言处理实证方法会议论文集，第 16051–16069 页，新加坡，2023年12月。计算语言学协会。doi：10.18653/v1/2023.emnlp-main.996。网址 https://aclanthology.org/2023.emnlp-main.996/。

[53] Ronghao Lin、Ying Zeng、Sijie Mai 和 Haifeng Hu。端到端以语义为中心的视频多模态情感计算。*arXiv* 预印本 *arXiv:2408.07694*，2024 年。18

[54] Zebang Cheng、Zhi-Qi Cheng、Jun-Yan He、Kai Wang、Yuxiang Lin、Zheng Lian、Xiaojiang Peng 和 Alexander G Hauptmann。Emotion-LLaMA：通过指令微调实现多模态情感识别与推理。载于第三十八届神经信息处理系统年会，2024 年。网址 https://openreview.net/forum?id=qXZVSy9LFR。18

[55] Jin Li、Shoujin Wang、Qi Zhang、Shui Yu 和 Fang Chen。以公平性生成：一种面向不完整多模态推荐的模态扩散反事实框架。载于*2025* 年美国计算机协会万维网会议论文集，WWW '25，第 2787–2798 页，纽约州纽约，美国，2025 年。美国计算机协会。ISBN 9798400712746。doi：10.1145/3696410.3714606。网址https://doi.org/10.1145/3696410.3714606。18，30

[56] Vittorio Pipoli、Alessia Saporita、Kevin Marchesini、Costantino Grana、Elisa Ficarra 和 Federico Bolelli。Im-fuse：一种用于不完整模态脑肿瘤分割的基于Mamba的融合块。载于 James C. Gee、Daniel C. Alexander、Jaesung Hong、Juan Eugenio Iglesias、Carole H. Sudre、Archana Venkataraman、Polina Golland、Jong Hyo Kim 和 Jinah Park 编，医学图像计算与计算机辅助干预 – *MICCAI 2025*，第 225–235 页，Cham，2025。Springer Nature Switzerland。ISBN 978-3-032-04984-1。18，30，31

[57] Paul Pu Liang、Amir Zadeh 和 Louis-Philippe Morency。多模态机器学习的基础与趋势：原则、挑战和开放问题。*ACM*计算综述，2024年4月。ISSN 0360-0300。doi：10.1145/3656580。网址https://doi.org/10.1145/3656580。已接受。 18

[58] Fei Zhao、Chengcui Zhang和Baocheng Geng。深度多模态数据融合。 *ACM*计算综述，56(9)，2024年4月。ISSN 0360-0300。doi: 10.1145/3649447。网址 https://doi.org/10.1145/3649447。 18

[59] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的。在 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Gar- nett（编辑）的神经信息处理系统进展第 30 卷。Curran Associates, Inc.，2017 年。网址 https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf。18

[60] Peng Xu、Xiatian Zhu 和 David A. Clifton。Transformer 多模态学习：综述。IEEE模式分析与机器智能汇刊，45(10)：12113–12132，2023 年。doi：10.1109/TPAMI.2023.3275156。18[61] Ruohong Huan、Guowei Zhong、Peng Chen 和 Ronghua Liang。Unimf：面向缺失模态和未对齐多模态序列的多模态情感分析统一多模态框架。IEEE多媒体汇刊，26：5753–5768，2024。doi：10.1109/TMM.2023。3338769。18[62] Jingjing Tang、Qingqing Yi、Saiji Fu 和 Yingjie Tian。不完整多视图学习：回顾、分析与展望。《应用软计算》，153：111278，2024。ISSN 1568-4946。doi：https://doi.org/10.1016/j.asoc.2024.111278。网址 https://www.sciencedirect.com/ science/article/pii/S1568494624000528。18[63] Changqing Zhang、Yajie Cui、Zongbo Han、Joey Tianyi Zhou、Huazhu Fu 和 Qinghua Hu。深度部分多视图学习。IEEE模式分析与机器智能汇刊，44(5)：2402–2415，2022 年。doi：10.1109/TPAMI.2020.3037734。18，25[64] Ravid Shwartz-Ziv 和 Naftali Tishby。通过信息打开深度神经网络的黑箱。3 2017。18[65] Kenji Kawaguchi、Zhun Deng、Xu Ji 和 Jiaoyang Huang。信息瓶颈如何帮助深度学习？载于国际机器学习会议（ICML），2023 年。18[66] Qi Wang、Claire Boudreau、Qixing Luo、Pang-Ning Tan 和 Jiayu Zhou。深度多视图信息瓶颈，第 37–45 页。doi：10.1137/1.9781611975673.5。网址 https:// epubs.siam.org/doi/abs/10.1137/1.9781611975673.5。18[67] Marco Federici、Anjan Dutta、Patrick Forré、Nate Kushman 和 Zeynep Akata。通过多视图信息瓶颈学习鲁棒表示。载于国际学习表征会议，2020 年。网址 https://openreview.net/forum?id=B1xwcyHFDr.[68] Zhibin Wan、Changqing Zhang、Pengfei Zhu 和 Qinghua Hu。多视图信息瓶颈表示学习。《美国人工智能学会会议论文集》，35(11)：10085–10092，2021 年 5 月。doi：10.1609/aaai.v35i11.17210。网址 https: // ojs.aaai.org/index.php/AAAI/article/view/17210。18[69] Diederik P. Kingma 和 Max Welling。自动编码变分贝叶斯。载于第二届国际学习表征会议，ICLR 2014，加拿大艾伯塔省班夫，2014 年 4 月 14–16 日，会议论文集，2014 年。18[70] Changhee Lee 和 Mihaela van der Schaar。用于多组学数据整合的变分信息瓶颈方法。载于 Arindam Banerjee 和 Kenji Fukumizu（编辑），第 24 届国际人工智能与统计会议论文集，《机器学习研究论文集》第 130 卷，第 1513–1521 页。PMLR，2021 年 4 月 13–15 日。网址 https://proceedings.mlr.press/v130/lee21a.html。18[71] Paul Pu Liang、Zihao Deng、Martin Q. Ma、James Y Zou、Louis-Philippe Morency 和 Ruslan Salakhutdinov。因子化对比学习：超越多视图冗余。载于 A. Oh、T. Naumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine（编辑），神经信息处理系统进展，第 36 卷，第 32971–32998 页。Curran Associates, Inc.，2023 年。网址 https://proceedings.neurips.cc/paper_files/paper/2023/file/6818dcc65fdf3cbd4b05770fb957803e-Paper-Conference.pdf。19[72] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于国际学习表征会议，2019 年。网址 https://openreview.net/forum? id=Bkg6RiCqY7。24[73] Soujanya Poria、Erik Cambria、Rajiv Bajpai 和 Amir Hussain。情感计算综述：从单模态分析到多模态融合。《信息融合》，37：98–125，2017 年。ISSN 1566-2535。doi：https://doi.org/10.1016/j.inffus.2017.02.003。网址 https://www.sciencedirect.com/science/article/pii/S1566253517300738。25

[74] Geetha A.V.、Mala T.、Priyanka D.与Uma E. 基于深度学习的多模态情感识别：进展、挑战与未来方向。《信息融合》，105:102218，2024。ISSN 1566-2535。doi: https://doi.org/10.1016/j.inffus.2023.102218。网址 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与Veselin Stoyanov。Ro{bert}a：一种经过稳健优化的{bert}预训练方法，2020。网址 https://openreview.net/forum?id=SyxS0T4tvS。 30[76] Pengcheng He、Jianfeng Gao与Weizhu Chen。DeBERTav3：采用梯度解耦嵌入共享的ELECTRA式预训练改进deBERTa。载于第十一届国际学习表征会议，2023。网址https://openreview.net/forum?id=sE7-XhLxHA。 30[77] Shicai Wei、Yang Luo、Yuji Wang与Chunbo Luo。通过表示解耦实现稳健的多模态学习。见Aleš Leonardis、Elisa Ricci、Stefan Roth、Olga Russakovsky、Torsten Sattler与Gül Varol编，计算机视觉——*ECCV 2024*，第38–54页，Cham，2025年。Springer Nature Switzerland。ISBN 978-3-031-72946-1。 30，31

A 相关工作

A.1 多模态表示学习

多模态表示学习旨在构建多模态系统，以开展多模态理解任务，包括多模态情感分析 [43, 46, 48–50]或情感识别[51–54], 多模态推荐 [55] 以及多模态分割 [56] 等。多模态学习的通用框架可分为两个主要模块：模态特定表示学习与多模态融合，旨在反映模态元素之间的异质性与相互关联 [57]。

一方面，在每种单独模态中有效提取任务相关特征，是多模态表示学习的基础 [47, 53]。针对不同模态，这些模态特定编码器着重提取关键特征并抑制无关噪声，从而缩短从原始数据到语义上有意义的单模态表示的路径。另一方面，多模态融合旨在弥合模态差距并利用模态间互补性[27, 57]。早期融合方案只是简单地拼接单模态表示，而晚期融合方法则通过加权投票或门控来组合单模态预测。混合融合结合早期融合和晚期融合，以进行分层且更精细的跨模态交互 [58]。

随着注意力机制的出现 [59], , 基于Transformer架构的多模态模型通过交错使用单模态表示与自注意力和跨模态注意力层，高效捕捉模态内和模态间动态 [25, 60]。然而，为了在多模态理解方面达到最先进的性能，先前的大多数多模态方法都假设训练和推理阶段的模态均成对且完整可观测。

A.2 缺失模态问题

当预训练多模态模型因缺少模态数据输入而在下游推理中出现严重性能下降时，就会产生缺失模态问题 [4, 61]。在真实世界中，由于传感器故障、硬件故障、隐私限制、环境干扰和数据传输问题等诸多原因，无法保证始终存在多模态输入 [1]。实证分析表明，多模态方法的效能关键取决于模态是否完整 [14], , 而基于Transformer的编码器对输入缺失尤其敏感 [2]。因此，解决缺失模态问题对于确保现实世界应用中的稳健性能至关重要，这一问题被称为不完整多模态学习。

从不完整多模态学习的另一种视角来看，缺失模态问题可视为多视图学习中特定视图信息缺失的问题 [62]。以往方法主要致力于对齐完整与不完整的多模态表示 [9, 63] , 或根据不完整多模态输入生成缺失信息 [17, 22, 24]。基于对齐的方法避免了显式生成的复杂性，但面对不确定且异质的信息不对称时，往往会导致推理效果不佳。相比之下，基于生成的方法仍容易受到单模态特征中的任务无关噪声和大量冗余信息的影响，这些因素可能轻易导致信息重建或生成过程不稳定。

此外，以往的方法需要针对特定的缺失场景进行调整，才能达到最佳性能，这限制了它们在现实世界不确定环境中的应用。此外，它们大多牺牲完整多模态学习的性能，以提高对模态缺失问题的鲁棒性。因此，如何实现一个能够联合优化完整与不完整多模态学习的单一统一框架，仍然是一项尚待解决的挑战。

A.3 基于信息瓶颈的深度学习

Tishby 等人最初提出的 [29], 信息瓶颈（IB）将表示学习表述为一种信息论方法，将任务相关表示定义为特征压缩与任务预测之间的权衡 [31]。由于表示学习具有出色的可解释性，且能够估计深度神经网络的泛化能力 [30, 64, 65],, 信息瓶颈在深度学习领域，尤其是在多视图问题中 [66–68], 持续受到关注。VIB [32] 引入变分近似，在信息流中生成瓶颈潜变量，其目标形式与用于生成的 VAE [69] 相似。DeepIMV [70] 利用乘积—

专家模型来将边缘特定视图表示整合为联合潜在表示。FactorCL [71] 将任务相关信息分解为共享信息和独有信息，并通过互信息定义多视图冗余。然而，它们既无法应用于不同的缺失情形，也无法实现最优的模态内和模态间信息提取性能。

在本论文中，我们采用信息瓶颈的原理，通过设计模态间特征的信息流以及语义引导，构建有效的信息潜在空间，从而为完整与不完整多模态学习中的性能提升铺平道路。

B 变分信息瓶颈的推导

在本节中，我们详细推导原始变分信息瓶颈以及所提出的词元级与标签级信息瓶颈的公式，如式 5、式 8和式11-12所示。此外，我们给出信息潜变量的可微采样过程，其呈现在式 6。

式 5 表示信息瓶颈损失 $\mathcal{L}_{ib}$，用于信息流，形式为（$S \to B \to T$），其中 S 表示源状态，B 表示瓶颈潜变量，T 表示目标状态。回顾可知，该损失首先由互信息的约束给出，其表述为：

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

其中 $\beta > 0$ 是用于平衡两项互信息的权衡参数。直接获得 $S/B/T$ 之间的互信息是不可实现的。因此，我们倾向于求得 $\mathcal{L}_{ib}(S,T)$ 的上界，从而将目标最小化问题转化为证据上界优化问题。

从第一项 $I(S;B)$ 开始，将其完整形式展开为：

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

考虑到瓶颈潜变量 $p(B) = \int dS p(B|S)p(S)$ 的边缘分布可能难以计算，令 $q(B)$ 为该边缘分布的变分近似。根据非负的 Kullback-Leibler（KL）散度 $KL(p(B) \parallel q(B)) \geq 0$，有 $\int dB\, p(B) \log p(B) \geq \int dB\, p(B) \log q(B)$，可推导出以下上界：

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

其中，$p_\theta(B|S)$ 可通过信息瓶颈编码器 E_S 学习得到：$S \mapsto B$。

然后，对于第二项 $I(B;T)$，以同样的完整形式表示为：

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

由于 $p(T|B)$ 难以计算，引入信息瓶颈解码器 D_T：$B \mapsto T$，令 $q_\phi(T|B)$ 为 $p(T|B)$ 的变分近似。类似地，根据 KL 散度 $KL(p(T|B) \parallel q_\phi(T|B)) \geq 0$，我们有 $\int dT\, p(T|B) \log p(T|B) \geq \int dT\, p(T|B) \log q_\phi(T|B)$。因此，$I(B;T)$ 的下界如下：

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

其中， $H(T)$ 是目标状态 T 的熵，仅由 T本身的分布决定，无论 T 是单模态令牌嵌入还是真实标签。因此， $H(T)$ 可以从损失函数中去掉。当源状态 S 与目标状态 T不同时，我们引入如下关系：给定 S 时， B 与 T 相互独立：

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

注意，当 $S = T$ 时，我们有 $p(T,B) = p(S,B) = \int p(S)p(B|S) = \int p(S)p(T|S)p(B|S)$，上述表达式仍然成立。此外， $p(T|S)$ 表示联合配对数据样本的已知分布。利用每个批次中的数据样本作为经验数据分布，我们有 $p(S,T) = p(S)p(T|S) = \frac{1}{N}\sum_{n=1}^{N}\delta_{S_n}(S)\delta_{T_n}(T)$。因此， $I(B;T)$ 通常可表示为：

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

结合 $I(S;B)$ 和 $I(B;T)$的上界，我们可以推导出 $\mathcal{L}_{ib}$ 的如下上界，见方程 5：

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

最后，变分信息瓶颈的优化目标可改写为：

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

现在，我们推导上述上界中最小化项和最大化项的实际表达式。

首先，为了最小化 $KL(p_\theta(B|S)\parallel q(B))$，在实践中，我们使用标准高斯分布$\mathcal{N}(0,\ 1)$作为 $q(B)$ 的先验分布，并将近似后验分布 $p_\theta(B|S)$ 建模为多元高斯分布 $\mathcal{N}(\mu,\ \varepsilon)$。其中， μ 和 ε 分别表示潜在高斯分布的均值向量和方差向量。由于 $q(B)$ 和 $p_\theta(B|S)$ 的每个维度相互独立，我们可以推导出 $KL(p_\theta(B|S)\parallel q(B))$ 的一维高斯分布情形，如下：

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

由于任意概率密度函数在其定义域上的积分均为 1，因此有 $\int\mathcal{N}(\mu_B,\ \sigma_B^2)\ db = 1$。根据高斯分布的方差和二阶原始矩的定义，有$\int\mathcal{N}(\mu_B,\ \sigma_B^2)(b-\mu_B)^2\,db = \sigma_B^2$ 以及 $\int\mathcal{N}(\mu_B,\ \sigma_B^2)$ $b^2\,db = \mu_B^2 + \sigma_B^2$。因此，解析解

$KL(\mathcal{N}(\mu_B,\ \sigma_B^2)\ \|\ \mathcal{N}(\mathbf{0},\ \mathbf{I}))$ 的计算结果为：

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

当多元高斯分布 $\mathcal{N}(\mu, \varepsilon)$ 中的每个维度 $d \in \{0, ..., C\}$ 相互独立时，我们有：

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

于是我们有：

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

其中， C 是瓶颈潜变量 B 的特征维度。

而为了最大化 $\mathbb{E}_{B\sim p(B|S)}\mathbb{E}_{S\sim p(S)}[\log\ q_\phi(T|B)]$,，我们根据特定目标状态 T 的属性，采用多种表达形式进行计算。本文将其分为词元级和标签级信息瓶颈。

B.1 词元级信息瓶颈的推导

对于作为目标状态 T 的单模态表示 $F_u = \{f_u\}|_{i=1}^{L} \in \mathbb{R}^{L\times C}$，我们采用信息瓶颈解码器 D_T： $B \mapsto T$，通过瓶颈桥 B 实现源表示与目标表示 F_S/F_T 之间的信息流。由于 F_S/F_T 的词元嵌入中各维度相互独立，因此我们有：

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

于是我们有：

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

为简化起见，假设每个维度的方差（σ_T^d）^{2}均保持一致并固定为常数 $d \in \{0, .., C\}$，信息瓶颈解码器只需输出投影词元嵌入的均值 $\mu^i = D_T$（b_S）。则有以下公式：

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

结合公式 36和公式 39，我们可以推导出公式 8作为 $S \rightarrow T$ 令牌级信息瓶颈的目标函数：

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

B.2 标签级信息瓶颈的推导

对于作为目标状态 T 的真实标签 $y_{gt} \in \mathbb{R}^1/\mathbb{R}^V$，根据监督信号是连续形式（回归值为 1）还是离散形式（V 为独热识别类别），我们采用两种形式的后验分布 $q_\phi(T|B)$。对于每个源模态 S，我们利用单模态预测器 P_S：$B_S \mapsto \hat{}y_S$ 来建模后验分布 $q_\phi(T|B)$，并基于单模态信息输出预测 $\hat{}y_S$。

对于回归任务中的连续标签 $y_{gt} \in \mathbb{R}^1$，我们将 $q_\phi(T|B)$ 视为一维高斯分布，计算如下：

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

接着，我们将后验分布 $q_\phi(T|B)$ 的最大化转化为：

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

Co结合公式 36 和公式 42，我们可以推导出公式 11 作为标签级信息瓶颈的目标函数 n bott颈：

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

在分类任务中，对于具有 V 个类别的独热离散标签 $y_{gt} \in \mathbb{R}^V$，我们将 $q_\phi(T|B)$ 视为多维伯努利分布，计算如下：

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

然后，我们将后验分布 $q_\phi(T|B)$ 的最大化转化为：

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

结合公式36 和公式45，我们可以推导出公式12，将其作为标签级信息瓶颈的目标：

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

B.3 信息瓶颈潜变量的重参数化技巧

由于信息瓶颈潜变量 B 是从 $p_\theta(B|S) \sim \mathcal{N}(\mu,\ \sigma^2)$ 中采样得到的，而采样过程不可微。为了通过梯度下降算法优化信息瓶颈编码器 E_S： $S \mapsto B$，我们需要确定 B 对编码器权重参数 θ 的梯度，根据链式法则表示如下：

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

其中， μ 和 σ 由编码器直接输出，因此 $\frac{\partial \mu}{\partial \theta}$ 和 $\frac{\partial \sigma^2}{\partial \theta}$ 易于求解。

为了根据采样得到的瓶颈潜变量 B计算 $\frac{\partial B}{\partial \mu}$ 和 $\frac{\partial B}{\partial \sigma^2}$，我们首先从非参数分布 $\mathcal{N}(\mathbf{0},\ \mathbf{I})$中采样一个随机向量$\mathbf{z}$，并利用变换函数 $g(\mathbf{z}\ \mu,\ \sigma^2)$得到瓶颈 B，如下所示：

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

由于从$\mathbf{z}\ \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$中采样的过程与参数 μ 和 σ无关，因此$\frac{\partial B}{\partial \mu}$和 $\frac{\partial B}{\partial \sigma^2}$的求导可转化为对 $\frac{\partial g_{(\cdot)}}{\partial \mu}$和 $\frac{\partial g_{(\cdot)}}{\partial \sigma^2}$的求导，而后者是可微的。

C 信息瓶颈与循环翻译相结合的理论基础

我们进一步阐释 CyIN 将信息瓶颈（IB）与跨模态循环翻译相结合的理论分析。除任务预测项外，式23 中信息瓶颈损失和循环翻译的优化目标推导如下：

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

在不失一般性的前提下，考虑具有两种模态的多模态学习，CyIN的理论建模可表述如下：

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

其中 S 表示源状态， T 表示目标状态。

• 带有 $S_1 \rightarrow B_1 \rightarrow T_1$ 和 $S_2 \rightarrow B_2 \rightarrow T_2$ 的左右两侧表示信息瓶颈链。请注意，源状态与目标状态在token级信息瓶颈中具有 $S/T \in \{F_{S_1}, F_{S_2}\}\{F_{T_1}, F_{T_2}\}$，或在标签级信息瓶颈中具有 $S \in \{F_{S_1}, F_{S_2}\}, T \in y$，二者均已在上文中得到理论推导，表述如下：

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

• 中间部分 $B_1 \rightleftharpoons B_2$ 表示使用 $B_1 \rightharpoonup B_{2\rightarrow 1}^{rec} \rightharpoonup B_1^{cyc} \sim B_1$ 和 $B_2 \rightharpoonup B_{1\rightarrow 2}^{rec} \rightharpoonup B_2^{cyc} \sim B_1$ 进行循环翻译，其中 $\rightharpoonup$ 表示翻译过程，而 $\sim$ 表示与原始瓶颈的对齐。将 Γ 视为翻译器网络，翻译过程可分为跨不同单模态潜变量的跨模态重构 B_1/B_2，以及从重构潜变量 $B_{1\rightarrow 2}^{rec}/B_{2\rightarrow 1}^{rec} = \Gamma_{1\rightarrow 2}(B_1)/\Gamma_{2\rightarrow 1}(B_2)$返回原始$B_2^{cyc}/B_1^{cyc} = \Gamma_{1\rightarrow 2}(B_{2\rightarrow 1}^{rec})/\Gamma_{2\rightarrow 1}(B_{1\rightarrow 2}^{rec})$的循环翻译，其形式如下：

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

通过最小化上述损失，我们可以为每种模态u获得信息丰富的单模态瓶颈潜变量 B_u，其中包含可用于有效多模态融合的模态间特征。与直接融合单模态表示 F_u 相比，在瓶颈潜变量 B_u 上实施的融合过程（记为 $F_M = D_M(B_1, B_2)$）能够利用信息瓶颈的压缩能力，从而降低跨模态转换中的重构难度。
当模态 u_1 与 u_2 同时出现时，即完整多模态学习，我们可以将循环翻译 $B_1 \rightleftharpoons B_2$ 视为一种跨模态交互，以增强模态共享动态的提取。此类跨模态协同效应有助于多模态融合过程 $F_M = D_M(B_1, B_2)$ 获得最终的判别性表征。

当其中一种模态缺失时，该框架需要进行不完整多模态学习。假设 u_1 缺失，我们可以从模态 u_2 获得瓶颈潜变量 B_2，以重构瓶颈潜变量 B_1^{rec}，记为 $B_2 \rightharpoonup B_1^{rec}$，并可进一步将其解码为多模态融合过程 $F_M = D_M(B_1^{rec}, B_2)$ 的补充信息。

上述双模态场景可以不受约束地推广到任意模态对，从而便于将CyIN高效扩展到涉及多种模态的任务。

D 超参数设置

由于缺失模态会在不同缺失场景下导致显著的性能波动，为确保各方法之间公平且一致的比较，我们使用固定的10个随机种子对每个实验进行十次评估。我们采用AdamW [72]作为训练优化器。

CyIN 的数据集划分和超参数设置见表6。

表 6：数据集划分及相应的超参数设置。

多模态任务	回归		分类	
数据集	MOSI	MOSEI	IEMOCAP	MELD
训练集	1,284	16,326	5,228	9,765
验证集	229	1,871	519	1,102
Test	686	4,659	1,622	2,524
训练轮数	50	30	50	50
批次大小	128	128	256	128
语言模型学习率	4e-5	1e-5	5e-6	3e-5
其他参数的学习率	1e-3	5e-4	1e-4	5e-4
权重衰减	1e-2	1e-5	1e-3	1e-4
单模态表示的维度 C_U	256	64	64	32
信息瓶颈编码器或解码器中的维度 C_{ib}	256	64	256	64
维度 C_B 瓶颈潜变量 B的	128	256	128	32
CRA中 $RA(\cdot)$的层数	8	16	8	4
CRA中 $RA(\cdot)$的维度	[64, 32, 16]	[128, 64, 32]	[128, 64, 32]	[128, 64, 32]
跨模态注意力的层数	2	8	2	4
注意力头 H的数量	8	2	4	8
β	16	32	4	4
γ	10	5	10	5
1^{st}:2^{nd} 训练阶段	1:9	3:7	3:7	1:4

E 关于评估协议的详细信息

在（1）完整多模态输入情况下的评估与其他最先进的基线方法保持一致，参见 $u \in \{l, a, v\}$，其中 l、a 和 v 分别表示语言、声学和视觉模态。

以下 [8, 17, 22, 24], 为评估模态缺失场景下的模型鲁棒性，我们采用两种缺失数据协议：(2) 固定缺失协议和 (3) 随机缺失协议。

(2) 在固定缺失协议中，每次迭代都会移除相同的模态子集。具体而言，我们会舍弃一个模态（例如 $u \in \{l,a\}/\{l,v\}/\{a,v\}$）或两个模态（例如 $u \in \{l\}/\{a\}/\{v\}$）。

(3) 相比之下，随机缺失协议通过在每个批次中为每个样本随机选择缺失模态组合来模拟现实场景，其中可能缺失一个或两个模态。按照 [22, 24], 随机缺失率（MR）定义如下：

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

其中，$|u_i|$ 表示第 i 个样本的可用模态数量，U 表示模态总数，N 表示样本总数。考虑到实际应用场景，每个样本至少保证有一个可用模态，从而确保 $|u_i| \in \{1, ..., U\}$，因此 $MR \leq \frac{U-1}{U}$。在包含 $U = 3$ 个模态的多模态回归和分类任务中，我们考虑缺失率 $MR \in \{0.0, 0.1, \ldots, 0.7\}$，其中 $MR = 0.7$近似表示可容忍的最大缺失程度。值得注意的是，$MR = 0.0$ 表示完整的多模态场景，即所有模态（$u \in l,\ a,\ v$）均存在。

F 基线方法

基线方法通过开源代码复现。按照 [27], , 我们对每个模型进行 50 次随机网格搜索，以找到最佳超参数。基线方法的描述如下。

CCA [11] 通过最大化配对模态之间的典型相关，将它们投影到共享低维空间中，为不完整数据提供一种经典的线性插补方法。

DCCA [12] 以深度神经编码器取代线性投影，从而实现模态之间的非线性对应关系。

DCCAE [13] 在 DCCA 的基础上加入基于自编码器的重建目标 [15], , 在学习最大典型相关的同时保留每种模态的内部结构。

CRA [34] 采用级联残差自编码器，利用部分输入逐步优化从完整输入中获得的表示的重构。

MCTN [16] 利用编码器—解码器循环神经网络，以循环翻译的方式在源模态和目标模态之间进行转换。

MMIN [17]通过级联残差架构扩展了循环一致性学习，根据观测视图想象缺失信息。

CPM-Net [63]利用部分多视图聚类来解决不完整多模态学习问题，将所有视图嵌入结构化潜在空间，并通过数据传递和聚类感知分类推断缺失特征。

GCNet [22] 引入了两个互补的图神经网络，对对话数据中的时间关系和说话人关系进行建模，然后根据学习到的图表示重建缺失特征。

IMDer [24] 采用基于分数的扩散模型，在迭 e
diff代扩散和去噪过程中学习缺失模态的分布，并以其余模态作为正则化依据。

LNLN [8] 采用主导模态校正模块，以确保主导模态表示的质量，并开展基于主导模态的多模态学习，从而抵御输入噪声。

G 评估指标

多模态回归和分类任务的评估指标公式及计算方法如下所示。

对于 MOSI 和 MOSEI 上的多模态回归任务，我们采用具有代表性的指标，包括**Acc7、F1 分数、平均绝对误差和相关系数**，来评估模型的性能 [73]:

Acc7（七分类准确率）用于衡量在 $[-3,+3]$ 的七个整数标签中预测正确的比例，其公式如下：

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

其中， N 是样本总数，$\hat{y}_n$ 是预测标签， y_n 是真实标签， $\lfloor\cdot\rceil$ 表示四舍五入到最近整数， $\mathbb{I}(\cdot)$是指示函数。给定 $k \in \{-3,-2,-1,0,1,2,3\}$，$\{n : y_n = k\}$ 表示真实标签为 k的样本集合； $\{n : y_n = k \wedge \hat{y}_n = k\}$ 是被正确预测为 k的样本子集。

F1（二元F1分数）通过将回归标签归并为两个类别（负类/正类），评估正负类的区分能力，其定义如下：

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

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

其中， C是类别数， w_c是类别 c的权重（通常为 $w_c = N_c/N$），而 TP_c、FP_c和 FN_c分别表示类别 c的真正例、假正例和假负例数量。

平均绝对误差（平均绝对误差）用于衡量预测分数与真实分数之间的平均绝对偏差，其定义如下：

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

Corr （皮尔逊相关系数）衡量预测值与标签之间的线性一致性，表示模型输出中的系统性偏差或偏斜，公式如下：

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

其中，$\bar{\hat{y}}$ 和 $\bar{y}$ 分别为预测分数和真实标签的均值。

对于 IEMOCAP 和 MELD 上的多模态分类任务，我们报告**准确率**和**加权 F1 分数**，以平衡各类别分数的权重 [74]。

准确率（二元准确率）表示正确预测的样本数占样本总数的未加权比例，公式如下：

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

wF1（加权 F1 分数）根据真实实例对加权精确率和召回率进行平衡 r
eac每个类别，分类中的加权 F1 分数公式与公式 54相同。

H 更多实验结果

H.1 各种模态缺失场景下的详细结果

采用固定缺失协议和随机缺失协议时，各种模态缺失场景下的详细实验结果分别列于表 7 和表 8中。需要注意的是，所报告的结果是基于 10 个固定随机种子的平均值。

在包含不完整设置的固定缺失协议下， $u \in \{l\}/\{v\}/\{a\}/\{l,a\}/\{l,v\}/\{a,v\}$，CyIN 在几乎所有设置中都展现出优异或具有竞争力的性能，在单模态和双模态缺失场景下均显著优于其他模型。从表 7中可以观察到，由于

Table 7：所提出的 CyIN 与基线方法在固定缺失例议包括不完整设置）$u \in \{l\}/\{v\}/\{a\}/\{l,a\}/\{l,v\}/\{a,v\}$。

固定 u	模型	MOSI				MOSEI				IEMOCAP		MELD	
		准确率7↑	F1↑	MAE↓	相关系数↑	准确率7↑	F1↑	MAE↓	相关系数↑	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**

表 8：所提出的 CyIN 与基线方法在随机缺失协议下的性能比较，缺失率为 $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$。

MR	模型	MOSI				MOSEI				IEMOCAP		MELD	
		准确率7↑	F1↑	MAE↓	相关系数↑	准确率7↑	F1↑	MAE↓	相关系数↑	Acc↑	wF1↑	Acc↑	wF1↑
0.1	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
	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
0.2	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
	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**
0.3	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
	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**
0.4	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
	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**
0.5	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
	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**
0.6	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
	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**
0.7	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
	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**

话语中丰富的语义和大量识别相关信息 [16, 43, 44]。尽管该主导模态的表现更好，CyIN 仍更加关注音频和视觉等次要模态。这一结果体现了所提出框架的泛化能力，因为 CyIN 并未像 LNLN [8]那样专门设计额外的精细模块，将训练重点放在主导模态上。此外，在一个模态缺失、以两个模态作为输入时，CyIN 能够充分整合双模态交互中的信息并取得更高的性能，验证了信息瓶颈潜变量在各种模态组合下的灵活性。

在缺失率为 $MR \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$的随机缺失协议下，随着缺失率上升，CyIN 的性能下降幅度较小。随机缺失情形主要用于模拟真实世界中多样的缺失状况，此时模态是否存在并不固定或已知。如表8所示，与基线方法相比，CyIN 在各种输入数据稀疏性下的鲁棒性凸显了通过令牌级和标签级信息瓶颈提取的信息瓶颈的优势，即使多模态输入在 $MR = 0.7$处严重缺失，也能确保性能稳定。此外，无论是多模态回归还是分类任务，CyIN 在 4 个数据集上均取得了最先进的性能。由于泛化能力有限，大多数基线方法在其中一项任务上的表现都不理想。

各种缺失场景下的实验结果清楚地表明，CyIN 在固定或随机缺失场景中都非常有效。通用框架使信息空间即使在输入模态缺失的情况下，也能保持强劲的性能，因此是适用于真实世界多模态应用的稳健且实用的解决方案。

H.2 性能稳定性比较

表 9：所提出的 CyIN 与基线方法的性能稳定性比较，展示了在 MOSI 和 IEMOCAP 数据集上的平均结果；其中固定缺失协议包括模态设置$u \in \{l\}/\{a, v\}$，随机缺失协议包括缺失率 $MR = 0.7$。

设置	模型	MOSI				IEMOCAP	
		准确率7↑	F1↑	MAE↓	相关系数↑	Acc↑	wF1↑
固定 $u \in \{l\}$	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
固定 $u \in \{a, v\}$	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
随机缺失 $MR = 0.7$	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

我们评估了模型在 MOSI 数据集的三种代表性场景下的性能稳定性，包括固定缺失协议 $u \in \{l\}/\{a, v\}$ 和随机缺失协议$MR = 0.7$。根据表 7 和表 8中报告的总体平均性能，我们进一步计算了每种缺失场景下相应结果在 10 次运行中的标准差，以展示现实场景中的性能变化程度。最终的稳定性指标见表 9。

在 $u \in \{l\}/\{a, v\}$ 等固定模态缺失协议下，所有方法的标准差都相对较小，这表明当缺失模态预先确定时，模型能够以更确定的方式学习如何进行补偿。相比之下，随机缺失协议会带来更大的方差，在 $MR = 0.7$ 这样的严重缺失场景下尤其如此。我们将其归纳为两个原因：首先是模态贡献不平衡（例如，在特定任务中，语言等主导模态所包含的语义比音频或视觉等较弱模态更多）；

其次，随机缺失设置更接近于模拟真实世界中不可预测的推理情形，从而暴露出各模型的真实鲁棒性。

在固定和随机缺失场景下，CyIN 不仅在大多数指标上取得了更优或具有竞争力的性能，而且波动相对较小，凸显了所构建信息空间在性能与稳定性之间的出色平衡。

H.3 不同语言模型上的泛化性能

为进一步验证泛化性能，我们使用不同规模的预训练语言模型训练并评估 CyIN，包括 BERT [40], RoBERTa [75], 和 DeBERTa-V3 [76]，如10所示。实验结果表明，扩大预训练语言模型骨干的规模显然有益：较大的模型（如 DeBERTa-V3）能够生成更优的文本表示，因此在所有模态均可用时具有更高的多模态性能。在最严重的缺失情形下进行评估时，也呈现出类似趋势 $MR = 0.7$，这表明 CyIN 具有有效的泛化能力。

表10：在完整和随机不完整多模态学习设置下，比较所提出的 CyIN 使用不同规模语言模型在 MOSI 数据集上的表现。

设置	模型变体（PLM）	MOSI			
		准确率7↑	F1↑	MAE↓	相关系数↑
完整 $u \in \{l, a, v\}$	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**
随机缺失 $MR = 0.7$	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 其他多模态任务的各种尺度

我们在表 11 -14中扩展了 CyIN 在 6 个具有 2–4 种模态和随机缺失率的非情感数据集上的实验，涵盖以下任务：**多模态推荐（3 种模态：视觉、音频、文本）** [55]（AmazonBaby、Tkitok、Allrecipes 数据集），**多模态人脸反欺骗（3 种模态：RGB、深度、红外）** [77]（CASIA-SURF 数据集），**多模态密集预测（2 种模态：RGB、深度）** [77]（NYUv2 数据集），以及**多模态医学分割（4 种模态：液体衰减反转恢复、T1、T1c、T2）** [56]（BraTS2023 数据集）。

表11：三个数据集（包括 Amazon Baby、Tiktok 和 Allrecipes 数据集）上的**多模态推荐**任务准确率和公平性表现（%）对比实验。模态设置采用随机缺失协议，缺失率为 0.4，与最先进的 [55] 论文一致。

数据集	模型	召回率 ↑		精确率 ↑		归一化折损累计增益 ↑		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	最先进 [55]	5.26	8.54	0.56	0.45	2.76	3.60	87.24	90.12	1.11	0.90
	CyIN（我们的方法）	**5.43**	**8.56**	**0.57**	0.45	**2.89**	**3.68**	**90.10**	**92.70**	**1.14**	0.90
TikTok	最先进 [55]	4.81	7.39	0.48	0.37	2.95	3.60	88.15	92.27	0.96	0.74
	CyIN（我们的方法）	**4.91**	**7.57**	**0.49**	**0.38**	**3.15**	**3.81**	**89.02**	**92.31**	**0.98**	**0.75**
Allrecipes	最先进 [55]	2.49	3.36	0.24	0.16	1.33	1.55	96.82	89.52	0.49	0.33
	CyIN（我们的方法）	**2.61**	**3.43**	**0.26**	**0.17**	**1.45**	**1.66**	**99.57**	**91.94**	**0.52**	**0.34**

实验结果表明，所提出的 CyIN 达到了与最先进（SOTA）方法相当的性能，进一步证明了 CyIN 在适应不同模态数量或多模态融合任务方面的有效性。此外，数据集的规模从 1,449 个样本（NYU v2）到 65,671 个样本（TikTok）、87k 个样本（CASIA-SURF）和 139,110 个样本（Amazon Baby）不等，这在一定程度上展现了所提出 CyIN 的规模性能。由于计算资源有限，我们将对更大规模数据集的探索留待未来工作。

表12：实验对比：**多模态人脸反欺骗**任务，平均分类错误率（ACER）（↓）指标，数据集为CASIA-SURF。模态设置按固定缺失协议设定，与最先进 [77]论文相同。

模态设置 RGB	深度	红外	最先进 [77]	**CyIN（本文）**
✓			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**
平均			3.84	**3.18**

表13：在 NYUv2 数据集上，使用平均交并比（↑）指标对**多模态密集预测**任务进行实验比较。模态设置采用与最先进 [77] 论文相同的固定缺失协议。

模态设置 RGB	深度	最先进 [77]	**CyIN（我们的方法）**
✓		**44.06**	43.93
	✓	41.82	**43.84**
✓	✓	49.89	**50.46**
平均值		45.26	**46.07**

表14：在 BraTS2023 数据集上，使用 Dice相似系数（DSC）%（↑）指标进行的**多模态医学分割**任务实验比较。结果为在固定模态设置下进行的平均结果，与原始最先进 [56] 论文中的设置一致。

类别	最先进[3]	**CyIN（我们的方法）**
增强肿瘤	74.6	**75.1**
肿瘤核心	85.0	**85.8**
全肿瘤	90.6	**91.2**

I 社会影响

致力于连接完整与不完整多模态学习，所提出的 CyIN 在医疗保健、教育、社交媒体分析、营销、广告和人机交互等真实世界领域展现出巨大潜力。通过整合文本、音频、图像和视频等多个来源的数据，CyIN 能够让多模态系统更细致地理解人类情绪和意图。

然而，使用丰富的多源数据会带来与隐私侵犯、未经授权的监控以及潜在的操纵滥用相关的风险。在缺失模态的场景中，重建数据的努力可能引入偏见、编造虚假信息，尤其会影响代表性不足群体并引发公平性方面的担忧。这些担忧凸显了在现实世界应用中谨慎、负责任地部署提出的方法的必要性。

神经信息处理系统大会论文清单

1. **主张**

   问题：摘要和引言中提出的主要主张是否准确反映了论文的贡献与范围?

   回答： [是]

   理由：摘要和引言中提出的主要主张清晰地反映了论文的贡献和范围。它们准确陈述了所提出的 CyIN 框架在联合处理完整与不完整多模态学习方面的设计和能力。

   指南：

   • 回答“不适用”意味着摘要和引言中未包含论文提出的主张。

   • 摘要和/或引言应明确阐述所提出的主张，包括论文的贡献以及重要的假设和局限性。对于此问题回答“否”或“不适用”，审稿人不会留下良好印象。

   • 所提出的主张应与理论和实验结果相符，并反映出预期这些结果在多大程度上可以推广到其他场景。

   • 将期望实现的目标作为动机的一部分是可以的，但前提是要明确这些目标并未由论文实现。

2. **局限性**

   问题：论文是否讨论了作者所开展工作的局限性?

   回答： [是]

   理由：局限性已在局限性章节中讨论。

   指南：

   • 答案“不适用”表示论文没有局限性，而答案“否”表示论文存在局限性，但未在论文中讨论这些局限性。

   • 建议作者在论文中单独设立“局限性”章节。

   • 论文应指出任何强假设，并说明结果对违反这些假设的情况有多稳健（例如，独立性假设、无噪声设置、模型设定正确，以及仅在局部成立的渐近近似）。作者应思考这些假设在实践中可能如何被违反，以及这会带来什么影响。

   • 作者应思考所提出主张的适用范围，例如，该方法是否仅在少数几个数据集上或通过少数几次运行进行测试。一般而言，实证结果往往依赖于隐含假设，作者应明确阐述这些假设。

   • 作者应反思影响该方法性能的因素。例如，人脸识别算法在图像分辨率较低或图像在弱光条件下拍摄时可能表现不佳。或者，语音转文字系统可能无法可靠地为在线讲座提供隐藏式字幕，因为它无法处理专业术语。

   • 作者应讨论所提出算法的计算效率，以及它们如何随数据集规模扩展。

   • 如适用，作者应讨论其方法在解决隐私和公平性问题方面可能存在的局限性。

   • 尽管作者可能担心，对局限性的完整坦诚会被审稿人用作拒稿理由，但更糟的结果可能是审稿人发现论文中未予承认的局限性。作者应运用最佳判断，并认识到支持透明度的个人行动在形成维护社区诚信的规范方面发挥着重要作用。审稿人将受到明确指示，不得因对局限性的诚实陈述而予以惩罚。

3. **理论假设与证明**

   问题：对于每个理论结果，论文是否提供了完整的假设集合以及完整（且正确）的证明?

回答： [是]
理由：论文已在附录中为所提及的方程提供了完整且正确的推导证明B。

指南：
- 答案“不适用”表示论文不包含理论结果。
- 论文中的所有定理、公式和证明都应编号并交叉引用。
- 所有假设都应清楚地陈述，或在任何定理的陈述中注明出处。
- 证明可以放在论文正文或补充材料中；如果放在补充材料中，建议作者提供简短的证明概要，以帮助读者建立直观理解。
- 反之，论文正文中提供的任何非正式证明都应由附录或补充材料中的正式证明加以补充。
- 证明所依据的定理和引理应当得到恰当引用。

4. **实验结果可复现性**

问题：论文是否充分披露了复现其主要实验结果所需的全部信息，以便评估这些结果对论文主要主张和/或结论的影响（无论是否提供代码和数据）？

回答： [是]
理由：论文在附录 D 中提供了可复现性所需超参数设置的详细信息。此外，正如摘要中所述，源代码已公开发布。

指南：
- 回答“不适用”表示该论文不包含实验。
- 如果论文包含实验，对本问题回答“否”将不会得到审稿人的好评：使论文可复现很重要，无论是否提供代码和数据。
- 如果贡献是数据集和/或模型，作者应描述为使其结果可复现或可验证所采取的步骤。
- 根据贡献的不同，可复现性可以通过多种方式实现。例如，如果贡献是一种新颖架构，充分描述该架构可能就足够了；如果贡献是特定模型及实证评估，则可能需要使他人能够使用相同数据集复现该模型，或提供对该模型的访问。一般而言，发布代码和数据通常是实现这一点的一种好方法，但也可通过详细说明如何复现结果、提供对托管模型的访问（例如大语言模型的情况）、发布模型检查点，或其他适合所开展研究的方式来提供可复现性。
- 虽然神经信息处理系统大会不要求发布代码，但会议要求所有投稿 - 都提供合理的可复现途径，具体方式可能取决于贡献的性质。例如

（a）如果贡献主要是一种新算法，论文应明确说明如何复现该算法。（b）如果贡献主要是一种新的模型架构，论文应清晰、完整地描述该架构。（c）如果贡献是一个新模型（例如大语言模型），则应提供一种访问该模型以复现结果的方式，或提供一种复现该模型的方式（例如，使用开源数据集，或提供如何构建数据集的说明）。（d）我们认识到，在某些情况下可复现性可能较为棘手，此时欢迎作者说明他们为实现可复现性所提供的具体方式。对于闭源模型，对模型的访问可能在某种程度上受到限制（例如仅限注册用户），但其他研究人员仍应有某种途径来复现或验证结果。

5. **开放获取数据和代码**

问题：论文是否提供对数据和代码的开放获取，并附有充分说明，以便按照补充材料中的描述忠实地复现主要实验结果?

回答： [是]

理由：本文使用的数据集可公开访问，也可应要求提供。此外，正如摘要中所述，源代码现已公开发布，从而确保透明度和可复现性。

指南：

- 答案“不适用”表示论文不包含需要代码的实验。
- 有关更多详情，请参阅神经信息处理系统大会的代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）。
- 虽然我们鼓励发布代码和数据，但我们理解这可能无法实现，因此“否”是可接受的答案。论文不能仅仅因为未包含代码而被拒稿，除非代码是贡献的核心（例如，针对新的开源基准）。
- 说明应包含复现结果所需运行的确切命令和环境。请参阅神经信息处理系统大会代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）以了解更多详情。
- 作者应提供关于数据访问和准备的说明，包括如何访问原始数据、预处理数据、中间数据和生成数据等。
- 作者应提供脚本，以复现新提出的方法及基线方法的全部实验结果。如果仅有一部分实验可复现，则应说明脚本中省略了哪些实验以及原因。
- 在投稿时，为保持匿名性，作者应发布匿名版本（如适用）。
- 建议在补充材料（附于论文之后）中尽可能提供全面的信息，但允许附上数据和代码的网址。

6. **实验设置/细节**

问题：论文是否说明了理解结果所需的所有训练和测试细节（例如，数据划分、超参数及其选择方式、优化器类型等）？

回答： [是]

理由：附录 D中说明了所有训练和测试细节，包括数据集划分、超参数、优化器类型等。此外，按照以往方法，每个模型的最佳超参数均通过进行五十次随机网格搜索选出。

指南：

- NA 表示论文不包含实验。
- 实验设置应在论文主体中以足够详细的程度呈现，以便读者理解结果并正确解读。
- 完整的细节可以随代码、在附录中或作为补充材料提供。

7. **实验的统计显著性**

问题：论文是否以适当且正确定义的方式报告误差条，或提供有关实验统计显著性的其他适当信息?

回答： [是]

理由：如附录 H.2所示，我们报告了性能变化，以证明提出的方法在各种缺失场景中的优越性。

指南：

- 答案 不适用表示该论文不包含实验 实验。

- 如果结果附有误差条、置信区间或统计显著性检验，作者应回答“是”，至少对于支持论文主要主张的实验如此。
- 误差条所反映的变异因素应予以明确说明（例如，训练/测试集划分、初始化、某些参数的随机抽取，或在给定实验条件下的整体运行）。
- 计算误差条的方法应予以说明（闭式公式、调用库函数、自助法等）。
- 所做的假设应予以给出（例如，正态分布误差）。
- 应明确误差条是标准差还是均值标准误。
- 报告1σ误差条是可以的，但应予以说明。若误差正态性假设未经验证，作者最好报告2σ误差条，而不是声称其具有96%置信区间。
- 对于非对称分布，作者应注意不要在表格或图中展示对称误差条，以免产生超出范围的结果（例如负的错误率）。
- 如果表格或图中报告了误差条，作者应在文本中说明其计算方法，并在文本中引用相应的图或表。

8. **实验计算资源**

   问题：对于每项实验，论文是否提供了足够的信息，说明复现实验所需的计算资源（计算工作节点的类型、内存、执行时间）？

   回答： [是]

   理由：我们在第 4 节的实现细节中报告了所需的计算资源。

   指南：
   - 答案“不适用”表示论文未包含实验。
   - 论文应说明计算工作节点的类型（CPU 或 GPU）、内部集群或云服务提供商，并列出相关的内存和存储信息。
   - 论文应提供每次单独实验运行所需的计算量，并估算总计算量。
   - 论文应说明完整研究项目所需的计算量是否超过了论文中报告的实验所需的计算量（例如，未纳入论文的预备实验或失败的实验）。

9. **伦理准则**

   问题：论文中的研究是否在各个方面都符合神经信息处理系统大会道德准则 https://neurips.cc/public/EthicsGuidelines?

   回答： [是]

   理由：本研究符合《神经信息处理系统大会道德准则》中的所有道德和伦理规范要求。

   指南：
   - 回答 NA 表示作者尚未审阅《神经信息处理系统大会伦理准则》。
   - 如果作者回答“否”，应说明需要偏离伦理准则的特殊情况。
   - 作者应确保匿名性（例如，如果其司法管辖区的法律或法规要求特别考虑）。

10. **更广泛的影响**

    问题：论文是否讨论了所开展工作潜在的积极的社会影响和消极的社会影响?

    回答： [是]

    理由：我们已在附录中讨论了正面和负面的社会影响I。

指南：

- 答案“不适用”表示所开展的工作没有社会影响。
- 如果作者回答“不适用”或“否”，则应说明其工作为何没有社会影响，或论文为何未涉及社会影响。
- 消极社会影响的例子包括潜在的恶意或非预期用途（例如传播虚假信息、生成虚假个人资料、监控）、公平性考量（例如部署可能对特定群体造成不公平决策的技术）、隐私考量和安全性考量。
- 会议预计，许多论文属于基础研究，与特定应用无关，更不用说实际部署了。不过，如果存在通向任何负面应用的直接路径，作者应当指出。例如，指出生成模型质量的提升可被用于生成深度伪造内容以传播虚假信息，是合理的。另一方面，无需指出用于优化神经网络的通用算法可能使人们能够更快地训练生成深度伪造内容的模型。
- 作者应考虑以下可能造成的危害：技术按预期使用且正常运行时可能产生的危害；技术按预期使用但给出错误结果时可能产生的危害；以及因（有意或无意）滥用技术而造成的危害。
- 如果存在消极的社会影响，作者还可以讨论可能的缓解策略（例如，受控发布模型、除了攻击之外还提供防御措施、监测滥用的机制、监测系统如何随时间从反馈中学习的机制，以及提高机器学习的效率和可及性）。

11. **防护措施**

问题：论文是否描述了为负责任地发布具有较高滥用风险的数据或模型而采取的防护措施（例如，预训练语言模型、图像生成器或抓取数据集）？

回答： [不适用]

理由：该论文使用了公开的数据集和预训练模型，不会带来此类风险。

指南：

- 答案“不适用”表示该论文不会造成此类风险。
- 发布具有较高滥用或双重用途风险的模型时，应配备必要的防护措施，以便对模型的使用进行管控，例如要求用户遵守使用指南，限制模型的访问权限，或实施安全过滤器。
- 从互联网抓取的数据集可能带来安全风险。作者应说明如何避免发布不安全图像。
- 我们认识到，提供有效的防护措施颇具挑战，而且许多论文并无此要求，但我们鼓励作者将这一点纳入考虑，并尽最大努力妥善处理。

12. **现有资产的许可证**

问题：论文中使用的资产（例如代码、数据、模型）的创作者或原始所有者是否得到了适当署名？许可证和使用条款是否已明确说明并得到妥善遵守？

回答： [是]

理由：所使用的数据已在论文中正确引用，并且许可证和使用条款均得到妥善遵守。

指南：

- 回答“不适用”表示该论文未使用现有资产。
- 作者应引用产生该代码包或数据集的原始论文。
- 作者应说明所使用的资产版本，并在可能的情况下提供网址。

- 每项资产都应注明许可证名称（例如知识共享署名 4.0）。
- 对于从特定来源（例如网站）爬取的数据，应提供该来源的版权信息和服务条款。
- 如果发布资产，应提供软件包中的许可证、版权信息和使用条款。对于热门数据集，paperswithcode.com/datasets为部分数据集整理了许可证信息。其许可指南有助于确定数据集的许可证。
- 对于重新打包的现有数据集，应提供原始许可证以及衍生资产的许可证（如有变更）。
- 如果无法在线获取这些信息，建议作者联系资产的创建者。

13. **新资产**

    问题：论文中引入的新资产是否有完善的文档说明，且是否随资产一并提供文档?

    回答： [是]

    理由：如摘要所示，源代码已公开发布。

    指南：
    - 答案“不适用”表示论文未发布任何新资产。
    - 研究人员应通过结构化模板，在投稿中说明数据集/代码/模型的详细信息。这包括训练、许可证、局限性等方面的详情。
    - 论文应讨论是否以及如何获得了使用其资产的相关人员的同意。
    - 提交时，请记得将资产匿名化（如适用）。您可以创建匿名网址，或附上匿名化的 ZIP 文件。

14. **众包和涉及人类受试者的研究**

    问题：对于众包实验和涉及人类受试者的研究，论文是否包含提供给参与者的完整说明文本（如适用，还包括屏幕截图），以及报酬（如有）的详细信息?

    回答： [不适用]

    理由：论文不涉及众包或涉及人类受试者的研究。

    指南：
    - 答案为“不适用”表示论文不涉及众包，也不涉及人类受试者研究。
    - 将这些信息纳入补充材料是可以的，但如果论文的主要贡献涉及人类受试者，则应尽可能在论文正文中提供详细信息。
    - 根据神经信息处理系统大会的《行为准则》，参与数据收集、整理或其他劳动的工作人员，其薪酬至少应达到数据收集者所在国家的最低工资标准。

15. **涉及人类受试者的研究所需的机构审查委员会（IRB）批准或同等审批 an 受试者**

    问题：论文是否说明了研究参与者可能面临的风险、是否向受试者披露了这些风险，以及是否获得了机构审查委员会（IRB）的批准（或根据所在国家或机构的要求取得了同等的批准/审查）？

    答案： [不适用]

    理由：这篇论文不涉及众包，也不涉及涉及人类受试者的研究。

    指南：
    - 答案“不适用”表示该论文不涉及众包，也不涉及涉及人类受试者的研究。

- 根据开展研究所在的国家，任何人体受试者研究可能都需要获得机构审查委员会批准（或同等批准）。如果您已获得机构审查委员会批准，应在论文中明确说明。
- 我们认识到，相关程序可能因机构和地点而有显著差异，并期望作者遵守 NeurIPS 道德准则及其所在机构的指南。
- 对于初次投稿，请勿包含任何会破坏匿名性（如适用）的信息，例如开展评审的机构。

16. **大语言模型使用声明**

问题：如果大语言模型是本研究核心方法中的重要、原创或非标准组成部分，论文是否描述了大语言模型的使用情况？请注意，如果大语言模型仅用于写作、编辑或格式排版，且不影响研究的核心方法论、科学严谨性或原创性，则无需声明。

回答： [不适用]

理由：本研究中核心方法的开发并未将大语言模型作为任何重要、原创或非标准组成部分。

指南：

- 答案“不适用”表示本研究的核心方法开发并未将大语言模型作为任何重要、原创或非标准的组成部分。
- 有关哪些内容应予以描述、哪些内容不应描述，请参阅我们的大语言模型政策（https://neurips.cc/Conferences/2025/LLM）。