arXiv:2611.32860v1 [cs.MM] 24 Nov 2026

MSRGF：面向不完整多模态情感分析的参考引导融合的多源重建

Yan Zhuang[1,2] Minhao Liu[1] Yanru Zhang[1] Li Ma[1]Weilong Chen[1] Xingyu Wu[2] Jiawen Deng[1*] Fuji Ren[1*]1电子科技大学2香港理工大学

{202211081370,202421080935}@std.uestc.edu.cn, xingy.wu@polyu.edu.hk{minhaoliu,yanruzhang,weilongchen,dengjw,renfuji}@uestc.edu.cn

摘要

不完整多模态情感分析（*MSA*）在部分观测缺失时，根据文本、声学和视觉特征进行情感预测。现有重建方法根据可用模态估计缺失特征，并将所得信息结合起来用于情感预测。然而，不同的可用模态可能对同一模态产生不同的估计，而直接融合往往未能充分利用这些估计之间的关系。我们提出***MSRGF***（***Multi-SourceReconstruction with Reference-Guided Fusion***），该方法保留独立估计，并在融合过程中利用它们之间的关系。具体而言，*MSRGF*包含三个组件：（*1*）逐源重建（***SWR***），使用模态特定查询和共享生成器，从每个可用模态为每个模态生成独立估计；（*2*）参考引导融合（***RGF***），利用各估计与可用观测之间的关系构建加权参考，根据它们的差异有选择地优化该参考，并在观测特征可用时将其纳入；（*3*）全局上下文调制（***GCM***），利用全局上下文调整模态摘要以进行情感预测。在三个数据集上的大量实验表明，*MSRGF*优于现有方法：在随机内部模态缺失情况下，*MOSEI*上的平均准确率提升*1.9%*，在随机整模态缺失下提升*0.6%*，在固定整模态缺失下提升*1.3%*。

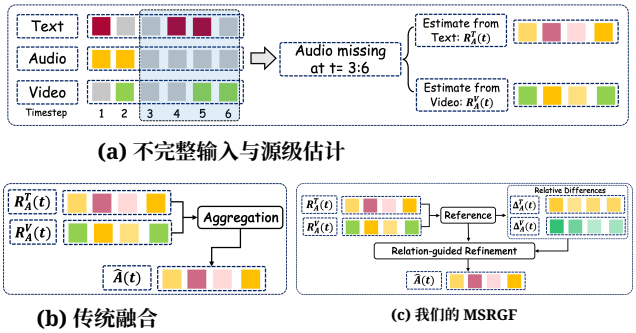


图1. MSRGF 的研究动机。(a) 在多模态观测不完整的情况下，现有的文本和视觉信息可对相同的缺失声学特征给出不同的估计。(b) 直接聚合将这些估计合并为单一表征，而不显式考察它们之间的关系。(c) MSRGF 保留源级估计，构建加权参考表示，并有选择地融入经变换的参考表示偏差，以细化目标表示。

1. 引言

多模态情感分析（MSA）融合文本、音频和视觉信息，以预测情感 [24, 25, 51]。它支持人机交互和视频理解等应用 [23]。在实际应用中，文-本记录可能不完整 [15, 28]，，语音片段可能不可用，视觉观测也可能因遮挡或跟踪失败而中断 [60]。此类干扰可能导致某一模态内出现缺失特征，或导致整个模态缺失，从而使情感预测更具挑战性 [8, 44, 55, 59]。

重建方法通过从剩余模态中估计不可用的表示 [22, 35, 38, 41, 42]，并将其与观测特征整合，以进行情感预测，从而弥补观测缺失。以往研究探索了跨模态转换 [29, 36]，、表征重建 [50, 55]，、分布建模 [41]，、扩散模型 [42]，和生成式提示 [8, 40]。当有多个模态可用时，每个模态都能提供对同一

*通讯作者。

模态的一种不同估计。这些估计可以结合起来构建恢复后的表征，但它们之间的关系通常未被明确用于引导融合过程。

如图 1所示，不同模态可能会随时间呈现不同的缺失模式。当声学信息部分缺失时，可用的文本和视觉观测能够分别估计缺失特征。这些估计在某些特征上可能一致，而在其他特征上有所不同。简单地将它们聚合为单一表征，可能无法充分利用彼此之间的差异。因此，我们选择保留这些独立估计，并学习它们之间的关系如何影响最终的目标表示。

为此，我们提出了**MSRGF**，一种面向不完整多模态情感分析的参考引导融合的多源重建框架。MSRGF由三个模块组成。首先，源级重建(SWR)采用模态特定查询和共享交叉注意力生成器，从每个可用模态中为每种模态生成独立估计，并保留这些估计以供后续融合。在此基础上，参考引导融合 (RGF)将这些估计相互比较，并与可用信息进行比较，以构建加权参考表示。随后，它通过源级加权和特征级门控，有选择地整合经过变换的差异；如果能够获取目标模态的观测特征，也会将其纳入其中。最后，全局上下文调制 (GCM)汇总增强后的模态表征，并利用全局上下文对其进行调整，以用于情感预测。这些模块共同使MSRGF能够在多种缺失情况下利用多源重建信息。

我们的主要贡献总结如下：

- 我们提出了面向不完整多模态情感分析的 MSRGF，该方法从可用模态生成独立估计，并利用它们之间的关系增强模态表示。
- 我们引入了 RGF，利用源级估计之间的关系构建加权参考，并有选择地整合其变换后的差值和观测到的特征，以增强模态表示。
- 在三个基准数据集上，针对固定的整体模态组合、随机整体模态缺失情况和随机模态内部缺失情况开展了大量实验，结果证明了 MSRGF 的有效性。

2. 相关工作

2.1. 不完整多模态情感分析

不完整多模态情感分析旨在文本、音频或视觉信息部分或全部不可用时，仍能维持情感预测 [4, 27, 31, 45–47]。现有方法通过知识蒸馏[19, 32, 57], 、表征增强 [58], 和缺失特征重建 [8]来应对这一挑战。知识蒸馏方法 [7, 16, 17, 19, 37, 39, 44, 49] 利用基于完整信息训练的教师模型，将信息迁移至基于不完整输入训练的学生模型。例如，CorrKD [19]引入样本级和类别级蒸馏，UMDF [17] 将自蒸馏与跨模态交互相结合，而 CMAD [57] 则通过模态感知优化迁移特征相关性。其他方法通过在 HRLF [18], 中进行层次化学习、在 LNLN [54], 中进行语言引导校正，以及在 HME [58] 和 IIE [60]中进行跨样本增强，来改善不完整输入的表征。

基于重建的方法从可用观测中恢复缺失模态表示。MCTN [36]通过循环翻译学习跨模态表示，而 MMIN [55]利用可用模态重建缺失表示。DiCMoR [41] 对跨模态分布迁移进行建模，IMDer [42]采用条件扩散，MPLMM [8]使用生成式提示进行特征恢复。包括 T^{2}DR[26] 和 API [40]，在内的较新方法，进一步探索了不完整输入下的特征预测与原型适配。这些方法以不同方式根据可用信息估计缺失表示。MSRGF 进一步考虑同一模态不同独立估计之间的关系，在重建过程中保留这些关系，并利用它们构建增强表示。

2.2. 多模态交互与融合

多模态交互与融合旨在结合来自不同模态的信息 [6, 48]。MISA [9] 学习模态不变表示和模态特定表示，而 ALMT [53]引入语言引导注意力以实现跨模态交互。CACR [43]纳入了感知冲突的跨模态重建和情感细化，而 EBMC [10] 将模态增强与平衡优化相结合，以改善多模态学习。对于不完整输入，融合方法还会考虑模态表示的可用性和质量。RECAP [14] 将情感恢复与互信息引导注意力相结合，而 DEAR [61]利用重建保真度和不确定性来调节恢复特征。ComP [13] 采用跨模态提示词和知识传播来增强表示。这些方法通过考虑跨模态交互、重建质量或不确定性来改善多模态融合。MSRGF关注由不同可用模态生成的同一模态独立估计之间的关系。RGF利用这些关系构建加权参考，并通过估计之间的差异对其进行选择性细化。

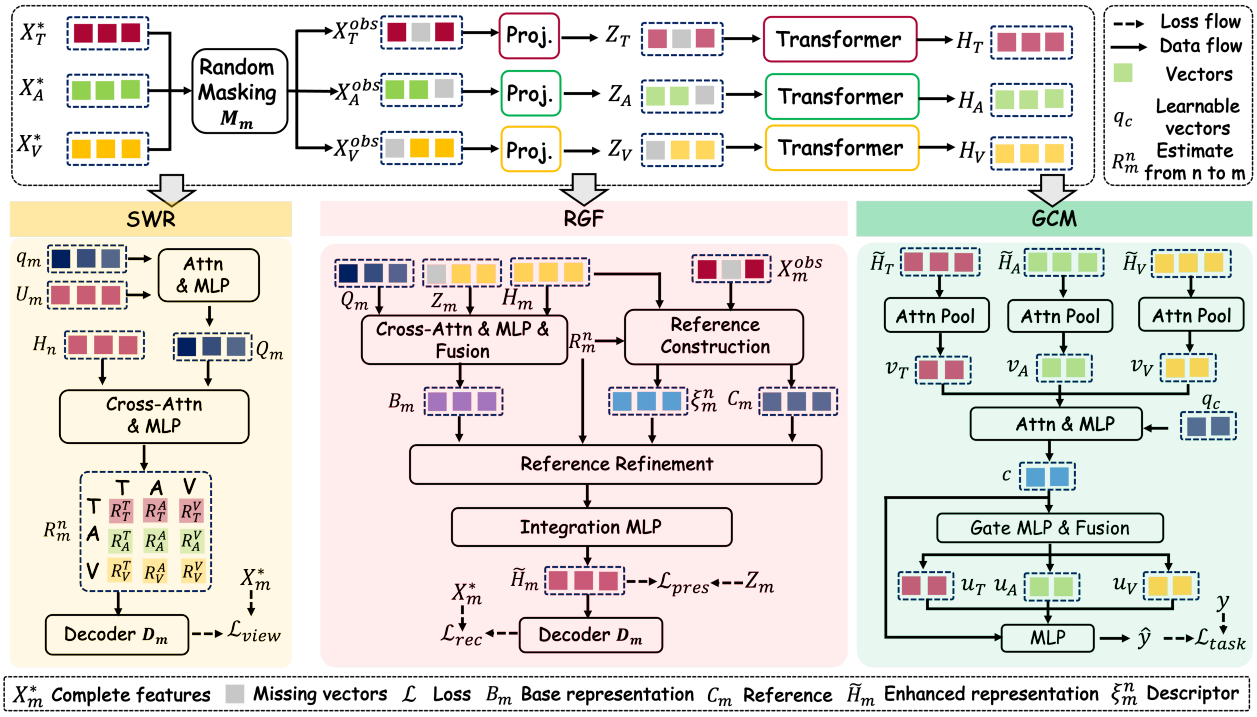


图 2. MSRGF 概览。SWR 根据每个可用模态生成各模态的独立估计。RGF 构建加权参考表示，利用估计之间的差异对其进行选择性细化，并将结果与基础表示相融合；当观测特征可用时，基础表示会纳入这些特征。GCM 利用全局上下文调整模态摘要，以进行情感预测。

3. 方法

3.1. 问题定义

我们研究包含文本、音频和视觉模态的不完整多模态情感分析，记为 $\mathcal{M} = \{T, A, V\}$。对于每种模态 $m \in \mathcal{M}$，$X_m^{\text{obs}} \in \mathbb{R}^{L \times d_m}$ 表示其观测特征序列，其中 L 为对齐序列长度，d_m 为其特征维度。观测掩码 $M_m \in \{0,1\}^L$ 标识观测位置，有效性掩码 $P_m \in \{0,1\}^L$ 则排除填充部分，其中 $M_m \le P_m$。缺失位置的特征设为零；$M_m = \mathbf{0}$表示模态 m完全缺失。每个样本至少保留一个观测位置，因此可用模态集合非空：$\mathcal{O} = \{n \in \mathcal{M} : \|M_n\|_1 > 0\}$。目标是根据$\{X_m^{\text{obs}}, M_m, P_m\}_{m \in \mathcal{M}}$预测情感 $\hat{y}$。在受控缺失情况实验中，完整特征 $X_m^{\star}$ 仅在训练期间保留为重建目标。

3.2. 框架概览

不同的可用模态可以对同一缺失信息提供不同的估计。为利用这些估计之间的关系，MSRGF 在融合之前始终将它们分开（图 2）。源级重建（SWR）首先根据每个可用模态生成各模态的估计。参考引导融合（RGF）随后比较这些估计以确定其贡献，形成加权参考表示，并利用它们的剩余差异来细化表征。估计模态的观测特征（如果存在）也会参与这一融合过程。最后，全局上下文调制（GCM）将三个增强表示结合起来，用于情感预测。

3.3. 源级重建

对于每个估计，SWR 需要回答两个问题：有哪些信息可用，以及应估计哪个模态和哪些时间位置。我们使用 $n \in \mathcal{O}$ 表示提供信息的可用模态，使用 $m \in \mathcal{M}$ 表示待估计的模态。

每种模态首先通过线性模态特定投影，将其观测特征映射到一个共同的隐藏维度 d，随后进行层归一化，记作 ϕ_m。然后，时序 Transformer E_m 读取观测位置，以生成其上下文表示：

$$\begin{aligned} Z_m &= M_m \odot \phi_m(X_m^{\text{obs}}), \\ H_m &= M_m \odot E_m(Z_m + e^{\text{pos}}; M_m). \end{aligned} \tag{1}$$

其中 e^{pos} 表示位置嵌入，M_m 则防止缺失位置参与注意力键或值的计算。编码特征 H_n 将提供信息

这些信息从可用模态 n中读取。我们还保留 Z_m ，以便模态 m 的观测特征能够直接进入 RGF。如果 $M_m = \mathbf{0}$，则跳过时间注意力，并将 H_m 设为零。

当模态 m 完全缺失时，该模态的特征无法定义重建位置。因此，SWR 使用模态身份e_m^{mod}、位置 e_t^{pos} 和观测状态 $e_{M_m}^{\text{mask}}{}_{(t)}$ 构造查询 Q_m 。在位置 t处，这三个嵌入构成： $q_m^0(t) = e_t^{\text{pos}} + e_m^{\text{mod}} + e_{M_m}^{\text{mask}}{}_{(t)}$ 。一个经过学习的提示词库 U_m 包含用于模态 m的额外可训练向量。初始查询关注该提示词库，共享多层感知机 f_Q 生成：

$$Q_m(t) = P_m(t) f_Q\big(q_m^0(t) + \text{Attn}(q_m^0(t), U_m, U_m)\big)\,. \tag{2}$$

因此， Q_m 规定了 SWR 将要估计的内容，而无需读取缺失的特征值； H_n 则提供了可据此推断估计结果的观测信息。

多源估计： 对于每个可用模态 n，查询 Q_m 通过共享的双模块交叉注意力生成器 $\mathcal{G}$读取 H_n 。其输出经过一个模态特定的多层感知机 f_m^{tar}：

$$R_m^n = P_m \odot f_m^{\text{tar}}\big(\mathcal{G}(Q_m + e_m^{\text{mod}}, H_n + e_n^{\text{mod}}; M_n)\big)\,. \tag{3}$$

这里， Q_m 提供查询，而 H_n 中的观测位置提供键和值。该生成器在所有模态对之间共享；对于固定的 m，每个可用模态也使用相同的输出多层感知机 f_m^{tar}。因此，估计结果 $R_m^n \in \mathbb{R}^{L\times d}$ 具有相同的目标位置和隐藏维度，同时保留每个估计结果由哪个可用模态生成的信息。对于每个模态 m，一个模态特定的线性解码器 D_m : $\mathbb{R}^d \to \mathbb{R}^{d_m}$ 在其估计结果与最终增强表示之间共享。它支持重建监督，以及在 RGF 中与观测特征进行比较。如果模态 m仅部分观测到，它还可以通过 $n = m$贡献一个估计结果。SWR 将所有$\{R_m^n\}_{n\in\mathcal{O}}$ 传递给 RGF。

3.4. 参考引导融合

SWR 为模态 m提供了多个估计，但这些估计不必等量贡献。立即将它们组合会丢失估计之间的区别，而平等对待它们则会忽略彼此差异。RGF 首先利用它们之间的关系形成一个参考表示，然后利用它们与该参考表示之间的偏差构建细化结果。当模态 m 具有观测特征时，这些特征也有助于确定如何使用这些估计。

对于部分观测的模态，其自身已观测到的特征能够提供无需从其他模态推断的信息。因此，RGF通过从 Z_m出发的直接路径构建基础表征 B_m 。为了补充未观测位置的特征，另一个交叉注意力块 $\mathcal{C}_{\text{anc}}$ 使用 Q_m读取 H_m ，随后接入一个共享的多层感知机f_{anc}:

$$a_m = f_{\text{anc}}(\mathcal{C}_{\text{anc}}(Q_m, H_m; M_m))\,. \tag{4}$$

基础表征将该结果与可用的观测特征相结合：

$$B_m = P_m \odot [M_m \odot (Z_m + \eta_a a_m) + (1 - M_m) \odot a_m] \tag{5}$$

这里的 η_a 是一个固定缩放系数。在观测位置，Z_m 直接进入，而 a_m 提供缩放后的更新；在其他位置， B_m 使用 a_m。当模态 m完全缺失时，将跳过交叉注意力更新，但查询残差和前馈操作仍会生成 a_m。因此，即使没有模态 m的观测特征， B_m 仍可引导后续加权。

参考构建： 分配给 R_m^n的权重不应仅考虑其在单个位置上的取值。因此，我们将该估计与生成它的模态、 m 的其他估计，以及在可能的情况下与 m的观测特征进行比较。这些比较为每个估计形成一个关系描述符 ξ_m^n 。

首先，我们在有效位置上进行 R_m^n ，并在观测位置上进行 H_n 。将这两个池化表征投影到相同维度后，分别记为 a_m^n 和 b_n。一个经过学习的多层感知机 f_{src} 利用表征本身、它们的乘积及其绝对差来衡量二者的交互：

$$\mathbf{c}_{m,n}^{\text{src}} = f_{\text{src}}([a_m^n, b_n, a_m^n \odot b_n, |a_m^n - b_n|])\,. \tag{6}$$

其次，我们通过对池化表征之间的余弦距离取均值，衡量 R_m^n 与其他估计之间的差异。第三，当模态 m被观测到且$n \neq m$时，我们在观测位置上使用平均余弦相似度，将 D_m（R_m^n）与 X_m^{obs}进行比较。这两个余弦度量被用作比较分数，并停止其梯度；源–估计交互仍可训练。当某项比较不可用时，其分数设为零，并由一个指示标记记录该比较未被执行。

这三项比较、 m 和n的观测覆盖率，以及它们的有效性指标共同构成描述符：

$$\xi_m^n = \big[c_{m,n}^{\text{tar}}, \mathbf{c}_{m,n}^{\text{src}}, c_{m,n}^{\text{div}}, \rho_m, \rho_n, v_{m,n}^{\text{tar}}, v_n^{\text{src}}, v_n^{\text{div}}\big]. \tag{7}$$

其中， c^{tar} 和 c^{div} 分别表示观测特征一致性和跨估计差异分数， $\rho_m = \|M_m\|_1/\max(\|P_m\|_1,1)$表示观测覆盖率。三个二值指示标记$v_{m,n}^{\text{tar}}$、$v_n^{\text{src}}$和 v_n^{div}分别说明对应的比较是否可用；该描述符针对每个估计计算一次，并在其时间位置之间复用。相关定义和具体的比较公式见附录。

随后，RGF 将该描述符与当前估计值 $R_m^n(t)$ 和基础表示 $B_m(t)$ 相结合，以计算一个

每个可用模态的权重：

$$\alpha_m^n(t) = \mathrm{softmax}_{n\in\mathcal{O}}\, f_\alpha([R_m^n(t), B_m(t), \xi_m^n] + e_m^\alpha)\,. \tag{8}$$

其中，e_m^α 是一种学习得到的模态嵌入，使共享评分多层感知机能够区分正在估计的模态。归一化权重为参考表示：

$$C_m(t) = \sum_{n\in\mathcal{O}} \alpha_m^n(t) R_m^n(t). \tag{9}$$

因此，C_m 会根据各项估计与可用信息之间的关系进行组合。原始的 R_m^n仍然保留，因此 RGF 仍可考察每项估计与该参考表示之间的差异。

参考表示细化：参考表示对各项估计进行总结，但并未明确表示其中任一项与它的偏差。因此，对于每个可用模态，RGF 会计算 $\Delta_m^n = R_m^n - C_m$，并通过共享非线性多层感知机 F_{diff}对其进行转换。由于这些偏差可能同时包含有用信息和重建误差，模型会学习每个转换后的偏差应贡献多少。

两个共享多层感知机（f_β 和 f_G）控制这一贡献。源级权重 β_m^n（t）从可用模态中进行选择，而特征级门控 G_m^n（t）则在每个转换后的偏差中选择维度：

$$\begin{aligned}\beta_m^n(t) &= \mathrm{softmax}_{n\in\mathcal{O}}\, f_\beta\big([R_m^n(t), B_m(t), \xi_m^n] + e_{m,n}^\beta\big)\,,\\ G_m^n(t) &= \sigma\big(f_G\big([C_m(t), B_m(t), R_m^n(t), \xi_m^n] + e_{m,n}^G\big)\big)\,.\end{aligned} \tag{10}$$

学习嵌入$e_{m,\,n}^\beta$ 和$e_{m,\,n}^G$ 可在两个共享多层感知机中区分不同的模态对。与用于构建参考表示的 α_m^n不同，β_m^n 控制偏差的贡献。门控 $G_m^n \in \mathbb{R}^{L\times d}$ 还利用参考表示本身调整各个特征维度。随后进行细化：

$$N_m = \mathbb{I}[|\mathcal{O}| > 1]\, P_m \odot \sum_{n\in\mathcal{O}} \beta_m^n \big[G_m^n \odot F_{\text{diff}}(\Delta_m^n)\big]. \tag{11}$$

通过在聚合前转换并加权偏差，RGF 可以构造出超越加权参考表示的更新。当只有一种模态可用时，$C_m = R_m^n$，细化设为 $N_m = 0$。

集成： RGF 现在包含三个作用各不相同的表示：B_m 在可用时保留观测特征，C_m 融合各个独立估计，N_m 则包含选定的基于差异的更新。增强后的表示以 B_m为起点，向 C_m移动，并通过学习得到的门控加入 N_m：

$$\widetilde{H}_m = P_m \odot \big[B_m + g_m^{\text{rec}} \odot (C_m - B_m) + g_m^{\text{enh}} \odot N_m\big]. \tag{12}$$

为获得这些门控，共享的多层感知机f_{int} 接收 B_m, C_m, N_m, M_m 以及一个关系摘要 $\bar{\xi}_m(t) = \sum_{n\in\mathcal{O}} \alpha_m^n(t)\xi_m^n$。它输出两个 Sigmoid 门控$[r_m^{\text{rec}}, r_m^{\text{enh}}] = \sigma(f_{\text{int}}([B_m, C_m, N_m, \bar{\xi}_m, M_m]))$。观测掩码随后对它们进行调整：

$$\begin{aligned} g_m^{\text{rec}} &= (1 - M_m) + \eta_g M_m \odot r_m^{\text{rec}},\\ g_m^{\text{enh}} &= (1 - M_m + \eta_g M_m) \odot r_m^{\text{enh}}. \end{aligned} \tag{13}$$

这里，η_g 是用于观测位置更新的固定缩放系数。在缺失位置，$g_m^{\text{rec}} = 1$：B_m 的直接贡献相互抵消，输出变为$C_m + g_m^{\text{enh}} \odot N_m$。在观测位置，来自$Z_m$ 的路径仍可通过 B_m保留，同时两种更新都按 η_g进行缩放。这样，RGF 在观测缺失时利用重建，在观测存在时保留直接观测路径。例如，如果音频完全缺失，而文本和视觉可用，那么两个估计 R_A^T 和 R_A^V 会先形成 C_A；随后，它们各自的偏差会共同作用于 N_A，以细化缺失的音频表示。对于仅部分缺失的模态，同样的计算也会保留直接观测路径。

3.5. 全局上下文调制

RGF 为每个模态生成一个增强序列 $\widetilde{H}_m$。为了获得样本级情感预测，GCM 首先对每个序列的有效位置进行摘要：

$$v_m = \mathrm{AttnPool}_m(\widetilde{H}_m; P_m). \tag{14}$$

随后，一个学习到的全局查询通过注意力 MHA$_c$关注三个摘要$\mathbf{H}_{\text{sum}} = \mathrm{Stack}(v_T, v_A, v_V)$，生成 $c = f_c(\mathrm{MHA}_c(q_c, \mathbf{H}_{\text{sum}}, \mathbf{H}_{\text{sum}}))$。其中，$f_c$ 是多层感知机，该上下文 c 提供了来自全部三个增强模态的信息。GCM 还使用相同的上下文调整每个单个摘要。它首先将 c 投影到 $s_m = W_m^{\text{back}} c$，并利用 v_m 与s_m 之间的关系来预测特征级门控：

$$\gamma_m = \sigma(f_\gamma([v_m, s_m, v_m - s_m, v_m \odot s_m] + e_m^\gamma))\,. \tag{15}$$

学习嵌入 e_m^γ 在共享门控多层感知机中标识该模态。调整后的摘要和最终预测为：

$$\begin{aligned} u_m &= v_m - \gamma_m \odot s_m,\\ \hat{y} &= f_{\text{pred}}([c, u_T, u_A, u_V]). \end{aligned} \tag{16}$$

因此，多层感知机 f_{pred} 接收联合上下文以及三个经过调整的模态摘要，并预测样本级情感 $\hat{y}$。全局上下文为情感预测提供联合表征，并为调整各个模态摘要提供参考。门控减法使每个模态摘要都能根据联合上下文进行修改，同时最终预测器可直接访问 c。

3.6. 学习目标

重建和融合阶段均以其生成的表征作为监督目标。在受控缺失情况实验中，$X_m^\star$ 是完整的训练目标，而 $\Omega_m = P_m \odot (1 - M_m)$ 用于标识缺失的有效位置。令 $\ell(U,\ V;\ W)$ 表示掩码 L_1 距离，该距离在选定位置和特征维度上取平均。记号 $\langle\cdot\rangle_\Omega$ 表示对至少存在一个缺失有效位置的样本-模态对取平均，而 $\langle\cdot\rangle_M$ 表示对存在观测位置的样本-模态对取平均；在每个配对内部，源误差按可用模态取平均。

SWR 对每个独立估计都接收直接监督，促使来自不同可用模态的估计重建完整的特征：

$$\mathcal{L}_{\text{view}} = \left\langle \frac{1}{|\mathcal{O}|} \sum_{n\in\mathcal{O}} \ell(D_m(R_m^n), X_m^\star; \Omega_m) \right\rangle_\Omega . \quad (17)$$

RGF 在各个估计完成融合后接收重建监督：

$$\mathcal{L}_{\text{rec}} = \left\langle \ell(D_m(\widetilde{H}_m), X_m^\star; \Omega_m) \right\rangle_\Omega . \quad (18)$$

最后，为限制观测位置的变化，我们鼓励增强表示保持接近投影后的观测特征：

$$\mathcal{L}_{\text{pres}} = \left\langle \ell(\widetilde{H}_m, Z_m; M_m) \right\rangle_M . \quad (19)$$

其中，Z_m 是可训练的，因此该项鼓励一致性，而不强制要求完全保留。情感目标 $\mathcal{L}_{\text{task}}$ 是 CMU-MOSI 和 CMU-MOSEI 上的 L_1 回归，以及 IEMOCAP 上四个二元情绪分类器的均值交叉熵。我们优化

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{rec}\mathcal{L}_{\text{rec}} + \lambda_{view}\mathcal{L}_{\text{view}} + \lambda_{pres}\mathcal{L}_{\text{pres}}. \quad (20)$$

4. 实验

4.1. 数据集与评估指标

我们在三个基准数据集上评估 MSRGF：CMU- MOSI [51], CMU-MOSEI [52], 和 IEMOCAP [2]。CMU- MOSI 和 CMU-MOSEI 分别包含 2,199 和 22,856 条话语，情感分数范围为 -3至 $+3$。遵循以往的研究 [58], 我们将情感预测和标签转换为二元类别（正面和负面），并报告准确率（Acc）和 F1 分数，同时排除中性样本。IEMOCAP 包含 4,453 个样本，涵盖四种二元情绪任务（开心、悲伤、愤怒和中性），我们报告其平均准确率和 F1 分数。

4.2. 实现细节

特征提取。在 CMU-MOSI 和 CMU-MOSEI 上，我们使用 DeBERTa [11, 12] 提取文本特征，并在 IEMOCAP 上使用 GloVe [34]。对于全部三个数据集，我们分别使用 OpenFace [1] 和 CO-VAREP [5], 提取视觉特征和声学特征。

缺失协议。我们考虑三种缺失设置。(1) 固定整模态缺失 [19, 57]评估七种固定模态组合：T、A、V、TA、TV、AV 和 TAV，其中 TAV 表示完整输入。(2)随机整模态缺失 [22, 58] 以 10% 至 70% 的比例随机移除整个模态，同时确保每个样本至少有一个模态可用。(3) 随机模态内缺失[20, 21, 61]以 10% 至 90% 的比例随机移除每个模态中的时间特征。

实验设置。所有模型均使用 PyTorch [33] 实现，并在 NVIDIA RTX A6000 GPU 上训练。我们将隐藏维度设为 128，并采用 AdamW 优化器 [30]。更多实现细节见附录。

4.3. 与最先进方法的比较

我们将 MSRGF 与具有代表性的不完整多模态情感分析方法进行比较，包括 DiCMoR [41], IMDer [42], LNLN [54],MPLMM [8], HME [58], CMAD [57], API [40], 和 TLRA[3]。对于随机模态内缺失，我们还纳入了 ALMT [53], P-RMF [56], DEAR [61],SemMSA [20], 和 TF-Mamba [21]。详细结果见附录。

固定整模态缺失。如表 1所示，MSRGF 在三个数据集上的平均准确率/F1值均达到最佳，在 MOSI 上为 77.4/77.1，在 MOSEI 上为 78.4/78.0，在 IEMOCAP 上为 79.7/78.2。与最强基线相比，这些结果分别提升了 1.0/0.9、1.3/1.4 和 0.5/0.4 个百分点，表明该方法在不同可用模态组合下均有效。

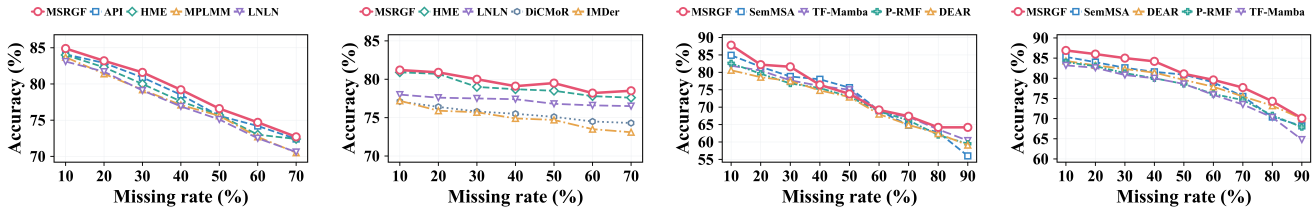
随机缺失。图 3(a,b) 展示了随机整模态缺失（10%–70%）情况下的结果，MSRGF 在 MOSEI 上的平均准确率/F1 分数为 79.0/78.3，在 IEMOCAP 上为 79.6/78.2，分别比最强基线高出 0.6/0.5 和 0.6/0.4 个百分点。在随机模态内缺失情况下，图 3(c,d) 显示，MSRGF 在 MOSI 上达到 74.1/74.1，在 MOSEI 上达到 80.5/80.2，分别比最强基线提升 1.7/1.6 和 1.9/1.0 个百分点。这些结果表明，MSRGF 在不同缺失类型和缺失率下均具有稳健性。

4.4. 消融研究

我们通过对 MOSI 和 MOSEI 上九种随机内部模态缺失率，以及 IEMOCAP 上七种随机整体模态缺失率的结果取平均，评估 SWR、RGF 和 GCM 的贡献。如表 2所示，移除 SWR 会导致 F1值分别下降 2.9、2.2 和 3.0

表1. 固定整模态缺失条件下的比较（准确率/F1值，%）。粗体和下划线分别表示最佳和次佳结果，并列时共享排名。 * 表示 HME 报告的结果 [58]； † 表示对六种可用组合取平均值。

方法	T	A	V	TA	TV	AV	TAV	Avg
CMU-MOSI								
DiCMoR* [41]	84.5/84.4	60.5/60.8	62.2/60.2	85.5/85.5	85.5/85.4	64.0/63.5	85.7/85.6	75.4/75.1
IMDer* [42]	84.8/84.7	62.0/62.2	61.3/60.8	85.4/85.3	85.5/85.4	63.6/63.4	85.7/85.6	75.5/75.3
LNLN* [54]	84.9/85.1	52.2/58.9	52.2/58.9	84.9/85.2	84.9/85.1	52.2/58.9	84.3/84.6	70.8/73.8
MPLMM* [8]	83.8/83.8	54.6/42.0	54.7/42.1	83.7/83.6	83.8/83.7	57.3/42.6	83.8/83.8	71.7/65.9
HME* [58]	85.7/85.6	62.7/62.2	63.6/63.4	85.8/85.7	85.8/85.8	64.6/64.5	86.4/86.3	76.4/76.2
API [40]	85.8/85.8	63.3/**63.2**	**64.5/64.4**	86.3/86.1	86.1/86.0	65.2/**65.2**	–	75.2/75.1†
TLRA [3]	85.2/85.2	62.5/62.0	62.7/62.2	85.8/86.0	85.5/85.5	61.3/61.3	86.3/86.2	75.6/75.5
MSRGF	**86.9/86.8**	**63.6**/62.9	63.0/61.9	**86.4/86.5**	**86.6/86.6**	**65.7**/64.9	**89.8/89.8**	**77.4/77.1**
CMU-MOSEI								
DiCMoR* [41]	84.2/84.3	62.9/60.4	63.6/63.6	85.0/84.9	84.9/84.9	65.2/64.4	85.1/85.1	75.8/75.4
IMDer* [42]	84.5/84.5	63.8/60.6	63.9/63.6	85.1/85.1	85.0/85.0	64.9/63.5	85.1/85.1	76.0/75.3
MPLMM* [8]	84.6/84.6	61.8/49.4	62.2/53.2	85.1/85.1	85.2/85.1	63.2/51.0	85.3/85.3	75.3/70.5
HME* [58]	85.5/85.4	64.1/59.7	64.4/62.9	85.6/85.4	85.7/85.7	65.4/63.7	86.2/86.2	76.7/75.6
CMAD [57]	86.1/86.0	63.0/60.8	65.7/64.4	86.3/86.2	86.4/86.4	65.6/64.8	86.1/86.2	77.0/76.4
API [40]	85.5/85.4	65.1/**65.5**	**66.7**/65.4	85.7/85.5	85.6/85.6	66.1/66.0	–	75.8/75.6†
TLRA [3]	85.1/85.1	**65.5**/63.7	66.0/**65.5**	85.9/85.8	84.8/84.8	66.6/65.6	85.9/85.9	77.1/76.6
MSRGF	**87.5/87.5**	65.2/64.2	66.4/**65.5**	**87.1/87.1**	**87.2/87.3**	**67.3/66.1**	**88.2/88.1**	**78.4/78.0**
IEMOCAP								
DiCMoR* [41]	75.2/68.2	76.0/68.7	73.6/68.8	78.7/76.5	74.4/72.1	76.3/75.2	78.7/76.5	76.1/72.3
IMDer* [42]	75.1/67.2	75.5/70.2	75.1/70.1	76.6/71.7	75.2/72.6	76.3/72.3	77.5/75.4	75.9/71.4
LNLN* [54]	76.8/74.4	77.1/74.6	74.5/69.9	77.9/75.4	77.0/75.5	77.3/74.9	78.4/77.4	77.0/74.6
MPLMM* [8]	69.2/69.3	59.8/59.7	57.6/57.0	76.0/75.4	74.7/74.5	67.3/67.4	–	67.4/67.2†
HME* [58]	79.1/78.1	78.9/76.8	75.5/70.8	80.7/80.0	79.7/78.9	78.1/76.3	81.1/80.6	79.0/77.4
CMAD [57]	**79.5/78.7**	79.2/76.7	73.6/70.6	**81.9/81.3**	79.7/79.1	79.0/77.3	**81.7/81.2**	79.2/77.8
MSRGF	79.3/78.6	**79.7/77.7**	**75.6/71.4**	81.2/80.6	**80.9/80.6**	**79.7/77.6**	**81.7**/80.8	**79.7/78.2**



(a) CMU-MOSEI (b) IEMOCAP (c) CMU-MOSI (d) CMU-MOSEI

图3. 随机缺失情况下的准确率表现。图(a、b)展示了缺失率为10%至70%的随机整模态缺失；图(c、d)展示了缺失率为10%至90%的随机模态内缺失。

个百分点，凸显了源级重建的重要性。将 RGF 替换为统一补全会使 F1值 下降 1.0–2.9 个百分点，证明了参考引导集成优于对源估计进行直接平均。移除 GCM 同样会使 F1值 下降 1.1–2.6 个百分点，说明了全局上下文调制的贡献。在 IEMO- 上进行的其他消融

CAP 进一步考察了固定和随机整模态缺失情况下的内部机制和辅助目标，详情见附录9 (表8)。

4.5. 超参数分析

我们还考察了在 IEMOCAP 的音频-视觉

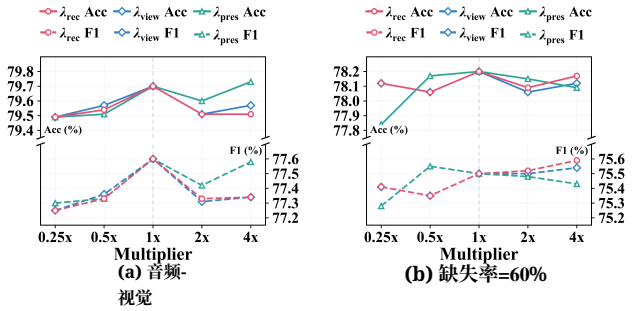


图4. IEMOCAP 上在 (a) 音频-视觉和 (b) 随机整模态缺失（缺失率=60%）情况下的损失权重敏感性。乘数分别独立缩放各个默认权重（0.1,0.1,0.05）。

表2. 组件消融结果，分别在 MOSI/MOSEI 的九种模态内缺失率和 IEMOCAP 的七种随机整模态缺失率上取平均。

方法	MOSI	MOSEI	IEMOCAP
不使用 SWR	71.2/71.2	77.9/78.0	77.6/75.2
不使用 RGF	73.1/73.1	78.9/78.1	77.7/75.3
不含 GCM	72.8/73.0	79.4/79.0	77.9/75.6
MSRGF	**74.1/74.1**	**80.5/80.2**	**79.6/78.2**

表3. IEMOCAP上的推理干预（F1值，%）。

		SWR	RGF 更新		
设置	Full	单一	Zero	随机	排列
TV→A	**80.61**	79.20	80.00	80.03	79.60
AV→T	**77.58**	77.03	77.33	77.28	76.88

组合和随机整模态缺失情况下，缺失率为=60%。如图4所示，每个系数均独立缩放为其默认值的 0.25 至 4 倍，其他系数保持固定。总体而言，不同乘数下的性能波动相对较小，默认配置在两种设置下均持续表现出有竞争力的结果。尽管某些替代权重带来了微小提升，但单独增大或减小某个系数并未呈现出一致的收益，这表明在所评估的范围内，MSRGF 对这些损失权重相对不敏感。

4.6. 机制分析

我们通过推理干预和重新训练的设计变体，探究源级估计和 RGF 的作用。

源级估计。我们首先通过推理干预和重新训练的设计替代方案，考察多个重建源的贡献。如

表4. IEMOCAP 上重新训练的 RGF 消融实验（准确率/F1，%）。

变体	固定	随机
均匀 α	78.38/76.12	78.05/75.74
早期聚合	78.43/76.16	78.05/75.74
普通残差	78.30/76.02	78.03/75.75
不使用差异细化	78.34/76.06	78.05/75.72
硬观测复制	77.96/75.76	77.83/75.38
不含观测融合	78.40/76.79	78.13/76.35
MSRGF	**79.73/78.17**	**79.63/78.25**

在表3中，与测试过的表现较强的单一源相比，在 TV 和音频-视觉设置下，同时保留两种源估计分别使 F1 值提高了 1.41 和 0.55 个百分点。我们还将 MSRGF 与重新训练的 T 优先、A 优先和 V 优先单源变体，以及融合源生成器进行比较。在固定和随机整模态缺失情况下，所有替代方案的平均准确率/F1 均较低（附录9，表 8）。这些结果表明，保留分别以源为条件的估计，有利于后续的参考表示引导融合。

有效差异更新。随后，我们通过将学习得到的更新 $U_m = g_m^{\text{enh}} \odot N_m$ 替换为零、等范数随机方向，或在样本之间进行置换的更新，来评估该更新。仅修改缺失的目标，而其他增强表示保持不变。原始更新在 TV 和 AV 下均优于所有替代方案（表3），表明学习得到的细化有助于预测。

参考构建与差异细化。接下来，我们将 RGF 与更简单的融合策略进行比较。均匀 α 融合以相等权重对源估计进行平均，而早期聚合则在 RGF 之前将它们合并。普通残差将源特定差异替换为$C_m - B_m$，而无差异细化则移除 N_m。在两种协议下，所有变体的准确率/F1值均低于 MSRGF（表 4），证明了学习得到的参考构建与源特定细化的优势。

观测特征整合。最后，我们考察 RGF 如何融合保留的观测。硬观测复制会将观测位置处的增强特征替换为 Z_m，而无观测融合则移除 B_m 的直接贡献及其向C_m的插值。两种变体都会导致性能下降，在固定/随机条件下，F1值分别降低 2.41/2.87 和 1.38/1.90 个百分点。这凸显了整合观测信息与重建信息的重要性，而不是依赖严格复制或移除观测融合路径。

更多消融实验和机制分析，包括逐源估计贡献、重建诊断——

、关系与细化干预、观测特征保留以及全局上下文调制，详见附录 9。

5. 结论

本文提出了 MSRGF，用于不完整多模态情感分析，以解决对源级估计之间互补信息利用不足的问题。具体而言，源级重建通过共享的重建生成器，从可用模态生成独立的目标估计。参考引导融合通过构建加权参考并细化源特定差异来整合这些估计，同时纳入保留的观测，以增强目标表示。全局上下文调制进一步利用全局上下文调整增强表示，从而进行情感预测。在 CMU-MOSI、CMU-MOSEI 和 IEMOCAP 上开展的大量实验表明，MSRGF 在各种缺失情况下均表现有效。

LLM 使用声明

在准备本研究的过程中，作者使用了 OpenAI 的 ChatGPT 进行语言润色。此外，作者还借助 AI 工具生成和改进实现代码，并协助讨论研究方向和实验流程。作者审阅并修订了所有 AI 辅助材料，并对本研究内容的准确性和完整性承担全部责任。

参考文献

[1] Tadas Baltrušaitis、Peter Robinson 和 Louis-Philippe Morency。Openface：一个开源面部行为分析工具包。载于 *2016 IEEE* 冬季计算机视觉应用会议（*WACV*），第 1–10 页。IEEE，2016 年。6[2]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。Iemocap：交互式情绪双人动作捕捉数据库。《语言资源与评估》，42：335–359，2008 年。6[3] Chenglizhao Chen、Yuchen Cao、Xinyu Liu、Mengke Song、Guisheng Zhang 和 Xiaomin Yu。利用缺失模态控制多模态情感分析中的决策漂移。载于第三十五届人工智能国际联合会议论文集，*IJCAI-26*，第 962–970 页。人工智能国际联合会议组织，2026 年。6，7[4] Junsong Chen、Jiyuan Liu、Suyuan Liu、Wei Zhang、Ao Li、En Zhu 和 Xinwang Liu。面向不完整多模态表示的样本特定模态诊断与跨模态增强。载于 *AAAI*人工智能会议论文集，第 20154–20162 页，2026 年。2

[5] Gilles Degottex、John Kane、Thomas Drugman、Tuomo Raitio 和 Stefan Scherer。Covarep——面向语音技术的协作式语音分析资源库。载于*2014 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 960–964 页。IEEE，2014 年。6[6] Xiang Deng、Feng Gao、Yong Zhang、Youxin Pang、Xu Xi-aoming、Zhuoliang Kang、Xiaoming Wei 和 Yebin Liu。U-Mind：用于视听生成的实时多模态交互统一框架。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第 10874–10886 页，2026 年。2[7] Nuno C Garcia、Pietro Morerio 和 Vittorio Murino。利用多流网络进行模态蒸馏以实现动作识别。载于 欧洲计算机视觉会议论文集（*ECCV*），第 103–118 页，2018 年。2

[8] Zirun Guo、Tao Jin 和 Zhou Zhao。面向情感分析和情绪识别的缺失模态多模态提示学习。载于 第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第1726–1736页，2024年。1，2，6，7

[9] Devamanyu Hazarika、Roger Zimmermann 和 Soujanya Poria。MISA：用于多模态情感分析的模态不变及特定表征。载于 第*28*届*ACM*国际多媒体会议，第1122–1131页，2020年。2

[10] Kang He、Yuzhe Ding、Xinrong Wang、Fei Li、Chong Teng 和 Donghong Ji。通过先增强后平衡的模态协作实现鲁棒多模态情感分析。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第30183–30193页，2026年。2[11] Pengcheng He、Jianfeng Gao 和 Weizhu Chen。DeBERTa-v3：使用 ELECTRA式预训练与梯度解耦嵌入共享改进 DeBERTa。*arXiv*预印本 *arXiv:2111.09543*，2021年。6[12] Pengcheng He、Xiaodong Liu、Jianfeng Gao 和 Weizhu Chen。DeBERTa：具有解耦注意力的解码增强型 BERT。在国际学习表征会议，2021年。6[13] Wen-Jue He、Xiaofeng Zhu 和 Zheng Zhang。用于平衡不完整多模态情绪识别的跨模态提示。在*AAAI*人工智能会议论文集，第17463–17471页，2026年。2[14] Huiting Huang、Tieliang Gong、Kai He、Wen Wen、Weizhan Zhang 和 Mengling Feng。恢复连贯的情感模式：解决多模态情感分析中的模态缺失问题。在*AAAI*人工智能会议论文集，第21957–21965页，2026年。2[15]Mimansa Jaiswal 和 Emily Mower Provost。用于情绪识别的隐私增强多模态神经表征。在*AAAI*人工智能会议论文集，第7985–7993页，2020年。1[16] Shanshan Lao、Guanglu Song、Boxiao Liu、Yu Liu 和 Yujiu Yang。UniKD：用于模仿同质或异质目标检测器的通用知识蒸馏。在*IEEE/CVF*国际计算机视觉会议论文集，第6362–6372页，2023年。2

[17] Mingcheng Li、Dingkang Yang、Yuxuan Lei、Shunli Wang、Shuaibing Wang、Liuzhen Su、Kun Yang、Yuzheng Wang、Mingyang Sun 和 Lihua Zhang。用于具有不确定缺失模态的多模态情感分析的统一自蒸馏框架。在 AAAI人工智能会议论文集，第10074–10082页，2024年。2[18] Mingcheng Li、Dingkang Yang、Yang Liu、Shunli Wang、Jiawei Chen、Shuaibing Wang、Jinjie Wei、Yue Jiang、Qingyao Xu、Xiaolu Hou 等。通过层次表示学习实现鲁棒不完整多模态情感分析。神经信息处理系统进展，37：28515–28536，2024年。2[19] Mingcheng Li、Dingkang Yang、Xiao Zhao、Shuaibing Wang、Yan Wang、Kun Yang、Mingyang Sun、Dongliang Kou、Ziyun Qian 和 Lihua Zhang。面向不完整模态多模态情感分析的相关性解耦知识蒸馏。在 IEEE/CVF计算机视觉与模式识别会议论文集，第12458–12468页，2024年。2、6[20] Wenhao Li、Zhibin Wu、Chong Xiao 和 Qiangchang Wang。SemMSA：具有不完整数据的潜在语义辅助鲁棒多模态情感分析。arXiv预印本 arXiv:2609.30238，2026年。6、7[21] Xiang Li、Xianfu Cheng、Dezhuang Miao、Xiaoming Zhang 和 Zhoujun Li。TF-Mamba：具有缺失模态的文本增强融合 Mamba，用于鲁棒多模态情感分析。在 EMNLP（Findings），第11252–11267页，2025年。6、7[22] Zheng Lian、Lan Chen、Licai Sun、Bin Liu 和 Jianhua Tao。GCNet：对话中的不完整多模态学习图补全网络。IEEE模式分析与机器智能汇刊，45：8419–8432，2023年。1、6[23] Zheng Lian、Licai Sun、Yong Ren、Hao Gu、Haiyang Sun、Lan Chen、Bin Liu 和 Jianhua Tao。MerBench：多模态情绪识别的统一评测基准。IEEE模式分析与机器智能汇刊，48(5)：5793–5810，2026年。1[24] Paul Pu Liang、Yiwei Lyu、Xiang Fan、Arav Agarwal、Yun Cheng、Louis-Philippe Morency 和 Ruslan Salakhutdinov。MultiZoo 与 MultiBench：用于多模态深度学习的标准化工具包。机器学习研究期刊，24(234)：1–7，2023年。1[25] Paul Pu Liang、Amir Zadeh 和 Louis-Philippe Morency。多模态机器学习基础与趋势：原则、挑战与开放问题。ACM计算调查，56(10)：1–42，2024年。1[26] Han Lin、Xiu Tang、Huan Li、Wenxue Cao、Sai Wu、Chang Yao、Lidan Shou 和 Gang Chen。T2DR：用于不完整多模态学习的双层抗缺失框架。在计算语言学协会论文集：ACL 2025 Findings，第8602–8616页，2025年。2[27] Ronghao Lin、Qiaolin He、Sijie Mai、Ying Zeng、Aolin Xiong、Li Huang、Yap-Peng Tan 和 Haifeng Hu。CYIN：用于连接完整与不完整多模态学习的循环信息潜在空间。神经信息处理系统进展，38：17605–17642，2026年。2[28] Ajian Liu、Zichang Tan、Jun Wan、Yanyan Liang、Zhen Lei、Guodong Guo 和 Stan Z Li。通过对抗性跨模态转换实现人脸反欺骗。《IEEE〈信息取证与安全汇刊〉》，16:2759–2772，2021年。1[29] Zhizhong Liu、Bin Zhou、Dianhui Chu、Yuhang Sun 和 Lingqiang Meng。在不确定缺失模态下基于模态转换的多模态情感分析。《信息融合》，101:101973，2024年。1[30] 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。6[31] Xiangyu Meng 和 Shicai Wei。通过动态质量感知实现通用不完整多模态学习。载于《欧洲计算机视觉会议》，第76–95页。施普林格，2026年。2[32] Yuchen Pan、Junjun Jiang、Kui Jiang 和 Xianming Liu。面向不完整多模态数据抑郁症识别的解耦多模态特权知识蒸馏。载于第32届ACM国际多媒体会议论文集，第5712–5721页，2024年。2[33]Adam Paszke、Sam Gross、Soumith Chintala、Gregory Chanan、Edward Yang、Zachary DeVito、Zeming Lin、Alban Desmaison、Luca Antiga 和 Adam Lerer。PyTorch中的自动微分。2017年。6[34] Jeffrey Pennington、Richard Socher 和 Christopher D Manning。GloVe：词表示全局向量。载于2014年自然语言处理实证方法会议（EMNLP）论文集，第1532–1543页，2014年。6 – 。[35] Hai Pham、Paul Pu Liang、Thomas Manzini、Louis-Philippe Morency 和 Barnab´as P´oczos。翻译中的发现：通过模态间循环转换学习鲁棒联合表征。载于AAAI人工智能会议论文集，第6892–6899页，2019年。1[36] Hai Pham、Paul Pu Liang、Thomas Manzini、Louis-Philippe Morency 和 Barnab´as P´oczos。翻译中的发现：通过模态间循环转换学习鲁棒联合表征。载于AAAI人工智能会议论文集，第6892–6899页，2019年。1、2[37] Jonathan Stroud、David Ross、Chen Sun、Jia Deng 和 Rahul Sukthankar。D3D：用于视频动作识别的蒸馏三维网络。载于IEEE/CVF冬季计算机视觉应用会议论文集，第625–634页，2020年。2。[38] Hu Wang、Yuanhong Chen、Congbo Ma、Jodie Avery、Louise Hull 和 Gustavo Carneiro。通过共享—特有特征建模实现缺失模态下的多模态学习。载于IEEE/CVF计算机视觉与模式识别会议论文集，第15878–15887页，2023年。1[39] Qi Wang、Liang Zhan、Paul Thompson 和 Jiayu Zhou。通过知识蒸馏实现不完整模态下的多模态学习。载于第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集，第1828–1838页，2020年。2[40] Xiaotao Wang、Yiyang Fang、Wenke Huang、Bin Yang、Guancheng Wan 和 Mang Ye。API：面向不完整多模态情感分析的自适应原型插补。载于

第43届国际机器学习会议论文集，第126107–126118页。PMLR，2026年。1、2、6、7、5[41] Yuanzhi Wang、Zhen Cui 和 Yong Li。不完整多模态学习中的分布一致模态恢复。载于2023年IEEE/CVF国际计算机视觉会议（ICCV），第21968–21977页。IEEE，2023年。1、2、6、7[42] Yuanzhi Wang、Yong Li 和 Zhen Cui。不完整多模态扩散情感识别。《神经信息处理系统进展》，36:17128，2023年，17117–。1、2、6、7。[43] Yan Wang、Fuyuan Cao 和 Xingwang Zhao。面向多模态情感分析的冲突感知自适应交叉重建。载于IEEE/CVF计算机视觉与模式识别会议论文集，第15880–15889页，2026年。2[44] Shicai Wei、Chunbo Luo 和 Yang Luo。Mmanet：面向不完整多模态学习的边际感知蒸馏与模态感知正则化。载于2023年IEEE/CVF计算机视觉与模式识别会议（CVPR），第20039–20049页。IEEE，2023年。1、2[45] Huangbiao Xu、Huanqi Wu、Xiao Ke、Junyi Wu、Rui Xu 和 Jinglin Xu。Mcmoe：利用专家混合补全不完整多模态动作质量评估中的缺失模态。载于AAAI人工智能会议论文集，第11241–11249页，2026年。2[46] Wenxin Xu、Hexin Jiang 和 Xuefeng Liang。利用模态专家知识进行不完整多模态学习。载于第32届ACM国际多媒体会议论文集，第438–446页，2024年。[47] Yingxue Xu、Fengtao Zhou、Chenyu Zhao、Yihui Wang、Can Yang 和 Hao Chen。面向不完整多模态生存预测的蒸馏提示学习。载于2025年IEEE/CVF计算机视觉与模式识别会议（CVPR），第5102–5111页。IEEE，2025年。2[48] Jie Yang、Junxiong Zhang、Kun Qian、Qingyu Yang、Weikai Li 和 Zhen Cheng。通过置信度感知多模态融合框架弥合药物-靶标相互作用预测中的模态可靠性差距。载于AAAI人工智能会议论文集，第27529–27537页，2026年。2 –。[49] Zhendong Yang、Ailing Zeng、Zhe Li、Tianke Zhang、Chun Yuan 和 Yu Li。从知识蒸馏到自知识蒸馏：一种结合归一化损失与定制软标签的统一方法。载于IEEE/CVF国际计算机视觉会议论文集，第17185–17194页，2023年。2[50] Ziqi Yuan、Wei Li、Hua Xu 和 Wenmeng Yu。用于鲁棒多模态情感分析的基于Transformer的特征重建网络。载于第29届ACM国际多媒体会议论文集，第4400–4407页，2021年。1[51] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency。MOSI：用于在线观点视频中情感强度与主观性分析的多模态语料库。arXiv预印本，arXiv:1606.06259，2016年。1、6

[52] AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency。在开放环境中的多模态语言分析：CMU-MOSEI 数据集与可解释的动态融合图。载于《计算语言学协会第56届年会论文集》（第1卷：长篇论文），第2236–2246页，2018年。6[53] 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征，用于多模态情感分析。载于《2023年自然语言处理经验方法会议论文集》，第756–767页，2023年。2、6、7[54] Haoyu Zhang、Wenbin Wang 和 Tianshu Yu。面向不完整数据的鲁棒多模态情感分析。神经信息处理系统进展，37：55943–55974，2024年。2、6、7[55] Jinming Zhao、Ruichen Li 和 Qin Jin。用于情绪识别的缺失模态想象网络：应对不确定的缺失模态。载于《计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集》（第1卷：长篇论文），第2608–2618页，2021年。1、2[56] Aoqiang Zhu、Min Hu、Xiaohua Wang、Jiaoyun Yang、Yiming Tang 和 Ning An。代理驱动的鲁棒多模态情感分析与不完整数据。载于《计算语言学协会第63届年会论文集》（第1卷：长篇论文），第22138–22123页，2025年。6、7[57] Yan Zhuang、Minhao Liu、Wei Bai、Yanru Zhang、Xiaoyue Zhang、Jiawen Deng 和 Fuji Ren。CMAD：面向相关性与模态感知的蒸馏，用于缺失模态下的多模态情感分析。载于《2025年 IEEE/CVF 国际计算机视觉大会》（ICCV），第4626–4636页。IEEE，2025年。2、6、7、5[58] Yan Zhuang、Minhao Liu、Wei Bai、Yanru Zhang、Wei Li、Jiawen Deng 和 Fuji Ren。面向缺失模态多模态情感分析的超模态增强。神经信息处理系统进展，38：142100–142129，2026年。2、6、7、5[59] Yan Zhuang、Minhao Liu、Yanru Zhang、Jiawen Deng 和 Fuji Ren。TMDC：面向缺失和噪声模态的多模态情感分析的两阶段模态去噪与补全框架。载于 AAAI人工智能会议论文集，第2281–2289页，2026年。1[60] Yan Zhuang、Yanru Zhang、Jiawen Deng 和 Fuji Ren。面向缺失模态多模态情感分析的样本内与模态内增强。IEEE多媒体汇刊，28：1847–1859，2026年。1、2[61] Shihao Zou、Wei Wei 和 Yongshuo Zhang。DEAR：面向缺失模态鲁棒多模态情感分析的分布误差感知可靠性。载于《计算语言学协会论文集：ACL 2026》的研究成果，第30344–30361页，2026年。2、6、7

MSRGF：参考引导融合的多源重建，用于不完整多模态情感分析

补充材料

6. 方法补充细节

本节介绍 MSRGF 的实现细节，说明从掩码多模态输入到最终预测的计算过程。缺失情况协议决定了逐源重建（SWR）可用的观测特征，其目标估计随后通过参考引导融合（RGF）进行组合。全局上下文调制（GCM）继而作用于增强后的模态表示，而本节末尾的网络配置则规定了相应的维度和参数共享规则。

6.1. 输入特征与缺失情况

MSRGF 处理对齐的文本、音频和视觉特征，其中观测掩码指定了 SWR 使用的输入，以及 RGF 融合过程中保留的位置。CMU-MOSI 的输入维度（d_T，d_A，d_V）为（1024，5，20），CMU-MOSEI 为（1024，74，35），IEMOCAP 为（300，74，35）。共同的隐藏维度为 $d = 128$。CMU-MOSI 和 CMU-MOSEI 使用文本特征；被选中进行掩码的词将被替换为[UNK]，然后进行上下文编码，其词级特征将从观测输入中排除。IEMOCAP 对全部三种模态使用预提取特征。完整特征 $X_m^\star$ 在训练期间提供重建监督，而预测网络接收 X_m^{obs}、M_m 和 P_m。

为区分人为掩码与缺乏有效监督的位置，令 A_m 表示掩码前存在可用重建目标的位置。结合主论文中定义的观测掩码 M_m 和有效性掩码 P_m，满足 $M_m \le A_m \le P_m$。因此，重建监督应用于：

$$\Omega_m = A_m \odot (1 - M_m). \tag{21}$$

对于受控缺失协议，$A_m = P_m$，因此 $\Omega_m = P_m \odot (1 - M_m)$。即使某个模态完全缺失，其对齐序列长度和有效性掩码仍会保留，使 SWR 能够在与完整序列相同的时间位置构建目标估计。

固定模态设置保留 $\{T, A, V\}$ 的每个非空子集。对于随机整模态缺失，会完整移除一个模态，同时确保每个样本至少保留一个模态。我们以 10% 为步长评估从 10% 到 70% 的缺失率；在三个模态的情况下，可达到的最大缺失率为 $2/3$，该缺失率用于 70% 条件。对于随机模态内缺失，时间位置在有效序列中按 10% 到 90% 的缺失率进行随机遮蔽。

6.2. 源级重建

根据掩码确定可用模态后，SWR 会针对每个观测源 $n \in \mathcal{O}$，分别构建目标模态 m 的独立估计 R_m^n。每条路径都将目标专属查询与该源的编码观测相结合，使共享生成器能够处理不同的源-目标对。具体而言，首先通过模态专属线性投影和层归一化（合称 ϕ_n），将观测特征映射至 Z_n；随后，由单层时间 Transformer 将其编码为 H_n。该 Transformer 使用四个注意力头、带 GELU 的 $128 \to 256 \to 128$ 前馈网络，以及 0.1 的 Dropout。其注意力会读取由 M_n 标记的位置；若某个模态完全缺失，则得到 $H_n = \mathbf{0}$。

为了在每个目标位置生成估计值，SWR 根据主论文中定义的位置、模态和观测状态嵌入构造 $q_m^0(t)$。目标专属的提示词库 $U_m \in \mathbb{R}^{4\times 128}$ 进一步通过对四个学习向量进行注意力计算来调整此查询：

$$\begin{aligned} p_m(t) &= \mathrm{softmax}\!\left(\frac{q_m^0(t) U_m^\top}{\sqrt{d}}\right) U_m, \\ Q_m(t) &= P_m(t) f_Q\big(q_m^0(t) + p_m(t)\big). \end{aligned} \tag{22}$$

这里，f_Q 是共享的 $128 \to 128 \to 128$ GELU 多层感知器，它将提示调整后的查询映射为 Q_m，随后在 SWR 生成器和 RGF 的上下文基础构建中重复使用。学习的位置嵌入表覆盖 512 个位置，而模态嵌入和观测状态嵌入分别包含三个和两个条目。

对于每个 $n \in \mathcal{O}$，共享生成器使用Q_m读取 H_n。它包含两个连续的交叉注意力块。对于查询特征 q、编码特征 h和观测掩码M，每个模块计算：

$$\begin{aligned} J &= \mathrm{MHA}\left(\mathrm{LN}_q(q), \mathrm{LN}_h(h), \mathrm{LN}_h(h); M\right), \\ x &= q + \mathrm{Dropout}(J), \\ q' &= x + \mathrm{Dropout}\left(\mathrm{FFN}(\mathrm{LN}(x))\right). \end{aligned} \tag{23}$$

每个交叉注意力块都为查询和观测特征使用独立的归一化层，包含四个注意力头，以及一个宽度为 256 的 GELU 前馈网络。尽管两个块的参数彼此独立，但都会在源-目标对之间重复使用。如果某次交叉注意力

调用没有接收到观测键，其注意力更新便为零，而残差和前馈操作仍然有定义；不含观测位置的源路径会从重建估计集合中排除。

生成器输出经由一个模态特定的两层 GELU 多层感知机 f_m^{tar}进行映射，其宽度为 $128 \to 128 \to 128$。生成器在各模态对之间共享，并且所有可用模态都使用同一个输出头来估计固定的 m。模态嵌入会添加到生成器的查询和观测特征输入中。所得独立估计 $R_m^n \in \mathbb{R}^{L\times 128}$ 与目标的时间位置共享。对于每种模态，一个独立的线性解码器 D_m： $\mathbb{R}^{128} \to \mathbb{R}^{d_m}$（带偏置）由其源级估计和最终增强表示共享。在重构监督以及 RGF 中的观测特征比较中，均使用同一个解码器；当推理阶段定义了该比较时同样如此。

6.3. 参考引导融合

SWR 生成的独立估计为目标模态提供了多个候选表示，但必须结合观测输入及各候选表示之间的关系来确定它们的贡献。RGF 首先构建观测条件化基表示 B_m 和关系描述符ξ_m^n，以引导学习得到的参考值 C_m。随后，它将各独立估计相对于该参考值$\tilde{}$的偏差转换为 N_m，再将这两项集成到 H_m中。相同的观测掩码控制该集成在观测位置和缺失位置上的行为。

基础表示。 RGF 从基础表示 B_m 开始，它在目标观测存在的位置保留来自Z_m 的直接贡献。一个与 SWR 生成器分别参数化的交叉注意力块，使用目标查询 Q_m读取 H_m；其输出经过共享的 $128 \to 128 \to 128$ GELU 多层感知器 f_{anc}，得到 a_m。在 $\eta_a = 0.1$ 的条件下，基础表示定义如下：

$$B_m(t) = \begin{cases} Z_m(t) + \eta_a a_m(t), & M_m(t) = 1, \\ a_m(t), & M_m(t) = 0,\ P_m(t) = 1, \\ \mathbf{0}, & P_m(t) = 0. \end{cases} \tag{24}$$

这个定义使观测位置能够直接从Z_m 获得信息，同时为缺失位置提供上下文基底。当目标模态完全缺失时，a_m 通过基础块的查询残差和前馈运算获得。由于 B_m 也会输入参考评分器、细化评分器和整合门控，因此所有重构路径都基于共同的目标条件表示进行评估。

估计之间的关系。 仅使用 B_m 无法区分由不同源生成的估计。因此，RGF通过三种互补比较构建ξ_m^n： R_m^n 与其源 n之间的关系、其相对于目标的其他估计的变化，以及其与可用的保留目标观测之间的一致性。

三个指示器用于指定哪些比较有效：

$$\begin{aligned} v_{m,n}^{\mathrm{tar}} &= \mathbb{I}[\|M_m\|_1 > 0,\ n \neq m], \\ v_n^{\mathrm{src}} &= \mathbb{I}[n \in \mathcal{O}], \\ v_n^{\mathrm{div}} &= \mathbb{I}[|\mathcal{O}| > 1]. \end{aligned} \tag{25}$$

此处$v_{m,\,n}^{\mathrm{tar}}$有助于与保留的目标观测达成一致，v_n^{src}用于识别可用的源槽位，而 v_n^{div}则允许对不同的源估计进行比较。由于仅针对 $n \in \mathcal{O}$计算描述符，源指示标记在有效路径上等于 1，但为适配固定槽位实现而予以保留。源–估计交互从掩码池化表示开始：

$$\bar{R}_m^n = \mathrm{Pool}(R_m^n, P_m), \quad \bar{H}_n = \mathrm{Pool}(H_n, M_n), \tag{26}$$

其中：

$$\mathrm{Pool}(X, W) = \sum_t W(t)X(t) / \max(\sum_t W(t), 1). \tag{27}$$

随后，将这些源摘要和目标摘要投影到64维，并与学习得到的模态嵌入相结合：

$$\begin{aligned} a_m^n &= \psi_m(\bar{R}_m^n) + e_m^{\mathrm{sem}}, \\ b_n &= \psi_n(\bar{H}_n) + e_n^{\mathrm{sem}}. \end{aligned} \tag{28}$$

共享多层感知机对它们的值、逐元素乘积和绝对差进行处理，以得到：

$$\mathbf{c}_{m,n}^{\mathrm{src}} = f_{\mathrm{src}}\big([a_m^n, b_n, a_m^n \odot b_n, |a_m^n - b_n|]\big). \tag{29}$$

网络 f_{src} 的宽度为 $256 \to 64 \to 16$，生成一个16维交互特征。

第二种比较衡量R_m^n 与模态 m的其他估计之间的差异。相应的分数为：

$$c_{m,n}^{\mathrm{div}} = v_n^{\mathrm{div}}\ \mathrm{sg}\left[\frac{1}{\max(|\mathcal{O}| - 1, 1)} \cdot \sum_{j\in\mathcal{O}\setminus\{n\}} \big(1 - \cos(\bar{R}_m^n, \bar{R}_m^j)\big)\right]. \tag{30}$$

其中余弦 表示余弦相似度，sg 表示阻止梯度通过该分数传播，可用性指示器定义见方程（25）。此比较使用同一 m的不同可用模态所得到的估计。

当模态 m包含观测位置且 $n \neq m$ 时，第三项比较衡量解码估计与观测特征之间的一致性：

$$c_{m,n}^{\mathrm{tar}} = v_{m,n}^{\mathrm{tar}}\ \mathrm{sg}\left[\frac{\sum_t M_m(t) \cos\big(D_m(R_m^n(t)), X_m^{\mathrm{obs}}(t)\big)}{\max(\sum_t M_m(t), 1)}\right]. \tag{31}$$

可用性条件见方程（25）；余弦得分的计算不涉及梯度。观测覆盖率进一步描述了各模态中可用信息的数量：

$$\rho_m = \frac{\|M_m\|_1}{\max(\|P_m\|_1, 1)}. \tag{32}$$

结合这些测量值及其有效性指标，可得：

$$\begin{aligned}\xi_m^n = \big[&c_{m,n}^{\mathrm{tar}}, \mathbf{c}_{m,n}^{\mathrm{src}}, c_{m,n}^{\mathrm{div}}, \rho_m, \rho_n,\\ &v_{m,n}^{\mathrm{tar}}, v_n^{\mathrm{src}}, v_n^{\mathrm{div}}\big] \in \mathbb{R}^{23},\end{aligned} \tag{33}$$

由此得到的 23 维 ξ_m^n 结合了 16 维源–估计交互、两个比较分数、观测覆盖率以及有效性指标。每个可用源仅计算一次，并在所有目标位置间共享，因此后续按位置加权的网络可以同时以局部特征和源级关系为条件。

参考与细化权重。 为了将源级比较转化为按位置的参考表示，参考评分器将 $R_m^n(t)$ 与 $B_m(t)$ 及 ξ_m^n 一同作为输入。组装好参考表示后，特征级细化门控还会接收 $C_m(t)$，从而能够依据共享参考表示来处理源特定差异。相应的输入为：

$$\begin{aligned}I_m^n(t) &= [R_m^n(t), B_m(t), \xi_m^n] \in \mathbb{R}^{279},\\ J_m^n(t) &= [C_m(t), B_m(t), R_m^n(t), \xi_m^n] \in \mathbb{R}^{407}.\end{aligned} \tag{34}$$

参考权重 α_m^n 由 I_m^n 通过一个共享的 $279 \to 128 \to 1$ GELU 多层感知器计算，并在可用模态上归一化。对 R_m^n 加权求和后得到 C_m，见方程(9)。细化评分器使用一个宽度相同的独立多层感知机，从 I_m^n中得到 β_m^n；与此同时，一个独立的共享 $407 \to 128 \to 128$ GELU 多层感知器根据 J_m^n 生成 G_m^n 特征级门控，并采用 sigmoid 激活。因此，评分器决定每个可用模态的贡献程度，而门控则控制相应的特征维度。

为在共享各路径评分网络的同时保留模态身份，输入中包含学习嵌入$e_m^{\alpha} \in \mathbb{R}^{279}$、 $e_{m,\ n}^{\beta} \in \mathbb{R}^{279}$和$e_{m,\ n}^{G} \in \mathbb{R}^{407}$，其索引分别对应目标或源–目标对。

一旦 C_m 形成后，会将每个估计值通过 $\Delta_m^n = R_m^n - C_m$与参考表示进行比较。共享的无偏置 $128 \to 256 \to 128$ GELU 多层感知器 F_{diff} 会在应用源权重 β_m^n 和特征门控 G_m^n 之前对每个差值进行变换。其两个线性层均无偏置，因此有 $F_{\mathrm{diff}}(\mathbf{0}) = \mathbf{0}$。随后将加权后的变换差值求和，得到 N_m 方程（11）中的 C_m 定义了公共参考表示，而 N_m 结合了最终整合中使用的源特定差异更新。

最终整合。参考表示和细化结果结合观测掩码，与基础表示进行整合。RGF 首先将这些关系概括为$\xi_m(t) = \sum_{n\in\mathcal{O}} \alpha_m^n(t)\,\xi_m^n$，因此用于构建 C_m 的信息也会进入整合门控。共享的$408 \to 128 \to 256$ GELU 多层感知器后接 Sigmoid 接收 $[B_m, C_m, N_m, \xi_m, M_m]$，并生成两个 128 维的门控 r_m^{rec}和 r_m^{enh}。根据方程（M_m 13），通过对它们的贡献进行调整，其中。这使得模型能够在每个位置分别控制参考更新和基于差异的细化。

将依赖观测的门控代入 RGF 积分规则，可得观测位置处的以下表达式：

$$\begin{aligned}\widetilde{H}_m(t) = Z_m(t) &+ \eta_a a_m(t)\\ &+ \eta_g r_m^{\mathrm{rec}}(t) \odot (C_m(t) - B_m(t))\\ &+ \eta_g r_m^{\mathrm{enh}}(t) \odot N_m(t).\end{aligned} \tag{35}$$

在有效缺失位置处，参考更新的权重为单位权重，因此表达式可简化为：

$$\widetilde{H}_m(t) = C_m(t) + r_m^{\mathrm{enh}}(t) \odot N_m(t). \tag{36}$$

应用 P_m后，填充位置仍为零。这两个表达式揭示了 RGF 的核心区别：观测特征通过 B_m 得以保留，并进行受控更新；而缺失位置则由学习得到的参考值和变换后的差值来表示。两条路径使用相同的源关系和加权网络。

可用模态。集合 $\mathcal{O}$ 决定了参考阶段和精炼阶段所用的估计值。当只有一个模态 n 可用时，其归一化参考权重为一，因此得到 $C_m = R_m^n$ 和 $N_m = \mathbf{0}$。当有两个或三个可用模态时，每个估计值都通过其学习得到的权重参与 C_m，而差值 $R_m^n - C_m$ 则构成 N_m的输入。例如，当文本和音频被观测到而视觉完全缺失时，R_V^T 和 R_V^A 会合并为 C_V；它们变换后的差值随后构成 N_V，并且这两个表征通过缺失位置门控生成$\widetilde{H}_V$。相同的计算也适用于部分观测模态，其中 B_m 还会保留一条从其观测特征出发的直接路径。

6.4. 全局上下文调制

GCM 通过将模态摘要与共享全局上下文相结合，把 RGF 生成的增强序列转换为样本级预测。它首先对每个 H_m 的有效位置 $\widetilde{}$ 进行池化，然后利用跨模态上下文在预测前调整模态特定摘要。具体而言，每个模态都有一个经过学习的池化查询 q_m^{pool}，该查询$\widetilde{}$通过四头注意力读取 H_m 的有效位置。

对于模态 m，池化操作为：

$$z_m = \mathrm{MHA}_m\left(q_m^{\mathrm{pool}}, \mathrm{LN}_m(\widetilde{H}_m), \mathrm{LN}_m(\widetilde{H}_m); P_m\right),$$
$$v_m = \Pi_m(W_m^{\mathrm{pool}} z_m + b_m^{\mathrm{pool}}). \tag{37}$$

$W_{\text{池化}_m}$ $z_m + b_m^{\mathrm{pool}}$和 Π_m都是将 128 维投影到 128 维的模态特定线性投影，使用不同的参数； Π_m 将池化输出映射到共享融合空间。一个独立的四头注意力层通过学习到的查询 q_c读取三个摘要 $[v_T, v_A, v_V]$，以生成全局上下文 c；后续的多层感知机 f_c 的宽度为 $128 \to 128 \to 128$，并使用 GELU。

对于每种模态， $s_m = W_m^{\mathrm{back}} c$ 将全局上下文投影到与 v_m 相同的宽度。共享多层感知机 f_γ 将 $[v_m, s_m, v_m - s_m, v_m \odot s_m]$ 与学习到的模态嵌入 $e_m^\gamma \in \mathbb{R}^{512}$ 结合，用于计算 γ_m。该共享网络的宽度为 $512 \to 128 \to 128$，后接 sigmoid，其最终偏置初始化为 -2。预测多层感知机 f_{pred} 接收 $[c, u_T, u_A, u_V]$，并使用宽度为 $512 \to 256 \to 128 \to d_y$的线性层，同时采用 GELU 和 Dropout 0。1。这里， $d_y = 1$对应两个情感数据集， $d_y = 8$ 对应 IEMOCAP 的四个双 logit 情绪输出。

6.5. 网络配置与参数共享

表5 汇总了上述介绍的架构配置，包括注意力和前馈组件的维度，以及源目标路径之间共享的参数。所有隐藏序列均使用 $d = 128$；当输入空间和输出空间不同时，则保留模态特定投影和输出映射。

7. 训练细节

训练目标将任务监督与针对各个 SWR 估计以及 RGF 生成的增强表征的损失相结合。由于这些量涉及特征宽度和观测数量各不相同的模态，我们先介绍辅助损失的归一化方式，再追踪其在网络中的梯度，并说明优化计划。

7.1. 损失归一化

视图、重建和保持目标作用于特征维度不同、以及观测位置或缺失位置集合大小不同的特征。在对参与计算的样本-模态对取平均之前，我们先分别在特征维度和符合条件的时间位置上对每个误差进行归一化。对于样本 i、模态 m和位置掩码 W_{im}，我们首先在特征维度 d_U 上对绝对差取平均，然后再对选定的时间位置取平均：

$$\ell_{im}(U, V; W) = \frac{\sum_t W_{im}(t)\, \|U_{im}(t) - V_{im}(t)\|_1 / d_U}{\max(\sum_t W_{im}(t), 1)}. \tag{38}$$

令 $\mathcal{E}_\Omega = \{(i,m) : \|\Omega_{im}\|_1 > 0\}$ 和 $\mathcal{E}_M = \{(i,m) : \|M_{im}\|_1 > 0\}$ 分别表示参与重建和保持的样本-模态对。两个重构损失遵循由 SWR 和 RGF 生成的表征：

$$\mathcal{L}_{\mathrm{view}} = \frac{1}{|\mathcal{E}_\Omega|} \sum_{(i,m)\in\mathcal{E}_\Omega} \frac{1}{|\mathcal{O}_i|} \cdot \sum_{n\in\mathcal{O}_i} \ell_{im}\big(D_m(R_{im}^n), X_{im}^\star; \Omega_{im}\big). \tag{39}$$

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{E}_\Omega|} \sum_{(i,m)\in\mathcal{E}_\Omega} \ell_{im}\big(D_m(\widetilde{H}_{im}), X_{im}^\star; \Omega_{im}\big). \tag{40}$$

对于 $\mathcal{L}_{\mathrm{view}}$，先在每个样本-模态对内对可用模态的误差取均值，再计算最终均值。保持损失使用模型观测到的位置：

$$\mathcal{L}_{\mathrm{pres}} = \frac{1}{|\mathcal{E}_M|} \sum_{(i,m)\in\mathcal{E}_M} \ell_{im}\big(\widetilde{H}_{im}, Z_{im}; M_{im}\big). \tag{41}$$

当任一符合条件的集合为空时，相应的损失设为零。投影特征 Z_m 在保持损失中仍可训练。情感损失和总目标函数遵循主论文第3.6 节。

7.2. 情感监督与梯度流

辅助目标在处理流程的不同位置监督重建和保持，而任务损失则训练由此产生的表示，以用于情感或情绪预测。在 CMU-MOSI 和 CMU-MOSEI 上，模型使用 L_1 对原始情感分数进行回归。IEMOCAP 使用四个二元情绪预测的平均交叉熵，每种情绪对应两个逻辑值。这些目标监督来自 GCM 的最终预测，其梯度会经过 RGF 中的增强表示以及 SWR 中的估计，传递至活跃的输入编码器。

辅助损失在同一计算流程的不同位置提供监督。 $\mathcal{L}_{\mathrm{view}}$ 通过各自的估计 R_m^n更新可用的重建路径和解码器 D_m。 $\mathcal{L}_{\mathrm{rec}}$ 监督增强表示~H_m，梯度经由 RGF 的权重和门控、SWR 的估计以及共享解码器传递。 $\mathcal{L}_{\mathrm{pres}}$ 作用于 H_m 的观测位置以及可训练的投影特征 $\widetilde{Z_m}$。因此，组合目标函数连接了源级估计、最终表示构建和情感预测。

在关系描述符中，梯度会通过学习得到的源–估计交互网络传播。用于衡量观测特征一致性和汇聚估计值之间差异的余弦分数采用停止梯度操作 sg(·)。这些值仍会输入后续的评分

表5. MSRGF中的网络配置与参数共享。

模块	网络	配置与参数共享
输入	投影与时序编码器	$d_m \to 128$ +层归一化；每个模态使用一个 4 头 Transformer
SWR	提示词库和查询网络	每个模态使用四个 128 维提示词；共享 $128 \to 128 \to 128$ MLP
SWR	重建生成器	两个 4 头交叉注意力块，每个块均带有 $128 \to 256 \to 128$ 前馈网络；在各路径间共享
SWR	输出多层感知机和解码器	三个模态专属的 $128 \to 128 \to 128$ GELU 多层感知器；三个模态专属的线性 $128 \to d_m$ 解码器（带偏置），在 RGF 中复用
RGF	基础表示	一个共享的交叉注意力块和 $128 \to 128 \to 128$多层感知机
RGF	关系比较	三个模态专属线性 $128 \to 64$ 投影；共享 $256 \to 64 \to$ 16 个 GELU 交互多层感知机；二值可用性指示器
RGF	参考与细化权重	两个独立的共享 $279 \to 128 \to 1$评分多层感知机
RGF	特征门和差分网络	共享 $407 \to 128 \to 128$ 门控；无偏置的 $128 \to 256 \to 128$ 多层感知机
RGF	集成	共享的 $408 \to 128 \to 256$多层感知机；$\eta_a = \eta_g = 0.1$
GCM	模态与全局注意力	三个模态特定4头池化器和全局4头注意力
GCM	模态投影	每种模态：独立的 $W_{\text{池化}m}$和 $\Pi_m(128 \to 128)$，以及无偏置的 W_m^{back} $(128 \to 128)$
GCM	上下文、门控和预测多层感知器	f_c：$128 \to 128 \to 128$ GELU；共享 f_γ：$512 \to 128 \to 128$ GELU + Sigmoid；f_{pred}：$512 \to 256 \to 128 \to d_y$ GELU +Dropout

网络；这些网络与差异变换和整合门控一同保持可训练。解码器还提供用于计算观测特征一致性得分的特征，并直接接收来自重建目标的梯度。

7.3. 实现细节

输入特征和缺失协议详见第 6.1 节，SWR、RGF 和 GCM 架构详见第 6.2–6.4 节，表 5。此处介绍其余训练和评估设置。

优化。我们使用 AdamW，(β_1, β_2) =(0.9, 0.999)，并采用 $\epsilon = 10^{-8}$。符合条件的矩阵参数的权重衰减为 0.01，偏置和归一化参数的权重衰减为 0。MSRGF 的默认学习率为 10^{-4}；文本编码器在 CMU-MOSI 和 CMU-MOSEI 上的默认学习率分别为 6×10^{-6}和 3×10^{-6}。默认批量大小为 32。每个优化阶段先进行 10% 的预热，然后线性衰减学习率，最低乘数为 0.1，并将全局梯度范数裁剪设为 1.0。主论文方程 20 中的默认辅助损失系数为 $(\lambda_{rec}, \lambda_{view}, \lambda_{pres})$ =(0.1, 0.1, 0.05)。

训练计划。 在 CMU-MOSI 和 CMU-MOSEI 上，文本编码器先冻结两个适配轮次，随后与 MSRGF 联合优化，最多训练 40 个轮次。阶段转换时会初始化一个新的优化器。编码器采用 DeBERTa-v3-large，词级特征通过对其子词表示取平均获得。在联合训练期间，编码器保持评估模式，以禁用 Dropout，同时保留梯度。

评估。 对于 CMU-MOSI 和 CMU-MOSEI，二分类准确率和加权 F1 分数按照先前研究的做法排除取值为零的情感标签 [57, 58],，这些标签仍用于回归训练；对于 IEMOCAP，准确率和加权 F1 分数按照 [57, 58] 的做法取四个二元情绪任务的平均值。

8. 补充实验结果

本节通过报告每个评估随机缺失率下的准确率/F1值，扩展了比较范围。随机整体模态缺失会改变 SWR 可用于构建估计的源，而随机模态内缺失则会改变提供给源编码器的时间观测。结果按两种协议分别呈现，平均值为所示各缺失率下的算术平均值；下一节将进一步考察导致这些预测的模型组件。

8.1. 随机整模态缺失

表6 评估了 CMU-MOSEI 和 IEMOCAP 上的随机整模态缺失情况：一个或多个模态可能完全不可用，但每个样本至少保留一个源。报告的七种缺失率涵盖 10%–70%，其中最高设置上限为三种模态下可达到的缺失率。

在 CMU-MOSEI 上，MSRGF 的平均准确率/F1 为 79.0/78.3，而 API [40],为 78.4/77.8；此外，MSRGF 在全部七种缺失率下的准确率均领先。在缺失情况为 50% 时，其准确率仍为 76.6%；在上限为 70% 的设置下，准确率为 72.7%。

表 6. 随机整模态缺失（准确率/F1值，%）。70%的设置是对2/3的近似。最佳和次佳结果分别以粗体和下划线标示；* 表示来自 HME 的结果 [58]。

方法	10%	20%	30%	40%	50%	60%	70%	Avg
CMU-MOSEI								
DiCMoR* [41]	83.2/83.1	81.2/80.8	79.0/77.9	76.9/75.2	73.7/71.7	71.1/70.4	70.6/68.9	76.5/75.4
IMDer* [42]	83.5/83.3	81.3/80.9	79.4/78.8	76.1/74.3	75.3/72.4	71.2/66.5	71.1/65.9	76.8/74.6
LNLN* [54]	83.1/83.0	81.7/81.0	79.1/78.1	76.9/75.6	75.1/72.4	72.5/68.6	70.6/65.2	77.0/74.8
MPLMM* [8]	83.7/83.5	81.4/80.8	79.2/78.3	77.1/76.1	75.6/73.8	72.7/70.6	70.5/69.2	77.2/76.0
HME* [58]	84.0/83.8	82.3/81.9	80.0/79.6	77.6/76.4	75.7/74.7	73.0/72.5	72.4/69.2	77.9/76.9
API [40]	84.1/84.1	82.9/82.8	80.9/81.1	78.5/78.3	75.6/75.3	74.2/**73.7**	72.4/69.3	78.4/77.8
MSRGF	**84.9/84.7**	**83.2/83.0**	**81.6/81.3**	**79.2/78.5**	**76.6/75.9**	**74.7**/73.5	**72.7/71.0**	**79.0/78.3**
IEMOCAP								
DiCMoR* [41]	77.1/73.7	76.4/72.5	75.8/71.6	75.5/69.0	75.1/68.5	74.5/68.0	74.3/67.5	75.5/70.1
IMDer* [42]	77.2/72.8	75.9/71.6	75.7/70.8	74.9/70.8	74.7/68.2	73.5/67.0	73.1/65.3	75.0/69.5
LNLN* [54]	78.0/76.3	77.6/75.7	77.5/75.4	77.4/74.0	76.8/72.9	76.6/72.6	76.5/70.2	77.2/73.9
MPLMM* [8]	76.1/70.8	75.7/67.9	75.6/67.4	75.5/66.6	74.0/66.2	73.2/66.0	72.4/65.7	74.6/67.2
HME* [58]	80.9/80.3	80.7/80.0	79.0/77.8	78.7/77.3	78.5/77.1	77.8/**76.1**	77.6/76.0	79.0/77.8
MSRGF	**81.2/80.5**	**80.9/80.2**	**80.0/79.4**	**79.1/77.7**	**79.5/77.6**	**78.2**/75.5	**78.5/76.8**	**79.6/78.2**

在 IEMOCAP 上，MSRGF 的平均准确率/F1 达到 79.6/78.2，较 HME [58]提高了 0.6/0.4 个百分点。

8.2. 随机模态内缺失

为测试在其他可用模态中的不完整观测，表 7报告了 CMU-MOSI 和 CMU-MOSEI 在 10% 至 90% 缺失率下相应的准确率/F1值结果。不同于整模态移除，该协议保留了分散的时间观测，这些观测仍可供 SWR 和 RGF 中的观测条件融合使用。

在九种模态内缺失率下，MSRGF 在 CMU-MOSI 上的平均准确率/F1值为 74.1/74.1，在 CMU-MOSEI 上为 80.5/80.2，分别比最强基线的平均准确率高出 1.7 和 1.9 个百分点。即使在高缺失情况下，性能仍有所提升：在 90% 掩码率下，MOSI 的准确率达到 64.2%，MOSEI 达到 70.1%。由于每个模态中仍保留了一些观测位置，该实验补充了整模态移除实验，评估源编码器和参考引导整合如何应对日益稀疏的时间输入。

9. 补充消融研究与机制分析

上述性能比较确立了 MSRGF 在不同缺失协议下的行为；剩余的问题是，其各项计算如何促成这些结果。

9.1. 重新训练的设计消融

我们首先考察 SWR 是否能生成有用的源条件估计，RGF 是否能改善其集成，以及 GCM 是否有助于情感预测。每种修改后的架构或目标都在 IEMOCAP 上独立重新训练，其中 Acc/F1 分别在七种固定模态组合和七种随机整体模态缺失率上取平均。除主要组件外，表8 还评估了参考构建、关系建模、差异细化、观测特征整合、上下文调整和辅助目标的内部设计。

主要组件与目标表示。 移除 SWR 后，重建视图将被绕过，并使用 $\sim H_m = B_m$，同时禁用 $\mathcal{L}_{\text{view}}$ 和 $\mathcal{L}_{\text{rec}}$。将 RGF 替换为均匀补全后，观测位置处的 Z_m 保持不变，而缺失位置处则对可用的源估计取平均。移除 GCM 后，模态池化和预测网络仍予以保留，但设置 $c = \mathbf{0}$ 和 $u_m = v_m$。在 Random 条件下，这些变体的 F1 分数分别为 75.17、75.27 和 75.63，而 MSRGF 为 78.25；在 Fixed 条件下也有类似的性能下降。我们进一步考察目标表示，方法是移除目标查询（$Q_m = \mathbf{0}$）或将上下文基底替换为 $B_m = Z_m$。这两种修改都会降低性能，表明目标特定查询和上下文基底有助于后续的重建与融合。

重建源组织。我们进一步考察，与单源重建或

表7。 随机模态内缺失（准确率/F1值，%）。最佳和次佳结果分别以粗体和下划线标出。

方法	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg
CMU-MOSI										
ALMT [53]	80.9/81.3	77.6/77.9	75.2/75.5	73.1/73.9	68.2/69.7	64.5/66.8	61.8/65.3	60.4/**65.5**	57.3/**64.9**	68.8/71.2
LNLN [54]	81.2/81.6	79.2/79.5	77.3/77.6	76.0/76.3	73.4/73.7	69.0/69.2	66.0/66.0	62.8/62.6	56.5/56.3	71.3/71.4
P-RMF [56]	82.6/82.9	79.6/81.0	76.8/79.3	75.5/77.7	73.0/73.3	68.8/68.6	66.2/64.7	62.0/60.8	59.5/56.7	71.6/71.7
DEAR [61]	80.6/82.5	78.6/80.4	77.5/78.4	74.8/76.1	72.9/74.1	68.0/**69.8**	64.9/67.1	62.4/63.4	59.1/60.4	71.0/72.5
SemMSA [20]	84.9/84.7	81.6/81.4	78.8/78.7	**78.0/77.9**	**75.6/75.5**	69.0/69.2	64.8/65.0	62.8/63.0	56.0/55.2	72.4/72.3
TF-Mamba [21]	81.9/81.9	80.5/80.6	77.7/77.9	76.1/76.2	75.0/75.2	68.6/68.8	67.2/**67.4**	63.6/63.8	60.4/60.6	72.3/72.5
MSRGF	**87.8/87.8**	**82.2/82.2**	**81.6/81.6**	76.4/76.4	73.8/73.9	**69.2**/69.4	**67.4/67.4**	**64.2**/64.3	**64.2**/64.2	**74.1/74.1**
CMU-MOSEI										
ALMT [53]	84.1/84.2	82.7/82.8	80.9/81.2	79.4/79.7	77.4/77.7	75.0/75.4	71.6/72.2	68.2/69.1	61.4/63.3	75.6/76.2
LNLN [54]	83.3/83.7	81.7/82.0	80.1/80.4	78.5/79.0	76.4/77.2	73.8/75.0	71.6/73.5	68.6/71.8	64.8/70.6	75.4/77.0
P-RMF [56]	84.0/84.4	83.0/83.4	81.3/81.9	80.0/80.7	78.6/79.5	76.1/77.6	74.6/75.9	70.8/73.0	67.9/**71.5**	77.4/78.7
DEAR [61]	84.3/84.9	83.2/83.3	82.3/82.9	81.4/80.9	79.6/79.8	77.9/76.8	75.5/74.7	73.2/71.8	70.0/69.6	78.6/78.3
SemMSA [20]	85.3/85.2	84.0/83.9	82.6/82.4	81.6/81.4	80.9/80.6	79.0/78.6	75.5/76.8	70.4/73.0	68.2/70.6	78.6/79.2
TF-Mamba [21]	83.2/83.0	82.6/82.4	80.8/80.6	80.0/79.8	78.6/78.3	75.9/75.6	73.5/73.3	70.3/70.2	64.8/64.9	76.6/76.5
MSRGF	**86.9/86.9**	**86.0/86.0**	**85.0/84.8**	**84.2/84.1**	**81.1/81.1**	**79.6/79.1**	**77.7/77.6**	**74.3/73.6**	**70.1**/69.0	**80.5/80.2**

联合源处理相比，分离的源条件估计是否更有益。三种优先单源变体按照固定的优先级顺序，选择第一个可用模态 $\pi_T = (T,\ A,\ V)$、$\pi_A = (A,\ T,\ V)$和 $\pi_V = (V,\ T,\ A)$。选定的源为每个目标模态提供唯一的重建估计，包括观测到的目标，同时保留原始观测输入及其直接集成路径。由于只有一个有效估计，参考表示等于该估计，因此差异细化不起作用。我们还评估了融合源生成器：它将编码后的模态特征与源嵌入拼接，并应用掩码联合注意力来生成单个目标估计。与对分别生成的估计取平均值的早期聚合不同，该变体在生成器内部融合源信息。所有变体均独立重新训练，并保留适用的重建目标以及下游融合和预测组件。如表8所示，优先单源变体在固定设置下的 F1 分数为 75.85–76.19，在随机设置下为 75.55–75.82；相比之下，MSRGF 分别为 78.17 和 78.25。融合源生成器的 F1 分数分别为 76.12 和 75.72，与早期聚合的 76.16 和 75.74 接近。这些结果表明，保留分离的源条件估计以供后续参考引导融合，比经过评估的单源和早期融合替代方案更有效。

参考构建与差异细化。接下来，我们考察 RGF 中的加权和细化机制。Uniform α （仅供参考）在计算 C_m 时，以均匀系数替代学习得到的参考权重；而 Uniform β 为细化权重赋予相同的值，并 $\beta = \alpha$ 将其与参考权重绑定。为考察源特定更新，w/o Difference Refinement 将 $N_m = \mathbf{0}$ 设为 ……，Ordinary Residual 用共享残差 $C_m - B_m$ 替代 $R_m^n - C_m$，而 Early Aggregation 则在细化前对估计取平均，从而消除其源特定差异。在 Fixed 和 Random 条件下，所有替代方法的 F1值均低于 MSRGF。这些结果表明，与所评估的替代方法相比，学习得到的参考加权和源特定差异细化能够提升集成效果。

关系描述符及其组成部分。我们进一步研究关系描述符 ξ_m^n 及其各项比较特征。w/o Relation Descriptor 变体移除 $\xi_m^n = \mathbf{0}$，另外三个变体则分别移除 16 维源–估计交互 $c_{m,\,n}^{\mathrm{src}}$、跨估计差异分数$c_{m,\,n}^{\mathrm{div}}$ 以及观测目标一致性分数 $c_{m,\,n}^{\mathrm{tar}}$。每次将一项特征设为零，同时保留描述符的其余条目、有效性指示量及后续融合操作。

移除整个描述符后，Fixed/Random 条件下的 F1值分别下降 2.01/2.44 个百分点；移除其中单项特征则会带来较小但一致的降幅。具体而言，移除 c^{src} 后，F1值下降 1.02/0.76 个百分点，表明源–估计交互有助于融合。移除 c^{div} 后，F1值下降 0.65/0.74 个百分点，尽管源特定差异 Δ_m^n 及其细化 N_m 仍然发挥作用。这表明，对源估计之间的差异进行显式度量，可为参考构建和细化提供额外信息。同样，移除 c^{tar} 后，F1值下降 0.70/0.68 个百分点，说明在可进行此类比较时，观测目标一致性有所贡献。综上，这些结果表明，关系描述符整合了互补的比较特征，从而改进学习得到的权重和门控。

观测特征整合。 随后，我们考察 RGF 如何整合保留的观测。Hard Observa-˜tion Copyforces $H_m(t) = Z_m(t)$ 会在整合后于观测位置强制进行复制，而 w/o Observed Fusion 使用 $H_m = C_m + g_m^{\mathrm{enh}} \odot N_m$，仅保留 B_m 用于加权和

表 8. 重新训练的 IEMOCAP 消融实验（准确率/F1，%）；固定和随机各对七种设置取平均值。

方法	固定	随机
主要组件		
不使用 SWR	78.09/75.98	77.64/75.17
不含 RGF	78.05/75.81	77.74/75.27
不含 GCM	78.25/76.00	77.94/75.63
源组织		
T优先级单源	78.38/76.09	78.08/75.82
A优先级单源	78.25/75.85	78.03/75.73
V优先级单源	78.37/76.19	77.86/75.55
融合源生成器	78.23/76.12	78.05/75.72
早期聚合	78.43/76.16	78.05/75.74
内部机制		
不使用目标查询	78.27/76.14	77.96/75.75
不使用上下文锚点	78.01/75.79	77.78/75.37
均匀 α (仅供参考)	78.38/76.12	78.05/75.74
均匀 β	78.29/76.02	78.03/75.72
$\beta = \alpha$	78.35/76.07	78.06/75.70
不含差异细化	78.34/76.06	78.05/75.72
普通残差	78.30/76.02	78.03/75.75
不含关系描述符	78.46/76.16	78.08/75.81
不含 c^{src}	78.68/77.15	78.88/77.49
不含 c^{div}	79.02/77.52	78.91/77.51
不含 c^{tar}	78.96/77.47	78.96/77.57
不含特征级门控	78.27/76.00	78.02/75.72
硬观测复制	77.96/75.76	77.83/75.38
无观测融合	78.40/76.79	78.13/76.35
不含上下文门	78.34/76.10	78.08/75.85
损失函数		
不含 $\mathcal{L}_{\mathrm{view}}$	78.20/75.99	78.07/75.79
不含 $\mathcal{L}_{\mathrm{rec}}$	78.20/75.99	78.08/75.81
不含 $\mathcal{L}_{\mathrm{pres}}$	78.38/76.44	78.22/75.98
MSRGF	**79.73/78.17**	**79.63/78.25**

门控。这两种变体都会导致性能下降，随机F1值分别降至 75.38 和 76.35。这些结果表明，通过学习到的融合机制整合观测信息与重构信息，相比严格复制或移除直接观测通路，能够改善预测效果。硬复制还会使复制位置的保持损失变为零，因此，该消融实验同时改变了表征及其训练约束。

上下文调整。除了移除完整的 GCM 通路外，我们还通过在不使用上下文门控的情况下固定$\gamma_m = \mathbf{1}$，同时保留全局上下文构建和预测通路，来考察其学习到的调制机制

表 9. 使用一个或两个源估计的 IEMOCAP 任务结果（准确率/F1，%）。

设置	源估计	单个估计	两个估计
TV	R_A^T	79.51/79.20	80.85/80.61
TV	R_A^V	79.18/78.70	80.85/80.61
AV	R_T^A	79.21/77.03	79.73/77.58
AV	R_T^V	77.32/73.80	79.73/77.58

。该修改使 Fixed/Random 条件下的 F1值降至 76.10/75.85，而 MSRGF 为 78.17/78.25。结果表明，学习到的上下文门控除了全局上下文表征之外，还对预测有所贡献。

辅助目标。最后，我们在保留相关模型计算的同时，考察这三个辅助目标。移除 $\mathcal{L}_{\mathrm{view}}$ 会取消对源级估计 R_m^n的直接监督，移除 $\mathcal{L}_{\mathrm{rec}}$ 会禁用对集成表征 H_m的重建监督，而移除 $\mathcal{L}_{\mathrm{pres}}$ 则会取消对观测位置的约束。在固定和随机设置下，这三种移除都会降低性能，随机F1值分别降至 75.79、75.81 和 75.98。这些结果补充了架构消融，通过考察用于学习重建和增强表征的训练目标，进一步完善了分析。

9.2. 源级估计的贡献

我们进一步考察单个源估计在训练后的 MSRGF模型中的贡献。组件消融揭示了 SWR 的重要性，但没有说明训练后如何使用不同的源估计。因此，对于完全缺失的目标，我们将推理限制为仅使用一个源估计，同时保持原始观测不变。如表9所示，结合两个估计后，在 TV 条件下，F1值从表现较好的单源值 79.20 提升至 80.61；在 AV 条件下，则从 77.03 提升至 77.58。此比较在不移除已观测模态的前提下，单独考察了在训练后的融合流程中保留多个估计的效果。

为考察这些任务变化背后的表征质量，我们使用模态特定解码器，将解码后的特征与完整的目标特征进行比较 D_m。表10 显示，当文本或音频是唯一源时，增强视觉表示的平均绝对误差更低、余弦相似度更高于观测基线。当有多个可用源时，表11 进一步区分了各个估计值与其参考表示和增强表示。对于缺失的文本和声学目标，在所展示的替代方案中，学习得到的参考值产生的解码均方误差最低，尽管最终的增强表示是针对重建和

表10。IEMOCAP 上解码后的视觉目标重建。

源	表征	MAE	MSE	余弦
T→V	B_V	0.583169	0.674870	0.060257
	$\widetilde{H}_V$	**0.365323**	**0.674498**	**0.621105**
A→V	B_V	0.584370	0.673633	−0.017629
	$\widetilde{H}_V$	**0.448677**	**0.635592**	**0.547529**

表11。缺失的 IEMOCAP 目标的解码后的重建；粗体表示每个目标中的最佳结果。

音频到视频 → 目标	Repr	平均绝对误差 ↓	均方误差 ↓	余弦 ↑
AV→T	B_T	0.779919	1.039006	−0.007759
	R_T^A	0.770692	1.029792	0.054224
	R_T^V	0.771048	1.030405	0.054383
	C_T	**0.770538**	**1.029438**	**0.055783**
	$\widetilde{H}_T$	0.770644	1.029687	0.055194
TV→A	B_A	0.676690	0.874101	0.011633
	R_A^T	0.645782	0.871201	**0.123966**
	R_A^V	0.666480	0.930592	0.042970
	C_A	**0.643271**	**0.867499**	0.118265
	$\widetilde{H}_A$	0.648512	0.881017	0.117184

任务目标联合优化的，并不必然使每个重建指标最小化。所有重建指标均在相同的目标模态内进行比较。

9.3. 参考与关系干预

在考察了 SWR 提供的估计后，接下来我们分析 RGF 中用于为这些估计赋权的关系信息。重新训练的描述符消融测试完整特征 ξ_m^n；而在此处，我们在推理时针对目标模态禁用其 16 维源–估计交互$\mathbf{c}_{m,n}^{\text{src}}$。其余关系成分、重建的估计和观测输入均予以保留，以便我们检查哪些目标特定交互会影响训练得到的融合过程。

在 IEMOCAP 上，显示出的最大 F1下降出现在 AV 条件下禁用缺失文本目标的交互时（0.93 个百分点）；而在 TV 条件下禁用音频目标交互会使 F1值下降 0.39 个百分点。在 MOSEI 上，当模态内缺失率为 90% 时，移除所有目标的交互会使准确率/F1值下降 0.88/0.86 个百分点。这些干预明确了关系项的影响，并补充了重新训练的消融实验；后者衡量的是在学习过程中不使用该信息所造成的后果。

9.4. 差异细化

参考表示 C_m 聚合源估计，而细化项使用经过单独变换的偏差 $\Delta_m^n = R_m^n - C_m$。为考察学习到的更新方向是否会影响预测，我们对 $U_m = g_m^{\text{enh}} \odot N_m$ 训练模型中完全缺失的目标进行干预，同时保持参考表示和其他增强模态不变。表12比较了原始更新及其置零、随机化、反向和样本置换的对应版本；随机化更新保留其范数，而置换会在样本之间交换更新。

对于 TV→A，反转更新方向会使 F1值从 80.61 降至 79.29，而替换为等范数随机方向后，得分为 80.03。对于 AV→T，也观察到了类似的排序，这表明在这些设置中，任务效果取决于学习得到的更新方向。随后，我们在 MOSEI 上评估模态内缺失率为 90% 时的解码重建，其中视觉目标呈现出最明显的效果：不进行细化时，其均方误差从 0.644883 增至 0.651519，而反转更新方向时则增至 0.666883。声学目标也呈现相同的排序，详见表15。

为检验视觉目标重建变化是否仅限于单一缺失率，我们比较 MOSEI 上所有九种模态内缺失率下的完整表征与零细化表征。表16 报告了每种缺失率下均为正值的解码均方误差降低，平均为 2。103×10^{-3}，且在缺失率为 90% 时达到最大值。这一逐缺失率结果将更新方向干预从单独的测试条件扩展到更广泛的观测覆盖率范围。

9.5. 观测特征保留

前面的实验探讨了缺失目标的重建，但 RGF 也需要保留已观测到的、富含信息的位置。为直接衡量这一方面的最终整合，我们比较 C_m 和 H_m 投影后的观测值 Z_m 在由 M_m选定的位置处的表现，并使用相同的 128 维表征空间。表13中的六组 IEMOCAP 设置—目标配对，区分了由源估计构建的参考表示与经观测条件融合后的最终表征。

在这六组配对中，相对于 Z_m 的平均均方误差从 的 1.647851 降至 C_m 的 0.009320 H_m，而平均余弦相似度则从 0.065341 升至 0.995687。这些测量结果刻画了观测位置处的最终整合效果；表*8*中的硬观测复制、无观测融合以及保留损失消融则分别考察了哪些计算与训练选择有助于提升任务表现。因此，对观测特征的高度吻合与预测有效性通过互补的测量方式进行评估。

9.6. 全局上下文调制

最后，我们考察增强后的模态表示构建完成后的预测阶段。重新训*w/o GCM* 消融在学习过程中移除了全局上下文构建与调制；而推理阶

表12。对 $U_m = g_m^{\text{enh}} \odot N_m$ 进行干预后的 IEMOCAP F1值（%）。

目标	Full	$U_m = 0$	随机	反向	置换
AV→T	**77.58**	77.33	77.28	76.85	76.88
TV→A	**80.61**	80.00	80.03	79.29	79.60

表13。IEMOCAP 与 Z_m 的观测位置相似度；均值为六组配对的平均值。

设置 / 目标	均方误差(C_m, Z_m)	MSE($\tilde{H}_m$, Z_m)	余弦(C_m, Z_m)	Cos($\tilde{H}_m$, Z_m)
音频-视觉 / A	2.033782	0.000106	0.018657	0.999952
音频-视觉 / 视觉	2.020127	0.000010	0.152937	0.999995
目标-音频 / 文本	1.155433	0.021737	0.031935	0.990000
目标-音频 / 音频	1.148216	0.011019	0.019081	0.995030
文本-视觉 / 文本	2.075091	0.022260	0.010071	0.989525
文本-视觉 / 视觉	1.454458	0.000786	0.159364	0.999619
Mean	1.647851	**0.009320**	0.065341	**0.995687**

表14。特定于目标的源–估计交互干预。

设置	目标	Δ准确率/F1值
MOSEI 90%	All	+0.88/+0.86
IEMOCAP 音频-视觉	T	+0.35/+0.93
IEMOCAP 文本-视觉	A	+0.35/+0.39

表15. MOSEI 在90%模态内缺失情况下的解码目标均方误差。

目标	Full	不含 N_m	随机	反向
A	**1.052734**	1.054066	1.054196	1.055929
V	**0.644883**	0.651519	0.653729	0.666883

表16。基于 MOSEI 的差异细化所解码的视觉目标均方误差降低。

模态内缺失率	均方误差降低（$\times 10^{-3}$）
10%	+3.700
20%	+1.821
30%	+1.359
40%	+1.819
50%	+1.529
60%	+0.222
70%	+0.740
80%	+1.105
90%	+6.636
Avg	**+2.103**

表17. IEMOCAP GCM干预结果。

设置	完整准确率/F1	Δ准确率/F1
TAV	81.70/80.81	+1.887/+2.457
AV	79.73/77.58	+0.656/+0.951

表18。在 IEMOCAP 上进行推理时路径干预。

禁用路径	固定	随机整体	随机内部
SWR	+3.71/+3.54	+4.54/+5.42	+2.10/+2.43
GCM	+0.41/+0.52	+0.21/+0.23	+0.10/+0.11

段的时间干预则在保留 RGF输出和原始预测器的同时，禁用训练后的网络中的相同贡献。表17 报告了各条件下的变化，其中 TAV 条件下 F1值降低 2.457 个百分点，AV 条件下降低 0.951 个百分点。

在每种缺失协议下，也对完整的 IEMO- CAP 测试集实施该干预，并结合 SWR路径干预，以区分组合效应与整体协议平均效应。表 18显示，在各种缺失协议下，禁用 SWR 的即时影响更大，而禁用 GCM 带来的变化较小，但准确率和 F1值均有所提升。这些结果补充了重新训练的组件比较，并完成了从逐源重建、融合到任务预测的分析。