Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MDRC: Mechanism-Decoupled Retrieval-Guided Conditional Recovery for Incomplete Multimodal Emotion Recognition

Qianxi Zhang
Xinjiang University
Shengcai Wang
Xinjiang University
Nengchao Wu
Xinjiang University
Junjie Bao
Xinjiang University
Xinyu Yang
Xinjiang University
Na Cui
Shineray Automobile Co.,Ltd
Rangxiang Zhao
Shineray Automobile Co.,Ltd
Qing Tao
Xinjiang University
Li Liu
dcsliuli@cqu.edu.cn

Xinjiang University

Research Article

Keywords: Missing modality, Retrieval enhancement, Conditional recovery, Multimodal learning

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10342113/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MDRC: Mechanism-Decoupled Retrieval-Guided Conditional Recovery for Incomplete Multimodal Emotion Recognition

Qianxi Zhang[1], Shengcai Wang[1], Nengchao Wu[1], Junjie Bao[1], Xinyu Yang[1], Na Cui[3], Rangxiang Zhao[3], Qing Tao[1], Li Liu[1,2*]

[1*]College of Mechanical Engineering, Xinjiang University, 666 Shengli Road, Urumqi, 830017, Xinjiang, China.
[2]School of Big Data and Software Engineering, Chongqing University, 174 Shazheng Street, Chongqing, 400030, China.
[3]Shineray Automobile Co.,Ltd, Chongqing, 400800, China.

*Corresponding author(s). E-mail(s): dcsliuli@cqu.edu.cn;
Contributing authors: 107552404373@stu.xju.edu.cn;
107552404177@stu.xju.edu.cn; wnc@stu.xju.edu.cn;
107552404289@stu.xju.edu.cn; yangxinyu@stu.xju.edu.cn;
13628321200@163.com; 271284387@qq.com; taoqing@xju.edu.cn;

Abstract

Incomplete multimodal emotion recognition aims to infer affective states from partially observed multimodal inputs. When one or more modalities are unavailable, cross-modal correlations are weakened, often leading to degraded multimodal representations and unreliable predictions. Existing methods usually address this problem through robust fusion, self-distillation, or feature-level reconstruction. However, the recovered representations may still lack task-discriminative affective information. Although retrieval evidence can provide useful external references for missing-modality recovery, its direct or unconstrained use may introduce semantic bias and affective inconsistency [1]. To address these challenges, we propose a **M**echanism-**D**ecoupled **R**etrieval-Guided **C**onditional Recovery framework (**MDRC**) for incomplete multimodal emotion recognition. MDRC first projects heterogeneous modalities into a unified representation space and formulates missing-modality recovery within a shared-private conditional variational framework, thereby modeling both cross-modal affective semantics and modality-specific information. To exploit retrieval evidence in a controlled manner, MDRC decouples its role into prior-level distribution guidance

and bounded residual refinement, which guide conditional recovery and supplement task-relevant information, respectively. In addition, a post-enhancement consistency regularization is introduced to align the enhanced recovered representations with the observed context and other recovered targets. Experimental results on multimodal emotion recognition benchmarks demonstrate that MDRC achieves effective and robust performance under diverse incomplete-modality settings.

Keywords: Missing modality, Retrieval enhancement, Conditional recovery, Multimodal learning

1 Introduction

Multimodal emotion recognition (MER) aims to infer human affective states by jointly modeling textual content, acoustic signals, and visual behaviors. Compared with unimodal methods, MER can exploit complementary cues from textual semantics, vocal patterns, and facial expressions, and has been widely studied in applications such as human–computer interaction, mental health assessment, eldercare assistance, and clinical support. With the development of deep representation learning and cross-modal interaction modeling, many methods have achieved competitive performance on public benchmarks [2]. For example, Tensor Fusion Network (TFN) [3] models high-order inter-modal interactions through tensor outer products, Multimodal Transformer (MulT) [4] captures dynamic dependencies among asynchronous modalities via cross-modal attention, and MISA [5] improves representation robustness by disentangling modality-invariant and modality-specific information.

Most existing MER methods are developed under the full-modality assumption, where textual, acoustic, and visual modalities are available during both training and inference. Although this setting facilitates the learning of stable cross-modal interactions, it is often inconsistent with real-world conditions. In practical scenarios, some modalities may be unavailable or corrupted because of environmental noise, device failure, sensor occlusion, transmission errors, or automatic transcription errors [6]. Such uncertain missingness can weaken cross-modal correlations, degrade multimodal representations, and substantially impair the reliability of conventional multimodal fusion models [7–9]. Therefore, incomplete multimodal emotion recognition has become an important extension of conventional MER [10].

Existing incomplete MER methods can be broadly categorized into non-recovery-based and recovery-based approaches. Non-recovery-based approaches aim to learn robust multimodal representations from incomplete inputs without explicitly reconstructing the missing modalities. For example, DCCAE learns correlated cross-modal representations to reduce modality discrepancy [11]. MCTN learns robust joint representations through cyclic translations and cycle-consistency constraints [12]. Compared with non-recovery-based approaches, recovery-based methods address modality absence more directly by estimating or completing missing modality representations before or during prediction. LNLN corrects missing information by modeling the

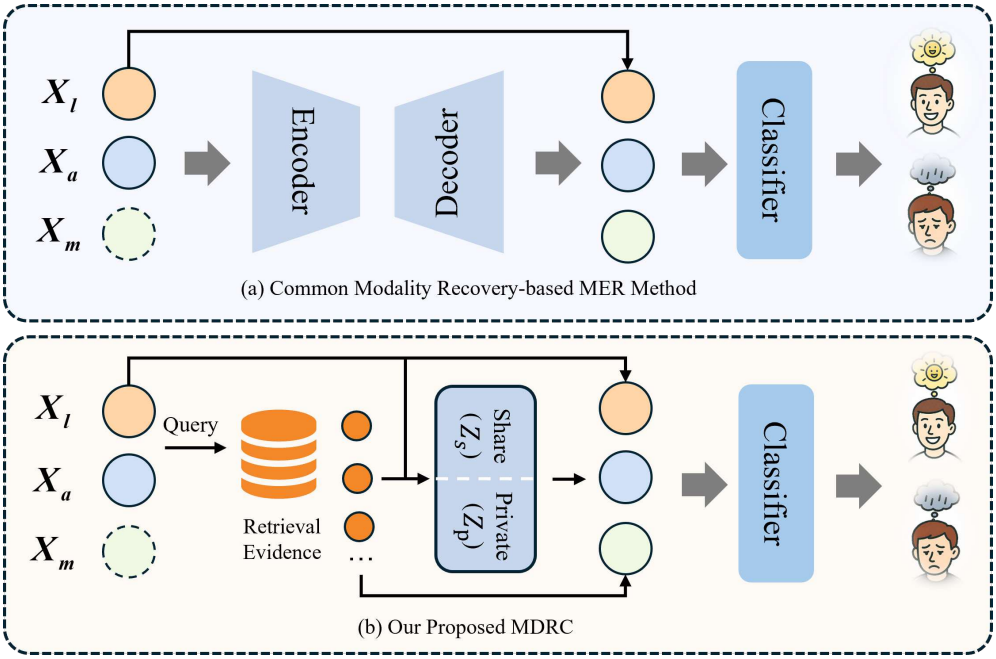


Fig. 1: Comparison between a representative recovery-based incomplete MER framework and the proposed MDRC. (a) A conventional recovery-based method first estimates missing modality representations from the available modalities and then performs prediction with the completed multimodal representations. (b) MDRC introduces retrieval evidence as external references and performs shared-private conditional recovery to enhance the recovered representations before prediction.

relationship between dominant and non-dominant modalities [13]. GCNet estimates missing information through graph-based completion [14], and MPLMM uses prompt learning to generate missing-modality-related features [15]. As shown in Fig. 1(a), representative recovery-based methods usually follow a "recover-then-classify" paradigm, where missing representations are first estimated from available modalities and then fed into the classifier. Although these methods improve robustness under incomplete inputs, they mainly focus on representation alignment, structural completion, or missing-pattern modeling. Less attention has been paid to whether the recovered representations contain affective information that is useful for downstream prediction [14, 15].

This issue is particularly evident when the available modalities provide only weak affective cues. The textual modality often carries more explicit and dense affective semantics, whereas acoustic and visual modalities tend to provide more implicit, sparse, and noise-sensitive cues [16, 17]. Therefore, although missing-modality recovery can be formulated for any target modality, it becomes particularly difficult when the textual modality is unavailable and only acoustic or visual observations are retained. In such cases, conditional recovery is weakly constrained by the observed modalities,

which may lead to ambiguous or over-smoothed representations with limited discriminative value for the downstream task. Existing recovery-oriented methods often rely on feature-level reconstruction, modality translation, or latent distribution alignment to reduce the discrepancy between recovered and complete representations [18]. Such objectives help preserve modality-level structure, but they do not by themselves ensure that the recovered representations contain sufficient affective cues for downstream prediction [19].

MDRC is built on two main insights from shared-private generative modeling and retrieval-augmented modeling. Shared-private representation learning suggests that cross-modal common factors and modality-specific information should be modeled separately for stable multimodal recovery [18]. Retrieval-augmented modeling further suggests that external memory can provide sample-level references beyond the current input, which is beneficial when the observed modalities contain limited affective information [20, 21]. However, retrieved samples may be inconsistent with the current instance in semantic content or affective polarity, and noisy or conflicting evidence may disturb the generation process [22]. Therefore, retrieval evidence should be incorporated with explicit reliability control rather than being directly injected into recovered representations. Following the residual learning principle, MDRC treats retrieval evidence as a bounded refinement of the base recovery, allowing task-relevant information to be supplemented while limiting retrieval-induced perturbations [22, 23].

Inspired by these observations, we introduce **M**echanism-**D**ecoupled **R**etrieval-guided **C**onditional Recovery (**MDRC**), a framework for incomplete multimodal emotion recognition. As illustrated in Fig. 1(b), MDRC introduces retrieval evidence as sample-level external references and regulates its influence through a mechanism-decoupled design. Specifically, MDRC first projects heterogeneous modalities into a unified representation space to reduce dimensional and distributional discrepancies. It then formulates missing-modality recovery within a shared-private conditional variational framework, where cross-modal affective semantics and modality-specific information are jointly modeled to provide a stable base recovery. To further improve the task discriminability of recovered representations, MDRC decouples retrieval evidence into prior-level distribution guidance and bounded residual refinement. The former guides the conditional recovery direction, while the latter supplements task-relevant information under a constrained residual budget. In addition, post-enhancement consistency regularization is introduced to keep the enhanced recovered representations consistent with the observed context and other recovered targets, thereby reducing the mismatch between missing-modality recovery and downstream affective prediction. A two-stage training strategy is further adopted to stabilize the optimization between conditional recovery and retrieval-guided enhancement. The main contributions of this paper are summarized as follows:

- We propose **MDRC**, a mechanism-decoupled retrieval-guided conditional recovery framework for incomplete multimodal emotion recognition. MDRC provides a unified formulation for recovering missing modality representations from available modalities under different incomplete-modality settings.
- We formulate missing-modality recovery within a shared-private conditional variational framework and integrate it with retrieval-guided enhancement. The recovery

module models both cross-modal affective semantics and modality-specific information, while the retrieval-guided enhancement module decouples retrieval evidence into prior-level distribution guidance and bounded residual refinement to improve recovery stability and task discriminability.
- We introduce post-enhancement consistency regularization and a two-stage training strategy to stabilize the interaction between conditional recovery and retrieval-guided enhancement. Extensive experiments under fixed and random missing-modality settings demonstrate the effectiveness and robustness of MDRC.

2 Related Work

2.1 Multimodal Emotion Recognition under Incomplete Modalities

Incomplete multimodal emotion recognition aims to maintain reliable affective prediction when one or more modalities are unavailable or corrupted [10]. One key challenge arises from the uncertainty of missing patterns, since different modality combinations may be absent during inference. Existing studies mainly address this problem from two perspectives. The first line focuses on robust representation learning, aiming to reduce the dependence of multimodal models on complete inputs and improve their adaptability to missing modalities [24]. The second line focuses on missing-information modeling, where unavailable modality information is estimated, completed, or implicitly modeled to narrow the information gap between incomplete and complete multimodal inputs [25, 26]. These methods provide important foundations for incomplete multimodal learning. However, most of them primarily emphasize representation robustness, missing-pattern adaptation, or structural plausibility, while less attention has been paid to whether the inferred or recovered representations contain discriminative information that is useful for downstream affective prediction.

2.2 Shared-Private Representation Learning and Generative Multimodal Recovery

Shared-private representation learning aims to disentangle cross-modal shared semantics from modality-specific information, thereby reducing modality heterogeneity and improving representation discriminability. In multimodal sentiment analysis, MISA disentangles shared and modality-specific representations to improve the robustness of affective representation learning. Generative multimodal modeling further provides a principled way to infer latent representations from incomplete modality observations [27]. For example, MoPoE-VAE combines product-of-experts and mixture-of-experts formulations within a generalized multimodal variational objective [28], while DMVAE models cross-modal shared factors and modality-specific information through separate shared and private latent variables. These studies demonstrate the effectiveness of shared-private latent modeling and multimodal variational inference for incomplete multimodal representation learning[29, 30]. However, they mainly focus on representation disentanglement, latent distribution modeling, or reconstruction quality, while the task discriminability of recovered representations is less explicitly considered.

Therefore, how to integrate shared-private generative modeling with task-oriented missing-modality recovery remains an important problem.

3 Method

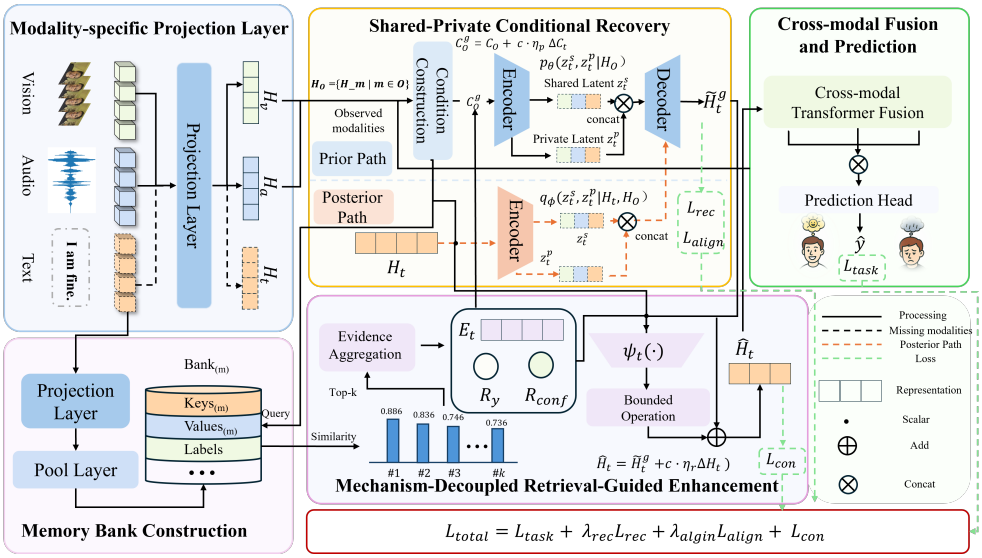


Fig. 2: Overall framework of the proposed MDRC.

This section presents the proposed MDRC, a mechanism-decoupled retrieval-guided conditional recovery framework for incomplete multimodal emotion recognition. As illustrated in Fig.2, MDRC consists of four main components: a modality-specific projection layer, a shared-private conditional recovery module, a mechanism-decoupled retrieval-guided enhancement module, and a cross-modal fusion and prediction module. Due to the dimensional and distributional discrepancies among textual, acoustic, and visual modalities, MDRC first maps heterogeneous modality features into a unified latent representation space through modality-specific projection layers. This provides a consistent representation basis for subsequent missing-modality recovery and retrieval-guided enhancement. The shared-private conditional recovery module then uses the observed modalities as conditions to infer missing modality representations in a generative manner, where cross-modal shared affective semantics and modality-specific information are jointly modeled. To further improve the task discriminability of recovered representations, the mechanism-decoupled retrieval-guided enhancement module introduces sample-level retrieval evidence and decomposes its effect into prior-level distribution guidance and bounded residual refinement, thereby supplementing task-relevant information while controlling retrieval-induced interference. Finally, the cross-modal fusion and prediction module receives both the observed representations and the enhanced recovered representations, and performs affective

prediction through cross-modal fusion. The details of each module are introduced in the following subsections.

Consider a multimodal sample with modality set $\mathcal{S} = \{t, a, v\}$, where t, a, and v represent the textual, acoustic, and visual modalities, respectively. For an arbitrary incomplete sample, the observed modality set and the missing modality set are denoted as $\mathcal{O} \subseteq \mathcal{S}$, $\mathcal{U} = \mathcal{S} \setminus \mathcal{O}$. Where, $\mathcal{O}$ denotes the modalities directly available under the current missing-modality scenario, while $\mathcal{U}$ denotes the missing modalities whose representations need to be recovered. Different from complete-modality emotion recognition, incomplete multimodal emotion recognition requires the model to capture latent correlations between observed and missing modalities and maintain stable affective discrimination when only partial modalities are available.

3.1 Modality-specific Projection Layer

Textual, acoustic, and visual modalities exhibit substantial differences in feature dimensions, representation distributions, and semantic granularity. Directly performing missing-modality recovery and cross-modal fusion on raw modality features may lead to representation-space mismatch, thereby increasing the difficulty of recovery modeling and subsequent multimodal fusion. To alleviate this issue, MDRC introduces modality-specific projection layers before the recovery and fusion modules, which map heterogeneous modality features into a unified latent representation space.

For each modality $m \in \mathcal{S}$, where $\mathcal{S} = \{t, a, v\}$ denotes the textual, acoustic, and visual modalities, the raw input feature is represented as $X_m \in \mathbb{R}^{B \times T_m \times d_m}$, where B is the batch size, T_m denotes the temporal length of modality m, and d_m denotes its original feature dimension. A modality-specific projection function $\phi_m(\cdot)$ is then used to obtain the projected representation:

$$\begin{aligned} H_m &= \phi_m(X_m), \\ H_m &\in \mathbb{R}^{B \times T_m \times d}. \end{aligned} \tag{1}$$

where all modalities are projected into a d-dimensional latent space. $\phi_m(\cdot)$ is instantiated as a modality-specific one-dimensional convolution, which transforms the feature dimension of each modality into the same hidden dimension while preserving its temporal structure.

3.2 Shared-private Conditional Recovery Module

In the unified latent representation space, MDRC performs conditional recovery for missing modality representations. Textual, acoustic, and visual modalities contain both cross-modal shared affective semantics and modality-specific information. Directly generating missing representations from the observed modalities may lead to over-smoothed recovery results or insufficient task-discriminative information. To address this issue, MDRC introduces a shared-private conditional recovery module, which decomposes missing-modality recovery into shared semantic modeling and private information modeling. The shared latent variable captures affect-related

semantics that are common across modalities, while the private latent variable preserves target-modality-specific information [31].

For generality, we denote the current missing target modality as $m^* \in \mathcal{U}$. In the framework figure, the textual-missing case is used as an example, where $m^* = t$. For each missing target modality m^*, the model first constructs a conditional context from the projected representations of the observed modalities:

$$C_{\mathcal{O}} = A_c\left(\{H_m \mid m \in \mathcal{O}\}\right), \tag{2}$$

where $A_c(\cdot)$ denotes the context aggregation function over the observed modality representations. When multiple modalities are observed, their conditional information is aggregated to form a unified conditional prior. Specifically, the prior distribution for recovering the missing target modality is formulated as

$$p_\theta\left(z^s_{m^*}, z^p_{m^*} \mid H_{\mathcal{O}}\right) = \mathrm{Agg}_{m\in\mathcal{O}}\left(p_\theta\left(z^s_{m^*}, z^p_{m^*} \mid H_m\right)\right), \tag{3}$$

where $z^s_{m^*}$ and $z^p_{m^*}$ denote the shared and private latent variables for recovering the missing target modality m^*, respectively. The aggregation function $\mathrm{Agg}(\cdot)$ is implemented in a product-of-experts style to integrate complementary information from the observed modalities.

The prior path generates the base recovered representation using only the observed modality conditions. Specifically, the latent variables are sampled from the aggregated conditional prior, and the conditional context $C_{\mathcal{O}}$, the shared latent variable $z^s_{m^*}$, and the private latent variable $z^p_{m^*}$ are fed into the target-modality decoder:

$$\widetilde{H}_{m^*} = D_{m^*}\left(C_{\mathcal{O}}, z^s_{m^*}, z^p_{m^*}\right), \tag{4}$$

where $D_{m^*}(\cdot)$ denotes the decoder corresponding to the missing target modality m^*, and $\widetilde{H}_{m^*}$ denotes the base recovered representation before retrieval-guided enhancement. This representation is used by the subsequent prediction branch and retrieval-guided enhancement module. Importantly, its generation depends only on the observed modality conditions and does not use the ground-truth missing modality.

During training, the ground-truth target representation H_{m^*} is used only to construct the posterior path and provide recovery supervision. The corresponding posterior distribution is defined as

$$q_\phi\left(z^s_{m^*}, z^p_{m^*} \mid H_{m^*}, H_{\mathcal{O}}\right). \tag{5}$$

The posterior path is not fed into the downstream prediction branch. Instead, it serves as a training-only reference to align the condition-only prior path with the latent distribution inferred from the ground-truth target modality. In this way, MDRC can exploit target-modality supervision during training while avoiding target information leakage during prediction.

The recovery module is optimized by a recovery loss and a latent alignment loss. The recovery loss constrains the feature-level consistency between the base recovered

representation and the ground-truth target representation:

$$L_{rec} = \sum_{m^* \in \mathcal{U}} \ell_{rec}\left(\widetilde{H}_{m^*}, H_{m^*}\right), \tag{6}$$

where $\ell_{rec}(\cdot)$ denotes the feature-level recovery error. The latent alignment loss constrains the prior path to approach the posterior path in the shared-private latent space:

$$L_{\text{align}} = \sum_{m\setminus t \in \mathcal{U}} D_{\text{KL}}\Big(q_\phi(z^s_{m\setminus t}, z^p_{m\setminus t} \mid H_{m\setminus t}, H_{\mathcal{O}}) \Big\| p_\theta(z^s_{m\setminus t}, z^p_{m\setminus t} \mid H_{\mathcal{O}})\Big). \tag{7}$$

Through this shared-private conditional recovery mechanism, MDRC obtains a stable base representation for each missing modality in the unified latent space. The output $\widetilde{H}_{m^*}$ is then passed to the retrieval-guided enhancement module, where external sample evidence is introduced to further improve its task-discriminative ability.

3.3 Mechanism-Decoupled Retrieval-Guided Enhancement Module

Although the shared-private conditional recovery module can generate base representations for missing modalities, the recovered representations may still contain insufficient affective semantics or limited task-discriminative information, especially when the observed modalities are incomplete or weak. To further enhance these representations without directly overwriting them, MDRC introduces retrieval evidence as sample-level external references. Since retrieved samples may be semantically mismatched or affectively inconsistent with the current instance, MDRC decouples the use of retrieval evidence into two controlled steps: prior-level guidance and bounded residual refinement.

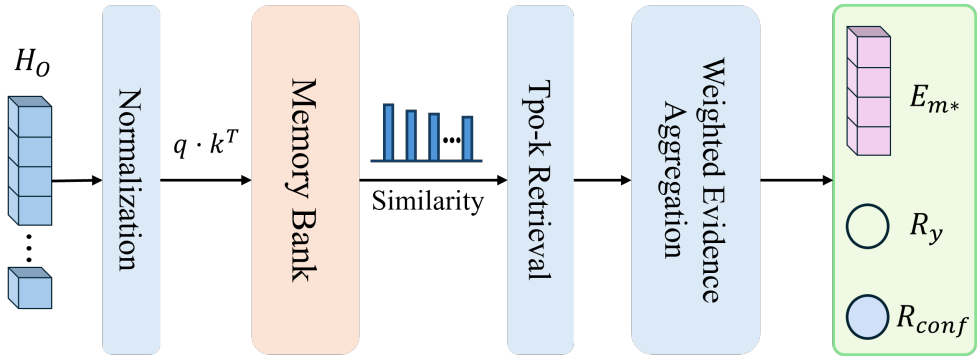


Fig. 3: Memory bank construction and retrieval evidence aggregation. The observed modalities are used to construct the retrieval query, while the retrieved values of the missing target modality provide feature-level evidence.

MDRC constructs a retrieval memory bank in the unified latent representation space. For the i-th training sample and modality m, the projected representation H_i^m is pooled and normalized as the retrieval key $k_i^m = \mathrm{Norm}(\mathrm{Pool}(H_i^m))$, while the projected representation itself is stored as the retrieval value $e_i^m = H_i^m$. The memory bank is denoted as $\mathcal{M} = \{(K_i, V_i, y_i)\}_{i=1}^N$, where $K_i = \{k_i^m \mid m \in \mathcal{S}\}$, $V_i = \{e_i^m \mid m \in \mathcal{S}\}$, and y_i is the affective label.

For the current missing target modality m^*, the query $q_{\mathcal{O}}$ is constructed from the observed modalities, and the retrieval key $k_i^{\mathcal{O}}$ of each memory sample is selected or aggregated under the same observed-modality condition. The similarity score and retrieval weight are computed as:

$$\begin{aligned} q_{\mathcal{O}} &= \mathrm{Norm}\left(A_q\left(\{\mathrm{Pool}(H_m) \mid m \in \mathcal{O}\}\right)\right), \\ s_i &= \mathrm{sim}\left(q_{\mathcal{O}}, k_i^{\mathcal{O}}\right), \\ \alpha_i &= \frac{\exp(s_i)}{\sum_{j \in \mathcal{N}_K} \exp(s_j)}, \qquad i \in \mathcal{N}_K. \end{aligned} \tag{8}$$

Where, $A_q(\cdot)$ denotes the query aggregation function and $\mathcal{N}_K$ denotes the set of top-K retrieved samples. Based on the retrieval weights, MDRC obtains the feature-level evidence, affective tendency, and retrieval reliability as:

$$\begin{aligned} E_{m^*} &= \sum_{i \in \mathcal{N}_K} \alpha_i e_i^{m^*}, \\ R_y &= \sum_{i \in \mathcal{N}_K} \alpha_i y_i, \\ R_{conf} &= \rho\left(\{s_i, y_i \mid i \in \mathcal{N}_K\}\right). \end{aligned} \tag{9}$$

Where, E_{m^*} provides feature-level evidence for the missing target modality, R_y summarizes the affective tendency of retrieved samples, and R_{conf} measures retrieval reliability by considering both similarity and label consistency.

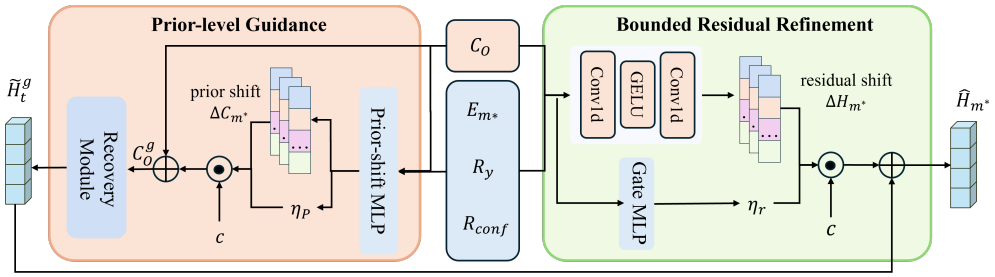


Fig. 4: Mechanism-decoupled retrieval-guided enhancement.

The first enhancement step performs prior-level guidance. Instead of directly replacing the base recovered representation, MDRC uses retrieval evidence to adjust

the conditional context of the recovery module. Given the conditional context $C_{\mathcal{O}}$, retrieval evidence E_{m^*}, affective tendency R_y, and reliability score R_{conf}, the prior shift and its sample-wise control coefficient are computed as:

$$\begin{aligned} \Delta C_{m^*} &= \mathcal{B}_p\left(\Gamma_{m^*}\left(C_{\mathcal{O}}, E_{m^*}, R_y\right); C_{\mathcal{O}}\right), \\ \eta_p &= R_{conf} \cdot g_p\left(\mathrm{Pool}(C_{\mathcal{O}}), R_y, R_{conf}\right). \end{aligned} \tag{10}$$

Where, $\Gamma_{m^*}(\cdot)$ denotes the prior-guidance function, $\mathcal{B}_p(\cdot)$ constrains the magnitude of the prior shift, and $g_p(\cdot)$ estimates the reliability-controlled coefficient η_p. The guided conditional context and the retrieval-guided recovered representation are then obtained as:

$$\begin{aligned} C^g_{\mathcal{O}} &= C_{\mathcal{O}} + c \cdot \eta_p \Delta C_{m^*}, \\ \widetilde{H}^g_{m^*} &= D_{m^*}\left(C^g_{\mathcal{O}}, z^s_{m^*}, z^p_{m^*}\right). \end{aligned} \tag{11}$$

Where c denotes a global retrieval-enhancement budget. In this way, retrieval evidence guides the recovery process at the conditional-prior level without directly overwriting the recovered representation.

The second enhancement step performs bounded residual refinement based on the retrieval-guided recovered representation. This step learns a bounded residual shift to supplement task-relevant affective information while preserving the stability of the recovered representation. The residual shift and its sample-wise control coefficient are computed as:

$$\begin{aligned} \Delta H_{m^*} &= \mathcal{B}_r\left(\psi_{m^*}\left(\widetilde{H}^g_{m^*}, C^g_{\mathcal{O}}, E_{m^*}, R_y\right); \widetilde{H}^g_{m^*}\right), \\ \eta_r &= R_{conf} \cdot g_r\left(\mathrm{Pool}(\widetilde{H}^g_{m^*}), R_y, R_{conf}\right). \end{aligned} \tag{12}$$

Where, $\psi_{m^*}(\cdot)$ denotes the residual refinement function, $\mathcal{B}_r(\cdot)$ constrains the residual magnitude, and $g_r(\cdot)$ estimates the reliability-controlled coefficient η_r. The final enhanced recovered representation is then obtained as:

$$\widehat{H}_{m^*} = \widetilde{H}^g_{m^*} + c \cdot \eta_r \Delta H_{m^*}. \tag{13}$$

With this bounded residual formulation, MDRC introduces task-relevant information from retrieved samples while reducing retrieval-induced bias and semantic drift.

To further encourage the enhanced recovered representation to remain consistent with the current observed context, MDRC introduces a post-enhancement consistency constraint. The semantic-level representations of the enhanced recovered modality and the observed context are obtained by pooling:

$$u_{m^*} = \mathrm{Pool}\left(\widehat{H}_{m^*}\right), \qquad u_{\mathcal{O}} = A_u\left(\{\mathrm{Pool}(H_m) \mid m \in \mathcal{O}\}\right), \tag{14}$$

where $A_u(\cdot)$ denotes the aggregation function for the observed context. The consistency loss is then defined as:

$$\mathcal{L}_{con} = \sum_{m^* \in \mathcal{U}} \left(1 - \cos\left(u_{m^*}, u_{\mathcal{O}}\right)\right). \tag{15}$$

Through the above mechanism-decoupled design, MDRC constrains the role of retrieval evidence at both the prior-guidance level and the residual-refinement level. The final representation $\widehat{H}_{m^*}$ is then used as the enhanced recovered representation for cross-modal fusion and prediction.

3.4 Fusion Prediction and Overall Optimization

After shared-private conditional recovery and retrieval-guided enhancement, MDRC obtains an enhanced recovered representation $\widehat{H}_m$ for each missing modality $m \in \mathcal{U}$. For downstream sentiment prediction, the observed modality representations and the enhanced recovered representations are assembled into a completed multimodal representation set:

$$\overline{H}_m = \begin{cases} H_m, & m \in \mathcal{O}, \\ \widehat{H}_m, & m \in \mathcal{U}, \end{cases} \tag{16}$$

where $\overline{H}_m$ denotes the final representation of modality m under the current incomplete-modality setting. Therefore, MDRC constructs the completed multimodal representation set $\{\overline{H}_m\}_{m\in\mathcal{S}}$, in which observed modalities retain their projected representations, while missing modalities are replaced by their enhanced recovered representations.

The completed modality representations are then fed into the cross-modal Transformer fusion module to model inter-modal interactions among textual, acoustic, and visual representations:

$$\mathbf{F} = \mathcal{F}_{fuse}\left(\{\overline{H}_m\}_{m\in\mathcal{S}}\right), \tag{17}$$

where $\mathcal{F}_{fuse}(\cdot)$ denotes the cross-modal fusion function, and $\mathbf{F}$ denotes the fused multimodal representation. Based on $\mathbf{F}$, the prediction head outputs the final sentiment category:

$$\widehat{y} = \mathcal{C}\left(\mathbf{F}\right), \tag{18}$$

where $\mathcal{C}(\cdot)$ denotes the prediction head and $\widehat{y}$ denotes the predicted sentiment category.

The task loss is defined as:

$$\mathcal{L}_{task} = \ell\left(\widehat{y}, y\right), \tag{19}$$

where y is the ground-truth sentiment label and $\ell(\cdot)$ denotes the classification loss.

3.5 Two-stage Training Strategy

The shared-private conditional recovery module and the retrieval-guided enhancement module play different roles in MDRC. The recovery module learns stable base representations for missing modalities from the observed modality conditions, while the retrieval-guided enhancement module further introduces sample-level external evidence to improve the task discriminability of the recovered representations. If retrieval evidence is introduced before the base recovery becomes stable, unreliable retrieved samples may disturb the recovery process and introduce affective bias or semantic drift. Therefore, MDRC adopts a two-stage training strategy to decouple base recovery learning from retrieval-guided enhancement.

The complete training objective of MDRC contains the task loss, recovery loss, latent alignment loss, and post-enhancement consistency loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{align}\mathcal{L}_{align} + \mathcal{L}_{con}. \tag{20}$$

This objective summarizes all loss components involved in MDRC, while the active loss terms are determined by the two-stage training strategy.

In the first stage, MDRC focuses on learning stable base recovery ability. The modality-specific projection layers, the shared-private conditional recovery module, and the cross-modal fusion and prediction module are jointly optimized, while the retrieval-guided enhancement module is not activated. For each missing target modality $m^* \in \mathcal{U}$, the model generates the base recovered representation $\widetilde{H}_{m^*}$ only from the observed modality conditions. The stage-one objective is defined as

$$\mathcal{L}_{stage1} = \mathcal{L}_{task}^{(1)} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{align}\mathcal{L}_{align}. \tag{21}$$

Where, $\mathcal{L}_{task}^{(1)}$ is the task loss computed with the base recovered representations, $\mathcal{L}_{rec}$ constrains the feature-level consistency between the base recovered representation and the ground-truth target representation, and $\mathcal{L}_{align}$ aligns the condition-only prior path with the posterior path in the shared-private latent space.

In the second stage, MDRC is initialized with the parameters learned in the first stage, and the retrieval-guided enhancement module is activated. The recovery module is frozen to preserve the learned base recovery ability. For each missing target modality m^*, the frozen recovery module first produces $\widetilde{H}_{m^*}$, after which prior-level guidance generates $\widetilde{H}_{m^*}^{g}$ and bounded residual refinement produces the final enhanced recovered representation $\widehat{H}_{m^*}$. Since the recovery module is frozen, $\mathcal{L}_{rec}$ and $\mathcal{L}_{align}$ are not optimized in this stage. The stage-two objective is therefore

$$\mathcal{L}_{stage2} = \mathcal{L}_{task}^{(2)} + \mathcal{L}_{con}. \tag{22}$$

Where, $\mathcal{L}_{task}^{(2)}$ is the task loss computed with the enhanced recovered representations, and $\mathcal{L}_{con}$ encourages the enhanced recovered representations to remain consistent with the observed context. The overall strength of retrieval-guided enhancement is controlled by the global enhancement budget c, while the sample-wise coefficients η_p and η_r regulate the influence of retrieval evidence at the prior-guidance and residual-refinement levels, respectively.

In both stages, the representation used by the downstream prediction branch is generated from the condition-only prior path. During training, the ground-truth target modality H_{m^*} is used only to construct the posterior path and provide supervision for $\mathcal{L}_{rec}$ and $\mathcal{L}_{align}$, rather than being directly fed into the prediction branch. This design prevents target information leakage and maintains consistency between training and inference.

Through this two-stage strategy, MDRC first establishes a stable shared-private conditional recovery basis and then introduces retrieval evidence in a controlled manner. The overall loss in the framework summarizes the complete set of objectives used

by MDRC, while the stage-wise objectives specify how these losses are activated during training.

4 Experiment

This section evaluates MDRC from three aspects: overall effectiveness, robustness under modality missingness, and component-level contribution. Experiments are conducted under both fixed and random missing-modality settings. In the fixed setting, different modality combinations are used as available inputs, where the remaining modalities constitute the recovery targets. In the random setting, missing patterns are generated according to predefined missing rates to examine robustness under different degrees of modality incompleteness. Ablation studies are further conducted to analyze the contributions of shared-private conditional recovery, retrieval-guided enhancement, post-enhancement consistency regularization, and the two-stage training strategy.

4.1 Datasets and Implementation Details

We evaluate MDRC on two widely used multimodal emotion recognition benchmarks, CMU-MOSI [32] and CMU-MOSEI [33]. Both datasets provide textual, acoustic, and visual modalities, with each sample associated with a continuous sentiment score in the range of -3 to 3. CMU-MOSI consists of 2,199 opinion video clips, partitioned into 1,284 training, 229 validation, and 686 test samples. CMU-MOSEI is a larger-scale benchmark containing 22,856 video clips, of which 16,326, 1,871, and 4,659 are allocated to the training, validation, and test sets, respectively.

For fixed missing-modality evaluation, we consider seven available-modality settings: $\{t\}$, $\{a\}$, $\{v\}$, $\{t, a\}$, $\{t, v\}$, $\{a, v\}$, and $\{t, a, v\}$. The available modalities are used as observed inputs, the missing modalities are inferred from the available modalities. This protocol evaluates the performance of MDRC under typical missing-modality patterns, especially in challenging cases where only weak acoustic or visual cues are available. For random missing-modality evaluation, missing patterns are randomly generated with missing rates ranging from 0.0 to 0.7. A missing rate of 0.0 corresponds to the full-modality setting, while a larger value indicates more severe modality incompleteness. Each sample preserves at least one observed modality, ensuring that conditional recovery remains feasible. The same missing-modality protocol is consistently adopted during training, validation, and testing to enable fair comparisons.

We compare MDRC with representative incomplete multimodal emotion recognition methods, including DCCAE, MCTN, LNLN, GCNET, and MPLMM. Following common practice, binary accuracy (Acc-2) and F1-score are adopted as the main evaluation metrics. Continuous sentiment scores are converted into positive and negative categories according to the standard non-zero evaluation protocol, where samples with positive scores are labeled as positive, whereas those with negative scores are labeled as negative.

Table 1: Performance comparison under the fixed missing-modality protocol on MOSI and MOSEI.

Dataset	Available	DCCAE	MCTN	LNLN	GCNET	MPLMM	Ours
MOSI	l	76.4/76.5	79.1/79.2	84.9/85.1	83.7/83.6	83.8/83.8	**84.9**/**84.9**
	v	52.6/51.1	55.0/54.4	52.2/58.9	56.1/55.7	54.7/42.1	**60.2**/**60.1**
	a	48.8/42.1	56.1/54.5	52.2/58.9	56.1/54.5	54.6/42.0	**59.6**/**59.7**
	l,v	76.7/76.8	81.1/81.2	84.9/85.1	84.3/84.2	83.8/83.7	**85.2**/**85.2**
	l,a	77.0/77.0	81.0/81.0	84.9/85.2	84.5/84.4	83.7/83.6	**85.2**/**85.2**
	v,a	54.0/52.5	57.5/57.4	52.2/58.9	**62.0**/**61.9**	57.3/42.6	61.3/61.2
	l,v,a	77.3/77.4	81.4/81.5	84.3/84.6	**85.2**/**85.1**	83.8/83.8	85.1/**85.1**
	Average	66.1/64.8	70.2/69.9	70.8/73.8	73.1/72.8	71.7/65.9	**74.5**/**74.4**
MOSEI	l	79.7/79.5	82.6/82.8	84.1/84.4	83.0/83.2	84.6/84.6	**85.0**/**84.9**
	v	61.1/57.2	62.6/57.1	62.9/**77.2**	61.9/61.6	62.2/53.2	**65.2**/63.1
	a	61.4/53.8	62.7/54.5	62.9/**77.2**	60.2/60.3	61.8/49.4	**64.6**/63.2
	l,v	80.4/80.4	83.2/83.2	84.4/84.7	84.3/84.4	85.2/85.1	**85.3**/**85.2**
	l,a	80.0/80.0	83.5/83.3	84.1/84.4	84.3/84.4	85.1/**85.1**	**85.1**/85.0
	v,a	62.7/59.2	63.7/62.7	62.9/**77.2**	64.1/57.2	63.2/51.0	**65.7**/63.5
	l,v,a	81.2/81.2	84.2/84.2	84.1/84.5	85.2/85.1	**85.3**/**85.3**	**85.3**/**85.3**
	Average	72.4/70.2	74.6/72.5	75.1/**81.4**	74.7/73.7	75.3/70.5	**76.6**/75.7

4.2 Comparison with State-of-the-art Methods

To evaluate the effectiveness of MDRC, we compare it with several representative incomplete multimodal emotion recognition methods, including DCCAE, MCTN, LNLN, GCNET, and MPLMM. These baselines address incomplete modalities through correlated representation learning, cross-modal translation, dominant-modality modeling, graph-based completion, or prompt-based missing-pattern modeling. In contrast, MDRC combines shared-private conditional recovery with retrieval-guided enhancement, aiming to obtain recovered representations that are both stable and task-discriminative.

Table 1 reports the results under the fixed missing-modality protocol. MDRC achieves competitive overall performance on both MOSI and MOSEI and obtains the best average Acc-2 results across different available-modality combinations. The improvements are more noticeable in several text-missing settings, such as acoustic-only, visual-only, and acoustic-visual inputs, where the available information is relatively weak. This suggests that the proposed recovery and retrieval-guided enhancement mechanisms can help alleviate the information insufficiency caused by missing textual modality. When textual information is available, MDRC maintains comparable performance to strong baselines, indicating that the proposed modules do not substantially disturb the already informative textual representations.

Table 2 presents the results under the random missing-modality protocol. As the missing rate increases, the performance of all methods generally decreases, showing that random modality missing weakens cross-modal complementarity. Compared with the baselines, MDRC maintains relatively stable performance across different missing rates and achieves the best average results on both MOSI and MOSEI. These results

Table 2: Performance comparison under the random missing-modality protocol on MOSI and MOSEI.

Dataset	Available	DCCAE	MCTN	LNLN	GCNET	MPLMM	Ours
MOSI	0.0	77.3/77.4	77.2/77.2	84.2/84.0	78.7/78.7	83.8/83.8	**85.1**/**85.1**
	0.1	74.5/74.7	72.2/72.2	81.4/81.4	76.5/76.5	80.6/80.7	**83.6**/**83.6**
	0.2	71.8/71.9	69.5/69.5	78.7/78.7	74.3/74.3	75.0/75.1	**80.7**/**80.7**
	0.3	67.0/66.7	69.0/69.0	74.7/74.8	71.3/71.4	74.7/74.8	**77.0**/**77.1**
	0.4	63.6/62.8	65.9/65.8	70.3/69.9	69.2/69.1	70.0/69.8	**73.5**/**73.5**
	0.5	62.0/61.3	62.0/61.1	67.1/66.3	66.8/66.9	67.4/66.6	**70.9**/**70.8**
	0.6	59.6/58.5	60.7/58.1	62.5/60.5	62.9/63.0	63.4/61.8	**68.8**/**68.5**
	0.7	58.1/57.4	57.1/53.7	61.0/58.3	61.1/61.1	59.2/56.2	**67.9**/**67.8**
	Average	66.7/66.3	66.7/65.8	72.5/71.7	70.1/70.1	72.1/71.5	**75.9**/**75.9**
MOSEI	0.0	81.2/81.2	84.2/84.2	85.1/85.1	85.2/85.1	**85.3**/**85.3**	**85.3**/**85.3**
	0.1	78.4/78.3	81.8/81.6	83.1/83.0	82.3/82.1	83.7/83.5	**84.0**/**83.7**
	0.2	75.5/75.4	79.0/78.7	81.7/81.0	80.3/79.9	81.4/80.8	**82.7**/**82.5**
	0.3	72.3/72.2	76.9/76.2	79.1/78.1	77.5/76.8	79.2/78.3	**80.4**/**80.1**
	0.4	70.3/70.0	74.3/74.1	76.9/75.6	76.0/74.9	77.1/76.1	**77.6**/**77.3**
	0.5	69.2/66.4	73.6/72.6	75.1/72.4	74.9/73.2	75.6/73.8	**75.7**/**74.9**
	0.6	67.6/63.2	73.2/71.1	72.5/68.6	**74.1**/72.1	72.7/70.6	74.0/**73.5**
	0.7	66.6/62.6	72.7/70.5	70.6/65.2	73.2/70.4	70.5/69.2	**73.5**/**72.9**
	Average	72.6/71.2	77.0/76.1	78.0/76.1	77.9/76.8	78.2/77.2	**79.2**/**78.8**

indicate that MDRC can adapt not only to fixed missing patterns, but also to more uncertain missing-modality conditions.

Overall, the comparison results suggest that MDRC provides a balanced solution for incomplete multimodal emotion recognition. The shared-private conditional recovery module provides a stable base representation for missing modalities, while retrieval-guided enhancement further introduces external evidence in a controlled manner. This combination contributes to the robustness of MDRC under both fixed and random missing-modality settings.

Quantitative analysis. To investigate the contribution of each key component in MDRC, we conduct ablation experiments on the MOSI dataset. The complete model is denoted as Full MDRC. We construct five ablation variants: w/o Recovery removes the shared-private conditional recovery module; w/o Retrieval Enhancement removes the retrieval-guided enhancement module; w/o Prior Guidance removes the condition-side prior guidance mechanism; w/o Residual removes the bounded residual enhancement mechanism; and w/o Two-stage removes the two-stage training strategy. Fig. 5 reports the performance drops of different ablation variants compared with Full MDRC under both fixed missing-modality scenarios and varying missing rates. A larger drop indicates that the corresponding component contributes more significantly to the final performance.

From the fixed missing-modality results, removing the recovery module or the two-stage training strategy usually leads to more evident performance degradation, especially in text-missing scenarios such as acoustic-only, visual-only, and acoustic-visual available settings. This indicates that the shared-private conditional recovery module is a fundamental component of MDRC, since it provides stable recovered

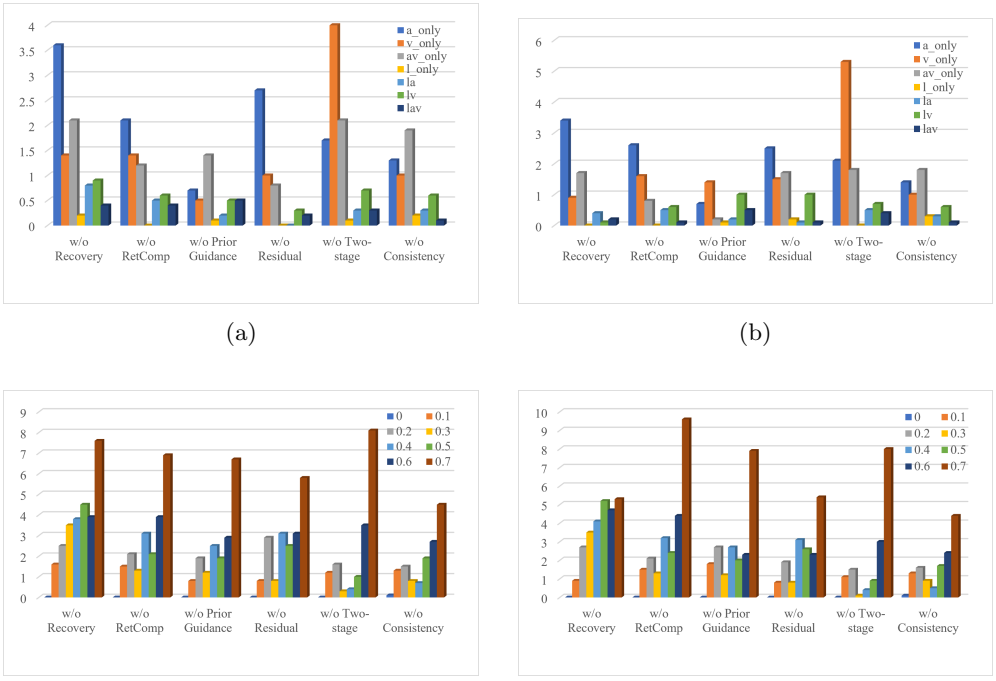


Fig. 5: Performance drop of ablation variants on MOSI. (a) Acc-2 drop under fixed missing scenarios; (b) F1 drop under fixed missing scenarios; (c) Acc-2 drop under varying missing rates; (d) F1 drop under varying missing rates.

representations for missing modalities. Meanwhile, the two-stage training strategy helps the model first learn reliable base recovery ability before introducing retrieval evidence, thereby reducing the interference caused by unstable retrieval guidance. In contrast, the performance drops under the text-only setting are relatively smaller, suggesting that the textual modality itself contains stronger affective semantics and the model relies less on recovery and retrieval enhancement in this scenario.

For retrieval-related variants, w/o Retrieval Enhancement, w/o Prior Guidance, and w/o Residual all cause performance drops to different extents, demonstrating the effectiveness of the retrieval-guided enhancement module and its two sub-mechanisms. The condition-side prior guidance adjusts the recovery direction by injecting retrieval evidence into the conditional context, while bounded residual enhancement further refines the recovered representation at the feature level. These two mechanisms complement each other, enabling MDRC to exploit external sample evidence while limiting the negative influence of unreliable retrieval results.

The random missing-rate results further show that the performance drops of most ablation variants increase as the missing rate becomes higher. This suggests that when modality incompleteness becomes more severe, the model depends more strongly

Table 3: Average performance across missing rates.

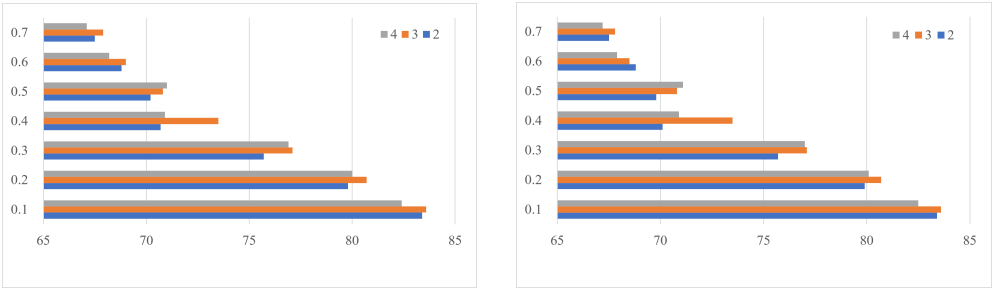
Top-k	Top-2	Top-3	Top-4
Acc-2	73.7	74.7	73.8
F1-score	73.6	74.6	73.8

on conditional recovery, retrieval enhancement, and staged optimization. In particular, w/o Recovery, w/o Retrieval Enhancement, and w/o Two-stage exhibit more obvious degradation under high missing rates, confirming the importance of base recovery, external retrieval evidence, and stable training strategy for robust incomplete multimodal emotion recognition.

Overall, the ablation results demonstrate that the performance gain of MDRC does not come from a single component, but from the collaboration of shared-private conditional recovery, retrieval-guided enhancement, and two-stage training. The recovery module provides stable base representations for missing modalities, the retrieval enhancement module enriches recovered representations with task-relevant external evidence, and the two-stage strategy stabilizes the interaction between recovery and enhancement. Therefore, Full MDRC achieves more robust emotion recognition performance under both fixed and random missing-modality settings.

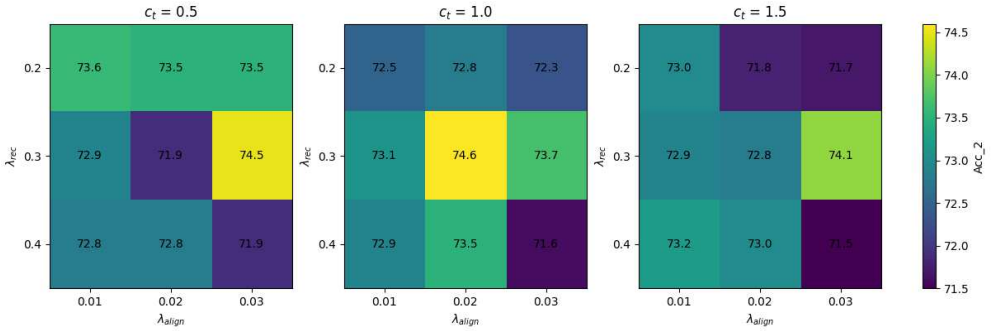
4.3 Parameter Analysis

Effect of retrieval top-K. To assess the sensitivity of MDRC to retrieval top-K, we compare three retrieval settings ($K = 2, 3, 4$) across varying missing rates on the MOSI dataset. These values were chosen to cover a reasonable range: smaller K may not provide sufficient external references, while larger K could introduce less relevant or noisy samples. As shown in Fig. 6 and table. 3, MDRC's performance remains relatively stable across these settings, indicating that the model is not highly sensitive to the number of retrieved samples. Empirically, $K = 3$ achieves a balanced trade-off between retrieval diversity and reliability, and is therefore adopted as the default in subsequent experiments.

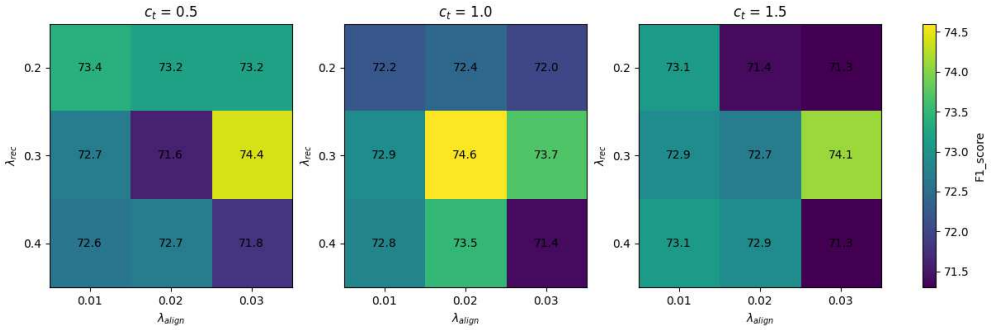


(a) Acc-2 (b) F1-score

Fig. 6: Effect of retrieval top-k under varying missing rates on MOSI.



(a) Acc-2



(b) F1-score

Fig. 7: Hyperparameter sensitivity analysis of MDRC on MOSI. The first row reports Acc-2 and the second row reports F1-score under different combinations of λ_{rec}, λ_{align}, and retrieval enhancement strength c.

Effect of loss weights and retrieval enhancement strength. After selecting the retrieval top-K, we evaluate the impact of three key hyperparameters: the recovery loss weight λ_{rec}, the latent alignment loss weight λ_{align}, and the retrieval enhancement strength c. Each parameter is tested with three candidate values, resulting in $3\times3\times3 = 27$ combinations.

As shown in Fig.7, MDRC's performance remains stable across different settings. Moderate values, specifically $\lambda_{rec} = 0.3$, $\lambda_{align} = 0.02$, and $c = 1.0$, achieve the best balance, providing sufficient recovery supervision, effective latent alignment, and stable retrieval-guided enhancement without overfitting to external evidence. These values are adopted as the default configuration in subsequent experiments.

5 Conclusion

This paper proposes MDRC, a mechanism-decoupled retrieval-guided conditional recovery framework for incomplete multimodal emotion recognition. MDRC addresses

the problems of insufficient missing-modality information and limited discriminability under weak-modality conditions. It first generates base recovered representations through a shared-private conditional recovery module, and then introduces external sample evidence through retrieval-guided enhancement to improve the affective relevance and task discriminability of the recovered representations. A two-stage training strategy is further adopted to separate base recovery learning from retrieval enhancement learning, improving the stability of the training process.

Experiments on CMU-MOSI and CMU-MOSEI demonstrate that MDRC achieves stable performance under both fixed and random missing-modality settings. Compared with existing methods, MDRC shows clearer advantages when only acoustic or visual modalities are available, indicating its effectiveness in alleviating the affective information insufficiency caused by textual modality missing or weak-modality-dominant conditions. Ablation studies further verify the contributions of the conditional recovery module, the retrieval enhancement module, and the two-stage training strategy.

Despite its effectiveness, MDRC still has certain limitations. When the observed modalities contain weak affective cues or strong noise, the recovered missing-modality representations may still be insufficient and may not fully approach the discriminative ability of the ground-truth modalities. In addition, although retrieval enhancement introduces external sample evidence, the retrieved samples may contain semantic mismatch or affective inconsistency, which can bring additional noise into the recovered representations. In future work, we will explore more reliable missing-modality recovery mechanisms, more fine-grained retrieval reliability estimation, and noise suppression strategies for generated or recovered representations, thereby improving model robustness and generalization across diverse missing-modality settings.

Declarations

- **Funding** This work is supported by the Chongging Municipal Economy and Information Technology Commission (grant No. YJX-2025001001008).

- **Conflict of interest/Competing interests** The authors declare no conflict of interest.
- **Ethics approval and consent to participate** Not applicable.
- **Consent for publication** Not applicable.
- **Data availability** No datasets were generated or analysed during the current study.
- **Materials availability** Not applicable.
- **Code availability** Not applicable.
- **Author contribution** Qianxi Zhang: Conceptualization, Methodology, Investigation, Writing, original draft preparation, Validation. Shengcai Wang: Validation, Writing, review and editing. Nengchao Wu: Methodology, Formal analysis, Writing, review and editing. Junjie Bao: Data curation, Investigation. Xinyu Yang: Data curation, Investigation. Qing Tao: Conceptualization, Methodology, Supervision. Na Cui: Data curation, Resources. Rangxiang Zhao: Data curation, Resources. Li Liu: Conceptualization, Methodology, Supervision, Funding acquisition, Writing, review and editing.

References

[1] Lang, J., Cheng, Z., Zhong, T., Zhou, F.: Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning (2025). https://arxiv.org/abs/2501.01120

[2] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Bagher Zadeh, A., Morency, L.-P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2247–2256. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1209

[3] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor Fusion Network for Multimodal Sentiment Analysis (2017). https://arxiv.org/abs/1707.07250

[4] Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal Transformer for Unaligned Multimodal Language Sequences (2019). https://arxiv.org/abs/1906.00295

[5] Hazarika, D., Zimmermann, R., Poria, S.: MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (2020). https://arxiv.org/abs/2005.03545

[6] Pipoli, V., Saporita, A., Bolelli, F., Cornia, M., Baraldi, L., Grana, C., Cucchiara, R., Ficarra, E.: Missrag: Addressing the missing modality challenge in multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3215–3224 (2025)

[7] Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2608–2618. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.203

[8] Chen, C., Zhang, P.: Modality-Collaborative Transformer with Hybrid Feature Reconstruction for Robust Emotion Recognition (2023). https://arxiv.org/abs/2312.15848

[9] Li, H., Han, H.: Multimodal emotion recognition from complete modality to missing modality based on text, audio, and visual: A review. Engineering Applications of Artificial Intelligence **170**, 114127 (2026) https://doi.org/10.1016/j.engappai.2026.114127

[10] Wu, R., Wang, H., Chen, H.-T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. Transactions on Machine Learning Research (2026). Survey Certification

[11] Wang, W., Arora, R., Livescu, K., Bilmes, J.: On Deep Multi-View Representation Learning: Objectives and Optimization (2016). https://arxiv.org/abs/1602.01024

[12] Pham, H., Liang, P.P., Manzini, T., Morency, L.-P., Póczos, B.: Found in translation: Learning robust joint representations by cyclic translations between modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6892–6899 (2019). https://doi.org/10.1609/aaai.v33i01.33016892

[13] Zhang, H., Wang, W., Yu, T.: Towards Robust Multimodal Sentiment Analysis with Incomplete Data (2024). https://arxiv.org/abs/2409.20012

[14] Lian, Z., Chen, L., Sun, L., Liu, B., Tao, J.: GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation (2023). https://arxiv.org/abs/2203.02177

[15] Guo, Z., Jin, T., Zhao, Z.: Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition (2024). https://arxiv.org/abs/2407.05374

[16] Weng, Y., Wang, H., Gao, T., Li, K., Niu, S., Du, J.: Enhancing Multimodal Sentiment Analysis for Missing Modality through Self-Distillation and Unified Modality Cross-Attention (2025). https://arxiv.org/abs/2410.15029

[17] Wu, S., Wang, X., Wang, L., He, D., Dang, J.: Enriching Multimodal Sentiment Analysis through Textual Emotional Descriptions of Visual-Audio Content (2024). https://arxiv.org/abs/2412.10460

[18] Lee, M., Pavlovic, V.: Private-Shared Disentangled Multimodal VAE for Learning of Hybrid Latent Representations (2020). https://arxiv.org/abs/2012.13024

[19] Lin, R., Hu, H.: Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. Transactions of the Association for Computational Linguistics **11**, 1686–1702 (2023) https://doi.org/10.1162/tacla00628

[20] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2021). https://arxiv.org/abs/2005.11401

[21] Fan, Q., Yuan, H., Zuo, H., Liu, R., Gao, G.: Leveraging Retrieval Augment Approach for Multimodal Emotion Recognition Under Missing Modalities (2024). https://arxiv.org/abs/2410.02804

[22] Wang, H., Prasad, A., Stengel-Eskin, E., Bansal, M.: Retrieval-augmented generation with conflicting evidence. In: Second Conference on Language Modeling (2025)

[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2015). https://arxiv.org/abs/1512.03385

[24] Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L.: A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 10074–10082 (2024)

[25] Liu, W., Zhan, H.: Multimodal Sentiment Analysis with Missing Modality: A Knowledge-Transfer Approach (2026). https://arxiv.org/abs/2401.10747

[26] Wang, Y., Li, Y., Cui, Z.: Incomplete multimodality-diffused emotion recognition. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems vol. 36, pp. 17117–17128. Curran Associates, Inc., Red Hook, NY, USA (2023)

[27] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (2022). https://arxiv.org/abs/1312.6114

[28] Sutter, T.M., Daunhawer, I., Vogt, J.E.: Generalized Multimodal ELBO (2021). https://arxiv.org/abs/2105.02470

[29] Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems vol. 31. Curran Associates, Inc., Red Hook, NY, USA (2018)

[30] Shi, Y., N, S., Paige, B., Torr, P.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems vol. 32. Curran Associates, Inc., Red Hook, NY, USA (2019)

[31] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems vol. 28, pp. 3483–3491. Curran Associates, Inc., Red Hook, NY, USA (2015)

[32] Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P.: MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (2016). https://arxiv.org/abs/1606.06259

[33] Bagher Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable

dynamic fusion graph. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1208