HiCTE: Hierarchical Cross-Expert Transferability Estimation for Incomplete Multimodal Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Incomplete multimodal representation learning aims to construct robust representations from partially observed inputs. Existing expert-based methods learn joint representations through cross-expert interaction, but typically transfer entire observed representations to different target experts. Such coarse-grained transfer overlooks target-specific transferability differences at the latent-dimension level and sample-specific transfer requirements. To address these limitations, we propose the Hierarchical Cross-Expert Transferability Estimation (HiCTE) framework. Specifically, HiCTE models each observed modality in a probabilistic latent space and derives reliability from the latent variance. A global estimator then provides initial transferability estimates for each target expert using the latent information and reliability. These estimates are further refined for each sample through local iterative optimization during training, with gradients propagated through the updates to train the global estimator. Based on these estimates, HiCTE constructs cross-expert representations for each target expert and a residual representation, then integrates them with the self representation through adaptive weighting. Extensive experiments on three multimodal benchmarks demonstrate the effectiveness of HiCTE under different missing protocols.

1 Introduction

Multimodal representation learning aims to integrate heterogeneous information from multiple modalities into effective representations for downstream prediction (Wang et al., 2026; Qian et al., 2026; Xiao et al., 2024; Dufumier et al., 2025). By exploiting complementary information across modalities, existing methods have achieved substantial progress in a wide range of multimodal tasks (Jiang et al., 2026; Xu et al., 2024). However, in real-world scenarios, multimodal data are often only partially observed due to sensor failures, data corruption, or privacy constraints (Jaiswal & Mower Provost, 2020; Wu et al., 2024). The absence of one or more modalities disrupts cross-modal interaction and often leads to substantial degradation in representation quality and downstream performance. Therefore, learning robust representations from incomplete multimodal inputs remains an important challenge in multimodal learning.

Existing approaches to incomplete multimodal learning mainly fall into two categories: missing-modality recovery and knowledge distillation. Recovery-based methods (Wang et al., 2023a;b; Zhao et al., 2021a) reconstruct unavailable modalities or latent representations before multimodal interaction, while distillation-based methods (Li et al., 2024a;c; Wei et al., 2023) typically exploit fully observed samples to guide representation learning from incomplete inputs. Despite their effectiveness, both paradigms often require fully observed multimodal samples during training, either as reconstruction targets or as sources of supervision. Such a requirement is difficult to satisfy when some modalities are already unavailable during training.

Accordingly, expert-based methods have emerged as a new paradigm for incomplete multimodal representation learning. These methods construct multimodal representations directly from observed modalities without explicitly recovering the missing ones (Xu et al., 2024). Through modality-specific experts, available modality features are processed by different experts, and the resulting representations are adaptively combined for prediction. However, existing expert-based methods

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

typically transfer an observed modality representation to different target experts as a whole. They do not explicitly distinguish which latent dimensions are useful for different target experts.

More importantly, availability does not imply transferability. Different latent dimensions within the same observed representation may exhibit different levels of transferability toward different target experts, while less reliable dimensions may introduce noise during cross-expert transfer. Moreover, even for the same source–target expert pair, the required transfer pattern may vary across individual samples. Such coarse-grained transfer therefore overlooks both target-specific differences in transferability and sample-specific variations in transfer requirements.

To address these limitations, we propose the Hierarchical Cross-Expert Transferability Estimation (HiCTE) framework. Specifically, HiCTE models each observed modality in a probabilistic latent space and derives reliability from the latent variance. Target-conditioned global estimators then use the latent information and reliability to provide initial transferability estimates toward different target experts. These estimates are further refined for each sample through K steps of local iterative optimization during training, using a sample-specific supervised transferability objective. The refinement steps are differentiably unrolled so that gradients through the local updates also train the global estimators. Based on these estimates, HiCTE constructs cross-expert representations for each target expert and a residual representation, then integrates them with the self representation through adaptive weighting for final prediction.

In summary, our main contributions are as follows:

- We propose HiCTE, a hierarchical cross-expert transferability estimation method. It selectively transfers latent dimensions to each target expert, rather than treating the entire observed representation as uniformly transferable.
- We propose a global–local iterative transferability estimation mechanism to capture sample-specific transfer requirements. The global estimator provides target-conditioned initial transferability estimates for each sample, while K-step local iterative refinement progressively adapts these estimates to individual samples during training. Gradients are backpropagated through the refinement steps to update the global estimator.
- Extensive experiments on three widely used benchmark datasets demonstrate the effectiveness of HiCTE under both fixed missing-modality and random missing-rate settings.

2 Related Work

2.1 Incomplete Multimodal Representation Learning

Incomplete multimodal representation learning aims to construct robust representations when only a subset of modalities is observed. Existing approaches can be broadly grouped into recovery-based and distillation-based approaches. Recovery-based methods infer missing modalities or reconstruct latent features from the available observations (Wang et al., 2023a;b; Zhao et al., 2021a; Zhuang et al., 2026; Guo et al., 2024). For example, TMDC (Zhuang et al., 2026) first extracts denoised modality-specific and modality-shared representations and then uses them to complement missing modalities. Distillation-based methods instead transfer knowledge from complete multimodal representations or teacher models to incomplete-input models (Li et al., 2024c;a; Wei et al., 2023; Li et al., 2024b). For example, MMANet (Wei et al., 2023) employs a complete-modality teacher network to transfer multimodal knowledge to the deployment network through margin-aware distillation. Although effective, both lines of work generally introduce additional dependence on complete multimodal information during training. This has motivated methods that operate directly on the available modalities without explicit missing-modality reconstruction.

2.2 Expert-Based Learning for Incomplete Multimodal Inputs

Expert-based methods provide an alternative approach to learning from incomplete multimodal inputs without explicitly reconstructing the missing modalities (Xu et al., 2024). By leveraging modality-specific experts, these methods directly construct representations from available modality features and combine the resulting expert representations for prediction. For example, MoMKE (Xu

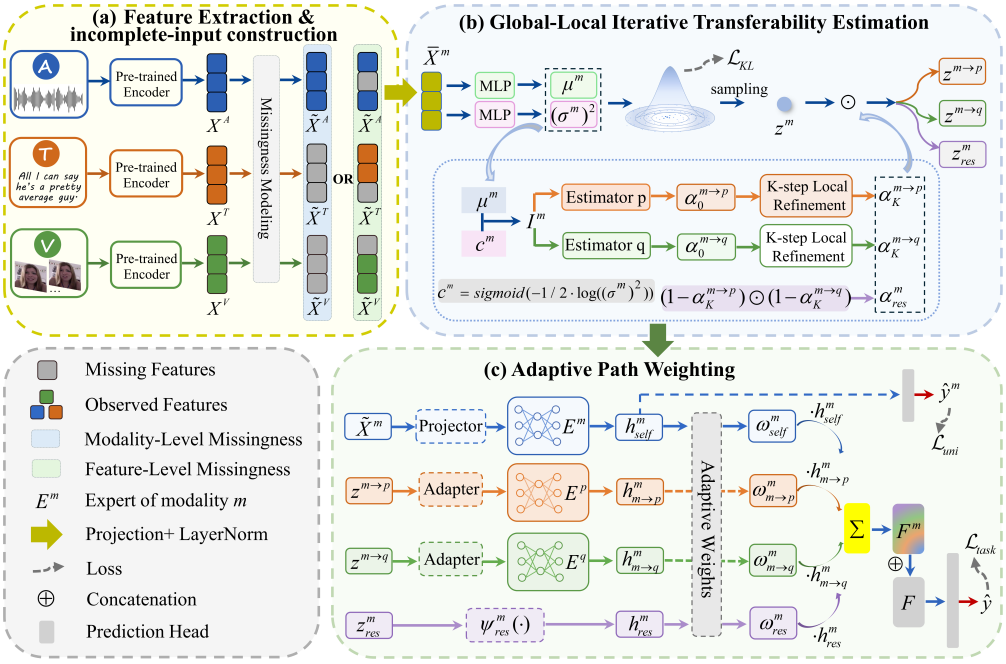


Figure 1: Overview of HiCTE. (a) Modality-specific feature extraction and incomplete-input construction. (b) Global–local iterative transferability estimation uses latent reliability to initialize target-conditioned, dimension-wise estimates and refines them locally during training. (c) Adaptive path weighting combines self, target-specific cross-expert, and residual representations for prediction. Panels (b)–(c) illustrate the audio-only condition $\mathcal{O} = \{A\}$, with $m = A$, $p = T$, and $q = V$; the same procedure applies to other observed source modalities.

et al., 2024) employs modality-specific experts and expert-level fusion to construct representations under different incomplete-input conditions.

However, existing expert-based methods typically perform routing and fusion at the representation level, without explicitly modeling dimension-wise transferability between source and target experts. This overlooks the fact that an observed representation may contain latent information that is useful for one target expert but less relevant to another. HiCTE addresses this limitation by explicitly estimating which latent dimensions of each observed modality should be transferred to different target experts. These target-specific transferability estimates are further refined for individual samples during training to capture sample-dependent transfer requirements.

3 METHODOLOGY

Motivation. Existing expert-based methods can directly exploit observed modalities, but their cross-expert interaction typically operates on entire source representations. Such interaction does not explicitly model differences in transferability across latent dimensions and target experts. The transfer pattern may also vary across samples. In addition, uncertainty in latent dimensions may impair cross-expert transfer. To address these issues, HiCTE derives dimension-wise reliability from the latent variance. Global estimators then combine latent information and reliability to produce initial transferability estimates for each target expert. These estimates are further refined for individual samples through local iterative optimization of a sample-specific transferability objective.

Overview. As illustrated in Fig. 1, HiCTE consists of three stages. First, modality-specific features are extracted using pretrained encoders (Appendix A.3), and incomplete inputs are constructed according to modality availability and feature-level masks. Second, for each observed modality, HiCTE models its representation in a probabilistic latent space and derives dimension-wise reliability. Target-conditioned global estimators then produce initial transferability estimates for different target experts. During training, the initial transferability estimates are refined through K differen-

tiably unrolled local optimization steps. The resulting gradients are propagated through the refinement process to update the global estimators. The resulting estimates separate the source latent representation into two target-specific transferred components and a residual component (Section 3.2). Third, the transferred components are mapped to their corresponding cross-expert representation paths and, together with the self and residual paths, are integrated through adaptive path weighting to construct the modality representation for prediction (Section 3.3). The overall training objective is described in Section 3.4.

3.1 Problem Formulation

Consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N multimodal sequences with utterance-level labels y_i. Let $\mathcal{M} = \{A, T, V\}$ denote the acoustic, textual, and visual modalities. Under complete observation, $x_i = \{X_i^m\}_{m\in\mathcal{M}}$, where $X_i^m \in \mathbb{R}^{S_i \times d_m}$ contains the utterance-level features of modality m. Here, S_i and d_m denote the utterance count and feature dimension, respectively.

We consider incompleteness at both the modality and feature levels. Let $\delta_i^m \in \{0, 1\}$ indicate whether modality m is observed, and let $R_i^m \in \{0, 1\}^{S_i \times d_m}$ denote the feature-level observation mask. The incomplete observation is defined as

$$\widetilde{X}_i^m = \delta_i^m \left(R_i^m \odot X_i^m\right), \quad m \in \mathcal{M}, \tag{1}$$

where $\odot$ denotes element-wise multiplication. The observed modality set is $\mathcal{O}_i = \{m \in \mathcal{M} \mid \delta_i^m = 1\}$. Under the fixed missing-modality setting, $R_i^m = \mathbf{1}$ for each observed modality, while δ_i^m determines modality-level availability. Under the random missing-rate setting, all modalities remain observed with $\delta_i^m = 1$, while R_i^m controls feature-level incompleteness.

Given the incomplete input $\widetilde{x}_i = \{\widetilde{X}_i^m\}_{m\in\mathcal{M}}$ and modality availability $\Delta_i = \{\delta_i^m\}_{m\in\mathcal{M}}$, our goal is to learn a robust multimodal representation

$$F_i = G(\widetilde{x}_i, \Delta_i), \qquad \hat{y}_i = g(F_i), \tag{2}$$

where G denotes the representation learning model and g denotes the downstream predictor.

3.2 Global–Local Iterative Transferability Estimation

3.2.1 Latent Reliability Modeling

For each observed modality $m \in \mathcal{O}_i$, we model its representation in a probabilistic latent space to characterize latent information and its reliability. The incomplete input feature sequence is first projected into a common latent dimension and normalized as $\bar{X}_i^m = \mathrm{LayerNorm}(\mathrm{Proj}^m(\widetilde{X}_i^m))$.

Two modality-specific linear layers then estimate the latent mean and log-variance:

$$\mu_i^m = \bar{X}_i^m (W_\mu^m)^\top + b_\mu^m, \qquad \log(\sigma_i^m)^2 = \bar{X}_i^m (W_\sigma^m)^\top + b_\sigma^m. \tag{3}$$

The corresponding variational posterior is $q(z_i^m \mid \widetilde{X}_i^m) = \mathcal{N}\left(\mu_i^m, \mathrm{diag}\left((\sigma_i^m)^2\right)\right)$. The latent variance characterizes dimension-wise dispersion and is used as a proxy for representation uncertainty. Accordingly, the reliability of each latent dimension is derived as

$$c_i^m = \mathrm{sigmoid}\left(-\frac{1}{2}\log\left((\sigma_i^m)^2\right)\right), \tag{4}$$

such that latent dimensions with larger variance receive lower reliability. The latent mean and reliability are concatenated along the latent dimension as $I_i^m = [\mu_i^m; c_i^m]$, which serves as the input to the transferability estimator.

During training, the latent representation is sampled using the reparameterization trick: $z_i^m = \mu_i^m + \sigma_i^m \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, where ϵ has independent standard normal entries. The latent mean μ_i^m is used at inference. To regularize the probabilistic latent space, we constrain the posterior distribution toward a standard Gaussian prior:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{m\in\mathcal{M}} \frac{1}{S_m} \sum_i D_{\mathrm{KL}}\left(q(z_i^m \mid \widetilde{X}_i^m) \,\|\, \mathcal{N}(0, I)\right). \tag{5}$$

All loss sums are restricted to valid observations in the current mini-batch. Here, S_m counts valid utterances with modality m observed, and each KL term sums divergences over valid positions within sequence i.

3.2.2 GLOBAL TRANSFERABILITY INITIALIZATION

We instantiate three modality-specific experts $\{E^A, E^T, E^V\}$, implemented as Transformer encoders. Modality-level supervision encourages each expert to retain modality-specific, task-discriminative information for target-conditioned transferability estimation. For each observed modality m, we supervise its self representation using the corresponding prediction head: $\hat{y}_i^m = g_m\Big(E^m\Big(\mathrm{Proj}^m(\widetilde{X}_i^m)\Big)\Big)$. The modality-level supervision is defined as

$$\mathcal{L}_{\text{uni}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{S_m} \sum_i \ell(\hat{y}_i^m, y_i), \tag{6}$$

where ℓ sums task-specific supervised losses over valid utterances in each sequence.

To strengthen the modality-specific experts at the early stage of training, the weight of $\mathcal{L}_{\text{uni}}$ follows

$$\lambda_{\text{uni}}^{(e)} = \begin{cases} 2, & 1 \le e \le 30, \\ 1, & e > 30. \end{cases} \tag{7}$$

This schedule uses stronger modality-specific supervision during early training and weaker supervision during subsequent joint optimization.

For an observed source modality $m \in \mathcal{O}_i$, the other two modality-specific experts are treated as target experts, denoted by E^p and E^q, where $\mathcal{M} \setminus \{m\} = \{p, q\}$. Given the reliability-aware input I_i^m, the target-conditioned estimator $\phi_\Theta^{m\to t}$, where $t \in \{p, q\}$, produces the initial transferability estimates toward target expert E^t:

$$\alpha_{i,0}^{m\to t} = \phi_\Theta^{m\to t}(I_i^m), \qquad \alpha_{i,0}^{m\to t} \in (0,1)^{S_i \times d_z}, \quad t \in \{p, q\}. \tag{8}$$

Each element of $\alpha_{i,0}^{m\to t}$ measures how transferable the corresponding latent dimension of source modality m is to target expert E^t. Because the estimator is conditioned on the target expert, the same source representation can yield different transferability patterns for different targets.

We define a common transferability objective for a candidate transferability estimate α:

$$\mathcal{J}_i^{m\to t}(\alpha) = \ell_{\text{pred}}\left(f^{m\to t}(\alpha \odot z_i^m), y_i\right) + \beta \left\|\alpha \odot (\sigma_i^m)^2\right\|_1 + \lambda_b \left(\frac{1}{d_z}\|\alpha\|_1 - 0.5\right)^2, \tag{9}$$

where $f^{m\to t}$ denotes the target-transfer prediction mapping through target expert E^t. Position indices are omitted, and all terms are averaged over valid positions within sequence i. The ℓ_1 norms are taken over latent dimensions, with the squared constraint evaluated position-wise before averaging. The three terms measure prediction loss, penalize transfer from high-variance latent dimensions, and discourage all-zero or all-one transferability estimates. β and λ_b control the regularization terms, with λ_b examined in Appendix C.4.

To provide good initial transferability estimates for subsequent local refinement, we optimize the globally estimated transferability using Eq. 9:

$$\mathcal{L}_{\text{init}} = \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{M}| - 1} \sum_{t \in \mathcal{M} \setminus \{m\}} \bar{\mathcal{J}}_0^{m\to t}. \tag{10}$$

Here, $\bar{\mathcal{J}}_0^{m\to t}$ averages the position-wise objectives evaluated at $\alpha_{i,0}^{m\to t}$ over valid utterances with source modality m observed in the mini-batch.

3.2.3 LOCAL ITERATIVE TRANSFERABILITY REFINEMENT

Although the global estimator provides target-conditioned initial transferability estimates, the transfer requirements for the same source–target pair may vary across individual samples. We therefore refine the initial estimates for individual samples through local iterative updates during training.

To keep the transferability estimates within $(0, 1)$, we optimize an unconstrained local variable. Specifically, we initialize $u_{i,0}^{m\to t} = \mathrm{logit}(\alpha_{i,0}^{m\to t})$ and obtain the transferability at step k as $\alpha_{i,k}^{m\to t} = \mathrm{sigmoid}(u_{i,k}^{m\to t})$. The corresponding transferred latent representation is

$$z_{i,k}^{m\to t} = \alpha_{i,k}^{m\to t} \odot z_i^m.$$

Let $\bar{\mathcal{J}}_k^{m\to t}$ denote the objective in Eq. 9 averaged over valid positions with source modality m observed in the mini-batch. We update

$$u_{i,k+1}^{m\to t} = u_{i,k}^{m\to t} - \eta_{\text{inner}} \nabla_{u_{i,k}^{m\to t}} \bar{\mathcal{J}}_k^{m\to t}, \qquad k = 0, \ldots, K-1, \tag{11}$$

where η_{inner} denotes the local refinement step size. After K refinement steps, the refined transferability is $\alpha_{i,K}^{m\to t} = \text{sigmoid}(u_{i,K}^{m\to t})$.

The K refinement steps are differentiably unrolled during training, allowing gradients through $\alpha_{i,K}^{m\to t}$ to update the global estimator. At inference, no label-dependent local refinement is performed, and the learned global estimator directly produces $\alpha_{i,0}^{m\to t}$. For simplicity, we denote the transferability used for subsequent representation construction as $\alpha_i^{m\to t}$, which corresponds to $\alpha_{i,K}^{m\to t}$ during training and $\alpha_{i,0}^{m\to t}$ during inference.

For the two target experts E^p and E^q, the target-specific transferred representations are

$$z_i^{m\to p} = \alpha_i^{m\to p} \odot z_i^m, \qquad z_i^{m\to q} = \alpha_i^{m\to q} \odot z_i^m. \tag{12}$$

The two target-specific transferability estimates are obtained independently. Information with low transferability toward both targets is retained through

$$\alpha_{i,\text{res}}^m = (1 - \alpha_i^{m\to p}) \odot (1 - \alpha_i^{m\to q}), \qquad z_{i,\text{res}}^m = \alpha_{i,\text{res}}^m \odot z_i^m. \tag{13}$$

3.3 Adaptive Path Weighting

The target-conditioned transferability estimates obtained in Section 3.2 are used to construct two cross-expert representations and a residual representation for each observed modality. Together with the self representation, these representations form four candidate paths for modality representation learning.

For an observed modality $m \in \mathcal{O}_i$, the self representation is $h_{i,\text{self}}^m = E^m\left(\text{Proj}^m(\widetilde{X}_i^m)\right)$. For the two target experts E^p and E^q, the transferred representations are mapped to their corresponding expert spaces as $h_{i,m\to p}^m = E^p(\psi^{m\to p}(z_i^{m\to p}))$, $h_{i,m\to q}^m = E^q(\psi^{m\to q}(z_i^{m\to q}))$, while the residual representation is $h_{i,\text{res}}^m = \psi_{\text{res}}^m(z_{i,\text{res}}^m)$.

Let $\mathcal{B}^m = \{\text{self}, m\to p, m\to q, \text{res}\}$ denote the resulting path set. The adaptive weighting function jointly considers the projected source representation and the residual latent representation to compute the path scores:

$$\mathbf{s}_i^m = \varphi^m\left(\text{Cat}\left(\text{Proj}^m(\widetilde{X}_i^m), z_{i,\text{res}}^m\right)\right), \qquad \mathbf{s}_i^m = [s_{i,b}^m]_{b\in\mathcal{B}^m}. \tag{14}$$

The corresponding normalized weight is

$$\omega_{i,b}^m = \frac{\exp(s_{i,b}^m)}{\sum_{b'\in\mathcal{B}^m}\exp(s_{i,b'}^m)}. \tag{15}$$

The modality representation is then obtained as $F_i^m = \sum_{b\in\mathcal{B}^m} \omega_{i,b}^m h_{i,b}^m$.

3.4 Training Objective

The modality representations obtained in Section 3.3 are concatenated as $F_i = \text{Cat}(F_i^A, F_i^T, F_i^V)$, where the representations of missing modalities are zeroed, and the final prediction is $\hat{y}_i = g(F_i)$.

The main task objective is

$$\mathcal{L}_{\text{task}} = \frac{1}{S}\sum_i \ell(\hat{y}_i, y_i),$$

where S is the total number of valid utterances in the current mini-batch.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{uni}}^{(e)}\mathcal{L}_{\text{uni}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{init}}\mathcal{L}_{\text{init}}, \tag{16}$$

where λ_{KL} and λ_{init} control the latent-space regularization and global transferability initialization, respectively. The sensitivity of the key hyperparameters is analyzed in Section 4.4. Further details are provided in Appendix C.7.

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

Method	Available Modalities						
	$\{A\}$	$\{T\}$	$\{V\}$	$\{A,V\}$	$\{A,T\}$	$\{T,V\}$	*Average*
				Results on CMU-MOSI			
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
MMIN	55.30/51.50	83.80/83.80	57.00/54.00	60.40/58.50	84.00/84.00	83.80/83.90	70.72/69.28
IMDer	62.00/**62.20**	84.80/84.70	61.30/60.80	63.60/63.40	85.40/85.30	85.50/85.40	73.77/73.63
GCNet	56.10/54.50	83.70/83.60	56.10/55.70	62.00/61.90	84.50/84.40	84.30/84.20	71.12/70.72
MMANet	42.22/59.38	85.67/82.85	54.73/59.30	52.74/61.44	85.67/82.91	85.21/82.14	67.71/71.34
MoMKE	62.03/59.33	86.65/86.59	62.20/61.98	61.89/61.99	86.59/86.56	87.01/86.90	74.40/73.89
HME	52.23/55.07	84.30/84.35	59.45/59.51	60.37/57.82	83.84/83.88	83.86/83.93	70.68/70.76
PaSE	60.01/58.79	84.70/84.23	61.43/61.50	63.35/63.32	86.71/86.79	87.14/86.99	73.89/73.60
TMDC	54.73/54.96	86.59/86.60	62.96/63.18	63.57/63.78	86.99/86.95	87.65/87.56	73.75/73.84
HiCTE	**62.50**/61.38	**87.25/87.17**	**63.56/63.60**	**66.01/65.82**	**87.30/87.31**	**88.29/88.25**	**75.82/75.59**
				Results on CMU-MOSEI			
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
MMIN	58.90/59.50	82.30/82.40	59.30/60.00	63.50/61.90	83.70/83.30	83.80/83.40	71.92/71.75
IMDer	63.80/60.60	84.50/84.50	63.90/63.60	64.90/63.50	85.10/85.10	85.00/85.00	74.53/73.72
GCNet	60.20/60.30	83.00/83.20	61.90/61.60	64.10/57.20	84.30/84.40	84.30/84.40	72.97/71.85
MMANet	64.47/69.46	84.95/86.18	63.59/68.49	67.47/71.88	85.20/86.37	85.17/86.59	75.14/78.16
MoMKE	72.02/71.46	86.35/86.40	62.71/63.13	72.81/72.29	86.10/86.11	86.70/86.64	77.78/77.67
HME	58.56/55.09	84.73/84.76	61.92/59.96	63.12/60.81	84.56/84.60	84.79/84.74	72.95/71.66
PaSE	69.04/68.56	84.36/84.08	68.69/68.74	72.03/71.90	86.47/86.42	86.73/86.45	77.89/77.69
TMDC	72.35/71.66	86.41/86.47	69.75/69.28	71.27/71.09	86.78/86.81	87.03/87.07	78.93/78.73
HiCTE	**73.01/72.80**	**87.31/87.29**	**71.04/70.23**	**73.76/73.08**	**87.36/87.30**	**87.52/87.46**	**80.00/79.69**
				Results on IEMOCAP			
	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA
MMIN	56.58/59.00	66.57/68.02	52.52/51.60	63.99/65.43	72.94/75.14	72.67/73.61	64.21/65.47
MMANet	50.60/52.78	62.61/65.34	57.09/48.16	67.45/64.87	69.38/72.26	69.94/71.49	62.85/62.48
MoMKE	69.52/70.21	77.45/78.25	56.81/51.82	68.75/66.96	79.70/79.30	75.64/74.90	71.31/70.24
PaSE	67.34/–	74.80/–	57.31/–	68.75/–	78.93/–	77.52/–	70.78/–
TMDC	68.74/70.19	76.19/76.99	57.99/53.78	69.86/68.95	78.96/80.15	77.43/77.52	71.53/71.26
HiCTE	**71.97/73.02**	**77.75/78.35**	**59.87/55.81**	**72.51/72.45**	**81.40/82.28**	**78.90/79.03**	**73.73/73.49**

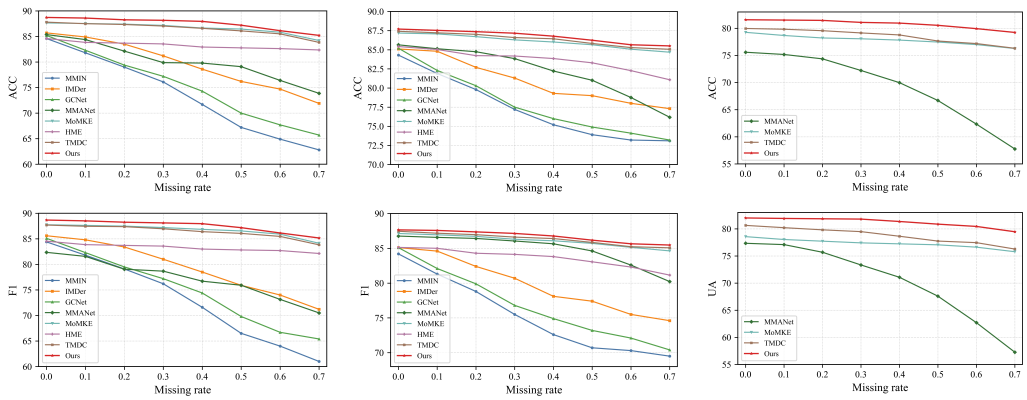
Note: $\{A\}$ denotes audio-only input; other combinations follow analogously. *Average* is computed over the six conditions. Best results are bold and second-best are underlined. "–" denotes not reported.

4 EXPERIMENTS

Datasets and Evaluation Metrics. We evaluate HiCTE on three widely used multimodal benchmarks: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). CMU-MOSI and CMU-MOSEI are used for multimodal sentiment analysis, where binary sentiment classification is evaluated using accuracy (ACC) and F1 score. For IEMOCAP, we consider four-class emotion recognition and report accuracy (ACC), unweighted accuracy (UA), and class-wise F1 scores. We further evaluate HiCTE on UR-FUNNY (Hasan et al., 2019) for cross-task validation on multimodal humor detection. The corresponding results are reported in Appendix B. Detailed dataset and metric descriptions are provided in Appendix A.1.

4.1 EXPERIMENTAL PROTOCOLS AND IMPLEMENTATION DETAILS

To evaluate HiCTE under incomplete-input conditions, we adopt two protocols: fixed missing-modality and random missing-rate. **Fixed missing-modality protocol.** During training, we consider six incomplete modality combinations: $\{A\}$, $\{T\}$, $\{V\}$, $\{A,V\}$, $\{A,T\}$, and $\{T,V\}$. For each training sample, only the modalities specified by the corresponding condition are retained. During validation and testing, the six conditions are evaluated separately with the corresponding modality combination fixed. **Random missing-rate protocol.** All three modalities remain observed, while each feature dimension at every valid time step is independently masked with probability r, following previous studies (Xu et al., 2024; Lian et al., 2023). We consider



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

Figure 2: Performance comparison under the random missing-rate protocol on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

$r \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and use $r_{\text{train}} = r_{\text{test}} = r$ for all modalities. Implementation details, hyperparameters, and feature extraction are provided in Appendices A.2 and A.3.

4.2 Comparison under Incomplete Multimodal Settings

We compare HiCTE with representative incomplete multimodal learning methods, including MMIN (Zhao et al., 2021a), GCNet (Lian et al., 2023), IMDer (Wang et al., 2023b), MMANet (Wei et al., 2023), MoMKE (Xu et al., 2024), HME (Zhuang et al., 2025), PaSE (He et al., 2026), and TMDC (Zhuang et al., 2026). Table 1 and Fig. 2 report the results under fixed missing-modality and random missing-rate protocols, respectively. Additional class-wise results on IEMOCAP are provided in Appendix C.1.

Inter-modality Missingness. As shown in Table 1, HiCTE achieves the best average performance on all three datasets across the six fixed missing-modality conditions. Compared with recovery-based methods such as TMDC (Zhuang et al., 2026), HiCTE directly exploits the observed modalities without explicitly reconstructing missing information, while compared with expert-based methods such as MoMKE (Xu et al., 2024), it models target-conditioned transferability at the latent-dimension level and further adapts it to individual samples through global–local iterative optimization. The strong performance under information-limited conditions such as $\{A\}$ and $\{A, V\}$ further indicates that HiCTE can effectively exploit the remaining observed information when only limited modality information is available.

Intra-modality Missingness. Fig. 2 shows that the performance of all methods generally decreases as the missing rate increases, reflecting the increasing difficulty caused by feature-level information loss. HiCTE consistently achieves the highest performance across the evaluated missing rates and maintains a clear advantage under severe feature-level missingness. These results show that HiCTE remains effective as feature-level information becomes increasingly incomplete.

4.3 Ablation Study

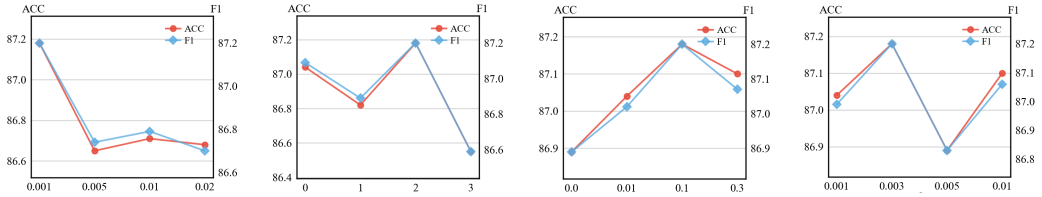
To evaluate the key components of HiCTE, we conduct ablation experiments under both the fixed missing-modality and random missing-rate protocols. For each variant, we report the average performance over the six fixed missing-modality conditions and eight random missing-rate settings, respectively. The results are summarized in Table 2. Specifically, we consider: (1) *w/o Latent Reliability Modeling*: the probabilistic latent representation is replaced with its deterministic mean, removing latent variance and the corresponding reliability information; (2) *w/o Local Refinement ($K = 0$)*: the training-time local iterative updates are disabled, and the initial transferability $\alpha_{i,0}^{m \to t}$ is directly used for representation construction; (3) *Stop-Gradient Refinement*: the same K-step forward refinement is retained, but $\alpha_{i,0}^{m \to t}$ is detached before refinement, blocking gradients through the refinement trajectory from reaching the global estimator; (4) *Target-Agnostic Shared Selector*: the

Table 2: Average ablation results under fixed missing-modality and random missing-rate protocols.

Variant	CMU-MOSI		CMU-MOSEI		IEMOCAP	
	Fixed Missing	Random Missing	Fixed Missing	Random Missing	Fixed Missing	Random Missing
HiCTE	**75.82/75.59**	**87.53/87.50**	**80.00/79.69**	**86.76/86.74**	**73.73/73.49**	**80.78/81.22**
w/o Latent Reliability Modeling	75.05/74.83	86.95/86.94	79.42/79.16	86.43/86.40	72.98/72.62	79.39/79.44
w/o Local Refinement ($K = 0$)	74.84/74.62	87.20/87.17	79.58/79.37	86.58/86.56	73.32/73.08	80.35/80.59
Stop-Gradient Refinement	75.25/74.89	87.22/87.21	79.64/79.35	86.51/86.45	73.40/73.05	80.59/80.81
Target-Agnostic Shared Selector	74.89/74.53	87.19/87.16	79.56/79.34	86.56/86.51	73.20/72.64	80.15/80.47

two target-conditioned estimators of each source modality are replaced with a single shared selector, which is jointly refined for both target branches, such that the same transferability estimate is used for both target experts.

Table 2 shows that all ablated variants underperform the full HiCTE model across the reported settings. Removing latent reliability modeling causes consistent performance degradation, particularly on IEMOCAP under feature-level missingness, indicating the importance of variance-derived reliability in estimating transferability under incomplete inputs. Disabling local refinement ($K = 0$) also reduces performance, while retaining the K-step refinement but stopping gradients from reaching the global estimator remains inferior to the full model. These results show that sample-specific local refinement is beneficial, and that allowing gradients to propagate through the refinement trajectory further improves the optimization of the global transferability estimator. Finally, replacing the two target-conditioned estimators with a shared selector consistently degrades performance, supporting the need to model transferability separately for different target experts. More detailed ablation analyses are provided in Appendix C.2.



(a) λ_{KL} (b) K (c) λ_{init} (d) β

Figure 3: Hyperparameter sensitivity on CMU-MOSI at random missing rate $r = 0.5$.

4.4 Hyperparameter Analysis

Figure 3 shows the sensitivity of HiCTE to four key hyperparameters on CMU-MOSI under the random missing-rate protocol with $r = 0.5$. Here, λ_{KL} controls latent-space regularization, K denotes the number of local refinement steps, λ_{init} weights the objective on the initial transferability estimate, and β controls the variance regularization in transferability estimation. The results show that HiCTE remains relatively stable across the tested hyperparameter ranges, while appropriate settings provide consistent gains. Specifically, $\lambda_{\mathrm{KL}} = 0.001$, $K = 2$, $\lambda_{\mathrm{init}} = 0.1$, and $\beta = 0.003$ achieve the best overall performance in their respective comparisons and are therefore selected as the default hyperparameter settings.

5 Conclusion

We propose HiCTE for incomplete multimodal representation learning, which models target-conditioned transferability at the latent-dimension level and further adapts it through sample-specific local refinement. Based on these transferability estimates, HiCTE constructs target-specific cross-expert and residual representations and integrates them through adaptive weighting. This enables effective representation learning from observed modalities without reconstructing missing ones. Experiments on CMU-MOSI, CMU-MOSEI, and IEMOCAP demonstrate consistent improvements under both fixed missing-modality and random missing-rate settings. In future work, we will investigate transferability modeling under modality conflicts and distribution shifts, and further evaluate the generalization of HiCTE to broader multimodal tasks.

AI USE STATEMENT

Generative AI tools were used for Chinese-to-English translation, language refinement, and code modification. All AI-assisted outputs were manually reviewed by the authors. Language revisions were checked for consistency with the intended meaning, and code modifications were reviewed and tested by the authors. The research design, methodological decisions, experimental validation, and scientific conclusions were determined by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We facilitate the reproducibility of HiCTE by providing detailed descriptions of the datasets, incomplete-input protocols, model optimization, and evaluation settings.

(1) Datasets and Features. The main experiments are conducted on three public multimodal benchmarks: CMU-MOSI, CMU-MOSEI, and IEMOCAP (four-class). We further evaluate HiCTE on UR-FUNNY for additional cross-task validation on multimodal humor detection. Detailed dataset statistics and evaluation settings are provided in Appendix A.1, while feature extraction and preprocessing details are provided in Appendix A.3. The additional UR-FUNNY experiments are reported in Appendix B.

(2) Incomplete-Input Protocols. For the main experiments, we evaluate HiCTE under fixed missing-modality and random missing-rate protocols. Under the fixed protocol, the six incomplete modality combinations are cycled in a randomly shuffled order during training, while evaluation is conducted separately for each fixed modality combination. Under the random protocol, intra-modality features are randomly masked at predefined missing rates, with the same missing rate used during training and testing. The additional UR-FUNNY experiment follows the fixed missing-modality protocol. Complete protocol definitions are provided in Section 4.1.

(3) Model and Optimization. The model architecture, latent reliability modeling, global transferability initialization, local iterative refinement, and adaptive path weighting are described in Section 3. Algorithm 1 provides the complete training procedure, including global initialization, differentiably unrolled local refinement, representation construction, and global parameter updates. Detailed hyperparameter and optimization settings are provided in Appendix A.2.

(4) Evaluation. We report ACC and F1 for CMU-MOSI and CMU-MOSEI, and ACC and UA for IEMOCAP. Class-wise F1 scores are additionally reported for category-level analysis on IEMOCAP. For UR-FUNNY, we report ACC and weighted F1 under each fixed missing-modality condition. For the main experiments, all runs use three fixed random seeds, $\{66, 67, 68\}$, and the reported results are averaged over the corresponding runs.

(5) Additional Analyses. The appendix provides an additional cross-task evaluation on UR-FUNNY, category-level results on IEMOCAP, detailed ablation studies, validation of latent reliability modeling and transferability modeling, a case study, computational-cost analysis, and the complete training algorithm of HiCTE. Together with the detailed experimental configurations, these materials support the reproducibility of the reported experiments.

REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, 2024.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2046–2056. Association for Computational Linguistics, 2019.

Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. PaSE: Prototype-aligned calibration and Shapley-based equilibrium for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30960–30968, 2026.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Mimansa Jaiswal and Emily Mower Provost. Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7985–7993, 2020.

Feng Jiang, Maolaaisha Aminanmu, Hehuan Ma, Yuzhi Guo, Tommaso Mansi, Junzhou Huang, Mangal Prakash, and Rui Liao. GRAM-DTI: Adaptive multimodal representation learning for drug–target interaction prediction. In *International Conference on Learning Representations*, pp. 107480–107509, 2026.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10074–10082, 2024a.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28515–28536, 2024b.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024c.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432, 2023.

Wei Luo, Mengying Xu, and Hanjiang Lai. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *MultiMedia Modeling*, pp. 411–422, 2023.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, pp. 152802–152824, 2026.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15880–15889, 2026.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22034, 2023a.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17117–17128, 2023b.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20039–20049, 2023.

Zhuojia Wu, Qi Zhang, Duoqian Miao, Kun Yi, Wei Fan, and Liang Hu. HyDiscGAN: A hybrid distributed cGAN for audio-visual privacy preservation in multimodal sentiment analysis. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6550–6558, 2024.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *International Conference on Learning Representations*, volume 2024, pp. 30072–30100, 2024.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 438–446, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618. Association for Computational Linguistics, 2021a.

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

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. In *Advances in Neural Information Processing Systems*, volume 38, pp. 142100–142129, 2025.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2281–2289, 2026.

APPENDIX

In this appendix, we provide detailed experimental settings and additional experimental results. The appendix is organized as follows:

A Experimental Details 13
A.1 Datasets and Evaluation Metrics 13
A.2 Implementation Details and Hyperparameter Settings 14
A.3 Feature Extraction 14
B Additional Evaluation on UR-FUNNY 15
C Additional Experimental Analysis 15
C.1 Category-Level Results on IEMOCAP 15
C.2 Additional Ablation Studies 16
C.3 Validation of Latent Reliability Modeling 17
C.4 Additional Validation of Transferability Modeling 17
C.5 Case Study 18
C.6 Computational Cost 19
C.7 Training Algorithm of HiCTE 19

A EXPERIMENTAL DETAILS

A.1 DATASETS AND EVALUATION METRICS

Datasets. Our main experiments are conducted on three widely used multimodal benchmarks: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). All three datasets contain textual, acoustic, and visual modalities. We further use UR-FUNNY (Hasan et al., 2019) for additional cross-task evaluation.

CMU-MOSI is a multimodal sentiment analysis dataset containing 2,199 opinion video segments. We follow the standard split with 1,284 samples for training, 229 for validation, and 686 for testing. Each segment is annotated with a sentiment score ranging from -3 to $+3$.

CMU-MOSEI is a large-scale multimodal sentiment analysis dataset. Following the standard split, we use 16,326 samples for training, 1,871 for validation, and 4,659 for testing. Each sample is annotated with a sentiment score in the range of $[-3, 3]$.

IEMOCAP is a multimodal conversational emotion recognition dataset collected from dyadic dialogue sessions. Following commonly adopted settings, we consider four emotion categories: happy, sad, neutral, and angry. We perform five-fold cross-validation, where one complete session is held out for testing in each fold and the remaining four sessions are used for training.

UR-FUNNY is a multimodal benchmark for binary humor detection, containing textual, acoustic, and visual information from conversational contexts and punchlines. Each sample is labeled as humorous or non-humorous. We use UR-FUNNY V2 with the official speaker-independent split. For each modality, we concatenate the mean-pooled context and punchline representations, resulting in acoustic, textual, and visual feature dimensions of 162, 600, and 742, respectively. The corresponding fixed missing-modality experiments are reported in Appendix B.

Evaluation Metrics. For CMU-MOSI and CMU-MOSEI, following previous multimodal sentiment analysis studies (Xu et al., 2024; Luo et al., 2023), we evaluate binary sentiment classification on non-zero sentiment samples, where positive and negative samples are distinguished according to the sign of the sentiment score. We report accuracy (ACC) and weighted F1 score. For IEMOCAP,

Table 3: Hyperparameter settings of HiCTE under the fixed missing-modality and random missing-rate protocols.

Hyperparameter	CMU-MOSI	CMU-MOSEI	IEMOCAP
Random Seeds	66, 67, 68	66, 67, 68	66, 67, 68
Batch Size	32	32	16
Learning Rate	0.0001	0.0001	0.0001
Hidden Dimension	256	256	256
Transformer Depth	4	4	4
Attention Heads	2	2	2
Dropout Rate	0.5	0.5	0.5
Weight Decay	1×10^{-5}	1×10^{-5}	1×10^{-5}
Training Epochs	400	400	400
KL Weight λ_{KL}	0.001	0.001	0.001
Initialization Weight λ_{init}	0.1	0.1	0.1
Variance Regularization Weight β	0.003	0.003	0.003
Transferability Constraint Weight λ_b	0.1	0.1	0.1
Local Refinement Step Size η_{inner}	0.003	0.003	0.003
Local Refinement Steps K (Fixed)	1	2	3
Local Refinement Steps K (Random)	2	2	3
Warm-up Epochs E_{warm}	30	30	30
Modality-Level Supervision Weight $\lambda_{\mathrm{uni}}^{(e)}$	$2 \rightarrow 1$	$2 \rightarrow 1$	$2 \rightarrow 1$

we report accuracy (ACC) and unweighted accuracy (UA) as the main evaluation metrics. In addition, class-wise F1 scores are reported for the four emotion categories to further examine category-level performance under different incomplete-input conditions (Appendix C.1). For UR-FUNNY, we report ACC and weighted F1 score for binary humor detection.

A.2 Implementation Details and Hyperparameter Settings

HiCTE is implemented in PyTorch and optimized using Adam on an NVIDIA A800 GPU with 80 GB of memory. All method-specific hyperparameters are kept the same across the two incomplete-input protocols, except for the number of local refinement steps K. The same model architecture and optimization procedure are used across all datasets, with dataset-specific settings limited to those listed in Table 3. No additional data augmentation is introduced beyond the incomplete-input construction defined by the corresponding experimental protocol. All experiments are conducted with random seeds 66, 67, and 68, and the results are averaged over the three seeds. The complete hyperparameter settings are summarized in Table 3.

A.3 Feature Extraction

For the three main benchmarks, we use textual, acoustic, and visual features extracted by pretrained modality encoders. These encoders transform the raw multimodal inputs into modality-specific feature representations, which are subsequently used as inputs to HiCTE. Since this work focuses on representation learning under incomplete multimodal inputs, the pretrained encoders are used only for initial feature extraction, while the proposed modules operate on the extracted representations.

For the textual modality, we employ pretrained DeBERTa-large (He et al., 2021) to encode the textual content and obtain 1024-dimensional features. For the visual modality, MTCNN (Zhang et al., 2016) is first used for face detection and alignment, followed by pretrained MA-Net (Zhao et al., 2021b) to extract 1024-dimensional visual features. For the acoustic modality, pretrained wav2vec-large (Schneider et al., 2019) is used to extract 512-dimensional acoustic features from speech signals.

After feature extraction, each sample is represented by three modality-specific feature sequences, X_i^A, X_i^T, and X_i^V, corresponding to the acoustic, textual, and visual modalities, respectively. Following the formulation in Section 3, modality-level and feature-level incompleteness are represented

Table 4: Fixed missing-modality results on UR-FUNNY. ACC and weighted F1 scores are reported as percentages (%).

Method	$\{A\}$	$\{T\}$	$\{V\}$	$\{A,V\}$	$\{A,T\}$	$\{T,V\}$	*Average*
Accuracy performance on UR-FUNNY							
MoMKE	58.45	63.78	60.26	61.17	66.00	64.19	62.31
HME	52.52	64.69	58.65	58.75	64.70	**66.30**	60.94
TMDC	59.05	62.98	60.06	61.77	64.39	64.49	62.12
HiCTE	**59.26**	**66.00**	**60.56**	**62.27**	**66.60**	**66.30**	**63.50**
F1 performance on UR-FUNNY							
MoMKE	58.33	63.78	60.21	61.16	66.00	64.14	62.27
HME	45.63	64.68	58.47	58.13	64.69	66.11	59.62
TMDC	59.03	62.90	60.00	61.75	64.38	64.48	62.09
HiCTE	**59.17**	**65.94**	**60.54**	**62.22**	**66.57**	**66.30**	**63.46**

by the modality availability indicator δ_i^m and feature-level observation mask R_i^m, respectively. The resulting incomplete input $\widetilde{X}_i^m$ is then fed into the subsequent latent reliability modeling, transferability estimation, and adaptive weighting modules.

B Additional Evaluation on UR-FUNNY

We further evaluate HiCTE under the fixed missing-modality protocol on UR-FUNNY to examine its generalization beyond sentiment and emotion recognition. We consider the same six incomplete modality combinations as in the main experiments: $\{A\}$, $\{T\}$, $\{V\}$, $\{A,V\}$, $\{A,T\}$, and $\{T,V\}$. All methods use the same feature representations, data split, and missing conditions. We compare HiCTE with representative incomplete multimodal learning methods, including MoMKE (Xu et al., 2024), HME (Zhuang et al., 2025), and TMDC (Zhuang et al., 2026). We report ACC and F1 scores under each missing-modality condition.

Table 4 reports the results on UR-FUNNY, where the task is extended from sentiment and emotion recognition to multimodal humor detection. HiCTE achieves the best average performance, obtaining 63.50% ACC and 63.46% weighted F1, outperforming the strongest baseline by 1.19 percentage points on both metrics. These results suggest that transferability estimation remains effective beyond sentiment and emotion recognition, providing additional evidence for the generalization of HiCTE to incomplete multimodal learning on a different downstream task.

C Additional Experimental Analysis

C.1 Category-Level Results on IEMOCAP

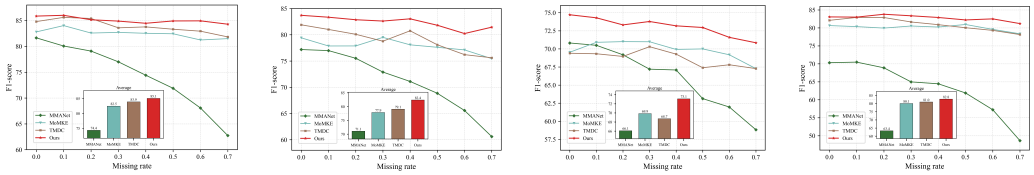
Beyond the fixed missing-modality and random missing-rate experiments reported in the main text on the three benchmark datasets, we further analyze the class-wise performance on IEMOCAP. As shown in Table 5, HiCTE achieves the highest average F1 score for all four emotion categories, outperforming the strongest baselines by 1.53, 1.95, 3.25, and 1.42 percentage points for Happy, Sad, Neutral, and Angry, respectively. HiCTE also achieves the best performance in 22 out of the 24 category-condition combinations. These results show that the advantage of HiCTE is consistently maintained across different emotion categories and fixed missing-modality conditions.

Fig. 4 further reports the category-level performance under increasing random missing rates. Averaged over the eight missing rates, HiCTE obtains F1 scores of 85.1, 82.4, 73.1, and 82.8 for Happy, Sad, Neutral, and Angry, respectively, achieving the best average performance for all four categories. Despite local fluctuations, HiCTE maintains relatively stable performance as feature-level informa-

Table 5: Class-wise F1 scores on IEMOCAP under six fixed available-modality conditions.

Method	Category	$\{A\}$	$\{T\}$	$\{V\}$	$\{A,V\}$	$\{A,T\}$	$\{T,V\}$	*Average*
MMANet	Happy	39.23	65.68	70.96	72.59	72.80	76.09	66.23
	Sad	62.31	61.35	23.26	66.29	72.35	68.41	59.00
	Neutral	49.19	62.17	50.38	62.04	64.84	67.17	59.30
	Angry	54.28	59.78	26.91	56.84	67.63	65.55	55.17
MoMKE	Happy	72.17	83.46	71.90	74.75	84.21	81.08	77.93
	Sad	73.26	79.78	31.62	66.88	83.03	72.35	67.82
	Neutral	60.70	66.34	50.06	60.42	69.19	68.16	62.48
	Angry	70.61	**77.91**	47.51	68.07	81.86	77.44	70.57
TMDC	Happy	71.93	82.49	71.70	74.87	83.83	84.39	78.20
	Sad	70.57	78.22	36.66	70.11	80.96	77.83	69.06
	Neutral	57.88	64.36	51.14	62.51	67.31	66.18	61.56
	Angry	**77.14**	77.69	51.33	68.76	82.89	78.62	72.74
HiCTE	Happy	**75.07**	**83.72**	**73.10**	**76.12**	**85.75**	**84.60**	**79.73**
	Sad	**73.45**	**80.20**	**37.65**	**71.83**	**83.85**	**79.05**	**71.01**
	Neutral	**62.55**	**67.96**	**54.01**	**66.99**	**72.05**	**70.84**	**65.73**
	Angry	76.76	77.88	**52.25**	**73.98**	**84.40**	**79.67**	**74.16**

Note: Results are reported as class-wise F1 scores. $\{A\}$, $\{T\}$, and $\{V\}$ denote acoustic-only, textual-only, and visual-only inputs, respectively, while the remaining combinations are defined analogously. *Average* denotes the mean over the six fixed conditions. The best result under each condition and category is boldfaced.



(a) Happy (b) Sad (c) Neutral (d) Angry

Figure 4: Class-wise F1 scores on IEMOCAP under different random missing rates. The inset bars report the average F1 score over all evaluated missing rates.

tion becomes increasingly incomplete. This further demonstrates that the improvements of HiCTE are not concentrated on a particular emotion category.

C.2 Additional Ablation Studies

In addition to the main ablations in Section 4.3, we conduct several additional experiments to further examine the individual design choices of HiCTE.

(1) *Whole Transfer.* Target-specific transferability estimation is removed, and the complete source latent representation is transferred to both target experts without selective transfer.

(2) *w/o Reliability Guidance.* The probabilistic latent representation is retained, while the variance-derived reliability information is removed from transferability estimation. The reliability input is replaced with a neutral constant, and the corresponding variance-based transferability guidance is disabled.

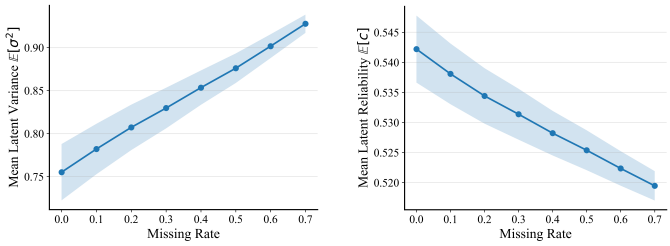
(3) *w/o Adaptive Weighting.* The adaptive path weighting module is removed, and the self, two target-specific cross-expert, and residual paths are assigned equal weights of 0.25.

As shown in Table 6, Whole Transfer consistently underperforms HiCTE, indicating that transferring the complete source latent representation is less effective than selectively allocating latent information to different target experts. Removing reliability guidance also degrades performance, supporting the use of variance-derived reliability in transferability estimation. In addition, replacing adaptive path weighting with uniform averaging leads to performance drops across the three datasets,

Table 6: Average ablation results under fixed missing-modality and random missing-rate protocols.

Variant	CMU-MOSI		CMU-MOSEI		IEMOCAP	
	Fixed Missing	Random Missing	Fixed Missing	Random Missing	Fixed Missing	Random Missing
HiCTE	**75.82/75.59**	**87.53/87.50**	**80.00/79.69**	**86.76/86.74**	**73.73/73.49**	**80.78/81.22**
Whole Transfer	74.50/74.32	87.03/87.02	79.49/79.24	86.59/86.54	72.63/72.00	78.53/78.51
w/o Reliability Guidance	75.27/74.83	87.01/86.99	79.84/79.59	86.60/86.58	72.95/72.58	79.97/79.93
w/o Adaptive Weighting	75.11/74.74	86.95/86.93	78.90/78.52	85.70/85.66	72.71/72.20	80.02/80.19

showing that the contributions of the self, cross-expert, and residual paths should be adjusted rather than treated equally.



(a) Mean latent variance (b) Mean latent reliability

Figure 5: Response of latent statistics to varying test missing rates on CMU-MOSI. For each random seed, HiCTE is trained with a fixed missing rate $r_{\text{train}} = 0.3$, after which the selected checkpoint is fixed and evaluated under $r_{\text{test}} \in \{0.0, 0.1, \ldots, 0.7\}$. Panels (a) and (b) show the mean latent variance and mean latent reliability, averaged over the audio, text, and visual modalities. Results are averaged over three random seeds, and shaded regions indicate standard deviations across seeds.

C.3 Validation of Latent Reliability Modeling

To examine whether the learned latent statistics respond to input degradation, we conduct an additional analysis on CMU-MOSI with $r_{\text{train}} = 0.3$. After training, the selected checkpoints are fixed and evaluated under $r_{\text{test}} \in \{0.0, 0.1, \ldots, 0.7\}$ without further parameter updates.

We report the mean latent variance $\mathbb{E}[\sigma^2]$ and the corresponding mean reliability $\mathbb{E}[c]$, averaged over the audio, text, and visual modalities and three random seeds. As shown in Figure 5, $\mathbb{E}[\sigma^2]$ increases from 0.7551 to 0.9277 as r_{test} increases from 0 to 0.7, while $\mathbb{E}[c]$ decreases from 0.5422 to 0.5195. These consistent trends indicate that the learned latent statistics respond systematically to increasing input missingness, with higher variance corresponding to lower estimated reliability.

C.4 Additional Validation of Transferability Modeling

We further conduct two experiments to examine the mechanisms underlying transferability modeling in HiCTE. The first evaluates whether the learned transferability estimates are associated with different target experts, while the second examines the effect of the transferability constraint on transferability modeling. The results are summarized in Table 7.

Target-selector correspondence. To examine whether HiCTE captures target-dependent transferability, we construct a HiCTE *Swapped-Selector* variant. For each source modality m, the two learned transferability selectors associated with its target experts p and q are exchanged, while the source representation, target-specific experts, adapters, and all learned model parameters remain unchanged. This intervention breaks the correspondence between each transferability selector and its intended target expert.

Swapping the selectors consistently reduces performance across all six fixed missing-modality conditions on the three datasets. These consistent degradations indicate that the learned selectors are not interchangeable across target experts and support the target-conditioned transferability modeling adopted in HiCTE.

Table 7: Additional mechanism validation under the fixed missing-modality protocol.

Method	Observed Modalities						
	$\{A\}$	$\{T\}$	$\{V\}$	$\{A,V\}$	$\{A,T\}$	$\{T,V\}$	*Average*
Results on CMU-MOSI							
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
HiCTE	**62.50/61.38**	**87.25/87.17**	**63.56/63.60**	**66.01/65.82**	**87.30/87.31**	**88.29/88.25**	**75.82/75.59**
HiCTE (Swapped-Selector)	61.47/59.42	86.43/86.33	61.89/62.10	65.16/64.87	86.78/86.59	87.65/87.56	74.90/74.48
w/o Transferability Constraint	59.65/57.91	87.14/87.10	62.80/62.82	65.40/65.07	87.20/87.20	88.14/88.09	75.06/74.70
Results on CMU-MOSEI							
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
HiCTE	**73.01/72.80**	**87.31/87.29**	**71.04/70.23**	**73.76/73.08**	**87.36/87.30**	**87.52/87.46**	**80.00/79.69**
HiCTE (Swapped-Selector)	72.51/72.30	86.90/86.86	69.46/68.79	72.98/72.09	87.26/87.23	87.07/87.03	79.36/79.05
w/o Transferability Constraint	72.93 / 72.43	87.19/87.19	69.61/69.15	73.62/72.84	87.34/87.29	87.48/87.43	79.70/79.39
Results on IEMOCAP							
	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA	ACC/UA
HiCTE	**71.97/73.02**	**77.75/78.35**	**59.87/55.81**	**72.51/72.45**	**81.40/82.28**	**78.90/79.03**	**73.73/73.49**
HiCTE (Swapped-Selector)	71.48/72.61	77.00/77.46	58.80/55.45	71.62/71.97	80.82/81.61	77.94/78.36	72.94/72.91
w/o Transferability Constraint	71.79/72.95	77.70/78.26	59.60/55.35	72.14/71.54	81.31/82.26	78.74/78.60	73.55/73.16

Note: The HiCTE *Swapped-Selector* variant exchanges the two target-conditioned selectors of each source modality. *w/o Transferability Constraint* sets $\lambda_b = 0$, removing the scale constraint used to stabilize transferability estimation. *Average* denotes the mean over the six fixed missing-modality conditions.

Transferability constraint. We further examine the transferability constraint in Eq. 9, $\lambda_b \left(\frac{1}{d_z} \|\alpha\|_1 - 0.5 \right)^2$. For the *w/o Transferability Constraint* variant, we set $\lambda_b = 0$ during training while keeping all other settings unchanged. The constraint regularizes the average transferability score around the reference level and discourages excessively large or small transferability estimates. Since the same transferability criterion is involved in global initialization and local refinement, the constraint regularizes the transferability estimates used in both stages.

Removing the constraint consistently decreases the average performance on all three datasets. The most pronounced degradation is observed under text-absent conditions, specifically $\{A\}$ on CMU-MOSI, $\{V\}$ on CMU-MOSEI, and $\{A, V\}$ on IEMOCAP. This pattern suggests that the constraint is particularly useful when prediction relies mainly on acoustic and/or visual evidence. Together with the variance regularization, the scale constraint provides additional regularization for cross-expert transfer under incomplete inputs.

Overall, the two experiments provide complementary evidence for the transferability design of HiCTE. The selector-swapping intervention shows that different target experts require distinct transferability estimates, whereas the constraint ablation demonstrates the effectiveness of the transferability constraint in supporting cross-expert transfer under incomplete multimodal inputs.

C.5 Case Study

To qualitatively examine how HiCTE constructs representations under incomplete inputs, we select the dialogue `Ses05F_impro08` from IEMOCAP under the text-only condition $\{T\}$. As shown in Fig. 6, we compare HiCTE with *Whole Transfer*, which directly transfers the complete textual latent representation to both target experts without selective transferability estimation. We also visualize the adaptive weights assigned by HiCTE to the text-self, $T \to A$, $T \to V$, and residual representation paths.

The dialogue contains several emotional transitions. Early utterances such as "This is ridiculous...", "I need a human being.", and "I know I paid last month..." express dissatisfaction, whereas later utterances such as "That would be wonderful..." and "you've made it much better..." convey more positive emotions. For the highlighted error cases, HiCTE produces predictions that match the ground-truth labels while *Whole Transfer* does not. This comparison suggests that target-conditioned transfer is more effective than directly transferring the complete textual representation.

The adaptive weights further show how HiCTE combines different representation paths. The text-self path receives the largest weight for most utterances, while the $T \to A$, $T \to V$, and residual paths receive varying non-zero weights across the dialogue. This variation shows that HiCTE adjusts the

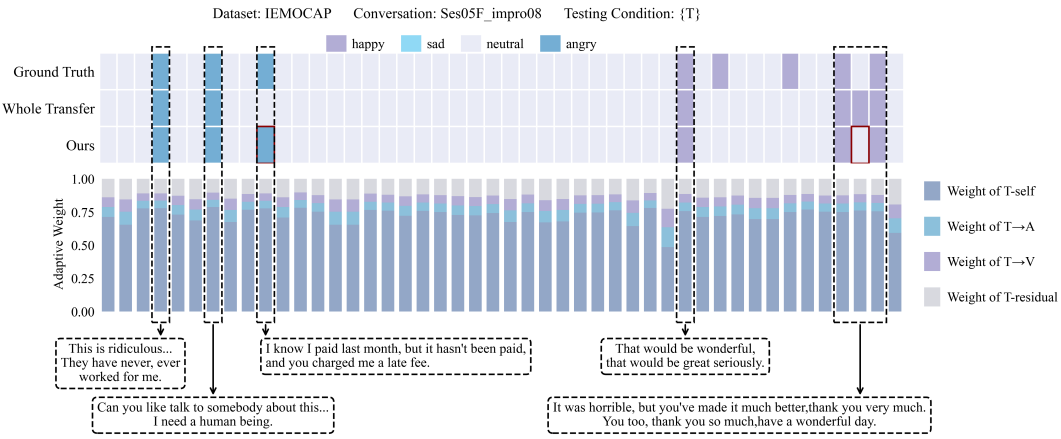


Figure 6: Case study on IEMOCAP under the text-only condition $\{T\}$. *Whole Transfer* transfers the complete textual latent representation to both target experts without selective transferability estimation. The upper panel compares utterance-level predictions, while the lower panel shows the adaptive weights assigned by HiCTE to different representation paths. Dashed boxes highlight selected utterances, and red boxes indicate utterances where HiCTE corrects prediction errors made by *Whole Transfer*.

contributions of different representation paths according to individual utterances rather than combining them uniformly. Overall, this case qualitatively illustrates how target-conditioned transfer and adaptive weighting jointly construct representations from the observed textual information.

C.6 Computational Cost

To evaluate the computational overhead of HiCTE, we compare its number of trainable parameters and training time with MoMKE, HME, and TMDC on CMU-MOSI. Training time is measured under the random missing-rate protocol with $r = 0.5$. All methods are trained for 400 epochs using a single NVIDIA A800 GPU with 80 GB of memory. The parameter counts include only trainable parameters of the downstream models and exclude the pretrained feature extractors.

Table 8: Computational cost comparison on CMU-MOSI under the random missing-rate protocol.

Method	# Parameters (M) ↓	Training Time (s) ↓
MoMKE	5.40	231
HME	120.41	2959
TMDC	5.78	319
HiCTE ($K = 0$)	8.16	354
HiCTE ($K = 2$)	8.16	516

As shown in Table 8, HiCTE contains 8.16M trainable parameters, independent of the number of local refinement steps K. Without local refinement ($K = 0$), HiCTE requires 354 s for training, whereas the default setting ($K = 2$) requires 516 s, corresponding to a 45.8% increase in training time. Compared with MoMKE and TMDC, HiCTE has more trainable parameters and higher training cost due to the additional transferability estimation and local refinement. Nevertheless, it remains substantially lighter than HME, using only 6.8% of its trainable parameters and 17.4% of its training time.

C.7 Training Algorithm of HiCTE

Algorithm 1 summarizes the complete training procedure of HiCTE, including global transferability initialization, local iterative refinement, adaptive representation construction, and end-to-end

parameter optimization. The K local refinement steps are differentiably unrolled during training. During training, the refined transferability is used for representation construction, and the prediction loss is back-propagated through the K refinement steps to update the global transferability estimator. Therefore, the local refinement not only adapts transferability for individual training samples, but also improves the global estimator used at inference. The corresponding gradient contribution through the refinement path can be expressed as

$$\left.\frac{\partial \mathcal{L}}{\partial \Theta}\right|_{\text{ref}} = \frac{\partial \mathcal{L}}{\partial \alpha_{i,K}^{m\to t}} \frac{\partial \alpha_{i,K}^{m\to t}}{\partial \alpha_{i,0}^{m\to t}} \frac{\partial \alpha_{i,0}^{m\to t}}{\partial \Theta}, \tag{17}$$

where Θ denotes the parameters of the global transferability estimator.

Algorithm 1 Training Algorithm of HiCTE

Require: Multimodal training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; modality set $\mathcal{M} = \{A, T, V\}$; number of epochs E; local refinement steps K; inner learning rate η_{inner}; β, λ_b, λ_{KL}, λ_{init}, λ_{warm}, and E_{warm}.
Ensure: Learned HiCTE parameters Ω.

for $e = 0, \ldots, E-1$ **do**
  Set $\lambda_{\text{uni}}^{(e)} \leftarrow \lambda_{\text{warm}}$ if $e < E_{\text{warm}}$, otherwise 1.
  for each mini-batch $\mathcal{D}_b \subset \mathcal{D}$ **do**
    Construct incomplete inputs $\widetilde{X}_i^m$ and determine $\mathcal{O}_i$ for each $(x_i, y_i) \in \mathcal{D}_b$.
    for each source modality m available in $\mathcal{D}_b$ **do**
  The following sample-indexed operations are performed jointly for all $(x_i, y_i) \in \mathcal{D}_b$ with $m \in \mathcal{O}_i$.
      Compute $\bar{X}_i^m = \text{LayerNorm}(\text{Proj}^m(\widetilde{X}_i^m))$ and estimate μ_i^m and $(\sigma_i^m)^2$.
      Compute the variance-derived reliability c_i^m and form $I_i^m = [\mu_i^m; c_i^m]$.
      Sample $z_i^m = \mu_i^m + \sigma_i^m \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$.
      Obtain the self representation $h_{i,\text{self}}^m$ and unimodal prediction $\hat{y}_i^m$.
      for each target expert $t \in \mathcal{M} \setminus \{m\}$ **do**
        Global initialization: $\alpha_{i,0}^{m\to t} \leftarrow \phi_\Theta^{m\to t}(I_i^m)$.
        Evaluate $\mathcal{J}_i^{m\to t}(\alpha_{i,0}^{m\to t})$ for the global initialization objective.
        if $K > 0$ **then**
          $u_{i,0}^{m\to t} \leftarrow \text{logit}(\alpha_{i,0}^{m\to t})$.
          for $k = 0, \ldots, K-1$ **do**
            $\alpha_{i,k}^{m\to t} \leftarrow \text{sigmoid}(u_{i,k}^{m\to t})$.
            Compute $\bar{\mathcal{J}}_k^{m\to t}$ by averaging the position-wise objectives at $\alpha_{i,k}^{m\to t}$ over valid positions in $\mathcal{D}_b$.
            Simultaneously update $u_{i,k+1}^{m\to t} \leftarrow u_{i,k}^{m\to t} - \eta_{\text{inner}} \nabla_{u_{i,k}^{m\to t}} \bar{\mathcal{J}}_k^{m\to t}$.
          end for
          $\alpha_i^{m\to t} \leftarrow \text{sigmoid}(u_{i,K}^{m\to t})$.
        else
          $\alpha_i^{m\to t} \leftarrow \alpha_{i,0}^{m\to t}$.
        end if
        $z_i^{m\to t} \leftarrow \alpha_i^{m\to t} \odot z_i^m$.
      end for
      Let $\{p, q\} = \mathcal{M} \setminus \{m\}$ and compute $z_{i,\text{res}}^m \leftarrow (1 - \alpha_i^{m\to p}) \odot (1 - \alpha_i^{m\to q}) \odot z_i^m$.
      Construct $h_{i,m\to p}^m$, $h_{i,m\to q}^m$, and $h_{i,\text{res}}^m$.
      Compute adaptive path weights $\omega_{i,b}^m$ for $b \in \mathcal{B}^m$.
      $F_i^m \leftarrow \sum_{b\in\mathcal{B}^m} \omega_{i,b}^m h_{i,b}^m$.
    end for
    for each $(x_i, y_i) \in \mathcal{D}_b$ **do**
      Set $F_i^m = 0$ for each $m \notin \mathcal{O}_i$.
      $F_i \leftarrow \text{Cat}(F_i^A, F_i^T, F_i^V)$, $\hat{y}_i \leftarrow g(F_i)$.
    end for
    Compute $\mathcal{L}_{\text{task}}$, $\mathcal{L}_{\text{uni}}$, $\mathcal{L}_{\text{KL}}$, and $\mathcal{L}_{\text{init}}$.
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \lambda_{\text{uni}}^{(e)} \mathcal{L}_{\text{uni}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{init}} \mathcal{L}_{\text{init}}$.
    Back-propagate $\mathcal{L}$ through the K unrolled local refinement steps and update Ω using Adam.
  end for
end for
return Ω