A Proofs

proof of Lemma 3. Note that the expected form of Eq. (5) is the right hand side of Eq. (3), since $\{x_i^{[M]}\}_i$ are i.i.d from joint distribution $U_{X^{[M]}}$. Therefore, from Lemma 1, one can immediately get the conclusion. □

Lemma 4. *Given a joint distribution $p(x^1, ..., x^M, y)$, where y is a discrete random variable, we can always find M independent random variables $z^1, ..., z^M$ such that $z^i \perp\!\!\!\perp y$ and $x^m = f_m(y, z^m)$, for $m \in [M]$.*

Proof. This proof is a generalization of the Proposition in [1]. For $z_1, ..., z_M \overset{i.i.d}{\sim} Uniform(0, 1)$, then from [1] we have $x^m|y = F^{-1}_{y,m}(z_m)$ where $F_{y,m}(t) = \mathrm{P}(x^m \leq t|y)$ where $\mathrm{P}(x^m \leq t|y)$ is the cumulative distribution function of $p(x^m|y)$. □

Lemma 5. *Given assumption 1, then the Marginal-joint ratio (definition 1) $R(x^1, ..., x^M)$ has*

$$R(x^1, ..., x^M) = \sum_{y \in \mathcal{C}} \frac{\Pi_m p(y|x^m)}{p(y)^{M-1}}$$

Further, the optimal g to make the equality holds in Lemma 1 has

$$g(x^1, ..., x^M) = 1 + \log \sum_{y \in \mathcal{C}} \frac{\Pi_m p(y|x^m)}{p(y)^{M-1}}.$$

Proof.

$$\begin{aligned} R(x^1, ..., x^M) &= \frac{p(x^1, ..., x^M)}{\Pi_m p(x^m)} \\ &= \frac{\sum_y p(x^1, ..., x^M, y)}{\Pi_m p(x^m)} \\ &= \frac{\sum_y p(x^1, ..., x^M|y)p(y)}{\Pi_m p(x^m)} \\ &= \frac{\sum_y \pi_m p(x^1|y)p(y)}{\Pi_m p(x^m)} \\ &= \sum_y \frac{\pi_m p(y|x^m)}{p(y)^{M-1}} \end{aligned}$$

One can immediately gets the conclusion for g from Lemma 1. □

Theorem 4 (Main theorem). *Define the expected total correlation gain $TCg(h^1, ..., h^M, p)$ as*

$$TCg(h^1, ..., h^M, p) = \mathbb{E}_{x_i^{[M]} \overset{i.i.d.}{\longleftarrow} U_{X^{[M]}}} \left(\mathcal{L}_{TC}(x_i^{[M]}; h^{[M]}, \mathbf{p}) \right)$$

Given the conditional independence assumption 1 and well-defined prior assumption 2, we have that

Ground-truth → Maximizer $(h_*^{[M]}, \mathbf{p}^*)$ *is a maximizer of* $\max_{\forall m, h^m \in H^m, \mathbf{p} \in \Delta_\mathcal{C}} TCg(h^1, ..., h^M, p)$, *in other words,* $\forall h^{[M]} \in H^{[M]}, \mathbf{p} \in \Delta_\mathcal{C}$,

$$TCg(h_\star^1, ..., h_\star^M, p^\star) \geq TCg(h^1, ..., h^M, p^\star)$$

Maximizer → (Permuted) Ground-truth *If the prior is well defined, then for any maximizer of* TCg, $(\tilde{h}^{[M]}, \tilde{\mathbf{p}})$, *there is a permutation* $\tilde{\pi} : \mathcal{C} \to \mathcal{C}$ *such that:*

$$\tilde{h}^m(x^m)_c = P(Y = \tilde{\pi}(c)|X^m = x^m),\ \tilde{\mathbf{p}}_c = P(Y = \tilde{\pi}(c))$$

Proof. We have

$$\begin{aligned} TCg(h^1, ..., h^M, p) = \mathbb{E}_{x^{[M]} \leftarrow U_{X^{[M]}}} (1 + \log \sum_{y \in \mathcal{C}} \frac{\Pi_m h_\star^m(x^m)}{\mathbf{p}^\star(y)^{M-1}}) \\ - \mathbb{E}_{x^{[M]} \leftarrow V_{X^{[M]}}} \sum_{y \in \mathcal{C}} \frac{\Pi_m h_\star^m(x_i^m)}{\mathbf{p}^\star(y)^{M-1}} \end{aligned} \tag{8}$$

Ground-truth → Maximizer From definition 1, i.e.,

$$h_\star^m(x^m)_c = \mathrm{P}(Y = c|x^m),\ (p^\star)_c = \mathrm{P}(Y = c)$$

Inspired by Lemma 3 and 1, we have that the $TCg(h_*^{[M]}, p^\star)$ can achieve the maximum value, which equals to $\mathrm{TC}(X^1, ..., x^M)$.

Maximizer → (Permuted) Ground-truth For any maximizer $(\tilde{h}^{[M]}, \tilde{\mathbf{p}})$, we have from Lemma 3 that

$$\mathcal{R}(\tilde{h}^{[M]}, \tilde{\mathbf{p}}) = 1 + \mathrm{PTC}(x^1, ..., x^M),$$

which means that the $(\tilde{h}^{[M]}, \tilde{\mathbf{p}})$ satisfies Eq. (4). With assumption 2, we have that there exists a permutation $\tilde{pi} : \mathcal{C} \to \mathcal{C}$ such that

$$\tilde{h}^m(x^m)_c = \mathrm{P}(Y = \tilde{\pi}(c)|X^m = x^m),\ \tilde{\mathbf{p}}_c = \mathrm{P}(Y = \tilde{\pi}(c)).$$

□

B Algorithm

We show the pipeline of TCGM in Alg 1.

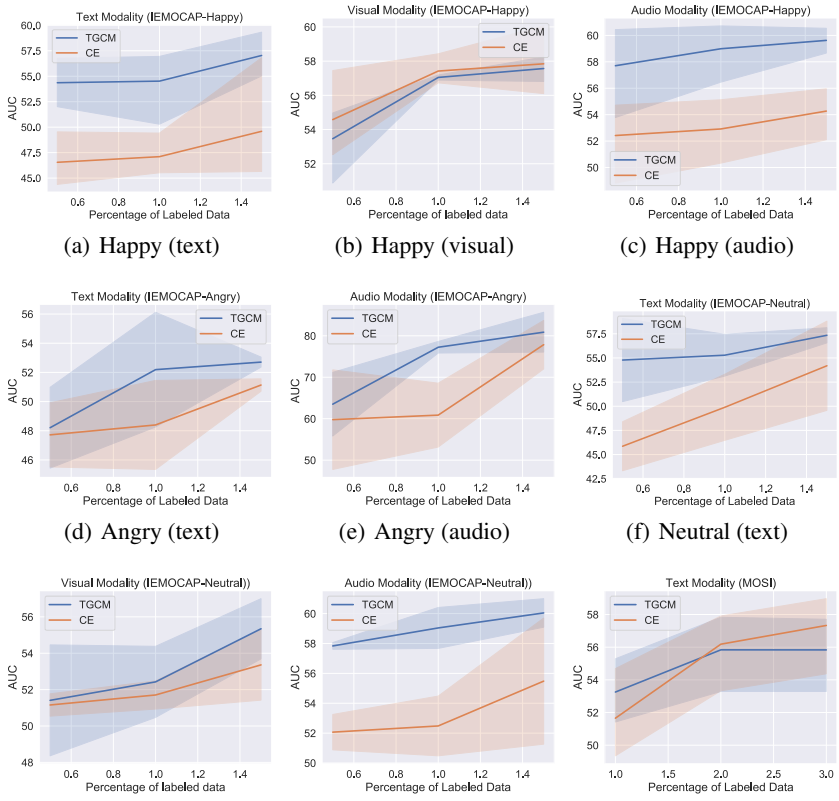
Time Complexity The overall loss function of our TCGM method is composed of $\mathcal{L}_{\text{CE}}$ which is $\mathcal{O}(M)$ since it is repeatedly implemented for M classifiers and $\mathcal{L}_{\text{TC}}^{(B)}$, as the addition of two terms, namely the term (a) ($\frac{1}{N}\sum_i \log \sum_{c \in \mathcal{C}} \frac{\Pi_m h^m(x_i^m)_c}{(p_c)^{M-1}}$) and the term (b) ($\frac{1}{N!/(N-M)!} \sum_{i_1 \neq i_2 \neq \cdots \neq i_M} \sum_{c \in \mathcal{C}} \frac{\Pi_m h^m(x_{i_m}^m)_c}{(p_c)^{M-1}}$) in Eq. (6). The term (a) with N samples generated from joint distribution $p(x^1, \ldots, x^M)$ for each modality; hence, the optimization of the term (a) is $\mathcal{O}(M)$. For term (b) with $N!/(N-M)!$ samples generated from marginal distribution $\Pi_m p(x^m)$, it is the sum of N terms by grouping terms with $h^m(x_i^m)$ for each $i \in [N]$ with $h^m(x_1^m), \ldots, h^m(x_N^m) form \in [M]$ calculated ahead (which is $\mathcal{O}(M)$), hence is linear scale with respect to M. Therefore, the time complexity for our loss function is $\mathcal{O}(M)$.

Algorithm 1 TCGM Optimization

Require: Unlabeled dataset $\mathcal{D}_u = \{x_i^{[M]}\}_i$, labeled dataset $\mathcal{D}_l = \{(x_i^{[M]}, y_i)\}_i$, M classifiers $\{h^m(\cdot;\Theta^m)\}_{m=1}^M$, epoch number T, learning rate γ_u, γ_l, batch size N and hyperparameter $\mathbf{p}$.
for epoch $t = 1 \to T$ **do**
  for $m = 1 \to M$ **do**
    for batch $b = 1 \to \lceil |\mathcal{D}_l|/B \rceil$ **do**
      Randomly sample a batch of samples:
      $\mathcal{B}_l = \{(x_i^{[M]}, y_i)\}_{i=1}^B$ from $\mathcal{D}_l$
      Compute the $\mathcal{L}_{\text{CE}}$ loss:
      $L \leftarrow \mathcal{L}_{\text{CE}}(\mathcal{B}_l; h^m(\cdot;\Theta^m))$
      Update Θ^m: $\Theta^m \leftarrow \Theta^m - \gamma_l \frac{\partial L}{\partial \Theta^m}$
    end for
  end for
  for batch $b = 1 \to \lceil (|\mathcal{D}_u| + |\mathcal{D}_l|)/B \rceil$ **do**
    Randomly sample a batch of samples:
    $\mathcal{B}_{u\cup l} = \{x_i^{[M]}\}_{i=1}^B$ from $\mathcal{D}_u \cup \mathcal{D}_l$
    Compute the $\mathcal{L}_{\text{TC}}$ loss:
    $L \leftarrow \mathcal{L}_{\textbf{TC}}(\mathcal{B}_{u\cup l}; \{h^m(\cdot;\Theta^m)\}_{m=1}^M, \mathbf{p})$
    Update $\Theta^{[M]}$: $\forall m, \Theta^m \leftarrow \Theta^m - \gamma_u \frac{\partial L}{\partial \Theta^m}$
  end for
end for

C Extended experiments results

We show the complete result of single modality classifiers on Emotion Recognition task in Figure 6.



(a) Happy (text) (b) Happy (visual) (c) Happy (audio)

(d) Angry (text) (e) Angry (audio) (f) Neutral (text)

(g) Neutral (video) (h) Neutral (audio) (i) MOSI (text)

Fig. 8. AUC of single modality classifiers by CE and TCGM.