Beyond Missing Modalities: Hypergraph Guided Diffusion for Uncertainty-Aware Multimodal Emotion Recognition

Supplementary Material

6. Theoretical Analysis

Assumption 1 ***(Smooth Evidence Mapping and Bounded Strength.)*** *For each modality $m \in \{t, v, a\}$ and each utterance i, the encoder $f_\theta^{(m)}$ outputs nonnegative evidence $\boldsymbol{e}_i^{(m)} \in \mathbb{R}_+^K$, with the Dirichlet parameter given by $\boldsymbol{\alpha}_i^{(m)} = \boldsymbol{e}_i^{(m)} + \mathbf{1}$. There exists $S_{\max} > 0$ such that the total strength $S_i^{(m)} = \sum_{k=1}^K \alpha_{i,k}^{(m)} \le S_{\max}$ almost surely. Moreover, $f_\theta^{(m)}$ is L_f–Lipschitz continuous in the input space:*

$$\|\boldsymbol{\alpha}_i^{(m)} - \boldsymbol{\alpha}_j^{(m)}\|_2 \le L_f \, \|u_i^{(m)} - u_j^{(m)}\|_2, \quad \forall i, j. \tag{23}$$

Assumption 2 ***(Sub-Gaussian Diffusion Reconstruction Error.)*** *Let $\epsilon_\theta(\mathbf{x}_t, C)$ be the diffusion model's conditional noise predictor and ϵ the true injected Gaussian noise. Define the per-sample squared error $u_i = \|\epsilon - \epsilon_\theta(\mathbf{x}_t, C)\|_2^2$. There exists $\sigma^2 > 0$ such that $u_i - \mathbb{E}[u_i]$ is sub-Gaussian with proxy σ^2, i.e.,*

$$\mathbb{E}(\exp\left(\lambda(u_i - \mathbb{E}[u_i])\right)) \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \quad \forall \lambda \in \mathbb{R}. \tag{24}$$

Assumption 3 ***(Safe and Isotone Fusion Operator.)*** *Let $\oplus$ denote the Dual Channel Evidence Fusion (DCEF) operator, which fuses two Dirichlet evidences $(\boldsymbol{\alpha}^{(a)}, m^{(a)}(\Omega))$ and $(\boldsymbol{\alpha}^{(b)}, m^{(b)}(\Omega))$ into $\boldsymbol{\alpha}^{(a\oplus b)}$. Assume:*

(i) ***Associativity:*** *$(\boldsymbol{\alpha}^{(a)} \oplus \boldsymbol{\alpha}^{(b)}) \oplus \boldsymbol{\alpha}^{(c)} = \boldsymbol{\alpha}^{(a)} \oplus (\boldsymbol{\alpha}^{(b)} \oplus \boldsymbol{\alpha}^{(c)})$.*

(ii) ***Conflict control:*** *The Dempster denominator satisfies $1 - \sum_{B\cap C=\emptyset} m_1(B)m_2(C) \ge \kappa > 0$.*

(iii) ***Isotonicity in the true class:*** *For the ground-truth class t, if $\alpha_t^{(a)} \ge \alpha_t^{(b)}$ and $\alpha_k^{(a)} \le \alpha_k^{(b)}$ for $k \ne t$, then $\alpha_t^{(a\oplus c)} \ge \alpha_t^{(b\oplus c)}$ and $\alpha_k^{(a\oplus c)} \le \alpha_k^{(b\oplus c)}$ for $k \ne t$.*

Lemma 1 (Dirichlet Moments) *Let $\mathbf{P} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ with $S = \sum_{k=1}^K \alpha_k$. For a one-hot target with $y_t = 1$,*

$$\mathbb{E}[P_t] = \frac{\alpha_t}{S}, \quad \mathbb{E}[P_t^2] = \frac{\alpha_t(\alpha_t+1)}{S(S+1)},$$
$$\mathbb{E}[P_j^2] = \frac{\alpha_j(\alpha_j+1)}{S(S+1)} \quad (j \ne t), \tag{25}$$

and

$$\mathbb{E}\left[\|\mathbf{y} - \mathbf{P}\|_2^2\right] = 1 - 2\frac{\alpha_t}{S} + \frac{1}{S(S+1)}\sum_{k=1}^K \alpha_k(\alpha_k+1). \tag{26}$$

***Proof* 1** *Standard Dirichlet identities yield $\mathbb{E}[P_k] = \alpha_k/S$ and $\mathbb{E}[P_k^2] = \alpha_k(\alpha_k+1)/[S(S+1)]$. Substituting into $\|\mathbf{y} - \mathbf{P}\|_2^2 = \sum_k (y_k^2 - 2y_kP_k + P_k^2)$ with $y_t = 1$ and $y_{j\ne t} = 0$.*

Lemma 2 (Monotonicity of the Dirichlet MSE Loss.) *Let $\ell(\boldsymbol{\alpha})$ denote the right-hand side of (26) for target class t. For fixed $S = \sum_k \alpha_k$, we have $\frac{\partial \ell}{\partial \alpha_t} < 0$ and $\frac{\partial \ell}{\partial \alpha_j} > 0$ for $j \ne t$. More generally, if $\boldsymbol{\alpha}^\star$ satisfies $\alpha_t^\star \ge \alpha_t$ and $\alpha_k^\star \le \alpha_k$ for $k \ne t$, then there exists $\delta \ge 0$ with $S^\star = S + \delta$ such that $\ell(\boldsymbol{\alpha}^\star) \le \ell(\boldsymbol{\alpha})$.*

***Proof* 2** *Differentiating (26) w.r.t. α_t with others fixed yields:*

$$\frac{\partial \ell}{\partial \alpha_t} = -\frac{2S - 2\alpha_t}{S^2} + \frac{(2\alpha_t+1)S(S+1) - \alpha_t(\alpha_t+1)(2S+1)}{S^2(S+1)^2}.$$

Algebraic manipulation shows this is negative for $\alpha_t > 0$ (omitted for brevity). For $j \ne t$, since the first term in (26) is independent of α_j and the second increases with α_j, we have $\partial\ell/\partial\alpha_j > 0$. When $S^\star \ne S$, expand and rearrange the difference $\ell(\boldsymbol{\alpha}^\star) - \ell(\boldsymbol{\alpha})$ to show it is bounded above by positive terms involving $(\alpha_t^\star - \alpha_t)$ and $-(\alpha_j - \alpha_j^\star)$ plus a decreasing function in $S^\star$, proving the claim.

Theorem 1 (Empirical Safety of HyperEF) *Let $\widehat{R}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\alpha}_i, y_i)$ be the empirical risk with ℓ defined in (26). Let $\boldsymbol{\alpha}_i'$ denote evidence before diffusion recovery/fusion, and $\widehat{\boldsymbol{\alpha}}_i$ the final evidence after MHGAT-conditioned diffusion and DCEF. Under Assumptions 1–3:*

$$\widehat{R}(\widehat{\boldsymbol{\alpha}}) \le \widehat{R}(\boldsymbol{\alpha}'). \tag{27}$$

***Proof* 3** *Fix i and let t be the true class. By Assumption 3(iii), DCEF fusion is isotone in the true class, implying $\widehat{\alpha}_{i,t} \ge \alpha'_{i,t}$ and $\widehat{\alpha}_{i,k} \le \alpha'_{i,k}$ for $k \ne t$ (or at least no disproportionate increase). Lemma 2 then gives $\ell(\widehat{\boldsymbol{\alpha}}_i, y_i) \le \ell(\boldsymbol{\alpha}'_i, y_i)$. Averaging over all i yields (27).*

Theorem 2 (Generalization Bound for HyperEF) *Let $\mathcal{H}$ be the hypothesis class mapping inputs to Dirichlet parameters via HyperEF, and let $\widehat{h} \in \arg\min_{h\in\mathcal{H}} \widehat{R}(h)$ denote the empirical minimizer. Assume:*

(i) The loss $\ell(\boldsymbol{\alpha}, y)$ is L_ℓ-Lipschitz in $\boldsymbol{\alpha}$ over the domain $\|\boldsymbol{\alpha}\|_2 \le B$;

(ii) The Rademacher complexity satisfies $\mathfrak{R}_n(\mathcal{H}) \le \mathcal{C}/\sqrt{n}$;

(iii) The diffusion error obeys Assumption 2 and contributes an additive term Δ_{diff} to the expected loss.

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$R(\widehat{h}) \le \widehat{R}(\widehat{h}) + 2L_\ell\,\mathfrak{R}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \Delta_{\text{diff}}. \quad (28)$$

where $R(\cdot)$ is the true risk. In particular, if $\mathfrak{R}_n(\mathcal{H}) \le C/\sqrt{n}$, then

$$R(\widehat{h}) \le \widehat{R}(\widehat{h}) + O\!\left(\frac{1}{\sqrt{n}}\right) + \Delta_{\text{diff}}. \quad (29)$$

Proof 4 *Consider the composed class $\ell \circ \mathcal{H} = \{(x,y) \mapsto \ell(h(x),y) : h \in \mathcal{H}\}$. By Talagrand's Lipschitz contraction lemma,*

$$\mathfrak{R}_n(\ell \circ \mathcal{H}) \le L_\ell\,\mathfrak{R}_n(\mathcal{H}) \le \frac{L_\ell\, C}{\sqrt{n}}. \quad (30)$$

Standard Rademacher complexity bounds (e.g., Bartlett & Mendelson) imply that, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}} \big|R(h) - \widehat{R}(h)\big| \le 2\,\mathfrak{R}_n(\ell\circ\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (31)$$

Let $\widehat{h} \in \arg\min_{h\in\mathcal{H}} \widehat{R}(h)$ be the empirical minimizer. Substituting $h = \widehat{h}$ and using the inequality above yields (28) except for the Δ_{diff} term. Finally, Assumption 2 bounds the additive bias Δ_{diff} contributed by the diffusion reconstruction error with high probability.

Theorem 3 ***(Semantic Consistency of MHGAT-Conditioned Diffusion)*** *Let $X \in \mathbb{R}^d$ be the ground-truth latent of a missing modality, C the MHGAT condition, and $\phi : \mathbb{R}^d \to \mathcal{S}$ a semantic map with Lipschitz constant L_ϕ (i.e., $\|\phi(x) - \phi(x')\|_{\mathcal{S}} \le L_\phi \|x - x'\|_2$). Consider a variance-preserving diffusion with classifier-free guidance (CFG) weight $w \ge 1$. Denote the true conditional score by $s_t^\star(\cdot \mid C) = \nabla \log p_t(\cdot \mid C)$, the learned conditional/unconditional scores by $s_t(\cdot \mid C)$ and $s_t(\cdot \mid \varnothing)$, and define the uniform conditional score error $\varepsilon_{score} := \sup_{t\in[0,T]} \Big(\mathbb{E}_{x\sim p_t(\cdot\mid C)} \|s_t(x \mid C) - s_t^\star(x \mid C)\|_2^2\Big)^{1/2}$. Sampling uses the CFG score $s_t^{\text{cfg}}(x \mid C) := s_t(x \mid \varnothing) + w\big(s_t(x \mid C) - s_t(x \mid \varnothing)\big)$. Let $\tilde{X} \sim q_0(\cdot \mid C)$ be the terminal sample produced by the reverse ODE/SDE with s_t^{cfg}, and let $p_0(\cdot \mid C)$ be the true conditional data law. Define the semantic target $\mu_\phi(C) := \mathbb{E}[\phi(X) \mid C]$ and the intrinsic conditional spread $\sigma_\phi(C) := \Big(\mathbb{E}\big[\|\phi(X) - \mu_\phi(C)\|_{\mathcal{S}}^2 \mid C\big]\Big)^{1/2}$. Assume: (A)* CFG tilt identity*: the ideal CFG score equals $\nabla \log\big(p_t(\cdot \mid C)^w p_t(\cdot)^{1-w}\big)$ up to an x-independent term [10]; (B)* Flow stability*: the reverse field is globally L-Lipschitz in state and score, the sampler is stable with constant K, and replacing the ideal score by an approximate one perturbs the terminal law by at most Ke^{LT} times the uniform L^2 score error along the path [33]; (C)* Density-ratio control*: along $t \in [0,T]$, the noised laws satisfy $r_t(x) := p_t(x)/p_t(x \mid C) \le R$ a.s. and have uniformly bounded second moments; (D)* Informative condition*: $\sigma_\phi(C) \le \epsilon_{\text{sem}}$. Then the semantic error $\Delta_\phi := \Big(\mathbb{E}\big[\|\phi(\tilde{X}) - \mu_\phi(C)\|_{\mathcal{S}}^2 \mid C\big]\Big)^{1/2}$ obeys $\Delta_\phi \le L_\phi\Big(K\,e^{LT}\,w\,\varepsilon_{\text{score}} + D\,(w-1)\Big) + \epsilon_{\text{sem}}$, for a constant $D = D(R, \text{schedule}) > 0$ depending only on the bounded density ratio and the diffusion schedule.*

Proof 5 ***(1) Terminal W_2 decomposition.*** *Let $\pi_0^{\text{cfg}}(\cdot \mid C)$ be the terminal law of the* ideal *reverse flow driven by the ideal CFG score from (A). By the triangle inequality in W_2,*

$$W_2\big(q_0, p_0\big) \;\le\; W_2\big(q_0, \pi_0^{\text{cfg}}\big) + W_2\big(\pi_0^{\text{cfg}}, p_0\big). \quad (32)$$

(2) Score approximation $\Rightarrow$ terminal W_2. *By (B),*

$$W_2\big(q_0, \pi_0^{\text{cfg}}\big) \;\le\; K\,e^{LT} \sup_t \Big(\mathbb{E}\|s_t^{\text{cfg}} - s_t^{\text{cfg}\star}\|_2^2\Big)^{1/2}, \quad (33)$$

where $s_t^{\text{cfg}\star} := s_t^\star(\cdot \mid \varnothing) + w\big(s_t^\star(\cdot \mid C) - s_t^\star(\cdot \mid \varnothing)\big)$. Using the triangle inequality and bounding the unconditional error by the same scale,

$$\sup_t \Big(\mathbb{E}\|s_t^{\text{cfg}} - s_t^{\text{cfg}\star}\|_2^2\Big)^{1/2} \;\le\; w\,\varepsilon_{\text{score}}, \quad (34)$$

hence $W_2(q_0, \pi_0^{\text{cfg}}) \le Ke^{LT} w\,\varepsilon_{\text{score}}$. ***(3) CFG tilt $\Rightarrow$ linear bias in $(w-1)$.*** *By (A), the ideal CFG at time t targets $\tilde{p}_t^{(w)}(x \mid C) \propto p_t(x \mid C)^w p_t(x)^{1-w} = p_t(x \mid C)\, r_t(x)^{1-w}$ with $r_t(x) = p_t(x)/p_t(x \mid C)$. Writing $w = 1+\delta$ and using (C), $|\log r_t(x)| \le \log R$, so a first-order expansion in δ plus stability of the probability flow under bounded log-density perturbations yields*

$$W_2\big(\pi_0^{\text{cfg}}, p_0\big) \;\le\; D\,(w-1), \quad (35)$$

with $D = D(R, \text{schedule})$; cf. transporting log-density/IPM perturbations to W_2 via optimal transport [36]. ***(4) From W_2 to semantics.*** *Couple $\tilde{X} \sim q_0(\cdot \mid C)$ and $X' \sim p_0(\cdot \mid C)$ optimally for W_2. By Lipschitzness of ϕ and Jensen,*

$$\Big(\mathbb{E}\|\phi(\tilde{X}) - \phi(X')\|_{\mathcal{S}}^2\Big)^{1/2} \;\le\; L_\phi\, W_2(q_0, p_0). \quad (36)$$

Decomposing around $\mu_\phi(C)$ and using (D),

$$\begin{aligned} \Delta_\phi &\le L_\phi\, W_2(q_0,p_0) + \sigma_\phi(C) \\ &\le L_\phi\Big(Ke^{LT} w\,\varepsilon_{\text{score}} + D(w-1)\Big) + \epsilon_{\text{sem}}. \end{aligned} \quad (37)$$

7. Algorithm Analysis

The model proposed in this paper can be primarily divided into two steps. The first step involves training the MHGAT-Conditioned Diffusion. The second step uses the pre-trained diffusion model to generate features for missing modalities to complete the dataset, followed by training the overall classification model.

In Algorithm 1, we present the pretraining procedure for the MHGAT-Conditioned Diffusion. Here we set epoch $P = 100$.

Algorithm 2 illustrates the training procedure of HyperEF, detailing how MHGAT-Conditioned Diffusion recoverys the missing features, and how DCEF evaluates the uncertainty and the overall objective function. We set epoch $E = 50$

Algorithm 1 Training of MHGAT-Conditioned Diffusion

Require: Dialogues U, masks $\mathcal{M}$, epochs P, steps T, layers L
Ensure: $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$

Initialize $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$
for epoch $= 1$ **to** P **do**
  for each mini-batch $\mathcal{B} \subset U$ **do**
    Build $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
    $v_i \leftarrow \text{Feat}(u_i) + \text{Emb}(M_i)$
    for $l = 1$ **to** L **do**
      $e_j^{l+1} = AGG_{v\to e}^{l}\left(e_j^l, \{v_i^l \mid v_i \in \mathcal{V}_j\}\right)$
      $v_i^{l+1} = AGG_{e\to v}^{l}\left(v_i^l, \{e_j^{l+1} \mid e_j \in \mathcal{E}_i\}\right)$
    end for
    $C \leftarrow \text{Concat}_{\text{modal}}(v^L)$
    for each missing feature x_0 **do**
      Sample $t \sim \mathcal{U}(1, T)$, $\epsilon \sim \mathcal{N}(0, I)$
      $x_t = \sqrt{\overline{\eta}_t} x_0 + \sqrt{1 - \overline{\eta}_t}\epsilon$
      $\hat{\epsilon} \leftarrow \text{U\_Net}(x_t, t, C)$
      $\mathcal{L}_{\text{diff}} \leftarrow \|\hat{\epsilon} - \epsilon\|_2^2$
    end for
    Update $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$ using $\nabla\mathcal{L}_{\text{diff}}$
  end for
end for
return $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$

Algorithm 2 Overall Training with HyperEF

Require: Pre-trained $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$, dialogues U, masks $\mathcal{M}$, labels Y
Ensure: Θ_{Evid}, fine-tuned $\Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$

Init Θ_{Evid}
for epoch $= 1$ **to** E **do**
  for mini-batch $\mathcal{B} \subset U$ **do**
    $C \leftarrow \text{MHGAT}(\mathcal{B}, \mathcal{M})$
    for missing feature x_0 **do**
      Sample $x_T \sim \mathcal{N}(0, I)$;
      for $t = T$ **to** 1 **do**
        $x_{t-1} \leftarrow \text{MHGATDIFF}(x_t, t, C)$
      end for
      $\tilde{x}_0 \leftarrow x_0$
    end for
    for $m \in \{t, v, a\}$ **do**
      $m^m(\Omega) = \gamma\, m_d^m(\Omega) + m_s^m(\Omega)$
      $m^m(k) = (1 - m^m(\Omega))P_k$
    end for
    Fuse $\{m^t, m^v, m^a\}$ by Dempster's rule
    $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1\mathcal{L}_{\text{KL}} + \lambda_2 Vac$
    Update all params via $\nabla\mathcal{L}$
  end for
end for
return $\Theta_{\text{Evid}}, \Theta_{\text{MHGAT}}, \Theta_{\text{Diff}}$

8. Detailed Network Structure

The UNet architecture in our diffusion model is a 1D convolutional encoder-decoder network with skip connections, tailored for sequential data. The input tensor of shape $[bs, seq, channel]$ is first projected to 128 channels via a 1D convolution, where bs is batch size, seq is sequence length. The downsampling path comprises four levels with channel multipliers $(1, 2, 2, 2)$, producing feature maps of 128, 256, 256, and 256 channels, respectively, with two residual blocks per level and halving the sequence length at each step (except the last). The bottleneck block maintains 256 channels. The upsampling path mirrors this structure, concatenating encoder features (skip connections) and reducing the channel dimensions stepwise back to 128 and finally $channel$.

Conditional embeddings are injected into every residual block via a cross-attention mechanism, where hidden states $[bs, C, L]$ are reshaped to $[bs, L, C]$ and attend to the conditional embedding $y \in \mathbb{R}^{bs\times 2\times 256}$, enabling context-aware modulation. Self-attention blocks at resolutions 8 and 16 (level 3 in the encoder and corresponding decoder levels) model long-range dependencies, while timestep embeddings modulate residual blocks to ensure temporal consistency throughout the denoising process.

9. Dataset Segmentation and Preprocessing

Dataset: As listed in Table 6, IEMOCAP4 includes four types of emotions: anger, happiness (where excitement is merged with happiness), sadness, and neutral. We assign 3290, 1000, and 1241 utterances for train, valid, and test. The six-class dataset IEMOCAP6 encompasses: anger, happiness, sadness, neutral, excitement, and frustration. We assign 4810, 1000, and 1623 utterances for train, valid, and test. MELD consists of 13780 utterances, where 9989, 1109, 2610 samples are set for train, valid, and test. Ut-

Dataset	Type	Train	Val	Test	Total
IEMOCAP4	Utterance	3,290	1,000	1,241	5,531
	Dialogue	90	30	31	151
IEMOCAP6	Utterance	4,810	1,000	1,623	7,433
	Dialogue	90	30	31	151
MELD	Utterance	9,989	1,109	2,610	13,708
	Dialogue	1,038	114	280	1,432

Table 6. Statistical information of utterances and dialogues on IEMOCAP4, IEMOCAP6, and MELD.

terances are labeled with emotions: anger, disgust, sadness, joy, surprise, fear, or neutral, and sentiment: positive, negative, or neutral.

Preprocessing: We apply different preprocessing pipelines to the IEMOCAP and MELD datasets.

IEMOCAP: For each modality, we employ the corresponding pre-trained network to perform feature extraction. 1) Language: Pre-trained DeBERTa [9] is employed as the language feature extractor. Motivated by its demonstrated superiority in natural language understanding and generation tasks, we leverage the DeBERTa-large variant to encode utterance sequences into 1024-dimensional representations. 2) Vision: The pre-trained MA-Net [46] serves as the visual feature extractor, utilizing global multi-scale and local attention mechanisms to handle occlusions and non-frontal poses. We first apply MTCNN to detect and align faces, followed by extracting facial features via pre-trained MA-Net. Frame-level features are then compressed into 1024-dimensional utterance-level representations through average encoding. 3) Acoustic: Pre-trained wav2vec [29] serves as the acoustic feature extractor, leveraging its multi-layer convolutional architecture trained on massive unlabeled speech data. Building on its demonstrated success in downstream applications like speech recognition , we adopt wav2vec-large to extract 512-dimensional acoustic features from utterances.

MELD: Inspired by [4], we adopt the previous approach for feature extraction. 1) Language: We utilize RoBERTa Liu et al. [22] to encode textual data into 1024-dimensional representations. 2) Vision: We use DenseNet [12] to encode vision data into 342-dimensional representations. 3) Acoustic: OpenSmile [30] is used to encode acoustic data into 300-dimensional representations.

To ensure consistency in the dimensionality of different modality features, we add a linear layer after each feature extractor to project the extracted features to a unified 1536-dimensional space.

10. Settings of MHGAT vs. Transformer encoder

Transformer encoder configuration: We use a compact Transformer encoder with two layers and pre layer normalization. Each layer applies global multi head self attention with eight heads, followed by a two layer MLP with GELU and expansion ratio four, with dropout 0.1 in both attention and MLP. Inputs are sequences of shape $[Batch\_size, Seq\_length, Dim]$, augmented with a learned absolute positional embedding. Notably, the Transformer requires concatenating the three modalities from the same dialogue along the Seq_length dimension. This concatenation enables self-attention to model arbitrary relations between utterances across modalities. no class token is used. Padding is handled through a key padding mask in attention. A final LayerNorm produces the output sequence, which is then passed to the shared classification head.

Time complexity: In both MHGAT and the Transformer, runtime is dominated by attention computation. Therefore, we focus our analysis on the time complexity of computing attention. Assume a dialogue contains N utterances and 3 modalities, with each utterance represented by a D dimensional feature. Because the Transformer concatenates the three modalities of the same sample along the sequence dimension, the sequence length becomes $3N$. The attention cost therefore scales quadratically with length, i.e., $O((3N)^2 D) = O(9DN^2)$. While in MHGAT, the number of hyeredge is $N + 3$. According to Eq. (1-5) in Sec. 3.1.1, the computational costs of the Node to Hyperedge and Hyperedge to Node updates are identical, i.e., $O(2((N + 3)3ND)) = O(6DN^2 + 18ND)$. Both models incur quadratic attention cost in sequence length, i.e.,$\Theta(DN^2)$. However, MHGAT has a smaller leading constant, yielding about 33% fewer leading-order operations under the same D. Thus, MHGAT is asymptotically equivalent but constant-factor cheaper.

11. Deep Analysis of Visualization Experiments

Figure 5 illustrates the distribution of features recovered by different modality recovery methods compared to the ground truth, under fixed modality-missing paradigms. To simulate scenarios with extreme data missingness, we established six fixed modality-missing paradigms. Under these paradigms, all data from the missing modalities are masked, and only conditional information extracted from the available modalities is used to guide the feature recovery process. All experiments were conducted on the IEMOCAP dataset. We can observe that, regardless of the modality incomplete condition, the distribution of the features recovered by HyperEF is closer to that of the ground truth compared to SOTA methods.

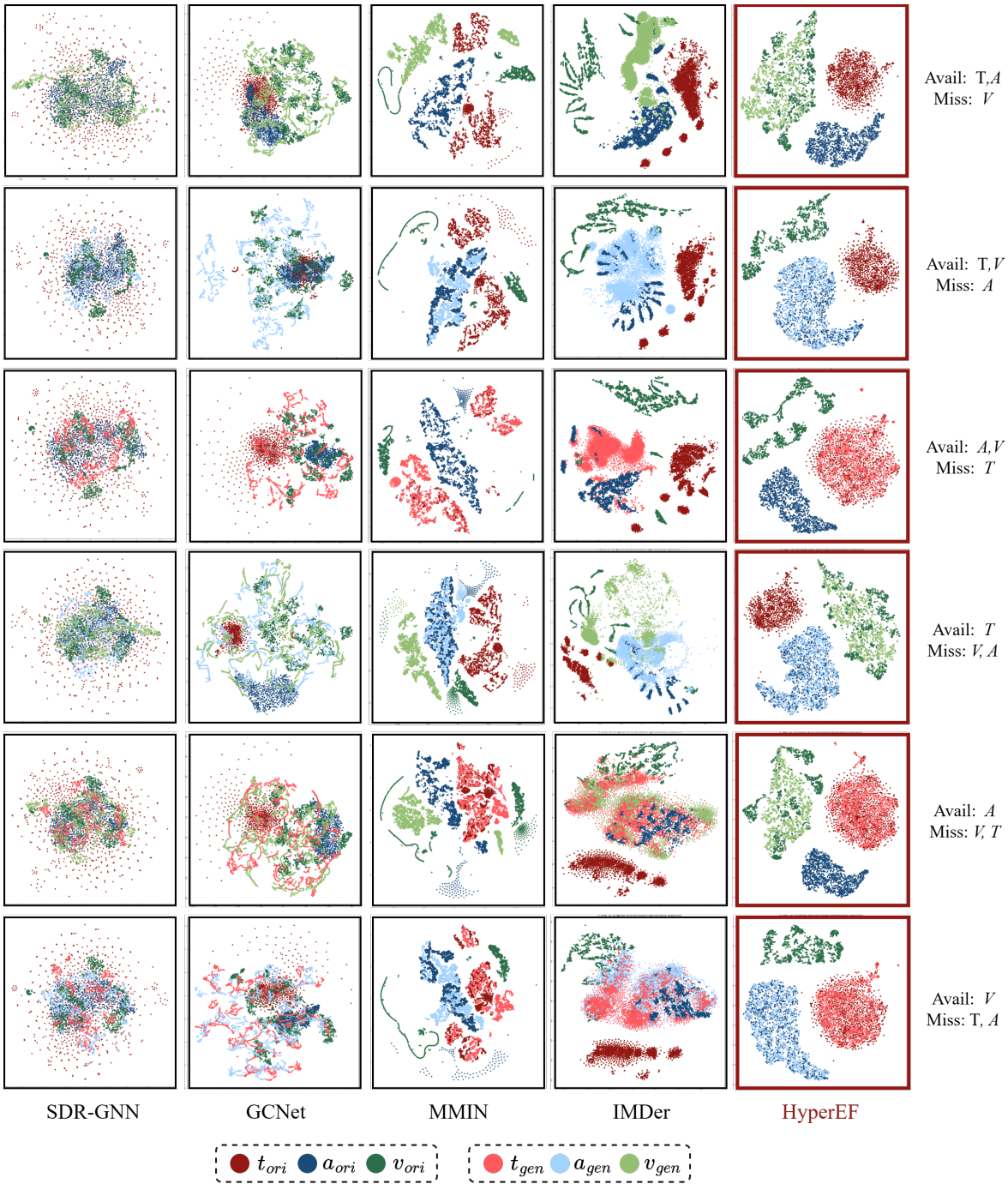

Figure 5. t-SNE visualization: comparison of recovered versus original feature distributions by different modality recovery methods under all modality incomplete conditions on IEMOCAP4.

12. Visualization of MHGAT's Conditional Effect

Figure 6 illustrates the t-SNE plots of the feature distributions for each modality after feature recovery using an unconditional diffusion model and the MHGAT-guided diffusion model, respectively. Inspired by [39], we randomly select 120 samples (20 per class) from the IEMOCAP6 test set and project both generated and original features into a 2D space via t-SNE. When recovering missing modalities with the unconditional diffusion model, the absence of semantic guidance results in recovered features that are only ap-

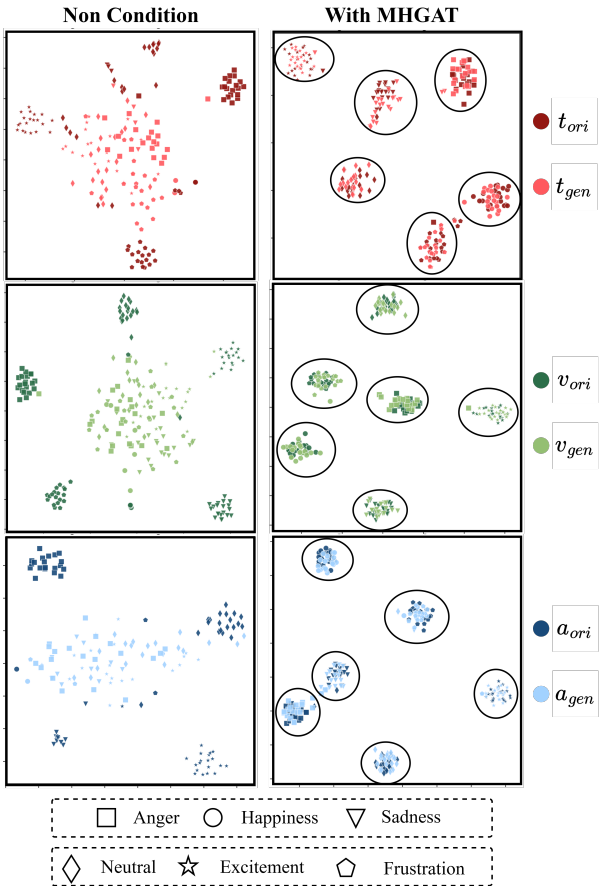

Figure 6. t-SNE visualization of feature distributions under different ablation settings on IEMCAP6

proximately similar to the original features in terms of overall data distribution. However, at the semantic level within individual modalities, the recovered features are inconsistent with the semantics expressed by the original features, leading to significant semantic ambiguity. In contrast, the features recovered by the MHGAT-guided diffusion model exhibit high semantic consistency with the original features, effectively mitigating the issue of semantic ambiguity in the recovered features.

13. Interpretable Ablation Experiments

Figure 7 reports how different uncertainty terms, when inserted into the objective function, affect accuracy, evidence, and output entropy. Experiments are conducted on IEMOCAP4 with a missing rate of 0.3. We denote the objective function described by Eq. 22 as 'Normal'.

Figure 7(a) illustrates the accuracy variation curves of the model under different target functions. It can be observed that removing the $\mathcal{L}_{KL}$ significantly degrades model performance, removing the vacuity term (Vac) slows down the model's convergence rate, and removing the dissonance

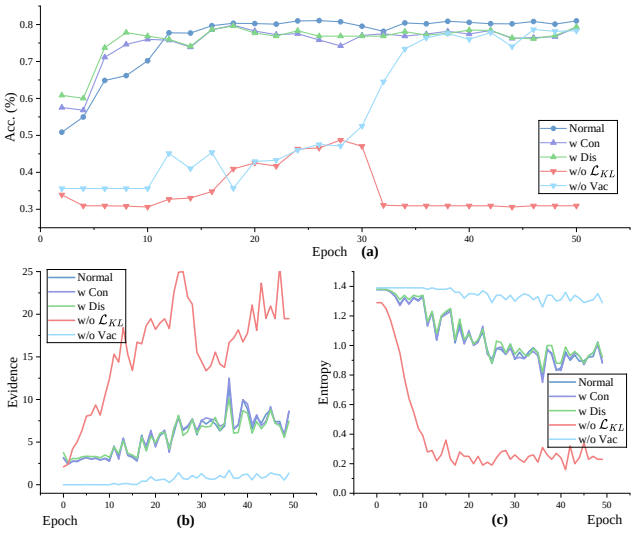

Figure 7. (a) the accuracy variation curves of the model under different regularization terms. (b) the average of evidence sum across different modalities under various target function conditions. (c) the average entropy under different regularization terms.

(Dis) and consonance (Con) terms has a smaller impact but still reduces the model's recognition accuracy to some extent.

Figure 7(b) records the average of evidence sum across different modalities during model training under various target function conditions. It can be observed that removing the $\mathcal{L}_{KL}$ leads to a sharp increase in the evidence output by the model, a phenomenon we term "evidence explosion." This occurs because, without the $\mathcal{L}_{KL}$ constraining incorrect categories, The Vac term biases the model toward producing larger output evidence, consistent with the results in Figure 7(a). Additionally, removing the vacuity term (Vac) significantly reduces the total evidence sum. For without the Vac constraint, the $\mathcal{L}_{KL}$ tends to drive every incorrect class evidence to zero. An excessively low evidence leads to ambiguous final discriminations, thereby slowing convergence speed, which also aligns with the results in Figure 7(a).

Figure 7(c) records the average entropy of the final output probability vectors across different modalities during the training process. After removing the $\mathcal{L}_{KL}$, the entropy sharply decreases. This is because, without the $\mathcal{L}_{KL}$, Vac amplifies interclass evidence gaps, pushing the model toward overconfidence. Conversely, when Vac is removed, the entropy remains very high, aligning with the prior analysis, because the evidence for each category is minimal, resulting in nearly equal probabilities across all categories. Excessively high entropy fails to provide effective category information.

From the above experiments, we observe that Dis and Con exert negligible influence on accuracy, evidence, and

Method	MELD								Time
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
DDPM	69.3	**67.2**	**64.8**	**63.8**	**61.8**	**60.4**	**58.6**	**57.1**	23.1s
DDIM	69.3	66.7	63.4	61.6	61.3	59.0	57.5	56.8	1.82s

Method	IEMOCAP4								Time
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
DDPM	82.9	**82.1**	79.1	**80.3**	77.6	**78.2**	**77.7**	**77.0**	22.4s
DDIM	82.9	81.1	**80.7**	79.5	**78.4**	77.6	76.5	76.9	1.77s

Table 7. Comparison of DDPM and DDIM on MELD and IEMOCAP4 under different missing rates, and their computation cost per batch.

entropy. To preserve training stability, we retain Vac and $\mathcal{L}_{KL}$ as the regularizers in our final objective function.

14. Analysis of computation cost

Considering that our method employs a diffusion model for feature recovery, computational cost becomes a crucial indicator of practical performance. In this section, we provide a detailed analysis of HyperEF's time cost. It is worth noting that our proposed feature recovery framework does not restrict the choice of generative model; any conditional generative model can be used. All main experiments in this paper adopt a DDPM-based conditional diffusion model with 300 timesteps. In the supplementary study, we replace the generator with a conditional DDIM using only 20 timesteps and conduct experiments on MELD and IEMOCAP across all missing rates. All remaining settings are held constant to isolate the effect of the generator and ensure a fair comparison. The results are shown in Table 7. The 'Time' column reports the average per-batch training time. Using DDIM with fewer timesteps requires only 7.9% of the DDPM time while achieving near-DDPM classification performance. On MELD, DDIM's restoration is slightly lower than DDPM. On IEMOCAP, DDIM and DDPM deliver comparable restoration quality.