Reliability-Coupled Manifold-Aware Diffusion for Missing-Modality Inference

Yiming Ren[1], **Yuhao Fang**[1], **Qing Zhou**[1], **Ming Li**[2],
Ye Zhang[1,3], **Zijian Wang**[4], **Yao Lu**[1], **Chun Li**[1*]
[1] Shenzhen MSU-BIT University, Shenzhen, China
[2] The Chinese University of Hong Kong, Shenzhen, China
[3] Beijing Institute of Technology, Beijing, China
[4] China Media Group, Beijing, 100859, China.
lichun2020@smbu.edu.cn

Abstract

In real-world multimodal time-series inference, a common practice is completion–inference decoupling: missing modalities are heuristically imputed and then passed to deterministic fusion and decision making. We show that this chained design yields unreliable predictions under missingness, noise, and temporally varying modality reliability, because imputation errors propagate without uncertainty feedback. Through theoretical and empirical analysis, we identify latent-consistent completion and evidential reliability modeling as two key ingredients for robust multimodal inference. Building on these insights, we propose MD2E-MI, a manifold-aware diffusion-based imputation and evidence-driven multimodal inference framework that couples diffusion-based completion with uncertainty-aware inference in a single pathway. Experiments demonstrate stable performance across diverse missingness and noise regimes, while providing interpretable, temporally varying reliability estimates.

1 Introduction

Multimodal temporal classifiers often receive incomplete text, audio, visual, or sensor measurements. A common response is to reconstruct missing inputs and then classify the completed sequence [1–6]. Such pipelines can be fragile: a plausible reconstruction need not be a reliable measurement, yet a downstream classifier may not distinguish the two. As missingness grows or persists across consecutive timesteps, completion errors can increasingly influence predictions.

Our question is therefore not only how to recover missing representations, but also how to estimate the confidence with which they should contribute to a decision. There are plausible alternatives to explicit completion–inference coupling: a variable-input model can classify directly from observed modality–timestep pairs, while end-to-end training or a post-hoc reliability head may improve inference without a shared evidential predictor. We compare these alternatives under controlled protocols.

We introduce MD2E-MI, which connects two stages through a *shared evidential predictor*. Stage A completes missing task latents using a fixed spherical diffusion backend with temporal regularization. At observed modality–timestep pairs, the predictor's Dirichlet mean determines divergence-dependent weights for denoising supervision. Stage B applies the same predictor to the completed latents, derives confidence from total Dirichlet evidence, and fuses the resulting predictions over valid timesteps. The contribution lies in jointly training this completion–inference interface, building on established diffusion, evidential-learning, and geometric components [7, 8].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We evaluate the framework under random and block-missing protocols on two sentiment and two physical-sensor datasets. Alongside comparisons with MD^{2}N and other baselines, controlled experiments assess completion-free inference, post-hoc reliability, ordinary end-to-end completion, and simpler Euclidean regularizers. The results help identify when the additional completion and geometric computation is justified. Our contributions are:

- **Shared reliability coupling.** A modality–timestep evidential pathway shapes denoising supervision and is reused to gate downstream fusion, linking completed representations with estimated trustworthiness.
- **Manifold-aware latent completion.** A fixed spherical backend and temporal regularization constrain the completion trajectories. Comparisons with Euclidean and unit-norm alternatives quantify the contribution of this geometric inductive bias.
- **Controlled evaluation.** Protocol-aligned comparisons and targeted ablations examine when completion is useful, how predicted confidence relates to completion error, and what accuracy–compute tradeoffs the method entails.

2 Related Work

Learning with missing modalities. Methods for incomplete multimodal learning either operate on the available observations or reconstruct missing representations before fusion [3, 5, 6, 9]. SMIL [10] studies severe modality missingness using Bayesian meta-learning. MCTN [11] learns cyclic cross-modal translations from a designated source modality, while MVAE [12] and MMVAE [13] support inference and generation with different observed-modality subsets. Our setting concerns time-varying modality–timestep masks and task-directed latent completion. We additionally evaluate attention- and evidence-based models that use only observed inputs.

Diffusion and geometric completion. Diffusion-based imputation includes CSDI and MissDiff [14, 15], alongside more recent work on missing-modality recovery and gradient-flow interpretations [16, 17]. Riemannian score-based modelling [8] develops generative processes on manifolds. In contrast, we use a fixed, discrete spherical backend as a constraint on task latents, without assuming spherical structure in the raw observations. Matched Euclidean and simple-regularizer baselines assess its empirical contribution.

Evidence and reliability-aware fusion. Evidential Deep Learning [7] provides the Dirichlet formulation, and related multi-view methods use uncertainty estimates for fusion [18–22]. MD2E-MI uses a shared trainable predictor to influence diffusion denoising supervision and to supply confidence for downstream classification. This coupling differs from a serial imputer–classifier and from a reliability head trained after freezing the completion backbone; the corresponding alternatives are evaluated directly.

3 Problem Setup

We study sample-level classification for a multimodal sequence with M modalities and T timesteps. A sample is

$$X = \{x_t^{(m)} \in \mathbb{R}^{d_m} : m = 1, \dots, M,\ t = 1, \dots, T\}, \tag{1}$$

with label $y \in \{1, \dots, K\}$. Missingness is specified by an observation mask $o_t^{(m)} \in \{0, 1\}$, where $o_t^{(m)} = 1$ means modality m is observed at timestep t.

For a modality with at least one observed timestep, define

$$T_{\text{valid}}^{(m)} = \{t : o_t^{(m)} = 1\}, \qquad |T_{\text{valid}}^{(m)}| > 0. \tag{2}$$

The sample-level valid-timestep union is

$$T_{\text{valid}} = \bigcup_{m=1}^{M} T_{\text{valid}}^{(m)}, \qquad |T_{\text{valid}}| > 0. \tag{3}$$

Algorithm 1 MD2E-MI training and inference

Require: Multimodal sequence X, observation mask o, label y

1: Encode observed pairs: $s_{0,t}^{(m)} = \mathrm{Enc}_m(x_t^{(m)})$ for $o_t^{(m)} = 1$
2: Sample diffusion time τ and tangent noise for observed pairs
3: Build cross-modal context from noised latents and masks
4: Call the shared evidential module on observed latents to obtain $\alpha_t^{(m)}$
5: Use $\alpha_t^{(m)}$ to compute the denoising weights in L_{diff}
6: Impute missing pairs with fixed-step reverse diffusion to obtain $\hat{s}_{0,t}^{(m)}$
7: Call the same evidential module on completed latents
8: Convert total evidence into confidence gates $\pi_t^{(m)}$, fuse Dirichlet evidence, and pool over T_{valid}
9: Optimize $L_{\mathrm{sup}} + \lambda_{\mathrm{diff}} L_{\mathrm{diff}} + \lambda_{\mathrm{PDE}} L_{\mathrm{PDE}} + \lambda_{\mathrm{geo_time}} L_{\mathrm{geo_time}} + \lambda_{\mathrm{geo_acc}} L_{\mathrm{geo_acc}}$

Prediction and supervision are pooled only over this union. If all modalities are missing at a timestep, that timestep is excluded from pooling and supervision; the method does not require every timestep to retain at least one observed modality. In the fixed-availability diagnostics, an entire modality may be absent, so $T_{\mathrm{valid}}^{(m)} = \varnothing$ for that modality. Per-modality loss summands that average over an empty set are defined to contribute zero; this is a convention for the normalized objectives, not an additional observed input. Sample-level prediction still requires $|T_{\mathrm{valid}}| > 0$.

Latent backend. Observed inputs are encoded as $s_{0,t}^{(m)} = \mathrm{Enc}_m(x_t^{(m)}) \in \mathcal{M}$. The reported main experiments use one shared spherical backend across all modalities for Exp, Log, PT, and $d_{\mathcal{M}}$. This fixes the geometric operators during evaluation and treats the manifold backend as a regularizer for severe missingness rather than as an empirical claim about the intrinsic geometry of every dataset.

Dirichlet evidence. For each modality and timestep, an evidential module outputs nonnegative evidence

$$e_{t,k}^{(m)} = \mathrm{Softplus}(g_{t,k}^{(m)}), \qquad \alpha_{t,k}^{(m)} = e_{t,k}^{(m)} + 1. \tag{4}$$

Let $\alpha_{t,0}^{(m)} = \sum_{k=1}^{K} \alpha_{t,k}^{(m)}$ and $\mu_{t,k}^{(m)} = \alpha_{t,k}^{(m)} / \alpha_{t,0}^{(m)}$. We define confidence by

$$c_t^{(m)} = \frac{\alpha_{t,0}^{(m)} - K}{\alpha_{t,0}^{(m)}} \in [0, 1). \tag{5}$$

4 Method

4.1 Overview

MD2E-MI has two coupled stages, summarized in Algorithm 1. Stage A completes missing latents using a diffusion denoiser and geometric regularization. Stage B predicts Dirichlet evidence, gates modalities by confidence, and pools over valid timesteps. The same evidential module $Evi_m(\cdot)$ is called in both stages but serves different roles and receives gradients through different losses.

4.2 Stage A: Manifold-Aware Diffusion Completion

For an observed pair (m, t), sample $\tau \sim \mathcal{U}(0, 1)$ and tangent noise $\varepsilon_t^{(m)} \sim \mathcal{N}(0, I)$ in $T_{s_{0,t}^{(m)}}\mathcal{M}$. The noised latent is

$$s_{\tau,t}^{(m)} = \mathrm{Exp}_{s_{0,t}^{(m)}}\left(\sigma(\tau)\varepsilon_t^{(m)}\right). \tag{6}$$

We build cross-modal context from other modalities and their masks,

$$C_t^{(m)} = \mathrm{Context}_{\phi}^{(m)}\left(\bigoplus_{m' \neq m} [s_{\tau,t}^{(m')}; o_t^{(m')}]\right), \tag{7}$$

and predict tangent noise

$$\hat{\varepsilon}_{\tau,t}^{(m)} = f_\theta(s_{\tau,t}^{(m)}, \tau, C_t^{(m)}) \in T_{s_{\tau,t}^{(m)}}\mathcal{M}. \tag{8}$$

Because injected and predicted noise live in different tangent spaces, the target is aligned by parallel transport:

$$\tilde{\varepsilon}_{\tau,t}^{(m)} = \mathrm{PT}_{s_{0,t}^{(m)} \to s_{\tau,t}^{(m)}}(\varepsilon_t^{(m)}). \tag{9}$$

Shared evidence in Stage A. At observed timesteps, Stage A calls the shared evidential module

$$\alpha_t^{(m)} = Evi_m(s_{0,t}^{(m)}, C_{t,0}^{(m)}), \tag{10}$$

but this output has no standalone classification target. It enters only through the denoising weight inside L_{diff}. Let $u_k = 1/K$ and $\mu_t^{(m)} = \alpha_t^{(m)}/\alpha_{t,0}^{(m)}$. We compute

$$D_{\mathrm{KL},t}^{(m)} = \sum_{k=1}^{K} \mu_{t,k}^{(m)} \log \frac{\mu_{t,k}^{(m)}}{u_k}, \tag{11}$$

and Sharma–Mittal divergence against the uniform reference,

$$D_{\mathrm{SM},t}^{(m)} = D_{\mathrm{SM}}^{(\rho,\delta)}(\mu_t^{(m)} \| u), \qquad \rho > 0,\ \rho \neq 1,\ \delta \neq 1. \tag{12}$$

The denoising weights depend on divergences of the predictive *mean*, rather than on evidence concentration. The same predictor's parameters also receive supervised gradients through Stage B, where total evidence is used to determine the fusion gates. The diffusion-stage attenuation is

$$g(\tau) = \frac{\sigma(\tau)^2}{\sigma(\tau)^2 + \sigma_{\min}^2}. \tag{13}$$

The nonnegative denoising weight is

$$w_{\tau,t}^{(m)} = g(\tau)\,\mathrm{Softplus}\left(1 + \lambda_{\mathrm{SM}} D_{\mathrm{SM},t}^{(m)} + \lambda_{\mathrm{KL}} D_{\mathrm{KL},t}^{(m)}\right). \tag{14}$$

The weighted denoising objective is

$$L_{\text{diff}} = \sum_{m=1}^{M} \frac{1}{|T_{\text{valid}}^{(m)}|} \sum_{t \in T_{\text{valid}}^{(m)}} w_{\tau,t}^{(m)} \left\| \hat{\varepsilon}_{\tau,t}^{(m)} - \tilde{\varepsilon}_{\tau,t}^{(m)} \right\|_2^2. \tag{15}$$

Reverse imputation. For observed pairs we set $\hat{s}_{0,t}^{(m)} = s_{0,t}^{(m)}$. For missing pairs, we run fixed-step reverse diffusion with K_{rev} equally spaced steps $\tau_i = i/K_{\text{rev}}$:

$$\hat{s}_{\tau_{i-1},t}^{(m)} = \mathrm{ReverseUpdate}_\theta(\hat{s}_{\tau_i,t}^{(m)}, \tau_i, \tau_{i-1}, X, o), \quad i = K_{\text{rev}}, \ldots, 1. \tag{16}$$

The default setting uses $K_{\text{rev}} = 20$.

4.3 Stage B: Evidence-Gated Fusion

After completion, the shared evidential module is applied to completed latents and context:

$$\alpha_t^{(m)} = Evi_m(\hat{s}_{0,t}^{(m)}, \hat{C}_t^{(m)}). \tag{17}$$

This path is trained through the supervised loss. Confidence is converted into modality gates,

$$\pi_t^{(m)} = \frac{\exp(\lambda c_t^{(m)})}{\sum_{j=1}^{M} \exp(\lambda c_t^{(j)})}. \tag{18}$$

Fused Dirichlet parameters and per-timestep class probabilities are

$$\alpha_t^{(\text{fuse})} = \sum_{m=1}^{M} \pi_t^{(m)} \alpha_t^{(m)}, \qquad p_{t,k} = \frac{\alpha_{t,k}^{(\text{fuse})}}{\alpha_{t,0}^{(\text{fuse})}}. \tag{19}$$

Because the fused Dirichlet parameters are normalized after evidence aggregation, the effective convex-mixture weight of modality m on the class probabilities is proportional to $\pi_t^{(m)} \alpha_{t,0}^{(m)}$, not simply $\pi_t^{(m)}$ (see Theorem 1 in Appendix A). The sample prediction pools over the valid-timestep union:

$$\bar{p}_k = \frac{1}{|T_{\text{valid}}|} \sum_{t \in T_{\text{valid}}} p_{t,k}, \qquad L_{\text{sup}} = -\log \bar{p}_y. \tag{20}$$

4.4 Temporal Regularization

On interior indices $2 \le t \le T-1$, we define a tangent residual between neighboring latent states:

$$r_t^{(m)} = \mathrm{Log}_{\hat{s}_{0,t}^{(m)}}(\hat{s}_{0,t+1}^{(m)}) - \mathrm{Log}_{\hat{s}_{0,t}^{(m)}}(\hat{s}_{0,t-1}^{(m)}), \tag{21}$$

and penalize

$$L_{\mathrm{PDE}} = \sum_{m=1}^{M} \frac{1}{|T_{\mathrm{valid}}^{(m)}|} \sum_{t \in T_{\mathrm{valid}}^{(m)}} \|r_t^{(m)}\|_2^2. \tag{22}$$

We also regularize geodesic velocity and the neighboring-state residual. Neighbor-dependent terms use only timesteps at which their required neighbors exist: $I_{\mathrm{vel}}^{(m)} \subseteq \{2, \dots, T\}$ and $I_{\mathrm{acc}}^{(m)} \subseteq \{2, \dots, T-1\}$. Boundary residuals are omitted and empty eligible sets contribute zero. The velocity penalty weights geodesic distances by confidence or gating:

$$L_{\mathrm{geo_time}} = \sum_{m=1}^{M} \frac{1}{|I_{\mathrm{vel}}^{(m)}|} \sum_{t \in I_{\mathrm{vel}}^{(m)}} \gamma_t^{(m)} d_{\mathcal{M}}(\hat{s}_{0,t}^{(m)}, \hat{s}_{0,t-1}^{(m)})^2, \tag{23}$$

with $\gamma_t^{(m)} = \gamma_0 c_t^{(m)}$ or $\gamma_0 \pi_t^{(m)}$. The second neighbor-residual term is

$$L_{\mathrm{geo_acc}} = \sum_{m=1}^{M} \frac{1}{|I_{\mathrm{acc}}^{(m)}|} \sum_{t \in I_{\mathrm{acc}}^{(m)}} \left\| \mathrm{Log}_{\hat{s}_{0,t}^{(m)}}(\hat{s}_{0,t+1}^{(m)}) - \mathrm{Log}_{\hat{s}_{0,t}^{(m)}}(\hat{s}_{0,t-1}^{(m)}) \right\|_2^2. \tag{24}$$

Both L_{PDE} and $L_{\mathrm{geo_acc}}$ use the displayed neighboring-state tangent difference, with separately weighted aggregations. Their shared integrand should not be interpreted as two independent second-order constraints. Despite the name of the latter term, this expression is not an intrinsic covariant second derivative; the label PDE denotes a regularization component rather than a separately solved physical governing equation.

The full objective is

$$L = L_{\mathrm{sup}} + \lambda_{\mathrm{diff}} L_{\mathrm{diff}} + \lambda_{\mathrm{PDE}} L_{\mathrm{PDE}} + \lambda_{\mathrm{geo_time}} L_{\mathrm{geo_time}} + \lambda_{\mathrm{geo_acc}} L_{\mathrm{geo_acc}}. \tag{25}$$

4.5 Supporting Properties

The fusion rule has three useful properties. First, the fused Dirichlet mean is a convex combination of modality-wise predictive means, so Jensen's inequality bounds its NLL by the corresponding weighted average of individual losses. Second, confidence $c^{(m)} = 1 - K/\alpha_0^{(m)}$ increases with total evidence, and the gate odds satisfy $\pi_i/\pi_j = \exp\{\lambda(c^{(i)} - c^{(j)})\}$. Third, under local nondegeneracy and Lipschitz conditions, prediction changes admit a local bound in terms of completion error. These properties characterize the proposed fusion mechanism; assumptions, proofs, and the scope of the local bound are given in Appendix A.

5 Experiments

5.1 Protocol

Datasets. We evaluate on CMU-MOSEI and CMU-MOSI for multimodal sentiment classification [23, 24], and on UCI-HAR and PAMAP2 for sensor-based human activity recognition [25, 26]. Sentiment datasets use text, audio, and vision modalities. Sensor datasets are used to stress long contiguous failures.

Missingness protocols. R1 Random Dropout masks modalities or channels independently with probability p_{miss}. Thus MR=0.9 means that each modality/channel is independently masked with probability 0.9; it does not mean that 90% of samples lose at least one modality. R2 Block Failure removes a contiguous segment covering a specified fraction of timesteps. For sensor experiments, Miss(0.8) denotes an R2 block covering 80% of timesteps.

Table 1: Performance on CMU-MOSEI and CMU-MOSI under varying missing rates. Results are Acc-2 / F1 / Acc-7 (%). Baseline entries marked † in the MR=0.9 column are unified-protocol re-evaluations/adaptations, not published numbers; Table 16 summarizes their source and protocol. The best, second-best, and third-best performances are highlighted in bold red, blue, and green, respectively.

Dataset	Method	Venue	0.0	0.3	0.5	0.7	0.9
CMU-MOSEI	MMIN [28]	ACL-IJCNLP'21	84.3 / 84.2 / 52.4	77.2 / 75.5 / 48.1	73.9 / 70.7 / 46.7	73.1 / 69.5 / 44.8	65.2 / 61.0 / 38.5†
	GCNet [29]	TPAMI'23	85.2 / 85.1 / 51.5	77.5 / 76.8 / 49.2	74.9 / 73.2 / 46.7	73.2 / 70.4 / 44.5	66.8 / 62.5 / 39.2†
	IMDer [30]	NeurIPS'23	85.1 / 85.1 / 53.4	81.3 / 80.7 / 51.3	79.0 / 77.4 / 49.2	77.3 / 74.6 / 47.6	70.5 / 67.2 / 41.5†
	SDR-GNN [31]	KBS'25	**85.9 / 85.8 / 53.9**	**82.1 / 81.5 / 51.8**	**80.2 / 78.6 / 49.8**	**78.5 / 75.8 / 48.2**	**72.1 / 69.0 / 42.8**†
	MD^{2}N [16]	ICCV'25	**86.1 / 85.8 / 54.1**	**82.4 / 81.6 / 52.0**	**81.2 / 78.8 / 50.1**	**79.4 / 77.6 / 48.6**	**73.4 / 70.8 / 43.1**†
	MD2E-MI	–	**86.6 / 86.2 / 54.5**	**82.6 / 82.2 / 52.4**	**81.2 / 80.1 / 50.8**	**80.4 / 78.8 / 49.5**	**77.1 / 75.2 / 46.8**
CMU-MOSI	MMIN [28]	ACL-IJCNLP'21	84.6 / 84.4 / 44.8	76.1 / 76.2 / 36.9	67.2 / 66.5 / 32.2	62.8 / 61.0 / 28.4	55.4 / 54.0 / 22.1†
	GCNet [29]	TPAMI'23	85.2 / 85.1 / 44.9	77.2 / 77.2 / 38.2	70.0 / 69.8 / 33.9	65.7 / 65.4 / 28.1	58.2 / 57.5 / 23.5†
	IMDer [30]	NeurIPS'23	85.7 / 85.6 / 45.3	81.2 / 81.0 / 42.5	76.2 / 75.9 / 37.9	71.9 / 71.2 / 33.4	64.5 / 63.8 / 28.6†
	SDR-GNN [31]	KBS'25	**86.2 / 86.0 / 45.8**	**82.0 / 81.8 / 43.1**	**77.5 / 76.8 / 38.5**	**73.0 / 72.1 / 34.0**	**66.2 / 65.5 / 29.5**†
	MD^{2}N [16]	ICCV'25	**86.4 / 86.1 / 46.0**	**82.3 / 82.0 / 43.9**	**77.6 / 77.0 / 39.4**	**73.8 / 72.4 / 34.3**	**68.0 / 67.2 / 31.5**†
	MD2E-MI	–	**87.2 / 86.6 / 46.7**	**83.2 / 83.0 / 44.5**	**79.8 / 79.2 / 41.5**	**76.5 / 75.8 / 37.2**	**71.8 / 70.5 / 33.0**

† Unified-protocol re-evaluation/adaptation under our R1 mask definition; not a published MR=0.9 number.

Train-test consistency. Unless stated otherwise, training and evaluation use the same missingness protocol and rate. All-missing timesteps are excluded from valid-timestep pooling and supervision.

Training details. All models are optimized with AdamW [27]. The default main setting uses one locked spherical backend for all manifold operators, an equidistant reverse schedule with $K_{\mathrm{rev}} = 20$, and fixed regularization weights. Stage A uses tangent-space Gaussian noise injection, Exp/Log/PT operators, and fixed-step reverse updates; it is not a fully intrinsic continuous-time Riemannian SDE solver.

Statistical reporting. The additional diagnostic experiments use the same data splits, missingness protocol, latent dimensionality, optimizer, training budget, and three random seeds as the main experiments. Compact main-text tables report three-seed means when space is limited; the factor of variability is the random seed, and supplementary tables record the corresponding run-level values and standard deviations. Appendix Tables 15 and 18 summarize the reporting and default training/evaluation contracts.

5.2 Baseline Source and Protocol Transparency

Appendix Table 16 separates values taken from published sources or official implementations from unified-protocol re-evaluations. Public random-missing reports for GCNet, IMDer, and SDR-GNN typically stop at MR≤ 0.7; therefore the MR=0.9 baseline entries in Table 1 are always treated as unified-protocol re-evaluations/adaptations, never as published numbers.

SDR-GNN's original masking procedure generates masks during train/validation/test and restricts the 3-modality mask rate to at most approximately 0.7. For MR=0.9 we align only the mask-generation and evaluation protocol while keeping the model, optimizer, and training schedule unchanged. Thus, the MR=0.9 SDR-GNN entry is an adapted unified-protocol re-evaluation, not a number quoted from the SDR-GNN paper. Under independent modality masking, jointly observed cross-modal interaction paths decrease rapidly as MR approaches 0.9, so the low-MR published trend is not a reliable extrapolation target.

5.3 Main Sentiment Results

Table 1 reports Acc-2, F1, and Acc-7 on CMU-MOSEI and CMU-MOSI across missing rates.

Appendix Figure 2 shows the robustness trends on sentiment and sensor benchmarks as the missing rate increases.

Table 2: Matched-protocol MD^{2}N comparison on CMU-MOSEI under high missingness (three-seed mean Acc-2/F1, %).

MR	Method	Acc-2	F1
0.70	MD^{2}N	79.4	77.6
0.70	MD2E-MI	80.4	78.8
0.75	MD^{2}N	78.8	76.9
0.75	MD2E-MI	80.3	78.6
0.80	MD^{2}N	77.8	75.6
0.80	MD2E-MI	80.2	78.4
0.85	MD^{2}N	75.9	73.5
0.85	MD2E-MI	78.8	76.9
0.90	MD^{2}N	73.4	70.8
0.90	MD2E-MI	77.1	75.2

Table 3: Ablation results on CMU-MOSEI at a missing rate of 0.8, averaged over three random seeds, with ECE reported where applicable.

Variant	Acc-2	F1	ECE
w/o imputation (zero-fill)	68.5	67.0	–
w/o manifold backend (Euclidean)	76.2	74.8	–
decoupled completion then fusion	**79.1**	**77.4**	–
w/o PDE/geo regularization	78.1	76.8	–
uniform fusion	77.5	76.1	0.109
entropy-gated fusion	**79.2**	**77.6**	**0.079**
attention fusion	78.8	77.3	**0.085**
Dirichlet evidence (full)	**80.2**	**78.4**	**0.058**

Table 4: Results on physical sensor datasets under full observation and R2 block failure Miss(0.8). Format: Acc / Macro-F1 (%), mean ± std over three seeds. The best, second-best, and third-best results are highlighted in bold red, blue, and green, respectively.

Method	Venue	UCI-HAR		PAMAP2	
		Full	Miss(0.8)	Full	Miss(0.8)
GRUD [32]	Sci. Rep.'18	93.5±0.5 / 93.2±0.6	74.2±2.0 / 71.5±2.1	88.5±0.7 / 86.2±0.8	62.5±2.0 / 58.0±2.1
CSDI [14]	NeurIPS'21	**95.2±0.4 / 94.8±0.5**	**81.5±1.8 / 79.2±1.9**	**91.2±0.6 / 89.5±0.7**	**72.8±1.7 / 69.5±1.8**
SDR-GNN [31]	KBS'25	**96.0±0.3 / 95.5±0.4**	**83.2±1.6 / 81.0±1.6**	**92.5±0.5 / 90.8±0.6**	**76.5±1.5 / 73.2±1.6**
MD2E-MI	–	**96.5±0.3 / 96.2±0.4**	**87.8±0.5 / 86.5±0.6**	**93.8±0.4 / 92.2±0.5**	**81.5±0.6 / 79.4±0.6**

5.4 Matched-Protocol MD^{2}N Comparison

MD^{2}N [16] is the most direct diffusion-based missing-modality comparator in our setting. In addition to the full-grid MD^{2}N row in Table 1, Table 2 focuses on the high-missingness CMU-MOSEI comparison using the same split, evaluator, missingness generator, and training budget as MD2E-MI. The comparison uses unified-protocol numbers rather than published MD^{2}N numbers under its original setting.

The margin increases from +1.0 Acc-2 and +1.2 F1 at MR=0.70 to +3.7 Acc-2 and +4.4 F1 at MR=0.90, consistent with the claim that reliability-coupled completion is most useful when reconstructed signals become more uncertain.

This comparison is a protocol-aligned reimplementation rather than a reuse of MD^{2}N's originally published setting. The main-text table uses means for readability; the corresponding supplement records standard deviations and per-seed values for both methods.

5.5 Sensor Block Failure Results

Under long block failures, MD2E-MI improves activity-recognition performance on both sensor datasets (Table 4). Figure 1 illustrates the recovered trajectory through a missing interval. These results evaluate the manifold-aware model as a whole; the contribution of simpler geometric constraints is examined separately in Section 5.9.

5.6 Ablations and Diagnostics

Table 3 collects component and diagnostic ablations on CMU-MOSEI at MR=0.8. All variants use matched splits and training budgets when applicable.

The decoupled variant trains completion first, freezes it, and then trains the same evidential predictor/fusion head. The full model improves by +1.1 Acc-2 and +1.0 F1, suggesting that end-to-end completion–reliability coupling contributes beyond architectural co-location. The uncertainty-surrogate variants show that the gain is not merely from adding any confidence signal: explicit Dirichlet evidence gives the best accuracy and calibration among the tested fusion choices. Appendix Figures 3 and 4 provide qualitative reliability and latent-space diagnostics.

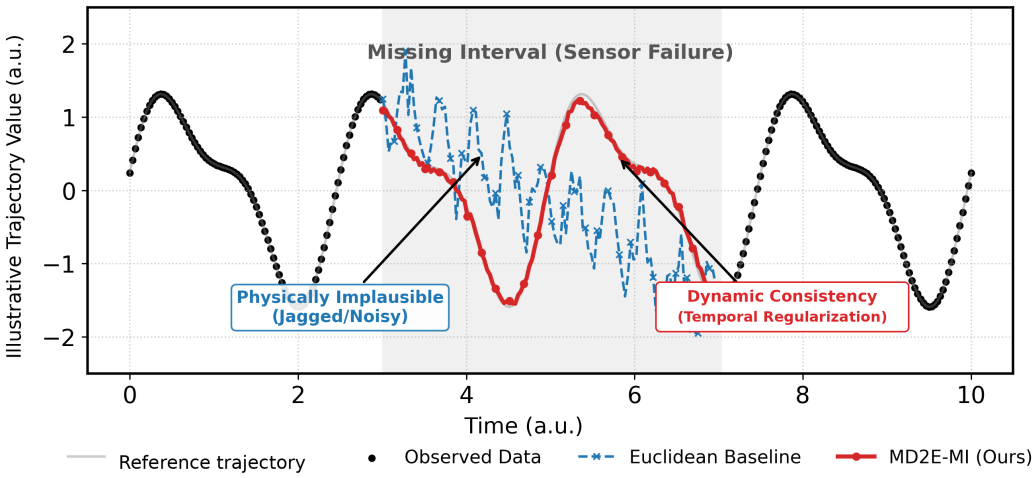


Figure 1: Trajectory completion under a long block-missing interval. The recovered trajectory is smoother during missing intervals while remaining consistent with the observed context before and after the block.

Table 5: Completion-free controls on CMU-MOSEI (Acc-2/F1, %). Random masking with matched three-seed protocols; not zero-fill ablations.

Setting	Obs.-only Attn.	Obs.-only Evidence	MD2E-MI
Full	**86.2 / 85.9**	**86.4 / 86.0**	**86.6 / 86.2**
MR=0.8	**78.8 / 77.2**	**79.2 / 77.6**	**80.2 / 78.4**
MR=0.9	**74.5 / 72.8**	**75.0 / 73.3**	**77.1 / 75.2**

5.7 Completion-Free Variable-Input Counterfactuals

A natural alternative to unreliable imputation is to omit completion entirely. *Observed-only Attention* consumes observed modality–time pairs with masks, modality/time embeddings and attention pooling. *Observed-only Evidence* is a stricter counterfactual retaining the same encoders, evidential predictor, fusion head and valid-timestep pooling, but excludes reverse diffusion and all imputed latents. Both share sample IDs, serialized masks, training/validation selection, and seeds with MD2E-MI; full results are in Appendix Table 8.

Full observation leaves a 0.2-point Acc-2 difference relative to Observed-only Evidence, while the difference reaches 2.1 Acc-2 and 1.9 F1 points at MR=0.9. Completion-free inference is thus competitive when sufficient observations remain; completion becomes more useful when the available subset is weak. Separately, the MD^{2}N versus MD2E-MI fixed-availability diagnostic (Appendix Table 13) shows smaller gains when text is available and larger gains when it is absent. That diagnostic is not an observed-only comparison.

5.8 Shared Reliability versus End-to-End and Post-hoc Alternatives

We also disentangle ordinary end-to-end learning from shared reliability. In an independently matched 90%/10% study, all backbones use the same fixed 90% D_{backbone}; only the post-hoc error head is fitted on a disjoint 10% D_{cal}, after freezing its backbone. Classification-only removes explicit confidence and completion regularizers; End-to-end No-Evidence keeps diffusion training but replaces evidence gating by matched attention. Full-90 jointly trains the shared evidential path on D_{backbone} and is independently retrained from scratch. Its reported mean happens to round to the same value as the full-training variant, but the training pools and seed-level outcomes differ. Test samples are not used to fit or select any variant.

Table 6: Reliability alternatives on CMU-MOSEI at MR=0.8 (matched 90%/10% protocol; three-seed means). Lower ECE is better.

Variant	Acc-2↑	F1↑	ECE↓
Classification-only	78.7	77.0	0.096
End-to-end No-Evidence	**79.1**	**77.5**	**0.086**
Post-hoc Reliability	**79.5**	**77.9**	**0.069**
Full-90 Shared Evidence	**80.2**	**78.4**	**0.058**

Table 7: Geometric constraints on CMU-MOSEI at MR=0.8 (Acc-2/F1, %) under the main full-training protocol.

Backend/constraint	Acc-2	F1
Plain Euclidean	76.2	74.8
+ Spectral normalization	78.1	76.7
+ L_2 trajectory	**78.6**	**77.1**
+ Unit-norm projection	**78.9**	**77.4**
Manifold-aware full	**80.2**	**78.4**

Full-90 improves Acc-2 by 0.7 ± 0.3 points across three paired seeds relative to post-hoc reliability estimation. At MR=0.9, the Acc-2 difference is 1.2 points; ECE was not evaluated at that rate (Appendix Table 9). These results favor shared evidence among the tested alternatives, although the margin over post-hoc reliability remains modest.

5.9 Manifold and Reverse-Step Sensitivity

Appendix Table 17 compares the Euclidean and manifold-aware backends across missing rates. We further evaluate spectral normalization, Euclidean L_2 trajectory regularization, and Euclidean unit-norm projection against the full manifold-aware model. Unit-norm projection recovers a substantial part of the difference from the plain Euclidean model, while the manifold-aware backend adds 1.3 Acc-2 points over unit-norm projection. The comparison evaluates the geometric components together rather than attributing the gain to a single operator. At MR=0.8, using $K_{\text{rev}} = 5, 10, 20$ gives Acc-2/F1 of 79.4/77.5, 79.8/78.0, and 80.2/78.4, respectively.

Reliability diagnostics. On CMU-MOSEI at MR=0.8, Spearman correlations between predicted confidence and latent completion error are -0.52 for text, -0.59 for audio, and -0.55 for vision; sample-clustered 95% intervals exclude zero. Confidence therefore tracks variation in completion error within each modality, although correlation alone does not establish calibration. Appendix Table 11 gives modality-specific errors, confidence scores, and intervals.

5.10 Compute Tradeoff

Compared with an equal-step Euclidean backend, the full manifold-aware model requires 24% more training time, about 15% more training memory, and 5.6 ms/sample additional inference time at $K_{\text{rev}} = 20$ under the same hardware and evaluation protocol. This overhead must be weighed against the accuracy gains, especially when missingness is mild or inference resources are limited.

6 Limitations

Our evaluation is limited to multimodal classification on four benchmarks; forecasting, control, and larger datasets with naturally occurring missingness remain untested. The high random-missing rates are controlled stress conditions, while block masking approximates sustained outages rather than their measured prevalence in deployment. The spherical backend is an inductive constraint on learned representations. Simpler unit-norm projection accounts for part of its benefit, and the full backend incurs additional computation. The improvement over post-hoc reliability estimation is

modest across three seeds, while confidence–error correlations measure association rather than causal reliability. High-missingness scores for MD^{2}N and SDR-GNN use protocol-aligned reimplementation or adaptation, as indicated in the comparison tables. Finally, the perturbation bound is local and may be uninformative for large completion errors.

7 Conclusion

We presented MD2E-MI, which uses a shared evidential predictor to connect latent diffusion completion with confidence-gated classification. Controlled comparisons show the clearest benefits when few reliable observations remain, while experiments against completion-free, post-hoc, and simple geometric alternatives clarify the contribution of each design choice. The additional accuracy under severe or persistent missingness comes with computational overhead, making the choice between completion-based and observed-only inference dependent on the operating conditions.

Acknowledgments and Disclosures

Funding. This work is supported by the National Key Research and Development Program of China (Grant No. 2025YFE0113400), General Project of Shenzhen Basic Research Special Natural Science Foundation (No. JCYJ20250604173228037), Special Projects in key areas of Guangdong Province (No.2022ZDZX1036), and Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011774).

Competing interests. The authors declare no competing interests.

References

[1] Brandon McKinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T Toshev. Robustness in multimodal learning under train-test modality mismatch. In *International Conference on Machine Learning*, pages 24291–24303. PMLR, 2023.

[2] SeungHyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*, pages 16654–16667. PMLR, 2023.

[3] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards Robust Multimodal Sentiment Analysis with Incomplete Data. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, 2024.

[4] Payal Mohapatra, Yueyuan Sui, Akash Pandey, Stephen Xia, and Qi Zhu. MAESTRO: Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series. In *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA (San Diego Convention Center) and Mexico City, MX (Hilton Reforma), Nov 30–Dec 7, 2025*, 2025.

[5] Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, and Xiaolu Hou. Toward Robust Incomplete Multimodal Sentiment Analysis via Hierarchical Representation Learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, volume 37, pages 28515–28536, 2024.

[6] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, 2025. doi: 10.1109/TPAMI.2024.3476487.

[7] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018. URL `https://arxiv.org/abs/1806.01768`.

[8] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/105112d52254f86d5854f3da734a52b4-Abstract-Conference.html.

[9] 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*, pages 22025–22034, 2023.

[10] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

[11] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019. doi: 10.1609/aaai.v33i01.33016892.

[12] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1802.05335.

[13] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1911.03393.

[14] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, Virtual Event, December 6–14, 2021*, pages 24804–24816, 2021.

[15] Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. MissDiff: Training Diffusion Models on Tabular Data with Missing Values. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=S435pkeAdT.

[16] Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517, 2025.

[17] Zhichao Chen, Haoxuan Li, Fangyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang. Rethinking the Diffusion Models for Missing Data Imputation: A Gradient Flow Perspective. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, 2024.

[18] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023.

[19] Zhipeng Xue, Yan Zhang, Ming Li, Chun Li, Yue Liu, and Fei Yu. Uncertainty Quantification for Incomplete Multi-View Data Using Divergence Measures. *IEEE Transactions on Image Processing*, 34(1):4328–4342, 2025.

[20] Yan Zhang, Ming Li, Chun Li, Zhaoxia Liu, Ye Zhang, and F. Yu. Uncertainty Quantification via Hölder Divergence for Multi-View Representation Learning. *IEEE Transactions on Multimedia*, 27(1):8263–8275, 2025.

[21] Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by flexible evidential deep learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=N6ujq5Yfwa.

[22] Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22953–22963, 2026.

[23] Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL `https://api.semanticscholar.org/CorpusID:51868869`.

[24] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis philippe Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31:82–88, 2016. URL `https://api.semanticscholar.org/CorpusID:1672698`.

[25] D. Anguita, Alessandro Ghio, L. Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition using Smartphones. In *The European Symposium on Artificial Neural Networks*, 2013. URL `https://api.semanticscholar.org/CorpusID:6975432`.

[26] A Reiss. UCI Machine Learning Repository: PAMAP2 Physical Activity Monitoring Data Set, 2016.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[28] 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2608–2618. Association for Computational Linguistics, 2021.

[29] 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.

[30] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[31] Fangze Fu, Wei Ai, Fan Yang, Yuntao Shou, Tao Meng, and Keqin Li. SDR-GNN: spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowledge-based Systems*, 309:112825, 2025.

[32] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

A Theoretical Analysis and Proof

We give the assumptions and proofs underlying the properties summarized in Section 4. The stability result relies on local smoothness and nondegeneracy assumptions.

Theorem 1 (Gated Dirichlet fusion gives a convex NLL upper bound)**.** *Under Assumption 1, let* $\alpha^{(\mathrm{fuse})} = \sum_m \pi_m \alpha^{(m)}$ *and* $p = \alpha^{(\mathrm{fuse})}/\alpha_0^{(\mathrm{fuse})}$*. Define*

$$\omega_m = \frac{\pi_m \alpha_0^{(m)}}{\sum_j \pi_j \alpha_0^{(j)}}. \tag{26}$$

Then $p_k = \sum_m \omega_m \mu_k^{(m)}$ *and*

$$-\log p_y \le \sum_m \omega_m (-\log \mu_y^{(m)}). \tag{27}$$

Theorem 2 (Confidence monotonicity)**.** *Under Assumption 1, for* $c^{(m)} = 1 - K/\alpha_0^{(m)}$*, confidence is strictly increasing in total evidence. Moreover,*

$$\frac{\pi_i}{\pi_j} = \exp\left(\lambda K \left(\frac{1}{\alpha_0^{(j)}} - \frac{1}{\alpha_0^{(i)}}\right)\right). \tag{28}$$

Theorem 3 (Completion-to-fusion stability)**.** *Under Assumptions 1 and 2, if* δ_m *bounds the local change in the full evidential input for modality* m *(including its cross-modal context), then*

$$\|p - p^\star\|_1 \le \frac{4}{a}\left(1 + \frac{\lambda K B}{2(K+\eta)^2}\right) \sum_{m=1}^{M} L_m \delta_m. \tag{29}$$

Assumption 1 (Valid evidential fusion)**.** *For each active modality,* $\alpha_k^{(m)} > 0$*,* $\alpha_0^{(m)} = \sum_{k=1}^K \alpha_k^{(m)}$*, and* $\mu_k^{(m)} = \alpha_k^{(m)}/\alpha_0^{(m)}$*. The fusion gates satisfy* $\pi_m \ge 0$ *and* $\sum_m \pi_m = 1$*. When confidence gating is used,* $\pi_m \propto \exp(\lambda c^{(m)})$ *with* $\lambda > 0$*.*

Assumption 2 (Local conditions for the stability bound)**.** *For the local comparison used in the stability statement, suppose there exist finite constants* $a, B, \{L_m\}_{m=1}^M$ *and* $\eta > 0$*. In the neighborhood considered, the predicted and reference fused total evidence are at least* a*, the reference modality evidence has* ℓ_1 *norm at most* B*, and each modality total evidence stays above* $K + \eta$*. Let* $z^{(m)} = (s^{(m)}, C^{(m)})$ *denote the full input to the* m*-th evidential predictor, which is assumed locally* L_m*-Lipschitz in a fixed norm on this input. The discrepancy* $\delta_m = \|z^{(m)} - z^{(m)\star}\|$ *therefore includes changes propagated through cross-modal context, not only changes in the* m*-th latent.*

Assumption 2 is a local nondegeneracy condition for the perturbation bound only. It is not a global assumption on the data distribution or on all possible network inputs; it is used to interpret how small completion errors can affect fused predictions near the evaluated latents.

A.1 Proofs

Proof of the convex NLL upper bound. By definition,

$$p_k = \frac{\sum_m \pi_m \alpha_k^{(m)}}{\sum_m \pi_m \alpha_0^{(m)}} = \sum_m \frac{\pi_m \alpha_0^{(m)}}{\sum_j \pi_j \alpha_0^{(j)}} \mu_k^{(m)} = \sum_m \omega_m \mu_k^{(m)}. \tag{30}$$

Since $-\log(\cdot)$ is convex, Jensen's inequality gives

$$-\log p_y = -\log\left(\sum_m \omega_m \mu_y^{(m)}\right) \le \sum_m \omega_m (-\log \mu_y^{(m)}). \tag{31}$$

□

Proof of confidence monotonicity. For $c(\alpha_0) = 1 - K/\alpha_0$,

$$\frac{dc}{d\alpha_0} = \frac{K}{\alpha_0^2} > 0. \tag{32}$$

Since $\pi_m \propto \exp(\lambda c^{(m)})$,

$$\frac{\pi_i}{\pi_j} = \exp(\lambda(c^{(i)} - c^{(j)})) = \exp\left(\lambda K\left(\frac{1}{\alpha_0^{(j)}} - \frac{1}{\alpha_0^{(i)}}\right)\right). \tag{33}$$

□

Proof of completion-to-fusion stability. Let $x = \alpha^{(\mathrm{fuse})}$ and $x^\star = \alpha^{(\mathrm{fuse})\star}$. By Assumption 2, the total fused evidence is bounded away from zero. Therefore ℓ_1 normalization is locally Lipschitz, so

$$\|p - p^\star\|_1 \le \frac{4}{a}\|x - x^\star\|_1. \tag{34}$$

Expanding the fused evidence difference gives a completion perturbation term and a gating perturbation term:

$$\|x - x^\star\|_1 \le \sum_m \pi_m \|\alpha^{(m)} - \alpha^{(m)\star}\|_1 + B\|\pi - \pi^\star\|_1. \tag{35}$$

Assumption 2 bounds the first term by $\sum_m \pi_m L_m \delta_m$, with context changes included in δ_m. Softmax Lipschitzness and the confidence Lipschitz bound for $\alpha_0 \ge K + \eta$ imply

$$\|\pi - \pi^\star\|_1 \le \frac{\lambda K}{2(K+\eta)^2} \sum_m L_m \delta_m. \tag{36}$$

Combining the inequalities yields the stated result. □

A.2 Additional Reporting Items

This appendix records additional reproducibility details: baseline source/protocol separation, MR definition, SDR-GNN high-MR adaptation, MD^{2}N matched-protocol comparison, the shared evidential module, the fixed spherical backend, divergence definitions, and compute sensitivity. The reported diagnostic tables use the same splits, missingness protocol, latent dimensionality, optimizer, training budget, and three random seeds as the main experiments; when standard deviations are not shown in compact tables, supplementary tables record the corresponding run-level values.

B Additional Controlled Experiments and Diagnostics

Table 8: Strictly completion-free controls versus full MD2E-MI. MOSEI: Acc-2/F1; sensor datasets: Acc/Macro-F1 (%). Values are mean ± standard deviation.

Setting	Observed-only Attention	Observed-only Evidence	Full
MOSEI, full	**86.2±0.4 / 85.9±0.4**	**86.4±0.4 / 86.0±0.4**	**86.6±0.4 / 86.2±0.4**
MOSEI, MR=0.3	**82.3±0.4 / 81.9±0.5**	**82.4±0.4 / 82.0±0.5**	**82.6±0.4 / 82.2±0.5**
MOSEI, MR=0.8	**78.8±0.5 / 77.2±0.6**	**79.2±0.5 / 77.6±0.6**	**80.2±0.4 / 78.4±0.7**
MOSEI, MR=0.9	**74.5±0.7 / 72.8±0.8**	**75.0±0.7 / 73.3±0.8**	**77.1±0.5 / 75.2±0.6**
UCI-HAR, block 0.8	**84.0±0.9 / 82.2±1.0**	**84.7±0.8 / 83.0±0.9**	**87.8±0.5 / 86.5±0.6**
PAMAP2, block 0.8	**77.4±1.0 / 74.3±1.1**	**78.2±0.9 / 75.2±1.0**	**81.5±0.6 / 79.4±0.6**

Table 9: Reliability alternatives on CMU-MOSEI under a matched 90%/10% training/calibration protocol. ECE is available only at MR=0.8.

Variant	MR=0.8			MR=0.9	
	Acc-2	F1	ECE	Acc-2	F1
Classification-only	78.7	77.0	0.096	74.9	72.9
End-to-end No-Evidence	**79.1**	**77.5**	**0.086**	**75.4**	**73.5**
Post-hoc Reliability	**79.5**	**77.9**	**0.069**	**75.9**	**74.1**
Full-90 Shared Evidence	**80.2**	**78.4**	**0.058**	**77.1**	**75.2**

Table 10: Additional geometric regularization controls on CMU-MOSEI under the main full-training protocol.

Backend	MR=0.3 Acc-2	MR=0.8 Acc-2/F1	MR=0.9 Acc-2
Plain Euclidean	81.9	76.2 / 74.8	73.8
+ Spectral norm	82.2	78.1 / 76.7	75.1
+ L_2 trajectory	**82.3**	**78.6 / 77.1**	**75.6**
+ Unit-norm projection	**82.4**	**78.9 / 77.4**	**75.8**
Manifold-aware full	**82.6**	**80.2 / 78.4**	**77.1**

Table 11: Modality-level imputation error and evidence confidence on CMU-MOSEI at MR=0.8 (diagnostic masked pairs; within-modality Spearman correlation).

Modality	Normalized MSE	Confidence	Spearman ρ	Clustered 95% CI
Text	0.44 ± 0.03	0.69 ± 0.02	-0.52	$[-0.58,-0.46]$
Audio	0.78 ± 0.05	0.46 ± 0.03	-0.59	$[-0.65,-0.52]$
Vision	0.71 ± 0.04	0.51 ± 0.03	-0.55	$[-0.61,-0.48]$

Gate temperature. At CMU-MOSEI MR=0.8, validation-only selection over $\lambda \in \{0, 1, 2, 5, 10\}$ gave validation F1 of 76.2, 77.4, 78.0, 78.5, and 77.9, respectively; $\lambda = 5$ was selected. This is the softmax-gating temperature, distinct from the diffusion-loss coefficient λ_{diff}. Large temperatures concentrate on the largest confidence but may amplify near-tie misordering. These validation observations do not establish a causal explanation.

Local sensitivity diagnostics. At MR=0.8, held-out MOSEI finite differences gave $a = K = 7$, $\hat{B} = 34.2$, $\hat{\eta} = 0.42$, and $\hat{L}_{\text{text/audio/vision}} = 2.31/2.74/2.58$. These are finite-sample diagnostics rather than certified global constants; the corresponding theoretical bound may be uninformative for larger perturbations.

Protocols and scope. In the observed-only matched runs, masks are fixed before fitting and shared across methods for each split/rate/seed. The post-hoc comparison is a distinct matched 90%/10% study; its results must not be interpreted as a single training-pool head-to-head against full-training variants. No naturally occurring missingness rate is inferred from these synthetic random or block tests.

Table 12: Run-level high-missingness comparison with MD^{2}N on CMU-MOSEI. Format: Acc-2 / F1 (%). All entries use the unified R1 random-dropout protocol, the same split, evaluator, missingness generator, validation-selection rule, and seeds. The best and second-best results are highlighted in bold red and blue, respectively.

MR	Method	seed 1	seed 42	seed 123	Mean ± std
0.70	MD^{2}N	**79.1 / 77.1**	**79.5 / 77.3**	**79.6 / 78.4**	79.4±0.3 / 77.6±0.7
0.70	**MD2E-MI**	**80.0 / 78.3**	**80.5 / 78.5**	**80.7 / 79.6**	80.4±0.4 / 78.8±0.7
0.75	MD^{2}N	**78.3 / 76.3**	**78.6 / 76.6**	**79.5 / 77.8**	78.8±0.6 / 76.9±0.8
0.75	**MD2E-MI**	**79.9 / 78.1**	**80.1 / 78.8**	**80.9 / 78.9**	80.3±0.5 / 78.6±0.4
0.80	MD^{2}N	**77.2 / 74.9**	**78.0 / 75.3**	**78.2 / 76.6**	77.8±0.5 / 75.6±0.9
0.80	**MD2E-MI**	**79.8 / 77.9**	**80.5 / 78.1**	**80.3 / 79.2**	80.2±0.4 / 78.4±0.7
0.85	MD^{2}N	**75.2 / 72.8**	**75.4 / 73.7**	**77.1 / 74.0**	75.9±1.0 / 73.5±0.6
0.85	**MD2E-MI**	**78.2 / 76.3**	**78.5 / 76.5**	**79.7 / 77.9**	78.8±0.8 / 76.9±0.9
0.90	MD^{2}N	**72.6 / 70.0**	**73.6 / 70.4**	**74.0 / 72.0**	73.4±0.7 / 70.8±1.1
0.90	**MD2E-MI**	**76.5 / 74.5**	**77.3 / 75.4**	**77.5 / 75.7**	77.1±0.5 / 75.2±0.6

Table 13: CMU-MOSEI modality-wise missingness diagnostics. This fixed-availability diagnostic complements the random R1 dropout protocol by controlling which modalities are available at evaluation time. T, A, and V denote text, audio, and vision. Format: Acc-2 / F1 / Acc-7 (%). The better method is highlighted in bold red, the baseline is highlighted in bold blue, and the gain over the baseline is highlighted in bold green.

Diagnostic condition	Available modalities	MD^{2}N	MD2E-MI	Gain over MD^{2}N
Full observation	T+A+V	**86.1 / 85.8 / 54.1**	**86.6 / 86.2 / 54.5**	**+0.5 / +0.4 / +0.4**
Vision missing only	T+A	**83.6 / 82.9 / 51.3**	**85.0 / 84.5 / 52.5**	**+1.4 / +1.6 / +1.2**
Audio missing only	T+V	**83.3 / 82.7 / 51.0**	**84.8 / 84.4 / 52.1**	**+1.5 / +1.7 / +1.1**
Text missing only	A+V	**73.2 / 71.0 / 43.1**	**76.1 / 74.0 / 44.8**	**+2.9 / +3.0 / +1.7**
Text only	T	**81.6 / 80.7 / 49.1**	**83.0 / 82.3 / 50.0**	**+1.4 / +1.6 / +0.9**
Audio only	A	**68.8 / 66.2 / 40.1**	**71.0 / 68.8 / 41.6**	**+2.2 / +2.6 / +1.5**
Vision only	V	**69.6 / 67.4 / 41.0**	**72.0 / 70.1 / 42.9**	**+2.4 / +2.7 / +1.9**

B.1 Supplementary Quantitative Tables

Table 14: Protocol details for MD^{2}N and SDR-GNN comparisons.

Item	Setting
MD^{2}N comparison type	Protocol-aligned reimplementation, not reuse of original published numbers.
MD^{2}N evaluation	Same CMU-MOSEI split, evaluator, R1 missingness generator, validation F1 selection rule, and seed list as MD2E-MI.
MD^{2}N high-MR range	Re-evaluated under R1 at MR=0.70, 0.75, 0.80, 0.85, and 0.90.
MD^{2}N tuning rule	Validation-selected under the same metric/evaluator; no test-set tuning.
SDR-GNN public coverage	Public random-missing reports typically cover up to approximately MR$\leq$ 0.7 in the released 3-modality masking setting.
SDR-GNN MR=0.9 source	Adapted unified-protocol re-evaluation.
SDR-GNN changed components	Mask generation and evaluation protocol only.
SDR-GNN unchanged components	Model architecture, optimizer, and training schedule.

Table 15: Default reporting contract for main diagnostic comparisons.

Comparison	Split/protocol	Seeds	Main-text reporting
Sentiment baselines	unified R1 where re-evaluated	3	mean metrics
MD^{2}N matched protocol	same split/evaluator/budget	3	mean metrics
Ablation variants	same split/budget	3	mean metrics, ECE where defined
Sensor block failure	unified R2	3	mean ± std
Reverse-step sensitivity	same MOSEI MR=0.8 setting	3	mean metrics

B.2 Original Baseline and Trajectory Diagnostics

Table 16: Baseline source and protocol summary. MR=0.9 baseline entries not available in prior publications are explicitly treated as unified-protocol re-evaluations/adaptations rather than published numbers.

Method	Public random-MR coverage	MR=0.9 entry	Notes
MMIN [28]	up to reported setting	unified-protocol re-eval	same split and metrics
GCNet [29]	MR$\leq$ 0.7	unified-protocol re-eval	same split and metrics
IMDer [30]	MR$\leq$ 0.7	unified-protocol re-eval	same split and metrics
SDR-GNN [31]	MR$\leq$ 0.7	unified-protocol adaptation	mask/evaluator aligned; model unchanged
MD^{2}N [16]	original setting differs	protocol-aligned reimplementation	matched split, budget, and evaluator

Table 17: Manifold-aware versus equal-step Euclidean completion on CMU-MOSEI (Acc-2, %).

Variant	MR=0.3	MR=0.5	MR=0.8	MR=0.9
Euclidean	**81.9**	**79.6**	**76.2**	**73.8**
Manifold-aware	**82.6**	**81.2**	**80.2**	**77.1**

B.3 Qualitative Diagnostics

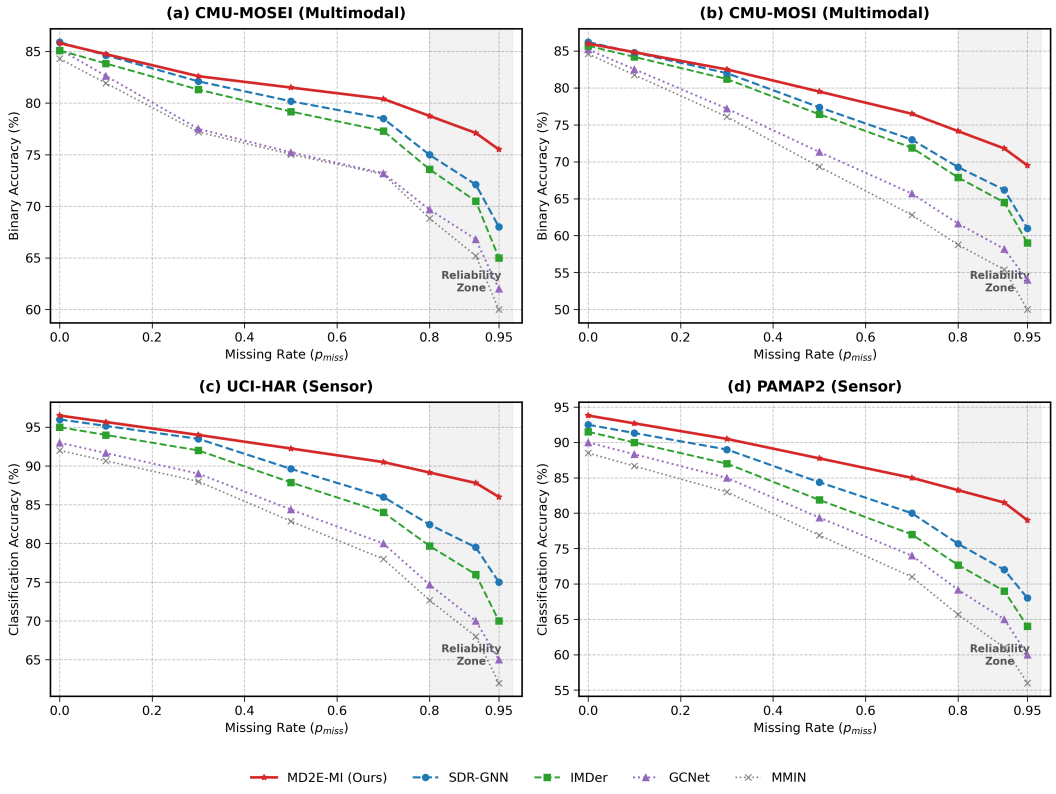


Figure 2: Robustness trends under increasing missingness on sentiment and sensor benchmarks. The curves show that performance differences are small under mild missingness and become more pronounced as the missing rate increases.

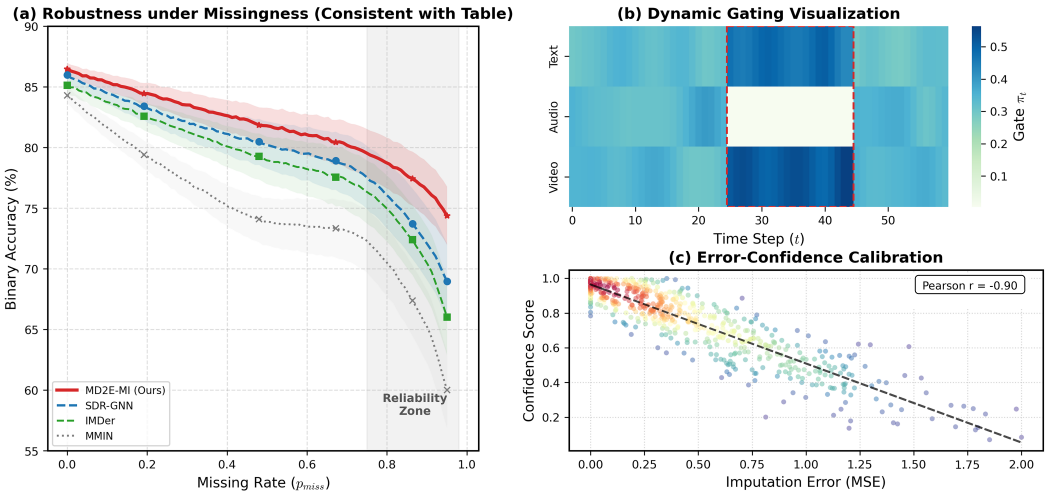


Figure 3: Reliability diagnostics on CMU-MOSEI. Evidence-derived confidence varies across modality–time pairs and is lower for less reliable completed inputs, supporting its use as a gating signal rather than a post-hoc score only.

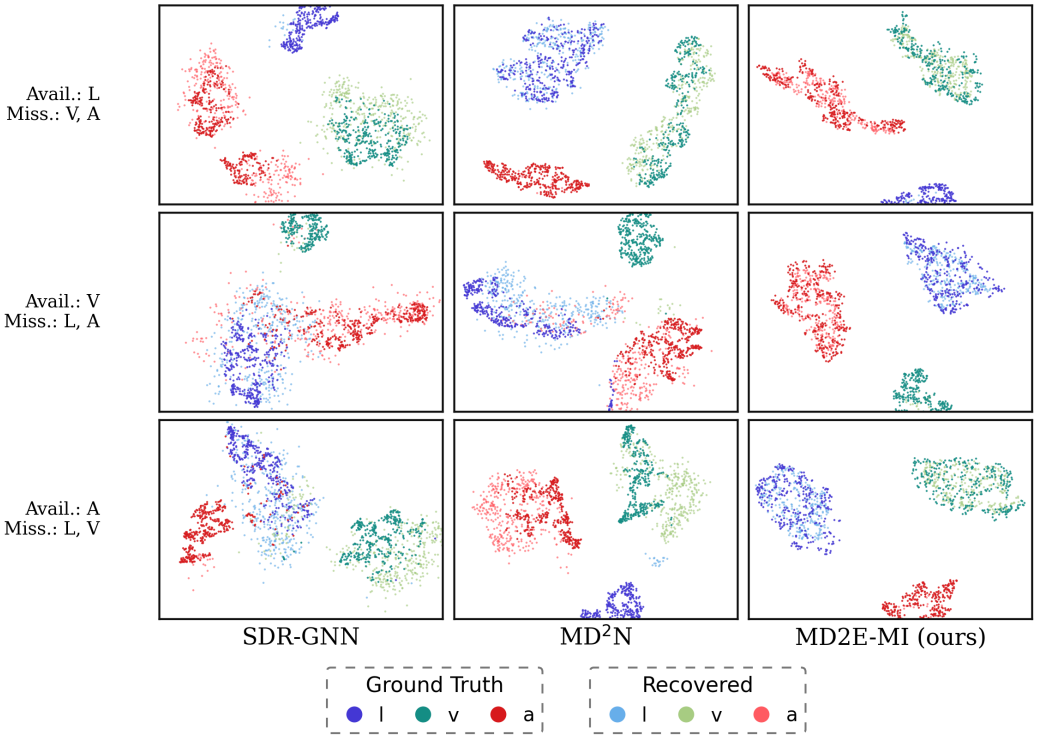


Figure 4: Latent representation visualization on CMU-MOSEI. The visualization provides a qualitative check that the completed latent space remains class-structured under severe missingness.

Table 18: Fixed default training and evaluation choices used unless explicitly ablated.

Item	Default setting
Optimizer	AdamW
Learning rate / weight decay	3×10^{-4} / 10^{-2}
Input feature dimensions	text 300, audio 74, visual 748
Encoder output dimensions	text 256, audio 128, visual 256
Classifier/shared hidden dimension	128
Residual projection hidden dimension	256
Maximum sequence length	80
Manifold backend	shared spherical backend
Geometric operators	fixed Exp, Log, PT, $d_{\mathcal{M}}$ backend
Reverse steps	$K_{\text{rev}} = 20$ with equidistant schedule
Diffusion schedule	50 diffusion steps with fixed schedule
Validation selection	best validation F1 checkpoint
Random seeds	1, 42, 123
Hardware	NVIDIA RTX 4090
Baseline tuning rule	same split/evaluator/missingness generator; validation-selected; no test-set tuning
Missing-pair handling	mandatory reverse-diffusion completion before fusion
Pooling	sample-wise valid-timestep union T_{valid}

Table 19: Default loss and divergence parameters. The static graph-geometric loss weight is disabled by default; the active temporal geometric terms are reported separately as $\lambda_{\text{geo_time}}$ and $\lambda_{\text{geo_acc}}$.

Item	Default setting
λ_{sup}	1.0
λ_{diff}	1.0
λ_{PDE}	0.1
$\lambda_{\text{geo_time}}$	0.1
$\lambda_{\text{geo_acc}}$	0.05
Static graph-geometric weight	0.0
D_{KL}	Dirichlet mean vs. uniform reference; no extra parameter
D_{SM}	Sharma–Mittal divergence against uniform reference
ρ / δ for D_{SM}	1.5 / 0.5
λ_{SM} / λ_{KL}	1.0 / 1.0
γ / ϵ_{conf} / $\sigma_{\min}$	1.0 / 10^{-3} / 10^{-3}

C Impact Statement

This work introduces MD2E-MI, a reliability-coupled framework for multimodal time-series classification when modality–time observations are missing. **Positive impacts.** By coupling latent completion with evidential reliability estimation, MD2E-MI can make multimodal systems less brittle under random dropout, long sensor failures, or incomplete text/audio/vision streams. This may benefit benign settings such as assistive sensing, human–computer interaction, multimodal sentiment analysis, activity recognition, and fault-tolerant data analysis, where one or more modalities may be unavailable at test time. The explicit confidence gates may also help practitioners audit which completed modalities are trusted, reducing the risk of treating imputed features as fully reliable measurements. **Potential negative impacts.** More robust inference under missing modalities could also make sentiment or activity-recognition systems easier to deploy in monitoring, profiling, workplace analytics, or surveillance contexts where some signals are absent, degraded, or intentionally withheld. If used in consequential settings, wrong predictions or miscalibrated confidence under severe missingness may amplify privacy, fairness, or safety harms. MD2E-MI does not introduce a new data-collection mechanism or a new generative content model, but it can lower the practical barrier to using incomplete multimodal data. We therefore recommend that downstream users apply the method only with informed consent, appropriate data governance, privacy safeguards, task-specific fairness and calibration audits, and human oversight; it should not be used as the sole basis for high-stakes decisions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as missing-modality multimodal sequence classification and match the reported theoretical, experimental, ablation, and diagnostic results. The claims are further qualified by the baseline-protocol discussion and the limitations section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated limitations section discussing task scope, manifold-backend interpretation, protocol-aligned baseline re-evaluation, MD2N reimplementation fidelity, compute overhead, and per-modality diagnostics.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical statements are given with explicit assumptions in the method/theory section, and complete proofs are provided in the appendix for the NLL bound, monotone evidence gating, and stability result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experiments describe datasets, R1/R2 missingness protocols, train-test consistency, optimizer, reverse-step setting, seeds, baseline source separation, MD2N matched-protocol comparison, SDR-GNN adaptation, and supplementary implementation contracts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in the paper are publicly available, and the submission materials provide implementation details, missingness protocols, hyperparameters, seeds, and reproduction instructions for the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments and appendix specify the data splits/protocol alignment, R1/R2 masks, AdamW optimizer, learning rate and weight decay, feature dimensions, hidden dimensions, reverse steps, diffusion schedule, validation-selection rule, seeds, and baseline tuning rule.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports three-seed means in compact main-text tables and gives run-level values with mean $\pm$ standard deviation in supplementary tables where space permits; the factor of variability is the random seed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the NVIDIA RTX 4090 hardware used for the experiments and provides compute tradeoff information, including relative training-time overhead, memory overhead, and inference-time overhead under the main evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses public benchmark datasets, reports matched protocols and uncertainty/statistical information, and does not involve deception, private data collection, deployment decisions, or high-risk model release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement discusses positive impacts for robust multimodal analysis under missing data and potential negative impacts involving monitoring, profiling, surveillance, privacy, fairness, and high-stakes use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk pretrained language model, image generator, scraped dataset, or other asset that would require special misuse safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The manuscript cites the original datasets and prior methods, uses publicly available benchmark assets, and follows their stated licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset or pretrained model asset; if an anonymized code package is added later, this answer should be changed to [Yes] with documentation details.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper uses existing public benchmark datasets and does not conduct new crowdsourcing experiments or new human-subject studies.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subject data collection or intervention is performed; the study only evaluates on existing public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method, experiments, and scientific claims do not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.