MEASURING CROSS-CONDITION GENERALIZATION UNDER MISSING MODALITIES: A CLOSED-FORM BOUND AND A CAUTIONARY CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing-modality robustness methods are typically evaluated at a single, matched train/test missing rate - a protocol that cannot distinguish a model that has genuinely learned to generalize across missingness patterns from one that has merely overfit to the pattern it was shown. We derive a closed-form, optimal-transport-solver-free upper bound on the risk gap between any two missing-rate conditions, exploiting a structural fact specific to the missing-modality setting: the label is invariant under masking, so no irreducible "distortion" term is required, unlike general conditional-distribution-shift bounds for generative models. We verify the transport inequality numerically on synthetic data and use its closed-form geometry to define descriptive sensitivity diagnostics on real trained multimodal networks. These diagnostics support a lightweight, pre-registered evaluation protocol for cross-condition robustness claims. Applying this protocol to a recent missing-modality defense (BALM) across two backbones and two datasets, and to a second, structurally different method (AMB-DSGDN), we find a heterogeneous pattern that no single evaluation setup reveals: (a) a real, previously undocumented artifact in BALM's evaluation masking (a positional correlation with the fixed test-set ordering); (b) on one backbone, BALM's central sensitivity-reduction claim appears confirmed at a small sample size but does not survive a properly powered, pre-registered multi-seed evaluation, and a second dataset shows the same null pattern; (c) on a second backbone, a distinct, slope-independent absolute-risk-reduction effect looks large and unanimous at 3 seeds, then at the full pre-registered 6-seed batch is significant by paired t-test but not by Wilcoxon - the same failure mode caught twice on two different axes; and (d) a structurally unrelated method (adaptive modality dropout) improves both slope and intercept together. No accuracy-at-a-few-rates table, the field's current norm, can distinguish these effects, and no single backbone/dataset/method combination reveals the full picture. We argue this decomposition, and the multi-seed, multi-axis protocol needed to see it, is a general need for the field, and offer both as a contribution independent of any specific verdict on any one method.

1 INTRODUCTION

Multimodal systems routinely lose inputs at test time: a microphone fails, a face leaves the frame, a transcript is unavailable. Nearly every recent multimodal emotion recognition (MER) method therefore reports robustness to missing modalities, usually as accuracy or weighted-F1 at a handful of missing-rate configurations, with the training and test rates matched at each point. Such a table measures how well a model fits the missingness pattern it was trained on. It does not measure how the model behaves when the test-time pattern differs from the training-time one, and reported results routinely conflate the two.

We ask two questions: for a fixed trained model, how much can risk change when the missing rate shifts from r to r', and how should a claim of improved cross-condition robustness be tested? For conditional generative models, bounds of this kind carry an irreducible distortion term, because different conditions define different target distributions (Section 2). Missing modalities are structurally simpler. Masking changes what is observed, never the label, so every rate r is a partial view of the

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

same draw $(x, y) \sim \mathcal{D}$. This gives a bound with no distortion term, whose transport term is available in closed form from per-modality second moments without an optimal transport solver.

The bound splits cross-condition risk into two parts: the risk at a reference rate (an *intercept*) and the growth of risk with the closed-form distance (a *slope*). We use this split to build an evaluation protocol and apply it to BALM (Nguyen et al., 2026), a recent method for imbalanced missing rates, on two backbones and two datasets, and to a structurally different method, AMB-DSGDN (Wang et al., 2026). The result is heterogeneous (Table 2). In the masking code used for BALM's evaluation, which utterances lose a modality depends on their position in the test set. On MMGCN, BALM's slope reduction is unanimous in a first batch of three seeds but does not survive corrected text features and a pre-registered six-seed evaluation. On MM-DFN, BALM lowers absolute risk instead, and whether that effect is significant depends on the test. AMB-DSGDN's adaptive modality dropout lowers slope and intercept together. None of these effects can be read off a matched-rate accuracy table; on MM-DFN, that table shows no difference at all (Table 2).

Our contributions are:

1. **A closed-form bound.** An OT-solver-free upper bound (Theorem 1) on the fixed-model risk gap between any two missing-rate conditions. Its transport term comes from a two-sided Wasserstein sandwich (Lemma 1) whose upper side is sharp: no bound that uses only per-modality second moments can be smaller (Proposition 1). We also verify it numerically on synthetic data (Appendix B.3).
2. **Descriptive diagnostics.** Slope and empirical-envelope statistics built on the same closed-form distance, which transfer to held-out seeds (Sections 4.1–5.2). They are not estimates of the theorem's global Lipschitz constant, and we do not use them as such.
3. **An evaluation protocol.** I.i.d. MCAR re-masking, an embedding-collapse check, a positional-MNAR check, and a pre-registered multi-seed decision rule (Section 6), which together caught failure modes that a matched-rate, single-seed evaluation cannot.
4. **A case study.** An application of the protocol to BALM across two backbones and two datasets, and to AMB-DSGDN, that surfaces a masking artifact and separates slope from intercept effects (Section 7).

2 Related Work

Missing-modality robustness. We reproduce BALM (Nguyen et al., 2026) on the MMGCN (Hu et al., 2021) and MM-DFN (Hu et al., 2022) backbones for MER. BALM belongs to a family of training-time mechanisms that rebalance per-modality features or gradients. DERL (Li et al., 2026) disentangles inputs into orthogonal experts with reconstruction supervision; ModalImmune (Fu et al., 2026) deliberately collapses modality information during training to inoculate against test-time loss; GMD/DS (Wang et al., 2024) decouples modality gradients; HARDY-MER (Liu et al., 2025) curriculum-weights hard samples; and CalMRL (Liu et al., 2026) calibrates representation-level anchor shift. AMB-DSGDN (Wang et al., 2026), our second method, instead learns a per-modality dropout probability through differential graph attention. All of these are robustness *mechanisms*. Our protocol does not depend on the mechanism and applies to any of them.

Evaluating robustness under missing modalities. MissBench (Pham et al., 2026), from the authors of BALM, standardizes shared and imbalanced missing-rate protocols over four datasets and introduces an inference-time equity metric (MEI) and a training-time gradient-imbalance metric (MLI). MissMAC-Bench (Lin et al., 2026) standardizes fixed and random missingness and summarizes models by Competence/Resilience ellipses, which are close in spirit to our intercept/slope split. Resilience, however, is a spread over conditions rather than a trend fitted against a distance between them. CIDer (Zhong et al., 2025) extends evaluation to random missingness at the feature level; our derivation covers that case by defining B_m per feature or token subset. None of these benchmarks bounds the risk gap between two conditions or supports extrapolation to an unseen rate. Outside MER, MMG-Ego4D (Gong et al., 2023) finds dropout-based fusion most robust when a modality seen in training is absent at test time; whether our diagnostics transfer there is untested.

OOD bounds for conditional generative models. Pham (2026) bounds out-of-distribution error in conditional flow matching by three terms: Interpolation Sparsity, Geometric Distortion, and In-

Distribution Fit. The distortion term is needed there because different conditions imply different target distributions with no guaranteed isometric embedding. Our setting removes it: the label is invariant under masking, so $(x, y) \sim \mathcal{D}$ is the same for every r (Appendix A.4). Classical domain-adaptation bounds (Ben-David et al., 2010) estimate a divergence between domains from samples. Here the conditions are related by a known masking transformation, so the population transport term is analytic, although its plug-in version on real data still carries moment and risk-estimation uncertainty.

Missing-data mechanisms. The MCAR/MAR/MNAR taxonomy (Rubin, 1976) frames Assumption 1. The positional artifact in Section 7.4 is a concrete MNAR mechanism created by evaluation code rather than by the data-generating process, and it differs from the usual informative-dropout story. ICYM2I (Choi et al., 2026) corrects MAR bias by inverse-probability weighting when the mechanism is unknown, which would complement our analysis if the MCAR assumption were relaxed.

3 Problem Setup

Let $\mathcal{X} := \prod_{m=1}^{M} \mathbb{R}^{d_m}$ be equipped with the block Euclidean norm $\|z\|^2 := \sum_{m=1}^{M} \|z^m\|_2^2$, and let $(X, Y) \sim \mathcal{D}$ with $X = (X^1, \ldots, X^M) \in \mathcal{X}$. For a missing-rate vector $r = (r_1, \ldots, r_M) \in [0, 1)^M$, let

$$E_r^m \sim \text{Bernoulli}(1 - r_m), \qquad \widetilde{X}_r^m := E_r^m X^m,$$

where the masks are independent across modalities and independent of (X, Y). Denote the feature marginal by $\nu_r := \text{Law}(\widetilde{X}_r)$ and the joint law by $P_r := \text{Law}(\widetilde{X}_r, Y)$. For a fixed parameter vector θ and per-example loss $g_\theta(z, y) := \ell(h_\theta(z), y)$ satisfying $\mathbb{E}|g_\theta(0, Y)| < \infty$, define the population risk

$$R(r; \theta) := \mathbb{E}_{P_r}[g_\theta(\widetilde{X}_r, Y)].$$

Goal. For one fixed model θ, control the change in its population risk between masking rates: given reference rates $r_1, \ldots, r_n$, bound $R(r^*; \theta)$ at a new rate r^*.

Assumption 1 (MCAR masking)**.** *For every evaluated rate r, the coordinates $E_r^1, \ldots, E_r^M$ are mutually independent and independent of (X, Y), and masking changes only the observed feature vector, not the label Y. This holds by construction for our i.i.d. evaluation masker; it remains an assumption for naturally occurring missingness.*

Assumption 2 (Global pointwise Lipschitz loss)**.** *For the fixed θ under consideration,*

$$L_\theta := \sup_{y \in \text{supp}(Y)} \sup_{\substack{z, z' \in \mathcal{X} \\ z \neq z'}} \frac{|g_\theta(z, y) - g_\theta(z', y)|}{\|z - z'\|} < \infty.$$

The theorem may equivalently use any certified upper bound on L_θ. A regression slope or a finite collection of observed risk–distance ratios is not such an upper bound without an additional certification argument.

Assumption 3 (Finite population second moments)**.** *For every modality,*

$$B_m^2 := \mathbb{E}_{\mathcal{D}}\|X^m\|_2^2 < \infty, \qquad \mu_m := \mathbb{E}_{\mathcal{D}}[X^m].$$

These are population moments; their sample analogues are plug-in estimates.

4 Theory

Full proofs are given in Appendix A.

Lemma 1 (Feature-marginal Wasserstein sandwich)**.** *Let W_p denote the p-Wasserstein distance on $\mathcal{X}$ with the Euclidean ground metric defined above. Under Assumptions 1 and 3, for any $r, r' \in [0, 1)^M$,*

$$\sqrt{\sum_{m=1}^{M} \|\mu_m\|_2^2 (r_m - r'_m)^2} \;\le\; W_1(\nu_r, \nu_{r'}) \;\le\; W_2(\nu_r, \nu_{r'}) \;\le\; d(r, r') := \sqrt{\sum_{m=1}^{M} B_m^2 |r_m - r'_m|}.$$

The distances are between feature marginals, so no metric on the label space is needed. The squared rate difference on the left is a mean-shift norm; the absolute rate difference on the right is the disagreement probability of a maximal Bernoulli coupling. The lemma is a population statement: substituting sample moments makes it computable without an OT solver, but certifying a population bound additionally requires simultaneous one-sided confidence bounds for the moments. The lower bound is vacuous for centered features ($\mu_m = 0$), which does not affect Theorem 1. The upper bound cannot be improved:

Proposition 1 (Sharpness). *If every X^m is almost surely constant, then $W_2(\nu_r, \nu_{r'}) = d(r, r')$ for all $r, r' \in [0,1)^M$; if moreover $M = 1$, the lower bound of Lemma 1 also holds with equality. Consequently, $d(r, r')$ is the smallest upper bound on $W_2(\nu_r, \nu_{r'})$ that depends on $\mathcal{D}$ only through $(B_1, \ldots, B_M)$.*

Theorem 1 (Fixed-model missing-rate risk-shift bound). *Under Assumptions 1 and 3, fix a parameter vector θ satisfying Assumption 2, and let d be the distance of Lemma 1. Then, for every $r, r' \in [0,1)^M$,*

$$\left|R(r;\theta) - R(r';\theta)\right| \le L_\theta d(r, r').$$

Consequently, for any target rate r^* and reference rates $r_1, \ldots, r_n$,

$$R(r^*;\theta) \le \min_{1 \le i \le n} \left\{R(r_i;\theta) + L_\theta d(r^*, r_i)\right\},$$

and more generally, for every $w \in \Delta^{n-1}$, $R(r^*;\theta) \le \sum_{i=1}^n w_i R(r_i;\theta) + L_\theta \sum_{i=1}^n w_i d(r^*, r_i)$.

Fixed-model scope. For separately trained models $\theta_1, \ldots, \theta_n$ the bound holds row-wise, $R(r^*;\theta_i) \le R(r_i;\theta_i) + L_{\theta_i} d(r^*, r_i)$; risks of different models cannot be combined into a bound for one model.

Scope for contextual dialogue models. In Section 5.1 an input is a dialogue $X = (X_t^m)_{t,m}$, masked independently for each valid utterance t and modality m. A formal application would treat each (t, m) as a block, with a padding-excluded sequence norm in Assumption 2 and in the second moments. We instead use utterance-level plug-in moments and an utterance-normalized distance d_{utt}. Since we do not certify Lipschitzness of the contextual backbones under that norm, d_{utt} and the slopes below are descriptive diagnostics, not certified applications of Theorem 1.

4.1 Descriptive cross-condition slopes

For a trained network θ_a with training (reference) rate r_a and an evaluation rate $r_j \neq r_a$, define

$$d_{aj} := d_{\mathrm{utt}}(r_a, r_j), \qquad \Delta_{aj} := \widehat{R}(r_j;\theta_a) - \widehat{R}(r_a;\theta_a),$$

where $\widehat{R}$ is held-out empirical risk. For a collection $\mathcal{K}$ of such within-model comparisons, the through-origin least-squares slope is

$$\widehat{\beta}_{\mathrm{LS}} := \arg\min_\beta \sum_{(a,j)\in\mathcal{K}} (\Delta_{aj} - \beta d_{aj})^2 = \frac{\sum_{(a,j)\in\mathcal{K}} d_{aj}\Delta_{aj}}{\sum_{(a,j)\in\mathcal{K}} d_{aj}^2}.$$

This is the average signed change in empirical risk per unit of distance on the observed grid; unlike a Lipschitz constant, it can be negative. Each Δ_{aj} compares two conditions for the same network, so pooling over a gives a method-level summary, not a cross-model instance of Theorem 1. We call $\widehat{\beta}_{\mathrm{LS}}$ the *slope* and the mean matched-rate risk $\widehat{R}(r_a;\theta_a)$ the *intercept*.

Finite-difference diagnostic. We also compute

$$\widehat{s}_{\mathrm{FD}}(\sigma) := \mathbb{E}_{(\widetilde{x},y),\varepsilon}\left[\frac{|g_\theta(\widetilde{x}+\varepsilon, y) - g_\theta(\widetilde{x}, y)|}{\|\varepsilon\|_{\mathrm{utt}}}\right], \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \text{ per valid utterance},$$

as a finite sample average, where $\|\cdot\|_{\mathrm{utt}}$ is the RMS Euclidean norm over valid, non-padding utterances, the same scale convention as d_{utt}. This is an average over random directions, not a supremum, and therefore neither an upper bound on nor an estimator of L_θ. We probe the three seed-12 MMGCN variants of Section 7 trained at $r = (0.5, 0.5, 0.5)$, at that rate, with 20 noise draws. At $\sigma = 1$, whose per-utterance perturbation norm (≈ 54) is comparable to the observed distance

range (5–45), $\widehat{s}_{\mathrm{FD}} \in [0.0016, 0.0026]$ across the three variants, the same order of magnitude as the pooled slope $\widehat{\beta}_{\mathrm{LS}} = 0.00288$ of Section 7.2. Sweeping σ moves $\widehat{s}_{\mathrm{FD}}$ from 0.0005 ($\sigma = 0.01$) to 0.008 ($\sigma = 3$, norm ≈ 163). The average diagnostic thus depends on the perturbation scale; L_θ, a single global supremum for a fixed model and norm, does not. The agreement in magnitude does not verify Assumption 2.

5 Empirical Diagnostics

We first checked Lemma 1 on a synthetic Gaussian mixture where the moments are known exactly: the upper bound, the term Theorem 1 uses, held for all 120 sampled rate pairs (Appendix B.3). The rest of this section examines the same geometry on real trained networks.

5.1 Cross-evaluation on trained networks

On MMGCN+IEMOCAP, for each of the three variants of Section 7 (baseline, FCM-only, full BALM), we train one model per missing-rate configuration (12 configurations; Appendix B.1) on IEMOCAP's official `trainVid` split, selecting checkpoints on a 10% development split of `trainVid`. Each model is evaluated on the held-out `testVid` utterances re-masked at all 12 configurations by the i.i.d. MCAR masker (Section 6.1); `testVid` is never used for training, hyperparameter choice, or checkpoint selection. Training uses the original masking code at the training rate. The two maskers agree in per-modality rates but not in joint patterns: the original code never removes every modality from an utterance, whereas i.i.d. masking does so with probability $r_{\mathrm{a}} r_{\mathrm{t}} r_{\mathrm{v}}$. The matched-rate (diagonal) risk is therefore matched in marginal rates; because every variant shares both maskers, paired comparisons between variants are unaffected. Distances use utterance-level plug-in training moments $\widehat{B}^2_m$, and from here on we write d for d_{utt}. We use two evaluation passes. The *primary pass* covers seeds 12–14 (3 variants $\times$ 3 seeds $\times$ 12 configurations $=$ 108 models). The *six-seed pass* re-evaluates these models with fresh re-masking draws and adds seeds 15–17 (216 models, 18 variant–seed matrices). Section 7 uses the six-seed pass throughout. Across the nine models both passes share, slopes differ by 0.00015 on average (max. 0.00041) and intercepts by 0.004, against seed-to-seed standard deviations of 0.0007–0.0011 and 0.05–0.08: training seeds, not re-masking draws, dominate the variance, which is why the decision rule of Section 6.3 budgets seeds.

In the six-seed pass, d_{aj} and Δ_{aj} are positively correlated in all 18 matrices, weakly in some (Pearson 0.04–0.41; pooled 0.25). This describes the observed grid only. Theorem 1 is an upper-envelope statement and predicts neither positive correlation nor linear or monotone growth of risk with distance.

5.2 Empirical envelope coverage

For an observed within-model pair (a, j) with $d_{aj} > 0$, let $q_{aj} := \Delta_{aj}/d_{aj}$. On a calibration set $\mathcal{C}$, define the empirical coefficients

$$\widehat{q}_{.95} := \mathrm{Quantile}_{.95}\{q_{aj} : (a, j) \in \mathcal{C}\}, \qquad \widehat{q}_{\max} := \max_{(a,j)\in\mathcal{C}} q_{aj}.$$

They describe an empirical upper envelope over the sampled networks, utterances, mask draws, and rate pairs; $\widehat{q}_{\max}$ covers its calibration pairs by construction. Neither is a certified bound on L_θ, so this calculation does not validate Theorem 1 on real networks.

The primary pass contains 3 variants $\times$ 3 seeds $\times$ 12 reference rates $\times$ 11 target rates $=$ 1188 within-model pairs. The slope $\widehat{\beta}_{\mathrm{LS}} = 0.00317$ covers 56.9% of them, as expected for a least-squares fit to the average trend. The 95th-percentile coefficient $\widehat{q}_{.95} = 0.01276$ covers 94.9% and the maximum $\widehat{q}_{\max} = 0.04053$ covers 100% by construction. We report all three to separate average sensitivity from the empirical upper envelope.

Held-out transfer. In the six-seed pass, $\widehat{q}_{.95} = 0.01274$ calibrated on the 1188 pairs from seeds $\{12, 13, 14\}$ covers 95.0% of the 1188 pairs from the unseen seeds $\{15, 16, 17\}$ (94.9% in-sample). Calibrated instead on the 1296 pairs whose endpoints both lie among 9 of the 12 configurations, $\widehat{q}_{.95} = 0.01128$ covers 89.8% of the 1080 remaining pairs, each of which involves one of the three

configurations that drop a single modality at rate 0.9. The envelope therefore transfers across seeds and degrades moderately under extrapolation to extreme configurations. These are finite-sample checks on the chosen splits, not distribution-free coverage guarantees.

6 A Protocol for Evaluating Cross-Condition Robustness Claims

6.1 Masking operator

The masking code shipped with the BALM codebase, which both backbones use, is a deterministic quota-fill rather than an i.i.d. Bernoulli sampler. It computes exact counts for each present/absent pattern from the target missing-rate vector, lays the patterns out in a fixed block order (all modalities present first), and assigns them to utterances in split order. This has three consequences. First, repeated evaluation at a fixed rate always produces the same mask, so the masking variance that an MCAR-consistent (Assumption 1) evaluation should average over cannot be estimated. Second, on an unshuffled split such as the held-out test set, the pattern an utterance receives is a deterministic function of its position: utterances early in the file keep every modality, later ones do not. This is a concrete way for MCAR to fail that generic "device dropout" stories do not capture. Third, the all-missing pattern is never produced; its mass is reassigned to single-modality patterns. We replace the masker at evaluation time with a drop-in i.i.d. MCAR masker that draws independent per-utterance, per-modality Bernoulli masks and is seedable for repeated, independent evaluation.

6.2 Diagnostics

Positional-correlation check. For a masker applied to an unshuffled evaluation split, we correlate each dialogue's position in the split with the fraction of its utterances missing a given modality, and assess the correlation with a permutation test. **Embedding-collapse check.** For frozen or pre-extracted features, we compare the mean cosine similarity of same-label and different-label pairs. Features that carry label information make same-label pairs more similar; magnitude statistics such as B_m^2 cannot detect the failure. Our initial frozen RoBERTa text features failed this check (Appendix B.2).

6.3 Pre-registered multi-seed decision rule

Seed-to-seed variation in these diagnostics is as large as the effects under study (Table 1), so an effect should be tested under a stopping rule fixed before the data are seen. We fix (i) a batch size of three seeds, (ii) a hard cap of two batches (six seeds), and (iii) the decision statistics: paired, two-sided t and Wilcoxon signed-rank tests over seeds at $\alpha = 0.05$, both reported. Seeds are added only in full batches and never beyond the cap, which rules out optional stopping (adding seeds until a desired result appears). At $n{=}3$ the smallest attainable two-sided Wilcoxon p is 0.25, so no effect can pass both tests on a single batch. Sections 7.2–7.3 show the rule changing the conclusion for two effects that were unanimous in their first batch.

7 Case Study: BALM

7.1 Setup

We reproduce BALM's Feature Calibration Module (FCM) and Gradient Rebalancing Module (GRM) from the official implementation on two dialogue backbones, MMGCN (Hu et al., 2021) and MM-DFN (Hu et al., 2022), using the official IEMOCAP training configuration of each, and compare three variants: the backbone alone (*baseline*), FCM without GRM (*FCM-only*), and FCM+GRM (*full BALM*). Each variant is trained at each of the 12 configurations of Appendix B.1 and cross-evaluated as in Section 5.1, with test NLL as the risk and three re-masking draws averaged per cell. Audio (1582-d) and visual (342-d) features match the original release; the 1024-d text features are reconstructed (Appendix B.2). For each variant–seed matrix we record the slope $\widehat{\beta}_{\mathrm{LS}}$ and the intercept (mean diagonal loss), and compare each BALM variant to the baseline with paired tests over seeds (Section 6.3).

Table 1: Slope $\widehat{\beta}_{\mathrm{LS}}$ per seed, MMGCN+IEMOCAP, reconstructed text features. The baseline has the *highest* mean, but per-seed values overlap and the paired difference (full BALM − baseline) is positive in the first batch (seeds 12–14) and negative in the second (15–17). No difference is significant (Section 7.2).

Variant	Seed 12	Seed 13	Seed 14	Seed 15	Seed 16	Seed 17	Mean
Baseline	0.00234	0.00314	0.00333	0.00156	0.00389	0.00391	0.00303
Full BALM	0.00276	0.00336	0.00352	0.00149	0.00293	0.00280	0.00281
FCM-only	0.00306	0.00279	0.00372	0.00148	0.00426	0.00156	0.00281

7.2 SLOPE: A UNANIMOUS THREE-SEED EFFECT THAT DOES NOT REPLICATE (MMGCN)

BALM's central claim is that FCM and GRM make a model less sensitive to the missing-rate pattern, which in our decomposition means a lower slope. Our first three-seed batch (seeds 12–14) was trained with frozen RoBERTa text features, and on it the claim looks confirmed: full BALM's $\widehat{\beta}_{\mathrm{LS}}$ is lower than the baseline's in all three seeds ($0.00287/0.00281/0.00323$ vs. $0.00407/0.00576/0.00587$; Table 3). Any single seed of this batch, which is how most results in the area are reported, would have confirmed the claim. Two parts of the protocol overturned it. First, the embedding-collapse check showed that the frozen text features do not separate emotion labels (Appendix B.2). With reconstructed features, the same three seeds were again unanimous, but in the opposite direction: full BALM's slope is *higher* in every seed (Table 1, seeds 12–14). Second, the pre-registered second batch (seeds 15–17) reversed the sign in all three new seeds, so the paired difference (full BALM − baseline) splits exactly $3/6$.

At $n{=}6$ neither BALM variant differs from the baseline. Paired t-tests give $t(5){=}-0.82$, $p{=}0.447$ for full BALM and $t(5){=}-0.48$, $p{=}0.650$ for FCM-only; Wilcoxon signed-rank tests give $W{=}9$, $p{=}0.844$ for both. Since neither comparison is significant even before correction, a Holm–Bonferroni adjustment cannot change the conclusion. The null is informative rather than merely underpowered. Full BALM's paired slope difference is -7% of the baseline mean, with a 95% confidence interval of $[-30\%, +15\%]$. Given the observed seed-to-seed spread, the six-seed design has 80% power to detect a 31% reduction. A large sensitivity reduction on this backbone is therefore ruled out; a small one is not. The pooled slope over all 18 variant–seed matrices is 0.00288 (pair-level bootstrap 95% CI $[0.00259, 0.00317]$, 2000 resamples). Intercepts agree. Full BALM's diagonal loss does not differ from the baseline's ($t(5){=}1.04$, $p{=}0.35$), while FCM-only raises it in all six seeds ($t(5){=}4.04$, $p{=}0.010$; Wilcoxon $p{=}0.031$; Table 3). An independent metric agrees as well: on our 12 MMGCN+IEMOCAP checkpoints, MissBench's MEI averages $0.259/0.254/0.231$ (baseline/FCM-only/full BALM), with no significant equity gain for full BALM (paired over configurations, $p{=}0.62$). MLI requires gradient logs we did not record.

7.3 INTERCEPT: AN ABSOLUTE-RISK EFFECT WHOSE SIGNIFICANCE DEPENDS ON THE TEST (MM-DFN)

On MM-DFN (216 models: 3 variants $\times$ 12 configurations $\times$ 6 seeds), the slope again shows no significant effect. Full BALM's $\widehat{\beta}_{\mathrm{LS}}$ is lower in five of six seeds, but by small amounts ($t(5){=}-1.90$, $p{=}0.12$; Wilcoxon $p{=}0.0625$). Absolute risk behaves differently. In the first batch, full BALM's diagonal loss is below the baseline's by 0.097–0.133 in all three seeds (paired $t(2){=}-10.70$, $p{=}0.009$; Wilcoxon $p{=}0.25$, the minimum attainable at $n{=}3$), a large and unanimous effect. The pre-registered second batch weakens it: seed 16 reverses sign (full BALM worse by 0.040), while seeds 15 and 17 do not. Over six seeds, the mean diagonal gap is -0.086 (s.d. 0.070). It is significant by paired t-test ($t(5){=}-3.00$, $p{=}0.030$) but not by Wilcoxon ($p{=}0.0625$), and the off-diagonal gap behaves the same way ($t(5){=}-3.45$, $p{=}0.018$; Wilcoxon $p{=}0.0625$). FCM-only shows no significant diagonal effect ($t(5){=}-1.54$, $p{=}0.19$), so the reduction requires GRM. The effect points the same way in five of six seeds, but whether it is significant depends on a normality assumption that six paired differences cannot check. Under our decision rule it is therefore *test-dependent*, not confirmed. This is the failure mode of Section 7.2 on a different axis, narrowed rather than resolved by more data.

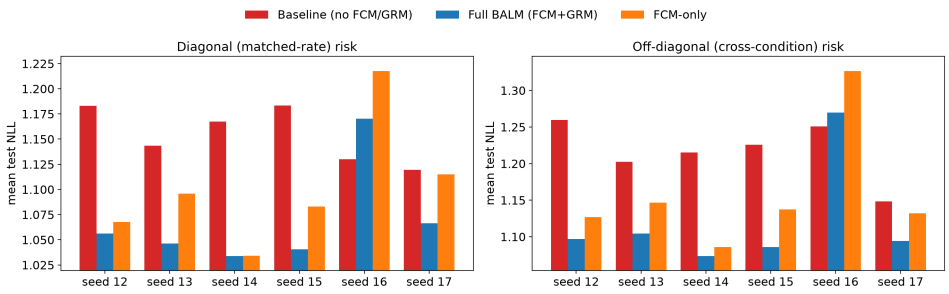

Figure 1: MM-DFN+IEMOCAP, six seeds: mean test NLL at the training rate (left, diagonal) and at the other 11 rates (right, off-diagonal). Full BALM lowers both in five of six seeds, and seed 16 reverses sign. The paired t-test is significant (p=0.030 and 0.018); the Wilcoxon test is not (p=0.0625). MMGCN shows no such effect (Section 7.2).

Table 2: Effects of each method relative to its own baseline, for every backbone/dataset/method combination tested. Parentheses give paired t-test / Wilcoxon signed-rank p-values over seeds; single-seed rows are descriptive. The last column is what a standard matched-rate results table reports (weighted-F1 difference in points). †Significant by t-test but not by Wilcoxon, i.e., test-dependent. The slope is a descriptive statistic, not an estimate of the Lipschitz constant in Theorem 1.

Method, backbone / data	Seeds	Slope	Intercept (NLL)	Matched-rate wF1
BALM, MMGCN / IEMOCAP	6	None (0.45/0.84)	None (0.35/0.44)	−1.1 (0.044/0.16)
BALM, MM-DFN / IEMOCAP	6	None (0.12/0.0625)	Lower† (0.030/0.0625)	+0.4 (0.26/0.44)
BALM, MMGCN / MELD	1	None	None (baseline lower)	−1.3
AMB-DSGDN, IEMOCAP	1	Lower	Lower	+0.5

Slope and intercept are separate axes. $\widehat{\beta}_{\mathrm{LS}}$ measures how fast empirical risk grows with distance from the training rate, and the intercept measures where that growth starts. The two move independently. On MMGCN, BALM moves neither; on MM-DFN, it moves only the intercept (Figure 1); for AMB-DSGDN, both move together (Section 7.5). A matched-rate accuracy table measures neither axis directly (Section 7.5).

7.4 A POSITIONAL ARTIFACT IN THE ORIGINAL MASKING

We apply the positional-correlation check to the masks that the original quota-fill code assigns to IEMOCAP's test split, whose dialogue order is never shuffled, at the configuration used in BALM's released IEMOCAP script, $(r_{\mathrm{a}}, r_{\mathrm{t}}, r_{\mathrm{v}}) = (0.5, 0.7, 0.3)$. Missingness tracks position: $\rho = 0.60$ for audio ($p < 0.005$, the resolution of the 200-permutation test), $\rho = 0.51$ for text ($p = 0.005$), and $\rho = 0.29$ for visual ($p = 0.16$, n.s.). The artifact does not depend on the configuration: at every one of our 12 grid configurations at least one modality is position-correlated at $p < 0.05$, and audio is correlated at $p < 0.005$ in 10 of 12 ($\rho = 0.34$–0.70; 10,000 permutations). Which dialogues lose a modality is therefore partly determined by where they sit in the test file, a deterministic MNAR mechanism that any evaluation with this masker inherits. The finding does not affect the slope and intercept results above, all of which are evaluated with our i.i.d. masker.

7.5 SYNTHESIS

Two further settings, detailed in Appendices B.4 and B.5, complete the picture. On a second dataset, MELD (one seed), MMGCN reproduces its IEMOCAP pattern: full BALM and the baseline have essentially the same slope (0.00402 vs. 0.00413), and the baseline has the lower diagonal loss (1.617 vs. 1.645). A structurally different method, AMB-DSGDN, whose adaptive modality dropout replaces feature and gradient rebalancing, lowers both the slope (0.00440 vs. 0.00496) and the diagonal loss at one seed. Its original training loop, which we kept for fidelity, selects checkpoints on the test split; both conditions share this rule, but together with the single seed it makes the result suggestive only. Table 2 collects all settings.

What a matched-rate table would report. The standard results table lists weighted-F1 at matched train/test rates, which is only the diagonal of our 12×12 matrices (last column of Table 2). It cannot register a slope, which lives in the off-diagonal entries, or the masking artifact. It also misses the one intercept effect we found. On MM-DFN, full BALM's matched-rate weighted-F1 is within 0.4 points of the baseline's (t: p=0.26; W: p=0.44), although its NLL intercept is lower in five of six seeds: the effect is a change in predictive confidence that accuracy does not see. On MMGCN, the same table shows full BALM 1.1 points worse (t: p=0.044; W: p=0.16), a test-dependent deficit that the NLL intercept does not show; off-diagonal, the deficit holds in all six seeds (-1.7 points; t: p=0.003; W: p=0.031). Table 4 lists all 20 paired comparisons (two variants, two backbones, five metrics). Four pass both tests at $\alpha = 0.05$, three of them harms on MMGCN, and none survives a Holm correction across the 20. With six seeds, per-comparison significance is fragile in both directions. Whether a mechanism helps, and along which axis, depends on the backbone, the method, and the metric. This heterogeneity is the main empirical finding of the case study.

8 Discussion and Limitations

What the theory certifies. Theorem 1 is a population statement for one fixed model under a global Lipschitz constant L_θ, which we assume and do not certify for any trained network. The quantities we compute ($\widehat{\beta}_{\mathrm{LS}}$, $\widehat{q}_{.95}$, $\widehat{q}_{\max}$, $\widehat{s}_{\mathrm{FD}}$) share the theorem's distance but are descriptive, and none substitutes for L_θ. They use held-out risks, training-sample moments, and, for dialogue backbones, an utterance-normalized distance (Section 4); without simultaneous confidence bounds, the reported envelopes carry no population coverage guarantee.

MCAR. Assumption 1 holds by construction for our i.i.d. masker, which produces every evaluation in Sections 5 and 7; BALM's backbones are trained with the original code, as in BALM itself. For naturally occurring missingness MCAR remains an assumption, and Section 7.4 shows it failing for the original deterministic masker through test-set position, a mechanism distinct from the device-dropout stories usually discussed.

Scale dependence. $d(r, r')$ weights modalities by raw B_m^2, so it is not invariant to feature rescaling, and $\widehat{\beta}_{\mathrm{LS}}$ inherits this dependence. Z-scoring every feature dimension does not weaken the distance–risk correlation (Appendix B.1), and we recommend it as a fixed convention.

Generality. The slope null holds on both backbones, while the intercept effect appears only on MM-DFN. MELD reproduces the MMGCN pattern, which suggests that the backbone, not the dataset, drives the difference, but that replication uses one seed and shows weak distance–risk correlations. The AMB-DSGDN result also uses one seed, with checkpoints selected on the test split by its original training loop. We hold it to the standard that overturned BALM's first-batch slope result: it is suggestive until replicated. All experiments use two modest-scale MER corpora with pre-extracted features; CMU-MOSEI and settings such as audio-visual speech recognition, retrieval, and vision-language models are untested.

9 Conclusion

Because masking never changes the label, the risk shift of a fixed model between two missing-rate conditions admits a distortion-free bound whose transport term is closed-form. The bound suggests evaluating robustness claims along two axes, slope and intercept, under a masking procedure that satisfies its assumptions and a stopping rule fixed in advance. Applied to BALM and AMB-DSGDN, this protocol found a masking artifact present at every configuration we tested, a slope effect that was unanimous at three seeds and absent at six, an intercept effect whose significance depends on the test and which accuracy does not register, and a method that improves both axes. We do not conclude that BALM is ineffective, or that AMB-DSGDN is effective, in general. We conclude that matched-rate, few-seed evaluation cannot tell these outcomes apart, and we offer the decomposition and protocol as tools for claims that should.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REPRODUCIBILITY STATEMENT

Appendix A contains complete proofs of Lemma 1, Proposition 1, and Theorem 1. Appendix B.1 lists all hyperparameters, the missing-rate grid, and the re-masking procedure; Appendix B.2 describes the text-feature reconstruction; Appendix B.6 reports per-seed results; and Algorithm 1 summarizes the full evaluation protocol.

REFERENCES

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Linus Bleistein, Aurélien Bellet, and Julie Josse. Optimal transport with heterogeneously missing data. *arXiv preprint arXiv:2505.17291*, 2025.

Young Sang Choi, Vincent Jeanselme, Pierre Elias, and Shalmali Joshi. ICYM2I: The illusion of multimodal informativeness under missingness. In *ICLR*, 2026.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

Rong Fu, WeiZhi Tang, Ziming Wang, Jia Yee Tan, Zijian Zhang, Zhaolu Kang, Muge Qi, Shuning Zhang, and Simon Fong. ModalImmune: Immunity driven unlearning via self destructive training. *arXiv preprint arXiv:2602.16197*, 2026.

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. MMG-Ego4D: Multi-modal generalization in egocentric action recognition. In *CVPR*, 2023.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP*, 2022.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *ACL-IJCNLP*, 2021.

Xiang Li, Xiaoming Zhang, Dezhuang Miao, Xianfu Cheng, Dawei Li, Honggui Han, and Zhoujun Li. Mixture of disentangled experts with missing modalities for robust multimodal sentiment analysis. *arXiv preprint arXiv:2602.01833*, 2026.

Ronghao Lin, Honghao Lu, Ruixing Wu, Aolin Xiong, Qinggong Chu, Qiaolin He, Sijie Mai, and Haifeng Hu. MissMAC-Bench: Building solid benchmark for missing modality issue in robust multimodal affective computing. *arXiv preprint arXiv:2602.00811*, 2026.

Rui Liu, Haolin Zuo, Zheng Lian, Hongyu Yuan, and Qi Fan. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities. In *ACM Multimedia*, 2025.

Xiaohao Liu, Xiaobo Xia, Jiaheng Wei, Shuo Yang, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. Calibrated multimodal representation learning with missing modalities. In *ICML*, 2026.

Phuong-Anh Nguyen, Tien Anh Pham, Duc-Trong Le, and Cam-Van Thi Nguyen. BALM: A model-agnostic framework for balanced multimodal learning under imbalanced missing rates. *arXiv preprint arXiv:2603.19718*, 2026.

Long H. C. Pham. Decomposing out-of-distribution error in conditional flow matching via Wasserstein geometry. In *ICML*, 2026.

Tien Anh Pham, Phuong-Anh Nguyen, Duc-Trong Le, and Cam-Van Thi Nguyen. MissBench: Benchmarking multimodal affective analysis under imbalanced missing modalities. *arXiv preprint arXiv:2603.09874*, 2026.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *AAAI*, 2024.

Yunsheng Wang, Yuntao Shou, Yilong Tan, Wei Ai, Tao Meng, and Keqin Li. AMB-DSGDN: Adaptive modality-balanced dynamic semantic graph differential network for multimodal emotion recognition. *arXiv preprint arXiv:2603.10043*, 2026.

Guowei Zhong, Ruohong Huan, Mingzhen Wu, Ronghua Liang, and Peng Chen. Towards robust multimodal emotion recognition under missing modalities and distribution shifts. *arXiv preprint arXiv:2506.10452*, 2025.

A Full Proofs

Throughout, let $\mathcal{X} = \prod_{m=1}^{M} \mathbb{R}^{d_m}$ carry the block Euclidean norm $\|z\|^2 = \sum_m \|z^m\|_2^2$, and let $(X, Y) \sim \mathcal{D}$. For $r \in [0,1)^M$,

$$\widetilde{X}_r^m = X^m E_r^m, \qquad E_r^m \sim \text{Bernoulli}(1 - r_m),$$

with masks independent across m and independent of (X, Y). Write $\nu_r = \text{Law}(\widetilde{X}_r)$ and $P_r = \text{Law}(\widetilde{X}_r, Y)$, and let

$$B_m^2 = \mathbb{E}_{\mathcal{D}}\|X^m\|_2^2, \qquad \mu_m = \mathbb{E}_{\mathcal{D}}[X^m].$$

A.1 Proof of Lemma 1 (feature-marginal Wasserstein sandwich)

Lower bound. Assumption 3 implies finite first moments, so Kantorovich–Rubinstein duality applies:

$$W_1(\nu_r, \nu_{r'}) = \sup_{\text{Lip}(f) \le 1} \left| \mathbb{E}_{\nu_r}[f] - \mathbb{E}_{\nu_{r'}}[f] \right|.$$

For any unit $u \in \mathcal{X}$, $f_u(z) = \langle z, u \rangle$ is 1-Lipschitz. Independence of the mask gives $\mathbb{E}[\widetilde{X}_r^m] = (1 - r_m)\mu_m$. Hence

$$\begin{aligned} W_1(\nu_r, \nu_{r'}) &\ge \sup_{\|u\|=1} \left| \left\langle \mathbb{E}[\widetilde{X}_r] - \mathbb{E}[\widetilde{X}_{r'}], u \right\rangle \right| \\ &= \left\| \mathbb{E}[\widetilde{X}_r] - \mathbb{E}[\widetilde{X}_{r'}] \right\| \\ &= \sqrt{\sum_{m=1}^{M} \|\mu_m\|_2^2 (r_m - r'_m)^2}. \end{aligned}$$

Upper bound. Draw $X \sim \mathcal{D}_X$ and independent uniforms $U^1, \dots, U^M \sim \text{Uniform}(0,1)$. Define

$$E_r^m = \mathbf{1}\{U^m < 1 - r_m\}, \qquad E_{r'}^m = \mathbf{1}\{U^m < 1 - r'_m\}.$$

Each mask has the required marginal and $\Pr(E_r^m \ne E_{r'}^m) = |r_m - r'_m|$. Thus $(\widetilde{X}_r, \widetilde{X}_{r'})$ is a coupling of $(\nu_r, \nu_{r'})$, with

$$\begin{aligned} \mathbb{E}\|\widetilde{X}_r - \widetilde{X}_{r'}\|^2 &= \sum_{m=1}^{M} \mathbb{E}\big[\|X^m\|_2^2 (E_r^m - E_{r'}^m)^2\big] \\ &= \sum_{m=1}^{M} \mathbb{E}\|X^m\|_2^2 \, \Pr(E_r^m \ne E_{r'}^m) \\ &= \sum_{m=1}^{M} B_m^2 |r_m - r'_m|. \end{aligned}$$

Therefore $W_2^2(\nu_r, \nu_{r'}) \le \sum_m B_m^2 |r_m - r'_m|$. Together with $W_1 \le W_2$, this proves Lemma 1. ■

A.2 Proof of Proposition 1 (sharpness)

Let $X^m = x^m$ almost surely, so that $B_m = \|x^m\|_2$ and $\mu_m = x^m$. By independence of the masks, $\nu_r = \bigotimes_m \nu_r^m$ with $\nu_r^m = (1 - r_m)\delta_{x^m} + r_m \delta_0$. Let (Z, Z') be any coupling of $(\nu_r, \nu_{r'})$. Its m-th block (Z^m, Z'^m) couples ν_r^m and $\nu_{r'}^m$, so $\Pr(Z^m \ne Z'^m) \ge \|\nu_r^m - \nu_{r'}^m\|_{\text{TV}} = |r_m - r'_m|$ when $x^m \ne 0$ (the m-th term below vanishes otherwise), and $\|Z^m - Z'^m\|_2 = \|x^m\|_2$ whenever $Z^m \ne Z'^m$. Because the squared cost is additive over blocks,

$$\mathbb{E}\|Z - Z'\|^2 = \sum_{m=1}^{M} \mathbb{E}\|Z^m - Z'^m\|_2^2 \ge \sum_{m=1}^{M} B_m^2 |r_m - r'_m| = d(r, r')^2.$$

Minimizing over couplings gives $W_2(\nu_r, \nu_{r'}) \ge d(r, r')$, and Lemma 1 gives the reverse inequality. For $M = 1$, the coupling in the proof of Lemma 1 moves mass $|r - r'|$ over distance $\|x\|_2$, so $W_1(\nu_r, \nu_{r'}) \le \|x\|_2 |r - r'| = \|\mu\|_2 |r - r'|$, which equals the lower bound. Finally, let f be any function with $W_2(\nu_r, \nu_{r'}) \le f(B_1, \dots, B_M, r, r')$ for every $\mathcal{D}$ satisfying Assumption 3. Applying it to the constant distribution with $\|x^m\|_2 = B_m$ gives $f \ge d$. ■

A.3 Proof of Theorem 1 (fixed-model missing-rate risk-shift bound)

Fix one parameter vector θ and rates r, r'. Use the coupling above, retaining the same label Y from the shared draw (X, Y). Then

$$\begin{aligned}
\big|R(r;\theta) - R(r';\theta)\big| &= \Big|\mathbb{E}\Big[g_\theta(\widetilde{X}_r, Y) - g_\theta(\widetilde{X}_{r'}, Y)\Big]\Big| \\
&\le \mathbb{E}\Big[\big|g_\theta(\widetilde{X}_r, Y) - g_\theta(\widetilde{X}_{r'}, Y)\big|\Big] \\
&\le L_\theta \mathbb{E}\|\widetilde{X}_r - \widetilde{X}_{r'}\| \\
&\le L_\theta \sqrt{\mathbb{E}\|\widetilde{X}_r - \widetilde{X}_{r'}\|^2} \\
&= L_\theta \sqrt{\sum_{m=1}^{M} B_m^2 |r_m - r'_m|} \\
&= L_\theta d(r, r').
\end{aligned}$$

The third line uses Assumption 2; the fourth is Cauchy–Schwarz. Setting $r = r^*$ and $r' = r_i$ gives, for every i,

$$R(r^*;\theta) \le R(r_i;\theta) + L_\theta d(r^*, r_i).$$

Taking the minimum over i proves the nearest-reference form. Multiplying each inequality by $w_i \ge 0$ and summing, with $\sum_i w_i = 1$, gives

$$R(r^*;\theta) \le \sum_{i=1}^{n} w_i R(r_i;\theta) + L_\theta \sum_{i=1}^{n} w_i d(r^*, r_i).$$

This proves Theorem 1. ■

Population versus empirical quantities. The proof uses population risks, population moments, and the global pointwise constant L_θ. Replacing them by held-out risks, sample moments, and observed slopes can understate the right-hand side. A numerical plug-in calculation is therefore descriptive unless it includes valid simultaneous confidence bounds for the moments and risks together with a certified upper bound on L_θ. If $\mathcal{D}$ is declared to be a fixed empirical distribution and all expectations are evaluated exactly, the theorem holds for that empirical distribution, not automatically for the source population.

A.4 Remark: why no distortion term

The three-term decomposition of Pham (2026) for conditional flow matching (Interpolation Sparsity, Geometric Distortion, and In-Distribution Fit) needs a distortion term because the map from condition to target distribution has no guaranteed isometric embedding into Wasserstein space: two conditions can imply arbitrarily different target geometries, and Distortion is the cost of that mismatch. In our setting, every r shares the *same* underlying $(x, y) \sim \mathcal{D}$ and only the observation changes. There is no geometric mismatch to pay for, and the bound has two terms: in-distribution fit $R(r_i;\theta)$ and a distance term $L_\theta d(r^*, r_i)$ that plays a role loosely analogous to Interpolation Sparsity. This follows from Assumption 1 and the invariance of the label under masking. It is not a weaker assumption set than that of Pham (2026); the two settings are not directly comparable in difficulty.

B Additional Experimental Details

B.1 Hyperparameters

Both backbones use the settings of BALM's released IEMOCAP training scripts; only the seed, the missing-rate configuration, and the ablation switches vary.

MMGCN+IEMOCAP (all variants): seeds $\{12, \ldots, 17\}$, 80 epochs, batch size 16, Adam, learning rate 3×10^{-4}, weight decay 3×10^{-5}, dropout 0.4, residual connections, $\tau = 0.3$, $\rho = 1.4$, FCM global-feature dimension $d_{\text{global}} = 737$, and normalized GRM modulation. Ablation switches:

full BALM (`use_fcm=1, use_grm=1`), baseline (`use_fcm=0, use_grm=0`), and FCM-only (`use_fcm=1, use_grm=0`). Checkpoints are selected by development F1 on a random 10% of `trainVid`.

MM-DFN+IEMOCAP: seeds $\{12, \ldots, 17\}$ (a first batch of three, extended to a second batch of three under the decision rule of Section 6.3), 80 epochs, batch size 16, speaker weights 3-0-1, 16-layer deep GCN, Adam, learning rate 3×10^{-4}, weight decay 10^{-4}, focal loss (`focal=0.5`), class weighting, dropout 0.4, $\tau = 0.4$, $\rho = 1.6$, $d_{\text{global}} = 737$. The three variants and checkpoint selection are as for MMGCN.

Missing-rate grid (both backbones, all of Section 7): 12 configurations $(r_{\text{audio}}, r_{\text{text}}, r_{\text{visual}}) \in \{ (0.1, 0.1, 0.1), (0.3, 0.3, 0.3), (0.5, 0.5, 0.5), (0.7, 0.7, 0.7), (0.5, 0.7, 0.3), (0.3, 0.5, 0.7), (0.7, 0.3, 0.5), (0.1, 0.1, 0.9), (0.1, 0.9, 0.1), (0.9, 0.1, 0.1), (0.9, 0.9, 0.1), (0.1, 0.9, 0.9) \}$, chosen to span uniform, permuted, and extreme configurations. Cross-evaluation matrices are the full 12×12 product of these configurations, re-masked with the i.i.d. MCAR masker (Section 6.1), with 3 independent re-masking draws averaged per cell to reduce masking noise.

Feature scaling. Because $d(r, r')$ weights each modality by its raw second moment B_m^2, a modality with large raw magnitude could dominate the distance for reasons unrelated to how informative it is. Recomputing B_m^2 after per-dimension z-scoring (B^2_{audio}: $1582 \to 1582$; B^2_{text}: $836 \to 1024$; B^2_{visual}: $34 \to 342$) does not weaken the pooled distance–risk correlation of the primary pass ($0.309 \to 0.342$), so the qualitative conclusion is not an artifact of raw feature scales. We nevertheless recommend standardizing as a matter of practice. Z-scoring is fixed and content-blind; *learned* per-modality weights, as in representation-calibration methods such as CalMRL (Liu et al., 2026), could adapt B_m toward discriminative content, at the cost of making the distance depend on a learned component.

B.2 Text feature reconstruction

The audio (openSMILE IS10, 1582-d) and visual (342-d) features are reused unmodified from the community-maintained MMGCN feature release. Our first reproduction used frozen, off-the-shelf RoBERTa-large mean-pooled text features, which are prone to embedding anisotropy when used uncalibrated. Their same-label and different-label cosine similarities were indistinguishable (0.9912 vs. 0.9914) despite unremarkable magnitude statistics ($B^2_{\text{text}} = 959$, comparable to audio's 1582), so a magnitude check alone cannot catch this failure. Whitening (BERT-flow/SimCSE-style) removed the geometric collapse but did not recover downstream accuracy. Task fine-tuning (RoBERTa-large as an emotion classifier for a few epochs) narrowed the reproduction gap to BALM's published Table 1 from 16–26 points to 3–8 points at matching configurations. A discriminative-content check is therefore necessary, and a magnitude check is not sufficient. The first MMGCN batch in Table 3 was trained on the frozen features.

The final text features were built as follows. RoBERTa-large was fine-tuned as a 6-way emotion classifier on IEMOCAP's training utterances only (class-weighted cross-entropy, 4 epochs, learning rate 2×10^{-5}, batch size 16, linear warmup over 6% of steps), reaching 60.4% validation accuracy on a held-out 10% of the training utterances. 1024-d features for all splits were then extracted by attention-masked mean pooling over the fine-tuned encoder's final hidden states. The reconstructed features pass the embedding-collapse check, including on the held-out test utterances the encoder never saw (same-label vs. different-label cosine similarity 0.34 vs. 0.17; 0.44 vs. 0.14 on training utterances), and have $B^2_{\text{text}} = 836$.

B.3 Synthetic check of Lemma 1

We check Lemma 1 numerically on a synthetic Gaussian mixture (3 modalities, 6 classes, total dimension 24), where B_m^2 and μ_m are known exactly from the generation parameters. Over 120 rate pairs sampled from a $5 \times 5 \times 5$ missing-rate grid, the upper bound, the term Theorem 1 uses, is never violated by the debiased W_2 estimates, correlates at 0.98 with the empirical W_2, and has mean tightness ratio $W_2/\text{bound} = 0.81$ (Figure 2). This checks the implementation of Lemma 1; it does not test the Lipschitz assumption for a trained network.

Naive empirical OT is biased upward by the curse of dimensionality (Fournier & Guillin, 2015): at d=24 and N=800, the empirical distance between two samples of the *same* distribution is already

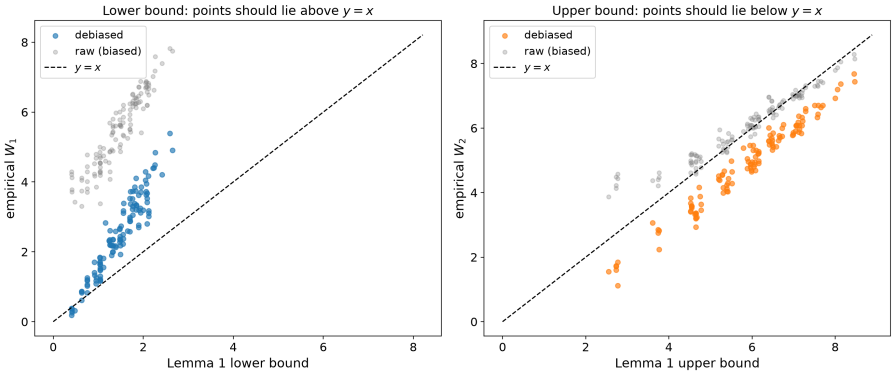

Figure 2: Synthetic check of Lemma 1 (n=120 rate pairs from a $5 \times 5 \times 5$ grid). Left: lower bound vs. empirical W_1 (points should lie above y=x). Right: upper bound vs. empirical W_2 (points should lie below y=x). Colored points are debiased and gray points raw. All 120 debiased W_2 estimates satisfy the upper bound; 7 debiased W_1 estimates at the smallest shifts fall below the lower bound because of the subtracted self-distance floor.

comparable to several of the bounds under test. We therefore subtract a measured self-distance floor $\Phi(r)$ before checking the bound and report both raw and debiased estimates. Debiasing is what makes the lower bound hard to check at small shifts. All raw W_1 estimates satisfy the lower bound, but 7 of 120 debiased ones fall below it. All seven are shifts of 0.2 in a single modality, where the lower bound is about 0.4 and the subtracted floor is about ten times larger. Recent work debiases empirical Bures–Wasserstein distances directly under MCAR missingness (Bleistein et al., 2025). That work is complementary: it estimates a debiased distance between masked samples, whereas our population expression is analytic in raw moments that must still be estimated on real data.

B.4 Dataset generality: MELD

We replicate the MMGCN protocol on MELD (1152 train / 280 test dialogues, 7 emotion classes, 9 speakers), using features from the same public release as our IEMOCAP audio and visual features. MELD's text features pass the embedding-collapse check without modification (same-label vs. different-label cosine similarity 0.880 vs. 0.773), so no reconstruction was needed. Training uses the MMGCN settings of Appendix B.1 with class weighting, $d_{\text{global}} = 310$, and seed 12.

At 1 seed, with 3 variants $\times$ 12 configurations, the correlation between $d(r_i, r_j)$ and risk change is positive but weak (0.076–0.099), well below IEMOCAP's pooled 0.25. Slopes are nearly identical for full BALM and the baseline (0.00402 vs. 0.00413; FCM-only 0.00499), and the baseline has the lower diagonal loss (1.617 vs. 1.645 for full BALM). This matches the MMGCN+IEMOCAP pattern rather than the MM-DFN intercept effect. The result is single-seed and descriptive.

B.5 A second method: AMB-DSGDN

We apply the protocol to AMB-DSGDN (Wang et al., 2026), an attention and differential semantic-graph architecture with an adaptive, performance-weighted stochastic modality dropout. Its hard-coded feature dimensions match our reconstructed IEMOCAP features, so we adapt only its dataloader to our features and i.i.d. MCAR masker (used here for both training and evaluation) and keep its architecture, training loop, and default seed (2094) unmodified. The baseline is the same architecture with adaptive dropout disabled. Two properties of the original training loop limit this comparison. It selects the checkpoint with the best test-split F1, so absolute numbers are optimistic; both conditions share this rule, and the cross-evaluation re-masks the test split with fresh draws. Its loss, which we also use as the risk, is a class-weighted NLL, so its values are not comparable with the unweighted NLL of the BALM rows; only within-method comparisons are meaningful.

The published default recipe (`warm_up_epochs=60` of `epochs=61`) leaves adaptive dropout nearly inactive: 10 of 12 configurations produced byte-identical checkpoints with and without it. We therefore reran the dropout-enabled condition with `warm_up_epochs=30`. At 1 seed, adaptive

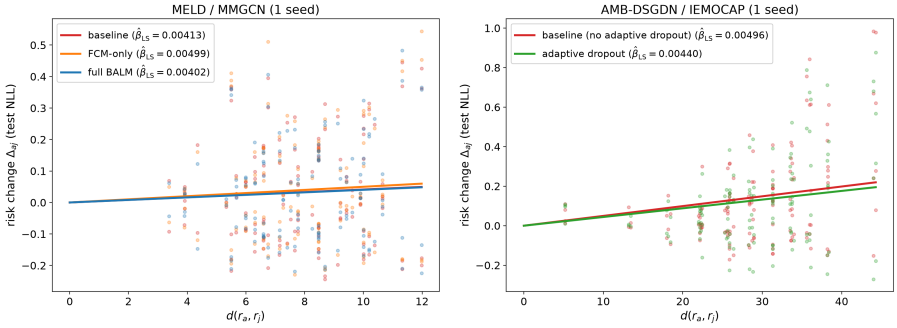

Figure 3: Left: MELD/MMGCN, one seed; weak distance–risk correlation, and the baseline has the lowest diagonal loss. Right: AMB-DSGDN/IEMOCAP, one seed; adaptive dropout lowers both slope and intercept. Lines are through-origin least-squares fits ($\widehat{\beta}_{\mathrm{LS}}$).

dropout improves both axes: $\widehat{\beta}_{\mathrm{LS}} = 0.00440$ vs. 0.00496, and diagonal loss 0.987 vs. 1.031 (off-diagonal 1.104 vs. 1.160); matched-rate weighted-F1 rises by 0.5 points. With one seed and test-split checkpoint selection, this result is suggestive, not confirmed.

B.6 Per-seed results

Table 3 lists the per-seed quantities behind Sections 7.2–7.5; the MMGCN slopes are in Table 1. Table 4 lists every paired comparison with the baseline.

Table 3: Per-seed intercepts (mean diagonal test NLL), MM-DFN slopes, and weighted-F1, six-seed pass. The bottom block is the first MMGCN batch, trained with the frozen RoBERTa text features that fail the embedding-collapse check (Section 7.2).

Variant	Seed 12	Seed 13	Seed 14	Seed 15	Seed 16	Seed 17
MMGCN, diagonal loss						
Baseline	1.227	1.151	1.092	1.127	1.204	1.151
Full BALM	1.167	1.145	1.093	1.169	1.309	1.215
FCM-only	1.270	1.257	1.108	1.310	1.331	1.284
MM-DFN, diagonal loss						
Baseline	1.183	1.143	1.167	1.183	1.130	1.119
Full BALM	1.056	1.046	1.034	1.040	1.170	1.066
FCM-only	1.068	1.096	1.034	1.083	1.217	1.115
MM-DFN, slope $\widehat{\beta}_{\mathrm{LS}}$ ($\times 10^{-3}$)						
Baseline	2.88	2.25	1.84	1.64	4.51	1.14
Full BALM	1.49	2.18	1.53	1.68	3.88	1.02
FCM-only	2.16	1.89	2.04	2.07	4.19	0.67
MMGCN, matched-rate weighted-F1 (%)						
Baseline	56.9	57.4	58.4	58.1	57.3	57.9
Full BALM	56.9	56.1	58.5	56.9	55.1	55.8
MMGCN, off-diagonal weighted-F1 (%)						
Baseline	54.6	54.1	55.0	55.9	53.5	54.0
Full BALM	53.4	52.1	54.4	54.5	50.6	51.8
MM-DFN, matched-rate weighted-F1 (%)						
Baseline	59.3	58.9	60.0	60.8	59.1	59.7
Full BALM	60.1	59.7	60.3	61.7	58.0	60.3
MMGCN, frozen RoBERTa text features: slope ($\times 10^{-3}$) / diagonal loss						
Baseline	4.07 / 1.239	5.76 / 1.203	5.87 / 1.196	–	–	–
Full BALM	2.87 / 1.500	2.81 / 1.474	3.23 / 1.353	–	–	–
FCM-only	5.09 / 1.364	5.80 / 1.343	4.41 / 1.286	–	–	–

Table 4: All paired comparisons with the baseline, six-seed pass: mean difference (variant − baseline) and paired t-test / Wilcoxon p-values over six seeds. *Both tests below 0.05. The smallest attainable two-sided Wilcoxon p at n=6 is 0.031, and no comparison survives a Holm correction across the 20 tests.

	Full BALM		FCM-only	
Metric	Diff.	p (t / W)	Diff.	p (t / W)
MMGCN				
Slope ($\times 10^{-3}$)	−0.22	0.447 / 0.844	−0.22	0.650 / 0.844
Diagonal NLL	+0.025	0.345 / 0.438	+0.102*	0.010 / 0.031
Off-diagonal NLL	+0.021	0.310 / 0.312	+0.096*	0.008 / 0.031
Matched-rate wF1 (pts)	−1.1	0.044 / 0.156	−0.1	0.703 / 0.844
Off-diagonal wF1 (pts)	−1.7*	0.003 / 0.031	−0.2	0.605 / 0.688
MM-DFN				
Slope ($\times 10^{-3}$)	−0.41	0.117 / 0.0625	−0.21	0.295 / 0.312
Diagonal NLL	−0.086	0.030 / 0.0625	−0.052	0.185 / 0.156
Off-diagonal NLL	−0.096	0.018 / 0.0625	−0.058	0.133 / 0.156
Matched-rate wF1 (pts)	+0.4	0.257 / 0.438	+0.6	0.053 / 0.094
Off-diagonal wF1 (pts)	+0.4	0.135 / 0.156	+1.0*	0.011 / 0.031

C Practitioner Checklist

Sections 4–6 interleave the theory, diagnostics, and protocol with our case study. Algorithm 1 consolidates them into a single procedure for applying the protocol to a new method or backbone.

Algorithm 1 Cross-condition robustness evaluation protocol

Require: Trained models $\{\theta_a\}$ with reference rates $\{r_a\}$; raw training features; held-out test split

1: Compute plug-in moments $\widehat{B}^2_m, \widehat{\mu}_m$ and label them as sample estimates
2: **Diagnose the masker:** run the positional-correlation check on the actual implementation; for an MCAR evaluation use an i.i.d. masker (Section 6.1)
3: **if** using frozen or pre-extracted features **then**
4: Run the embedding-collapse check; if it fails, fix the features before proceeding
5: **end if**
6: Use the synthetic check (Appendix B.3) only to test the implementation of Lemma 1
7: **for** each fixed model θ_a and target rate r_j **do**
8: Compute $\widehat{d}_{aj} = \sqrt{\sum_m \widehat{B}^2_m |r_{a,m} - r_{j,m}|}$
9: Re-mask held-out utterances at r_j in ≥ 3 independent draws and record $\widehat{R}(r_j; \theta_a)$
10: Form $\Delta_{aj} = \widehat{R}(r_j; \theta_a) - \widehat{R}(r_a; \theta_a)$
11: **end for**
12: Fit the descriptive slope $\widehat{\beta}_{\mathrm{LS}}$; do not call it a Lipschitz estimate
13: If reporting an envelope, calibrate $\widehat{q}_{.95}$ or $\widehat{q}_{\max}$ on a stated set and test transfer on a disjoint set
14: Report slope and intercept separately, next to the matched-rate metric (Table 2)
15: **Pre-register:** fix the seed-batch size, the decision tests, and a hard cap on seeds in advance; spend budget on training seeds before re-masking draws (Section 5.1)
16: **if** claiming a theorem-certified numerical bound **then**
17: Provide a global pointwise certificate for L_θ and population quantities or valid simultaneous confidence bounds
18: **end if**