DACE: DIVIDE-AND-COMPOSE EVIDENCE FOR MISSING-MODALITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal recognition must remain robust when modalities are unavailable at test time. Existing methods either reconstruct entire missing representations or fuse modalities without distinguishing transferable from modality-specific information. We introduce Divide-and-Compose Evidence (DACE), a lightweight recognition head that achieves robustness through learned evidence separation and selective cross-modal transport. For each observed modality, a complementary router divides its encoded feature into two branches: specific evidence that captures modality-unique cues and remains local, and shared evidence that expresses transferable cross-modal information. When a modality is missing, DACE recovers only its shared slot by transforming and averaging the shared branches from observed sources, leaving modality-specific content unestimated. The recovered and observed evidence is then composed and refined through availability-conditioned attention and observed-pair agreement. Experimental results demonstrate consistent robustness: across three benchmarks and 27 missingness conditions, DACE achieves the highest score in every comparison, and a single trained model maintains strong performance when test rates shift from 10% to 90%. On naturally incomplete medical data, DACE reaches 84.30% accuracy and 75.60% AUROC, demonstrating robustness on real-world incomplete medical data.

1 INTRODUCTION

Multimodal learning combines heterogeneous observations to form a more complete account of an event or object. Images provide appearance and spatial context, text supplies explicit semantics, and audio or clinical measurements contribute signals that may not be recoverable from either source alone (Baltrušaitis et al., 2019; Poria et al., 2017). Large-scale vision–language pretraining has strengthened this complementarity by placing modalities in related representation spaces (Radford et al., 2021). Most downstream systems, however, are developed under a modality-complete assumption that rarely survives deployment.

Inputs can be absent because of sensor failure, privacy restrictions, acquisition cost, or asynchronous collection. A classifier trained on complete multimodal inputs may therefore encounter incomplete observations at test time, and performance degrades severely when the missing source contains class-defining evidence (Ma et al., 2022). The history of missing-modality learning reflects three broad responses to this challenge. Early approaches address severe incompleteness through meta-learning or infer the absent latent representation from available modalities (Ma et al., 2021; Zhao et al., 2021). A second line learns shared and modality-specific spaces intended to remain useful under incomplete observations (Wang et al., 2023; Zhang et al., 2023). More recent methods specialize pretrained multimodal transformers for different availability patterns through missing-aware prompts and parameter-efficient adaptation (Lee et al., 2023; Hu et al., 2024; Zhang et al., 2025; Zhao et al., 2026). While these developments improve robustness, they leave open a fundamental design question: which evidence should be retained locally, which part is safe to transport across modalities, and how should these components be recombined?

Figure 1 provides examples for the above-mentioned issues. The examples expose two distinct consequences of a missing input. Evidence in an observed modality may be decisive for classification yet impossible to infer from other sources. At the same time, some shared evidence about the absent modality may be recoverable from available inputs. Feature fusion directly mixes modality-specific

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.

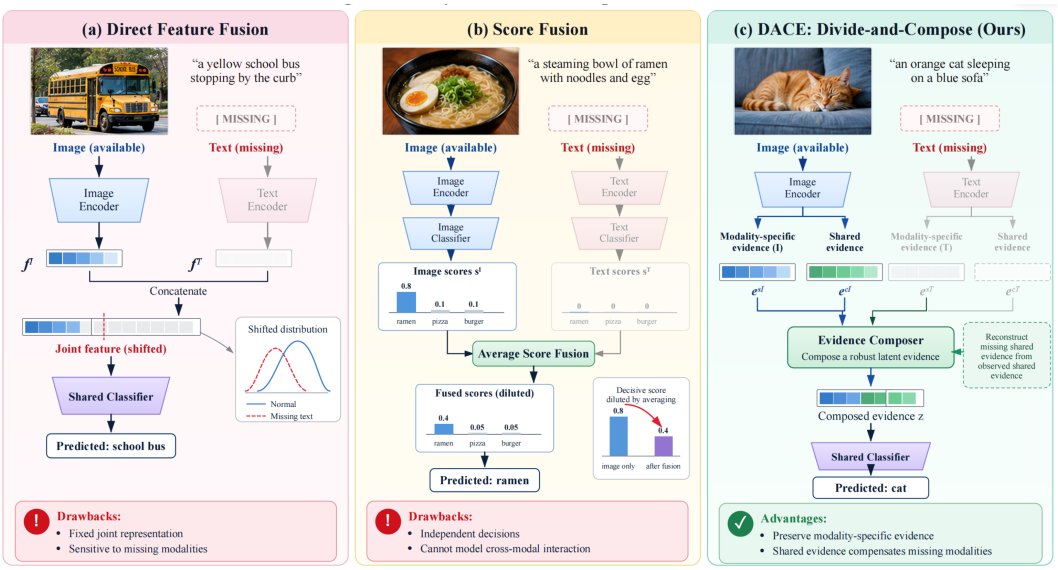


Figure 1: Motivation. Feature fusion entangles specific and shared information, while score fusion treats modalities as independent decisions. DACE divides every observed feature into complementary evidence, reconstructs only transferable shared evidence, and composes a representation conditioned on the observed set.

and shared information, so removing one input changes both the support and geometry presented to the classifier. Reconstructing a complete missing feature conflates these two cases and forces the model to estimate private content without a reliable source. Score fusion avoids this entanglement by keeping separate heads, but it combines decisions only after class prediction and discards cross-modal interactions when a modality is missing. The central question is therefore not how to fabricate a fully observed pair, but rather what to preserve from observed inputs and what can be reliably estimated from them.

DACE addresses this question through a *divide-and-compose* strategy. It divides each observed representation into complementary specific and shared evidence, recovers only the shared component of a missing modality through learned cross-modal transport, and composes the result according to the availability mask. Specifically, a complementary router divides each observed feature into specific evidence that remains local and shared evidence that can be transported across modalities. When a modality is missing, learned cross-modal maps transform and average the shared branches from observed sources to recover only its shared component, leaving modality-specific content unestimated. The composer combines observed specific branches with all shared evidence into a unified representation, giving the classifier access to both local cues and completed shared information. Mask-conditioned attention and observed-pair agreement add a bounded correction when more evidence is available. The resulting head can exploit observed private cues without treating cross-modal estimates as if they were observed. The complete pipeline is shown in Fig. 2. Our main contributions are summarized as follows:

- We propose DACE, a divide-and-compose recognition head that learns complementary evidence separation into specific and shared branches, and recovers only the transferable shared component of missing modalities through learned cross-modal transport, avoiding unreliable reconstruction of private content.
- We introduce availability-conditioned composition with norm-bounded refinement from mask-conditioned attention and observed-pair agreement, and establish conservation, missing-input invariance, and residual-boundedness properties.
- We evaluate DACE across 27 matched missingness conditions on three benchmarks, cross-rate single-model transfer from 10% to 90% test rates, and naturally incomplete medical data, demonstrating consistent robustness and advantages over fusion controls throughout.

2 Related Work

2.1 Multimodal fusion

Multimodal classifiers commonly combine representations before prediction or combine decisions from separate unimodal heads. Gated multimodal units learn feature-level weights (Arevalo et al., 2017), while Tensor Fusion Networks model explicit cross-modal products (Zadeh et al., 2017); low-rank fusion reduces the cost of such interactions (Liu et al., 2018). These designs describe how available inputs should interact, but their behavior can change sharply when an expected source disappears. Robustness studies of multimodal transformers confirm that complete-input training alone does not ensure reliable incomplete-input predictions (Ma et al., 2022). Feature- and score-level fusion are therefore useful reference points, but neither specifies which part of an observed representation should stand in for a missing source.

2.2 Learning with missing modalities

Work that explicitly targets incomplete inputs either adapts to missingness or estimates information from the surviving modalities. SMIL uses Bayesian meta-learning for severe missingness (Ma et al., 2021), and MMIN predicts missing representations for emotion recognition (Zhao et al., 2021). Shared–specific representation methods instead organize evidence into factors that can remain useful when the observed set changes (Wang et al., 2023; Zhang et al., 2023). These approaches highlight that not all information has the same transferability. For recognition, however, predicting a full absent representation is a stronger requirement than necessary: private content may not be identifiable from the available input. DACE keeps that content local and transports only a learned shared branch, judging the estimate by its contribution to classification.

2.3 Adapting pretrained multimodal models

Prompt and low-rank methods offer another route when strong pretrained encoders are available. CoOp, MaPLe, and DePT develop parameter-efficient prompting for vision–language recognition (Zhou et al., 2022; Khattak et al., 2023; Zhang et al., 2024). Missing-aware extensions condition prompts on availability: MMP uses missing-modality prompts (Lee et al., 2023), DCP models correlated prompts (Hu et al., 2024), and SyP combines static and dynamic prompts (Zhang et al., 2025). MoRA applies modality-common and modality-specific low-rank adaptation (Zhao et al., 2026). These methods adapt the pretrained representation; DACE instead leaves the encoders frozen and organizes their outputs into specific, transferable, and composed evidence within a downstream head. The distinction concerns where missingness is handled, not whether pretrained features are useful.

3 Method

3.1 Problem formulation and evidence encoding

Let $\mathcal{M} = \{1, \dots, M\}$ index the modalities, $\mathbf{o} \in \{0,1\}^M$ record their availability, and $\mathcal{O} = \{m : o_m = 1\} \neq \varnothing$ be the observed set. We predict a label from the available inputs $\{X_m : m \in \mathcal{O}\}$. For modality m, a frozen encoder E_m produces a feature $f_m \in \mathbb{R}^{D_m}$. To combine features with different dimensions and magnitudes, we first rescale them and learn projections into a common d-dimensional space:

$$\bar{f}_m = \sqrt{D_m}\frac{f_m}{\|f_m\|_2}, \qquad e_m = \phi(A_m \bar{f}_m) \in \mathbb{R}^d. \tag{1}$$

Here $A_m \in \mathbb{R}^{d \times D_m}$ is the projection matrix and ϕ is GELU. Projection is evaluated only for observed inputs, so absent modalities contribute no encoded evidence.

3.2 Divide: complementary evidence routing

An observed feature can contain both cues that transfer across modalities and cues specific to its source. Transporting the entire feature treats both kinds alike, even though missing modality-specific

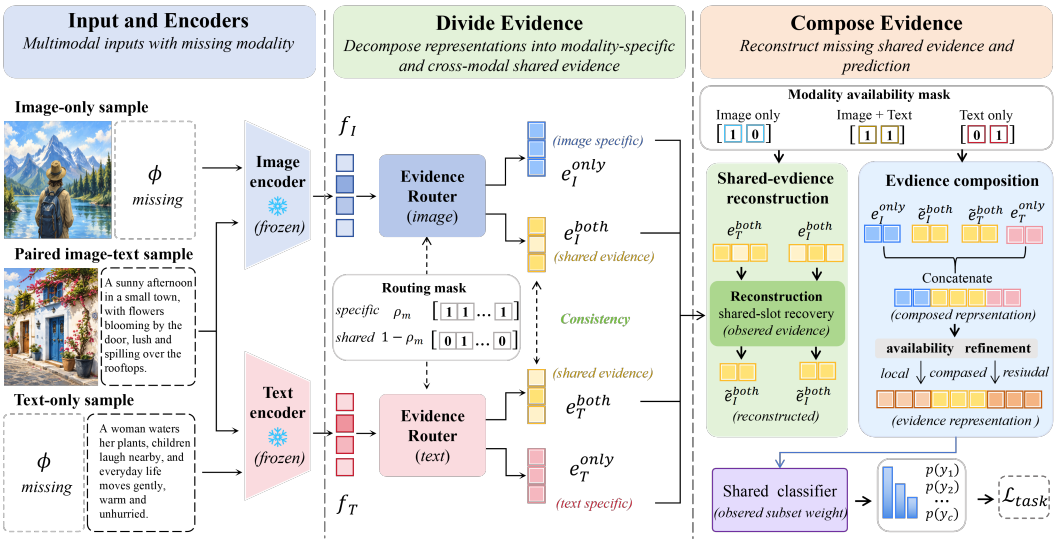


Figure 2: Overall Framework of Divide-and-Compose Evidence (DACE)

information may not be inferable. DACE therefore learns a complementary division: one branch retains specific evidence for prediction, while the other supplies shared evidence for cross-modal estimation. A modality-specific router assigns a gate to each feature coordinate:

$$\rho_m = \sigma(U_m \phi(V_m e_m)), \qquad e_m^{\mathrm{sp}} = \rho_m \odot e_m, \quad e_m^{\mathrm{sh}} = (1 - \rho_m) \odot e_m. \tag{2}$$

The learned matrices $V_m \in \mathbb{R}^{r \times d}$ and $U_m \in \mathbb{R}^{d \times r}$ compress the feature to bottleneck width $r < d$ and return d gating values. The elementwise sigmoid σ produces the routing mask $\rho_m \in (0,1)^d$, and $\odot$ denotes elementwise multiplication. The resulting e_m^{sp} and e_m^{sh} are the specific and shared branches. The router's output layer starts at zero, initially allocating half of each coordinate to each branch. Their sum remains exactly e_m throughout learning, preserving the input evidence without assuming a unique semantic decomposition.

3.3 Compose: Shared-Slot Recovery

After division, a missing modality leaves its shared-evidence slot empty. We estimate that slot from observed shared branches, leaving modality-specific evidence with its source. Because different modalities need not express shared cues in the same coordinates, a learned map $\mathcal{T}_{n \to m}(u) = u + C_{n \to m} \phi(D_{n \to m} u)$ transforms a shared vector $u \in \mathbb{R}^d$ from source n to target m. Here $D_{n \to m} \in \mathbb{R}^{r_t \times d}$ and $C_{n \to m} \in \mathbb{R}^{d \times r_t}$ form a bottleneck of width $r_t < d$; initializing $C_{n \to m}$ to zero makes the map start at the identity. Averaging the source estimates gives

$$\widehat{e}_m^{\mathrm{sh}} = \frac{1}{\max(1, |\mathcal{N}_m|)} \sum_{n \in \mathcal{N}_m} \mathcal{T}_{n \to m}(e_n^{\mathrm{sh}}), \tag{3}$$

where $\mathcal{N}_m = \mathcal{O} \setminus \{m\}$ contains the observed sources other than the target. When m is missing, this is the entire observed set. The completed slot $\widetilde{e}_m^{\mathrm{sh}}$ averages observed and estimated shared evidence when both exist and otherwise uses the available source:

$$\widetilde{e}_m^{\mathrm{sh}} = \begin{cases} \widehat{e}_m^{\mathrm{sh}}, & m \notin \mathcal{O}, \\ \frac{1}{2}(e_m^{\mathrm{sh}} + \widehat{e}_m^{\mathrm{sh}}), & m \in \mathcal{O},\ \mathcal{N}_m \neq \varnothing, \\ e_m^{\mathrm{sh}}, & m \in \mathcal{O},\ \mathcal{N}_m = \varnothing. \end{cases} \tag{4}$$

The last case retains an observed shared branch when it has no other source. To preserve the distinction between observed specific cues and completed shared cues, we pool them separately and concatenate the results:

$$e_{\mathcal{O}}^{\mathrm{cmp}} = \left[\frac{1}{|\mathcal{O}|} \sum_{m \in \mathcal{O}} e_m^{\mathrm{sp}} \,;\, \frac{1}{M} \sum_{m=1}^{M} \widetilde{e}_m^{\mathrm{sh}} \right] \in \mathbb{R}^{2d}. \tag{5}$$

3.4 EVIDENCE REFINEMENT AND PREDICTION

Uniform averaging gives a representation of fixed size, but treats eligible modalities equally even when their usefulness changes with the observed set. DACE therefore adjusts their contributions and adds pairwise evidence. To retain specific cues and provide a reference scale for this correction, we first form a local readout with learned maps $B_m \in \mathbb{R}^{r_\ell \times d}$, where r_ℓ is its width:

$$e_{\mathcal{O}}^{\mathrm{loc}} = \frac{1}{|\mathcal{O}|} \sum_{m \in \mathcal{O}} \phi(B_m e_m^{\mathrm{sp}}) \in \mathbb{R}^{r_\ell}. \tag{6}$$

To adapt modality contributions to the available inputs, a learned linear map of $\mathbf{o}$ produces two d-dimensional attention queries. One pools observed specific vectors into $a_{\mathcal{O}}^{\mathrm{sp}}$; the other pools all completed shared slots into $a_{\mathcal{O}}^{\mathrm{sh}}$. We add learned modality and branch embeddings to the vectors and apply softmax to query–vector inner products scaled by $1/\sqrt{d}$. The departure from uniform composition, $\delta = [a_{\mathcal{O}}^{\mathrm{sp}}; a_{\mathcal{O}}^{\mathrm{sh}}] - e_{\mathcal{O}}^{\mathrm{cmp}}$, supplies the correction $e_{\mathcal{O}}^{\mathrm{att}} = \mathrm{Bnd}((\delta_{1:r_p} + \delta_{d+1:d+r_p})/\sqrt{2})$. This combines the first $r_p \le d$ coordinates of each block; $\mathrm{Bnd}(v) = v/\max(1, \|v\|_2)$ limits the correction norm.

To also expose agreement between observed modalities, learned maps $P_m \in \mathbb{R}^{r_p \times D_m}$ project their features before elementwise multiplication. Averaging over the observed pairs $\mathcal{P}(\mathcal{O}) = \{(m,n) : m < n,\ m,n \in \mathcal{O}\}$ gives

$$e_{\mathcal{O}}^{\mathrm{pair}} = \mathrm{Bnd}\left(\frac{1}{\max(1, |\mathcal{P}(\mathcal{O})|)} \sum_{(m,n) \in \mathcal{P}(\mathcal{O})} \phi(P_m \bar{f}_m) \odot \phi(P_n \bar{f}_n) \right). \tag{7}$$

This term is zero for a singleton input. We combine the attention and pair corrections and limit their magnitude relative to local evidence:

$$e_{\mathcal{O}}^{\mathrm{res}} = \mathrm{Bnd}\left(\frac{e_{\mathcal{O}}^{\mathrm{att}} + e_{\mathcal{O}}^{\mathrm{pair}}}{1 + \mathbb{I}[|\mathcal{O}| > 1]} \right) \mathrm{sg}\big(\|e_{\mathcal{O}}^{\mathrm{loc}}\|_2\big), \tag{8}$$

where $\mathbb{I}$ is the indicator and sg stops gradients through the scale. The composed, local, and residual evidence form the final evidence representation $z_{\mathcal{O}}$, from which a shared classifier produces logits:

$$z_{\mathcal{O}} = [e_{\mathcal{O}}^{\mathrm{cmp}}; e_{\mathcal{O}}^{\mathrm{loc}}; e_{\mathcal{O}}^{\mathrm{res}}], \qquad \ell_{\mathcal{O}}(x_i; \theta) = b + W z_{\mathcal{O}}(x_i; \theta). \tag{9}$$

For K classes, $W \in \mathbb{R}^{K \times (2d + r_\ell + r_p)}$ and $b \in \mathbb{R}^K$ is the fixed empirical label prior. The local and residual columns of W start at zero.

For N labeled image–text inputs (x_i, y_i), we learn all head parameters θ with the task loss weighted by the probability $\mu_\eta(\mathcal{O})$ of each observed subset at missing rate η:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathcal{O}: \mu_\eta(\mathcal{O}) > 0} \mu_\eta(\mathcal{O}) \mathcal{L}_{\mathrm{task}}(\ell_{\mathcal{O}}(x_i; \theta), y_i). \tag{10}$$

$\mathcal{L}_{\mathrm{task}}$ is binary cross-entropy for multilabel prediction and cross-entropy for single-label prediction. Appendix B derives evidence conservation, missing-input invariance, and the bounds on shared-slot error and residual magnitude.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and metrics. We evaluate image–text recognition on MM-IMDb, UPMC Food-101, and Hateful Memes (Arevalo et al., 2017; Wang et al., 2015; Kiela et al., 2020). Their metrics are macro-F1 for 23-label genre prediction, top-1 accuracy for 101-way food classification, and AUROC for binary hateful-meme detection, respectively. We additionally study 12-month survival prediction on the naturally incomplete MMIST-ccRCC cohort (Mota et al., 2024), using CT, WSI, and 17 clinical fields. For this task we report accuracy, AUROC, positive-class AUPRC, and macro precision, recall, and F1 (mAP, mAR, mF1). The latter three are classification averages, not ranking average precision or recall. Split sizes and processing are given in Appendix C.

Missingness protocol. For a named missing case, an η fraction of examples contains only the retained modality and $1-\eta$ remains complete. In the mixed case, $\eta/2$ is image-only, $\eta/2$ is text-only, and $1-\eta$ is complete. The main comparison trains a separate model for every dataset, case, rate $\eta \in \{50\%, 70\%, 90\%\}$, and seed in $\{42, 43, 44\}$, then evaluates on the matched test condition. Validation scores select the trained models; test labels are not used for architecture or epoch selection.

Baselines. The published columns reproduce MoRA's comparison—CoOp (Zhou et al., 2022), MMP (Lee et al., 2023), MaPLe (Khattak et al., 2023), DePT (Zhang et al., 2024), DCP (Hu et al., 2024), and MoRA (Zhao et al., 2026)—and add reported SyP results (Zhang et al., 2025). Our two simple controls are *Feature Fusion* (FF), which concatenates normalized, zero-filled image–text features before a linear head, and *Score Fusion* (SF), which averages logits from available unimodal heads. On MMIST, FF concatenates the original zero-filled medical features. FF and SF use one sampled incomplete mask per training example; DACE evaluates its declared subset risk exactly. All image–text methods follow the same benchmark missingness protocol, while published and local scores retain their original experimental provenance. Local results are arithmetic means over seeds 42, 43, and 44; bold and underline mark the best and second-best displayed values per comparison, not significance.

Implementation details. For image–text data, the local methods use frozen OpenAI CLIP ViT-B/16-QuickGELU features (Radford et al., 2021). DACE uses $d = 512$, router rank 32, local-readout width 128, and pair rank 64; the Full head has fewer than one million trainable parameters. AdamW uses learning rate 10^{-3} and weight decay 10^{-2} (Loshchilov & Hutter, 2019). We train for at most 50 epochs with patience 3 and batch sizes 64/256/32 for MM-IMDb/Food-101/Hateful Memes. MM-IMDb uses unweighted binary cross-entropy and threshold 0.5; the others use cross-entropy. We run the experiments with CUDA on an NVIDIA A100-SXM4-80GB GPU.

For MMIST, we follow the feature organization of AdaCoMed (Chen et al., 2025) and use balanced training-split class weights. Adam runs with learning rate 10^{-4}, batch size 64, at most 100 epochs, patience 10, and seeds 42/43/44. Mean validation AUROC over natural and leave-one-modality-out views selects the trained model; Appendix C gives split and availability details.

4.2 Main Comparison

Table 1 presents the comparison results under 3 missing patterns and 3 missing ratios across 3 different image-text datasets. DACE achieves the highest displayed score in every condition. Its nine-condition means are 54.98, 85.42, and 72.68 on MM-IMDb, Food-101, and Hateful Memes, respectively. Relative to the stronger of FF and SF on each dataset, the gains are 3.86, 1.07, and 3.43 points. The gap is largest on Hateful Memes, where cross-modal interactions are critical for detecting hateful content that neither modality reveals independently. Food-101 shows smaller gains because its visual category cues enable strong image-only predictions. These results quantify the benefit of DACE's divide-and-compose strategy over direct fusion under shared frozen features.

Against published methods, DACE's mean exceeds MoRA's by 0.99, 1.02, and 2.93 points, respectively. The improvements are most pronounced under severe missingness. At mixed/both 90%, DACE scores 52.50 macro-F1, 82.81 accuracy, and 71.80 AUROC, versus the next best 51.64, 81.50, and 68.93. The modality asymmetry reveals task-dependent characteristics. At 90% missingness, removing text from Food-101 gives 78.89 accuracy while removing images gives 87.15, confirming that food classification relies heavily on visual appearance. For Hateful Memes, the corresponding AUROCs are 72.84 and 68.93, indicating more balanced modality contributions where text provides crucial context for interpreting potentially hateful imagery.

Performance degradation from 50% to 90% missingness further exposes task structure. Under text missingness, DACE falls by 4.54 points on MM-IMDb, 6.51 on Food-101, and 1.54 on Hateful Memes. Under image missingness, the corresponding drops are 4.33, 2.88, and 4.64 points. Food-101 degrades sharply without images but remains robust without text, while Hateful Memes shows similar sensitivity to either modality. The margin over prior methods also varies. DACE exceeds the strongest published score by 4.54 points at mixed 50% on Hateful Memes, but only 0.15 points at image-missing 90%, where the surviving text modality provides limited discriminative information. This pattern suggests that DACE's advantage is largest when both modalities contribute complementary evidence that can be effectively separated and recombined.

Table 1: Matched-condition recognition scores (%): MM-IMDb macro-F1, Food-101 accuracy, and Hateful Memes AUROC. I/T denotes the observed image/text percentage.

Dataset	η	I/T	CoOp	MMP	MaPLe	DePT	DCP	SyP	MoRA	FF	SF	DACE
MM-IMDb	50%	100/50	48.06	48.88	49.58	50.64	52.13	53.90	54.62	52.04	51.43	**55.97**
		50/100	49.89	51.46	52.32	52.78	54.32	56.27	57.61	54.47	52.91	**58.67**
		75/75	48.37	49.32	49.56	50.87	52.32	55.02	55.88	52.84	50.71	**56.84**
	70%	100/30	44.13	45.64	45.52	46.38	48.52	51.37	52.56	50.69	48.54	**53.81**
		30/100	48.82	50.52	50.64	52.13	53.14	54.20	56.39	53.09	51.77	**56.83**
		65/65	46.84	48.12	49.16	50.32	51.42	52.90	52.97	50.50	49.44	**54.42**
	90%	100/10	44.76	45.32	46.84	47.56	49.26	50.21	50.67	47.79	46.51	**51.43**
		10/100	48.32	49.12	50.13	50.88	52.22	53.72	53.57	50.67	50.24	**54.34**
		55/55	44.12	44.87	45.12	46.54	48.04	49.63	51.64	47.94	47.66	**52.50**
Food-101	50%	100/50	77.45	77.89	79.64	80.16	82.11	83.20	84.41	84.70	84.75	**85.40**
		50/100	87.02	87.16	87.35	82.14	89.12	89.64	89.63	89.34	89.23	**90.03**
		75/75	81.24	81.72	82.34	83.12	85.24	86.17	86.68	86.69	86.58	**87.94**
	70%	100/30	76.34	76.52	77.02	77.34	78.87	79.56	80.85	81.33	81.26	**82.15**
		30/100	84.78	85.64	85.89	86.12	87.32	88.67	88.01	87.70	87.68	**88.96**
		65/65	78.87	79.12	79.84	81.46	81.87	82.45	83.77	83.91	84.06	**85.44**
	90%	100/10	71.87	73.14	73.46	74.12	75.26	76.33	78.41	78.02	77.98	**78.89**
		10/100	81.67	82.14	83.12	83.56	85.78	86.41	86.77	85.92	86.05	**87.15**
		55/55	76.46	76.58	77.85	78.12	79.87	81.03	81.09	81.50	81.43	**82.81**
Hateful Memes	50%	100/50	60.56	60.31	60.87	61.87	62.32	68.25	70.66	70.86	69.99	**74.38**
		50/100	62.41	62.35	63.13	63.88	64.46	66.80	71.58	69.67	69.24	**73.57**
		75/75	64.87	65.84	65.46	65.86	66.02	68.16	69.58	68.94	68.52	**74.12**
	70%	100/30	60.74	61.12	61.26	61.56	62.82	68.94	69.43	70.21	69.73	**73.95**
		30/100	62.74	63.24	63.14	63.48	64.12	66.98	70.68	69.04	68.36	**71.26**
		65/65	64.82	65.04	65.23	65.48	66.08	68.42	70.15	69.52	68.94	**73.23**
	90%	100/10	60.03	57.21	60.74	61.14	62.08	69.70	68.52	69.71	69.68	**72.84**
		10/100	61.46	61.52	61.87	62.42	63.87	64.54	68.78	67.55	67.12	**68.93**
		55/55	64.32	63.34	64.85	65.37	66.78	68.93	68.37	67.70	67.68	**71.80**

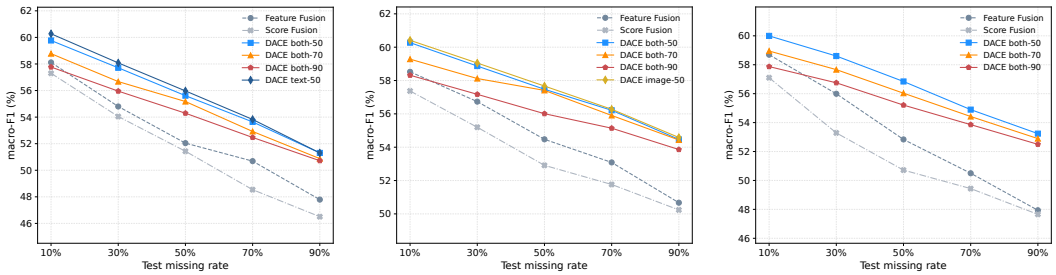
Table 2: Single-model transfer (%). Using a single model trained with 70% both/mixed missingness, compare the performance of different missingness conditions with that of models trained at their matched missingness rates. $\Delta = \text{Single} - \text{Matched}$, computed from the displayed scores.

		MM-IMDb			Food-101			Hateful Memes		
Test case	η_{test}	Matched	Single	Δ	Matched	Single	Δ	Matched	Single	Δ
Text	50%	55.97	55.17	−0.80	85.40	85.89	+0.49	74.38	74.81	+0.43
Image	50%	58.67	57.81	−0.86	90.03	90.24	+0.21	73.57	72.97	−0.60
Mixed	50%	56.84	56.03	−0.81	87.94	87.93	−0.01	74.12	73.46	−0.66
Text	70%	53.81	52.92	−0.89	82.15	82.26	+0.11	73.95	74.25	+0.30
Image	70%	56.83	55.90	−0.93	88.96	88.60	−0.36	71.26	71.10	−0.16
Mixed	70%	54.42	54.42	0.00	85.44	85.44	0.00	73.23	73.23	0.00
Text	90%	51.43	51.28	−0.15	78.89	78.96	+0.07	72.84	73.73	+0.89
Image	90%	54.34	54.43	+0.09	87.15	86.82	−0.33	68.93	68.88	−0.05
Mixed	90%	52.50	52.89	+0.39	82.81	82.80	−0.01	71.80	71.78	−0.02
Mean	–	54.98	54.54	−0.44	85.42	85.44	+0.02	72.68	72.69	+0.01

4.3 Robustness and Cross-Condition Transfer

Table 2 evaluates one model trained on mixed 70% missingness against models trained separately for each test condition, demonstrating DACE's remarkable cross-condition robustness. Reusing one model changes the nine-condition mean by only -0.44 macro-F1 on MM-IMDb, $+0.02$ accuracy on Food-101, and $+0.01$ AUROC on Hateful Memes, showing that a single DACE model achieves nearly equivalent performance to condition-specific training. On MM-IMDb, the single model actually improves by 0.39 points under mixed 90% missingness, demonstrating that DACE's learned evidence decomposition generalizes effectively to severe incompleteness. Food-101 shows exceptional stability, with all deviations within ± 0.49 points across nine conditions, highlighting DACE's ability to maintain robust performance regardless of which modality is missing. On Hateful Memes, the single model achieves substantial gains of 0.89 points for text missing at 90%, confirming DACE's capacity to leverage learned cross-modal transport across different availability patterns.

Figure 3 demonstrates DACE's superior robustness by holding each MM-IMDb model fixed while varying the test missing rate from 10% to 90%. DACE maintains a clear advantage over both fusion



(a) Text missing. (b) Image missing. (c) Both missing.

Figure 3: MM-IMDb macro-F1 under changing test missing rates. Each DACE curve uses one model trained at the rate indicated by its legend. "Both missing" means "Mixed missing"

Table 3: MMIST-ccRCC full test set ($n = 121$): natural and leave-one-modality-out evaluation (%).

Condition	Method	Acc	AUROC	AUPRC	mAP	mAR	mF1
Natural	FF	75.76	73.10	93.21	60.20	66.61	61.09
	SF	66.12	68.73	93.71	58.64	64.58	55.41
	DACE	**84.30**	**75.60**	**94.01**	**66.35**	**67.11**	**66.60**
Without CT	FF	60.06	72.76	93.27	58.04	67.27	52.62
	SF	44.63	64.21	90.68	55.47	60.15	40.70
	DACE	**76.58**	**75.62**	**95.03**	**61.74**	**69.73**	**62.98**
Without WSI	FF	80.17	73.35	93.25	63.22	64.73	62.73
	SF	65.56	69.03	93.45	57.01	62.50	54.24
	DACE	**82.92**	**73.53**	**93.70**	**65.80**	**69.85**	**67.33**
Without clinical	FF	55.10	60.81	90.43	53.00	56.47	46.51
	SF	70.80	62.44	90.88	52.28	54.92	50.83
	DACE	**72.73**	**66.25**	**91.90**	**55.02**	**56.91**	**53.98**

controls across all patterns and test rates, with the performance gap widening as missingness becomes more severe. At 90% missingness, DACE achieves approximately 51.4 macro-F1 for missing text, 54.5 for missing image, and 53.2 for mixed conditions, consistently outperforming FF and SF by substantial margins. Models trained at 50% missingness often achieve the strongest performance even at drastically different test rates, confirming that DACE's divide-and-compose strategy enables effective cross-rate transfer. This practical reuse capability represents a significant advantage over methods requiring retraining for each deployment scenario.

Naturally incomplete medical data. DACE achieves the highest accuracy, AUROC, AUPRC, and macro-F1 across all four conditions of Table 3, demonstrating robust performance on real-world incomplete data. Its AUROC exceeds the stronger fusion control by 2.50 points under natural availability, 2.86 without CT, 0.18 without WSI, and 3.81 without clinical data, with particularly strong advantages when critical modalities are absent. Clinical information proves most influential, yet DACE maintains reasonable performance even when it is removed, achieving 66.25 AUROC and 53.98 macro-F1 compared to 75.60 and 66.60 with all modalities. Notably, DACE's macro-F1 remains at 67.33 when WSI is removed, nearly matching the natural-availability result and demonstrating effective compensation through its divide-and-compose mechanism. Appendix E provides additional CT subgroup analysis.

4.4 Ablation Study

Table 4 follows the implemented progression under mixed 50% missingness, revealing the contribution of each DACE component. Direct averages projected evidence without routing. Divide introduces complementary specific and shared branches. Compose additionally transports shared evidence into missing slots, and Full adds mask-conditioned attention and observed-pair agreement. The results confirm that evidence division is the key innovation. Introducing the division delivers the largest gain on every dataset, improving Direct by 4.28, 0.46, and 5.24 points on MM-IMDb,

Table 4: Mixed-50% ablation.Right: modules active in the DACE variants.

Performance

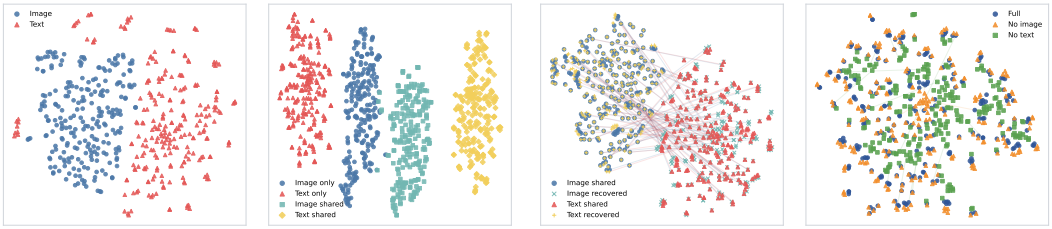
Method	MM-IMDb	Food-101	Hateful
Feature Fusion	52.84	86.69	71.13
Score Fusion	50.71	86.58	68.52
Direct	51.63	87.23	67.77
Divide	55.91	87.69	73.01
Compose	56.25	87.83	73.40
DACE (Full)	**56.84**	**87.94**	**74.12**

Active components

Variant	Divide	Recover	Refine
Direct	–	–	–
Divide	✓	–	–
Compose	✓	✓	–
Full	✓	✓	✓

Food-101, and Hateful Memes respectively. This validates DACE's core principle of preserving local cues while restricting cross-modal transport to shared evidence.

The later stages provide smaller but consistent gains, demonstrating that each component contributes to DACE's overall superiority. Compose improves Divide by 0.34, 0.14, and 0.39 points, while Full adds another 0.59, 0.11, and 0.72 points. The larger gains on Hateful Memes reflect greater value from availability-conditioned refinement when cross-modal interactions are critical, while Food-101's modest increments align with its strong image-only performance that limits potential for cross-modal recovery.



(a) Encoded evidence. (b) Specific/shared split. (c) Shared recovery. (d) Composed evidence.

Figure 4: Food-101 Visualization of evidence at different stages.

4.5 Evidence Visualization

Figure 4 visualizes four stages of Food-101 evidence using t-SNE. The encoded image and text features form separated clouds (Fig. 4a). After division, each modality produces distinct specific and shared evidence (Fig. 4b), confirming successful complementary decomposition. Recovered shared representations fall close to observed shared regions (Fig. 4c), demonstrating effective cross-modal transport. Most notably, the final composed representation (Fig. 4d) reveals DACE's key achievement: complete, no-text, and no-image observations form a tightly integrated distribution, contrasting sharply with the separated clusters in Fig. 4a. The three conditions intermingle as if from a single unified modality, demonstrating that DACE successfully bridges the modality gap and normalizes representations across availability conditions, enabling robust classification regardless of which modalities are observed.

5 Conclusion

We present DACE, a divide-and-compose recognition head for robust multimodal learning under missing modalities. DACE learns a complementary division of each observed feature into specific evidence that remains local and shared evidence that can be transported; only the shared slot of a missing modality is recovered through learned cross-modal maps, avoiding unreliable reconstruction of private content. The composed representation is refined with mask-conditioned attention and observed-pair agreement. Experimental results demonstrate consistent robustness: across 27 conditions on three image–text benchmarks, DACE ranked first in every comparison. On naturally incomplete medical data, DACE achieved 84.30% accuracy and 75.60% AUROC with advantages of 2.50 to 3.81 AUROC points over fusion controls across leave-one-modality-out conditions, establishing robustness within the evaluated tasks and missingness patterns.

AI USE STATEMENT

During manuscript preparation, we use a large language model as a general writing assistant for grammar checking, sentence restructuring, and improving the readability of technical and experimental descriptions. It does not originate the core research idea, experimental design, or technical innovations. The authors independently check the final text, cited sources, mathematical arguments, and reported results, and take responsibility for the accuracy and completeness of all claims and artifacts in this submission.

ETHICS STATEMENT

The Hateful Memes benchmark contains harmful language and imagery; performance on it should not be taken as evidence of fair or deployment-ready content moderation. MMIST-ccRCC contains sensitive medical information, and our small-cohort analysis is not a clinical validation or a recommendation for patient care. We use the datasets only for the stated research evaluations and do not release patient-level data or new personal information.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and objective, and Section 4.1 describes datasets, missingness protocols, metrics, training settings, validation-based model selection, and the provenance of published and local results. Appendices A–F provide implementation details, complete derivations, dataset descriptions, parameter counts, and additional robustness results. Reported local results are arithmetic means over the stated three seeds.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. URL https://arxiv.org/abs/1702.01992.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Wanyi Chen, Zihua Zhao, Jiangchao Yao, Ya Zhang, Jiajun Bu, and Haishuai Wang. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30763–30773, 2025.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 67446–67466, 2024. doi: 10.52202/079017-2153.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

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, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18156–18165. IEEE, 2022.

Tiago Mota, M. Rita Verdelho, Diogo J. Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. MMIST-ccRCC: A real world medical dataset for the development of multi-modal systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2395–2403, 2024.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. doi: 10.1016/j.inffus.2017.02.003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6, 2015. doi: 10.1109/ICMEW.2015.7169757.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. DePT: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2024.

Yunhua Zhang, Hazel Doughty, and Cees Snoek. Learning unseen modality interaction. volume 36, pp. 54716–54726, 2023.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1881–1890, 2025.

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*, pp. 2608–2618, 2021. doi: 10.18653/v1/2021.acl-long.203.

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. Mora: Missing modality low-rank adaptation for visual recognition. In *International Conference on Learning Representations*, volume 2026, pp. 68635–68646, 2026.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision–language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A Implementation details of evidence composition

The router in Equation (2) uses $V_m \in \mathbb{R}^{r\times d}$ and $U_m \in \mathbb{R}^{d\times r}$. Each transport map is $\mathcal{T}_{n\to m}(u) = u + C_{n\to m}\phi(D_{n\to m}u)$ and starts at the identity. When $\mathcal{N}_m = \varnothing$, the only possible nonempty observation is modality m itself, for which Equation (4) retains the observed shared branch.

The two mask queries start at zero, so attention initially assigns equal weight within each eligible pool. The modality/type embeddings distinguish token identity and branch type even when the mask changes. In Equation (7), $P_m \in \mathbb{R}^{r_p\times D_m}$ is learned for each modality and the pair mean is zero when no pair is observed. The outer clipping in Equation (8) limits the combined direction to unit norm before multiplication by the detached local-evidence norm.

B Detailed theoretical analysis

This appendix makes explicit the deterministic guarantees used in Section 3. We use the Euclidean norm for vectors and the induced spectral norm for matrices. The statements hold for any learned parameter values; they do not require statistical independence between the routed branches.

B.1 Division conserves evidence

Proposition 1 (Conservative division)**.** *For every e_m and $\rho_m \in [0,1]^d$, Equation* (2) *reconstructs e_m exactly and obeys*

$$\tfrac{1}{2}\|e_m\|_2^2 \le \|e_m^{\mathrm{sp}}\|_2^2 + \|e_m^{\mathrm{sh}}\|_2^2 = \|e_m\|_2^2 - 2\sum_j \rho_{m,j}(1-\rho_{m,j})e_{m,j}^2 \le \|e_m\|_2^2. \tag{11}$$

Proof of Proposition 1. For every coordinate j,

$$e_{m,j}^{\mathrm{sp}} + e_{m,j}^{\mathrm{sh}} = \rho_{m,j}e_{m,j} + (1-\rho_{m,j})e_{m,j} = e_{m,j}, \tag{12}$$

which proves exact reconstruction. The routed energy is

$$\|e_m^{\mathrm{sp}}\|_2^2 + \|e_m^{\mathrm{sh}}\|_2^2 = \sum_j \left[\rho_{m,j}^2 + (1-\rho_{m,j})^2\right]e_{m,j}^2 \tag{13}$$

$$= \|e_m\|_2^2 - 2\sum_j \rho_{m,j}(1-\rho_{m,j})e_{m,j}^2. \tag{14}$$

Because $0 \le \rho(1-\rho) \le 1/4$ for $\rho \in [0,1]$, each coefficient $\rho^2 + (1-\rho)^2$ lies in $[1/2, 1]$. Multiplication by the nonnegative $e_{m,j}^2$ and summation yield both inequalities. □

Corollary 2 (Balanced start)**.** *At the prescribed initialization $\rho_m = \frac{1}{2}\mathbf{1}$, $e_m^{\mathrm{sp}} = e_m^{\mathrm{sh}} = \frac{1}{2}e_m$ and the total routed energy equals $\frac{1}{2}\|e_m\|_2^2$.*

This initialization is symmetric: neither branch receives a larger signal before the data update the router. The energy reduction concerns the concatenated branch norm, not information loss, because their sum still reconstructs e_m exactly.

B.2 Composition averages transport error

Proposition 3 (Stable shared-slot recovery)**.** *Let s_m be an ideal shared target and suppose $\|\mathcal{T}_{n\to m}(e_n^{\mathrm{sh}}) - s_m\|_2 \le \epsilon_{n\to m}$. For nonempty $\mathcal{N}_m$, the missing-slot error is at most the mean source error. If m is observed, the reconciled-slot error is at most one half of that mean plus $\frac{1}{2}\|e_m^{\mathrm{sh}} - s_m\|_2$.*

Proof of Proposition 3. Let $r_{n\to m} = \mathcal{T}_{n\to m}(e_n^{\mathrm{sh}}) - s_m$. If m is missing, Equations (3)–(4) give

$$\|\widetilde{e}_m^{\mathrm{sh}} - s_m\|_2 = \left\|\frac{1}{|\mathcal{N}_m|}\sum_{n\in\mathcal{N}_m} r_{n\to m}\right\|_2 \tag{15}$$

$$\le \frac{1}{|\mathcal{N}_m|}\sum_{n\in\mathcal{N}_m}\|r_{n\to m}\|_2 \le \frac{1}{|\mathcal{N}_m|}\sum_{n\in\mathcal{N}_m}\epsilon_{n\to m}, \tag{16}$$

where the first inequality is the triangle inequality. If m is observed and $\mathcal{N}_m$ is nonempty, subtracting s_m from Equation (4) and applying the same argument gives

$$\|\widetilde{e}_m^{\mathrm{sh}} - s_m\|_2 \le \tfrac{1}{2}\|e_m^{\mathrm{sh}} - s_m\|_2 + \frac{1}{2|\mathcal{N}_m|}\sum_{n\in\mathcal{N}_m}\epsilon_{n\to m}. \tag{17}$$

Thus the error bound depends on the average source error and does not grow linearly with the number of available sources. □

Corollary 4 (Composed-representation error). *Let $e_{\mathcal{O}}^{\mathrm{cmp},\star}$ replace each shared slot in Equation (5) by its ideal target s_m while leaving the specific block fixed. If $\delta_m = \|\widetilde{e}_m^{\mathrm{sh}} - s_m\|_2$, then*

$$\|e_{\mathcal{O}}^{\mathrm{cmp}} - e_{\mathcal{O}}^{\mathrm{cmp},\star}\|_2 \le \frac{1}{M}\sum_{m=1}^{M}\delta_m, \qquad \|W_{\mathrm{cmp}}(e_{\mathcal{O}}^{\mathrm{cmp}} - e_{\mathcal{O}}^{\mathrm{cmp},\star})\|_2 \le \frac{\|W_{\mathrm{cmp}}\|_2}{M}\sum_{m}\delta_m. \tag{18}$$

The first inequality follows from the shared-slot mean and the triangle inequality; the second follows from submultiplicativity. This result clarifies the role of Compose: it offers controlled aggregation of transferable shared evidence, not exact reconstruction of an unobserved modality.

B.3 Mask correctness and set order

Proposition 5 (Missing-input invariance). *Fix* $\mathbf{o}$. *If two examples agree on all f_m with $m \in \mathcal{O}$, DACE produces identical logits regardless of values stored outside $\mathcal{O}$.*

Proof of Proposition 5. Consider two inputs that coincide on every feature f_m with $m \in \mathcal{O}$. Projection and division are evaluated only for these observed inputs, hence their specific and shared rows coincide. Every missing shared slot is computed solely from $\{e_n^{\mathrm{sh}} : n \in \mathcal{O}\}$; the local mean and specific attention contain only observed rows; the shared attention contains observed or recovered slots; and Equation (7) contains only pairs in $\mathcal{P}(\mathcal{O})$. The mask query is fixed because $\mathbf{o}$ is fixed. Following the recovered slots through composition, refinement, and the shared classifier therefore gives identical intermediate representations and logits. □

Proposition 6 (Enumeration-order invariance). *Reordering the internal enumeration of observed modalities or observed pairs, while retaining each modality label and its parameters, leaves the DACE logits unchanged.*

Proof. The local readout, recovery operator, shared-slot mean, and pair term are sums or means and hence commute with a permutation of their operands. In dot-product attention, jointly permuting key–value rows only permutes the softmax weights; their weighted sum is unchanged. Modality tokens and maps remain attached to their labels, so all subsequent concatenated blocks are unchanged. □

B.4 Bounded refinement and loss perturbation

Proposition 7 (Bounded refinement). *For $W = [W_{\mathrm{cmp}}, W_{\mathrm{loc}}, W_{\mathrm{res}}]$, Equation (8) guarantees $\|e_{\mathcal{O}}^{\mathrm{res}}\|_2 \le \|e_{\mathcal{O}}^{\mathrm{loc}}\|_2$ and $\|W_{\mathrm{res}}e_{\mathcal{O}}^{\mathrm{res}}\|_2 \le \|W_{\mathrm{res}}\|_2\|e_{\mathcal{O}}^{\mathrm{loc}}\|_2$.*

Lemma 8 (Non-expansive clipping). *$\mathrm{Bnd}(v) = v/\max(1, \|v\|_2)$ is the Euclidean projection onto the closed unit ball. Consequently, $\|\mathrm{Bnd}(v)\|_2 \le 1$ and $\|\mathrm{Bnd}(u) - \mathrm{Bnd}(v)\|_2 \le \|u - v\|_2$.*

Proof. If $\|v\|_2 \le 1$, the closest point in the unit ball is v; otherwise it is the radial point $v/\|v\|_2$. Thus Bnd is the metric projection onto a nonempty closed convex set. Euclidean projections onto closed convex sets are non-expansive, which yields the second claim; the first is immediate from the definition. □

Proof of Proposition 7. Let $u_{\mathcal{O}}$ denote the argument of the outer Bnd in Equation (8). Lemma 8 gives

$$\|e_{\mathcal{O}}^{\mathrm{res}}\|_2 = \|\mathrm{Bnd}(u_{\mathcal{O}})\|_2\,\mathrm{sg}(\|e_{\mathcal{O}}^{\mathrm{loc}}\|_2) \le \|e_{\mathcal{O}}^{\mathrm{loc}}\|_2. \tag{19}$$

Applying $\|Av\|_2 \le \|A\|_2\|v\|_2$ with $A = W_{\mathrm{res}}$ proves the logit bound. Stop-gradient changes derivatives but not the forward value, so it does not alter either inequality. □

Corollary 9 (Task-loss perturbation). *Let $\ell_{\mathcal{O}}^{-\mathrm{res}}$ be the logits with the residual block removed. If $\mathcal{L}_{\mathrm{task}}(\cdot, y)$ is L_y-Lipschitz in logits on the evaluated domain, then*

$$\left|\mathcal{L}_{\mathrm{task}}(\ell_{\mathcal{O}}, y) - \mathcal{L}_{\mathrm{task}}(\ell_{\mathcal{O}}^{-\mathrm{res}}, y)\right| \le L_y \|W_{\mathrm{res}}\|_2 \|e_{\mathcal{O}}^{\mathrm{loc}}\|_2. \tag{20}$$

This follows by the Lipschitz property and Proposition 7. The bound is conditional on L_y over the realized logit region; it is not a universal boundedness claim for unconstrained cross-entropy logits.

B.5 Exact subset risk and missingness shift

For a fixed parameter vector, define the empirical loss under subset $\mathcal{O}$ as $g_\theta(\mathcal{O}) = N^{-1} \sum_i \mathcal{L}_{\mathrm{task}}(\ell_{\mathcal{O}}(x_i; \theta), y_i)$ and let $R_\mu(\theta) = \sum_{\mathcal{O}} \mu(\mathcal{O}) g_\theta(\mathcal{O})$.

Theorem 10 (Subset-risk identity and shift bound). *Equation* (10) *equals $\mathbb{E}_{\mathcal{O}\sim\mu_\eta}[g_\theta(\mathcal{O})]$ exactly. Wherever the finite terms are differentiable, $\nabla R_{\mu_\eta} = \mathbb{E}_{\mathcal{O}\sim\mu_\eta}[\nabla g_\theta(\mathcal{O})]$. Moreover, for any test mask law ν on the same nonempty subsets, if $0 \le g_\theta(\mathcal{O}) \le B$, then*

$$|R_\nu(\theta) - R_{\mu_\eta}(\theta)| \le B \ \mathrm{TV}(\nu, \mu_\eta), \qquad \mathrm{TV}(\nu, \mu_\eta) = \tfrac{1}{2} \sum_{\mathcal{O}} |\nu(\mathcal{O}) - \mu_\eta(\mathcal{O})|. \tag{21}$$

Proof. The expectation identity follows by expanding the expectation over the finite support of μ_η, yielding Equation (10). The gradient identity follows by differentiating this finite weighted sum. For the shift bound, set $\Delta(\mathcal{O}) = \nu(\mathcal{O}) - \mu_\eta(\mathcal{O})$ and let $S_+ = \{\mathcal{O} : \Delta(\mathcal{O}) \ge 0\}$. Since both laws sum to one, $\sum_{S_+} \Delta = \sum_{\mathcal{O}\notin S_+} (-\Delta) = \mathrm{TV}(\nu, \mu_\eta)$. Hence the positive and negative parts of $\sum_{\mathcal{O}} \Delta(\mathcal{O}) g_\theta(\mathcal{O})$ each lie in $[0, B\,\mathrm{TV}(\nu, \mu_\eta)]$, and their difference has magnitude at most $B\,\mathrm{TV}(\nu, \mu_\eta)$. □

The theorem distinguishes objective fidelity from transfer: enumerating the declared training law removes mask-sampling error, but a large train–test mask shift can still change risk. This is why Section 4.3 evaluates a single trained model under missingness patterns without per-pattern retraining.

C Datasets and preprocessing

MM-IMDb. The movie-genre benchmark uses 15,552 training, 2,608 validation, and 7,799 test movie IDs. We predict the 23 genres specified by the benchmark configuration with independent binary outputs and report the unweighted mean of per-label F1 scores. Each available plot entry is encoded as a text example paired with its movie poster; thus a movie with multiple plot entries contributes multiple feature rows while its split assignment and genre vector remain fixed. This distinction between movie IDs and plot rows matters when reproducing cached-feature training.

UPMC Food-101 and Hateful Memes. Food-101 contains 61,127/6,845/22,716 train/validation/test image–text instances over 101 food categories. An image is paired with its associated recipe text; accuracy is the fraction of correct top-1 predictions. Hateful Memes contains 8,500/500/1,000 train/validation/test image–text memes. We use its binary labels and report positive-class AUROC. For all three image–text tasks, images are converted to RGB and passed through the frozen CLIP image encoder; text is tokenized at CLIP's 77-token context length. Encoder outputs are cached before training the recognition heads. The source splits are kept fixed, and missingness masks are generated only within those splits.

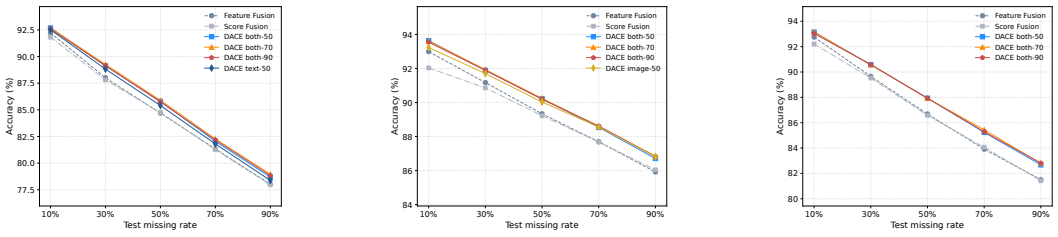
MMIST-ccRCC. The dataset contains 618 patients and a published 497/121 train/test split. We split the 497 training patients into 397 fit and 100 validation patients using a fixed stratified split. Our selected modalities are a 512-dimensional CT feature, a 1,024-dimensional WSI feature, and one-hot encodings of 17 clinical fields; the latter vocabulary is fitted on the fit split only. WSI and clinical information are present for all patients, while the selected CT index covers 239. Evaluation uses the natural availability mask and leave-one-modality-out masks. The fixed CT-observed test cohort has 58 patients (54 positive, four negative) and is evaluated both with and without CT. These small and imbalanced groups motivate reporting accuracy, AUROC, AUPRC, and macro precision/recall/F1 together.

Table 5: Trainable parameters of the frozen-feature heads (millions). Encoders are excluded; each dataset uses the same head design with its own number of output classes.

Method	MM-IMDb	Food-101	Hateful Memes
FF	0.023575	0.103525	0.002050
SF	0.023598	0.103626	0.002052
Direct	0.027968	0.122816	0.002432
Divide	0.749888	0.844736	0.724352
Compose	0.815424	0.910272	0.789888
DACE (Full)	0.886080	0.980928	0.860544

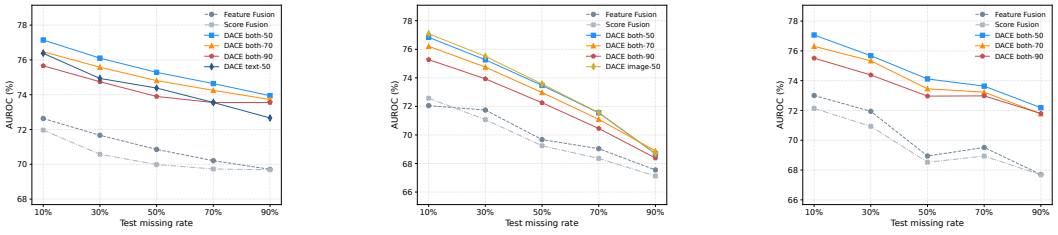
Table 6: MMIST-ccRCC CT-observed cohort ($n = 58$): original inputs versus CT withheld from the same patients (%).

Condition	Method	Acc	AUROC	AUPRC	mAP	mAR	mF1
CT observed	FF	92.53	71.14	94.88	46.53	49.69	48.06
	SF	89.08	70.68	**97.42**	50.02	55.56	51.44
	DACE	**92.95**	**72.22**	95.59	**59.53**	**60.96**	**60.03**
CT withheld	FF	59.77	77.93	98.16	56.40	74.54	48.37
	SF	44.25	72.38	97.66	54.84	66.20	37.38
	DACE	**75.86**	**84.88**	**98.82**	**60.20**	**83.18**	**59.75**



(a) Text missing. (b) Image missing. (c) Both missing.

Figure 5: Food-101 accuracy across test missing rates for models trained at fixed rates.



(a) Text missing. (b) Image missing. (c) Both missing.

Figure 6: Hateful Memes AUROC across test missing rates for models trained at fixed rates.

D TRAINABLE PARAMETER COMPARISON

Table 5 counts trainable head parameters under the same frozen 512-dimensional image and text features used in the three image–text experiments. The counts use 23, 101, and two output classes, respectively, and exclude the frozen encoders. With C output classes, FF has $1025C$ parameters (one classifier over concatenated features), while SF has $1026C$ (two modality-specific classifiers, each with a bias). The Direct, Divide, Compose, and Full counts are $1216C$, $721920 + 1216C$, $787456 + 1216C$, and $858112 + 1216C$, respectively. These count parameters marked trainable

in each implemented variant, including classifier columns whose inputs are inactive in Direct; they describe allocated trainable parameters, not effective degrees of freedom or inference cost. The table is an implementation-matched capacity reference for the ablation, not a parameter-matched comparison with methods that adapt pretrained encoder layers.

E SUPPLEMENTARY MMIST-CCRCC SUBGROUP EXPERIMENT

Table 6 evaluates the 58 test patients with observed CT both with their original inputs and after CT is withheld. DACE has the highest accuracy, AUROC, macro precision, macro recall, and macro-F1 in both conditions; SF has the highest AUPRC when CT is observed, while DACE leads after CT is withheld. For DACE, removing CT raises AUROC from 72.22 to 84.88 and macro recall from 60.96 to 83.18, but lowers accuracy from 92.95 to 75.86; macro-F1 changes only from 60.03 to 59.75. Ranking and threshold-based metrics therefore move in different directions on this subgroup. The 54-to-4 class imbalance makes these results sensitive to a few patients, so they characterize missing-input behavior rather than clinical performance.

F ADDITIONAL ROBUSTNESS AND TRANSFER RESULTS

Figures 5 and 6 extend cross-rate evaluation to Food-101 and Hateful Memes; the corresponding single-model comparisons appear together in Table 2. Missing images cost more than missing text on Food-101, while Hateful Memes changes the ranking of training rates across test patterns. The curves support reuse over the tested rates, not a universally best training rate.