EASE: EFFECT-AWARE REPRESENTATION EDITING FOR MISSING-MODALITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing modalities typically result in irreversible information loss, significantly impairing the reasoning and decision-making capabilities of Multimodal Large Language Models (MLLMs). While existing methods largely focus on recovering missing information or adapting models to incomplete inputs, a fundamental question remains: how exactly does modality absence influence MLLM decision-making, and can this influence be leveraged to mitigate the resulting performance degradation? Across 12 settings involving four MLLMs and three emotion recognition and sentiment analysis datasets, we find that interventions in a low-rank subspace approximate the label-score changes produced by restoring complete-input representations at the same site. Based on these findings, we propose Effect-Aware Subspace Editing (EASE), which learns sample-specific corrections from available inputs within a fixed low-dimensional space constructed from paired training representations, while keeping the backbone frozen. Training combines task supervision with utility-gated effect matching and magnitude regularization. Experiments on IEMOCAP-4, CMU-MOSI, and CMU-MOSEI show improvements over the frozen backbone in all nine missing-modality settings. On IEMOCAP-4, the full objective improves mean weighted F1 (W-F1) by 2.59 points over cross-entropy-only training with the same editor, supporting the combined objective beyond task-only adaptation.

1 INTRODUCTION

Missing modalities can lead to the loss of task-relevant information and consequently impair model predictions. Mitigating the effects of this information loss is therefore important for reliable prediction from incomplete inputs. Recent work on learning with missing modalities has pursued two complementary goals: recovering information absent from the observed inputs and adapting models to predict effectively from incomplete inputs. Information recovery methods reconstruct task-relevant latent representations through cross-modal translation (Lin et al., 2025), exploit retrieved examples to recover missing content and generate contextual prompts (Lang et al., 2025), or use large multimodal models for training-free modality completion (Ke et al., 2025). Adaptation methods instead improve the use of available information through dynamic modality experts (Li et al., 2025), shared and modality-specific low-rank updates (Zhao et al., 2026), and instance-conditioned prompts incorporating modality priors (Lang et al., 2026).

These methods provide effective mechanisms for information compensation and predictive adaptation. However, based solely on the gains in predictive performance, it remains unclear how the internal changes triggered by a missing modality affect decision-making, or whether the effects of attempting to reconstruct that modality can directly substitute for its actual function. Addressing this question requires distinguishing between representational shifts, output effects, and task performance gains. Following modality loss, the representational shift observed in natural representations does not necessarily reflect the impact on model output, and the influence of a modality on decision accuracy can be either positive or negative. These distinctions give rise to two interrelated questions: how to identify changes that genuinely influence decision-making, and how to leverage changes that benefit the decision process.

We investigate these questions through controlled hidden-state interventions, drawing on representation inspection and activation steering (Ghandeharioun et al., 2024; Rimsky et al., 2024), in emo-

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.

tion recognition and sentiment analysis. On held-out samples, mean-removal analyses and sample-matched, shuffled, and orthogonal-complement controlled intervention experiments show that the subspace captures variation beyond the shared training mean and preserves sample-dependent score effects. Applying only a low-rank projection of the internal difference between complete and incomplete inputs can produce score changes similar to applying the entire difference at the same layer. Including more directions further improves the match to the reference score changes (Figure 1(a)), although the additional classification gains are less consistent (Figure 1(b)). In our IEMOCAP diagnostic, replaying the full hidden displacement at one position leaves an average normalized mean squared error (NMSE) of 0.305 relative to the complete–missing score effect (Appendix A.2). These findings motivate learning compact corrections guided by both reference effects and classification performance.

Building on these findings, we propose EASE. We train a lightweight editor to predict sample-specific interventions from incomplete inputs within a fixed low-dimensional subspace derived from paired complete–incomplete training representations, while keeping the backbone frozen. Training combines task supervision with utility-gated matching of decision effects and a soft magnitude penalty when matching is inactive to discourage excessive score shifts.

Our contributions are:

- We use controlled replay to examine how intervention subspaces and sample correspondence shape decision effects, distinguishing representation reconstruction, effect fidelity, and task utility.
- We show that accurately reproducing reference score changes does not consistently improve classification. This finding motivates corrections guided by task benefit, with effect matching used selectively to support prediction.
- We developed EASE, a lightweight editing framework. It addresses the information loss caused by missing modalities from a novel perspective by performing interventions that simulate, in a low-dimensional space, the decision shifts induced by missing modalities.

2 Related Work

Recent work on incomplete multimodal learning develops several ways to maintain prediction quality when some input sources are unavailable. Prompt-based approaches adapt representations to different patterns of modality availability. MPLMM (Guo et al., 2024) introduces prompts for generating missing features and encoding missing-modality conditions, while RAGPT (Lang et al., 2025) uses retrieved examples to support missing-modality generation and construct dynamic prompts. Other methods improve how available information is used: MoMKE (Xu et al., 2024) combines knowledge from unimodal experts with joint representations, and ComP (He et al., 2026) combines cross-modal prompting with gradient modulation and adaptive weighting of modality outputs. Alongside these architectural choices, complete inputs available during training can provide supervision for prediction from incomplete observations.

Knowledge distillation uses this supervision to transfer features, relationships, and prediction structure. CorrKD (Li et al., 2024) transfers correlations among samples and categories, together with response information that guides the student's decision boundaries. CMAD (Zhuang et al., 2025) combines feature and cross-sample correlation alignment with regularization that accounts for the difficulty of different modality combinations. These approaches motivate supervision that captures both representation structure and its relevance to prediction. Our work focuses on the changes in class-score contrasts between complete and incomplete inputs as a reference for learning corrective interventions. Applying this supervision to a frozen backbone also requires identifying where and along which directions its internal representations should be modified.

Recent work on representation interventions provides tools for studying and learning such modifications. Patchscopes (Ghandeharioun et al., 2024) interprets hidden representations by inserting them into other model computations. Contrastive Activation Addition (Rimsky et al., 2024) constructs steering vectors from averaged activation differences between paired examples of contrasting behaviors. ReFT (Wu et al., 2024) learns hidden-state interventions on frozen language models, including interventions in low-rank subspaces. Building on these ideas, we propose EASE, which learns sample-specific edits from incomplete inputs in representational space.

3 COMPACT CARRIERS FOR DOWNSTREAM EFFECT REALIZATION

We analyze how complete–incomplete hidden-state changes affect class scores in four MLLMs (Emotion-LLaMA-v2 (Cheng et al., 2024), Qwen2.5-Omni (Xu et al., 2025), AffectGPT (Lian et al., 2025), and MiniCPM-o (Cui et al., 2026)). The main results cover IEMOCAP-4 (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Bagher Zadeh et al., 2018). An editing carrier is a low-dimensional subspace defining the directions available for hidden-state edits.

Can a Compact Representation-Derived Space Preserve Downstream Effects? We first examine whether a low-dimensional projection of a complete–incomplete hidden-state change can preserve its effect on the model's class scores. Let $h_{i,\text{full}}, h_{i,m} \in \mathbb{R}^d$ denote the answer-position hidden states of sample i under complete input and with modality m missing, respectively, where d is the hidden dimension. Their paired displacement is

$$\delta h_{i,m} = h_{i,\text{full}} - h_{i,m}. \tag{1}$$

For each missing condition, we stack the training displacements as rows of A_m and apply truncated singular value decomposition (SVD):

$$A_m = L_m \Sigma_m V_m^\top, \qquad U_{m,k} = V_m[:, 1\!:\!k], \qquad P_{m,k} = U_{m,k} U_{m,k}^\top. \tag{2}$$

The columns of $U_{m,k}$ span the rank-k representation-derived subspace $\mathcal{V}_{m,k}^{\text{REP}} = \text{span}(U_{m,k})$, which we refer to as REP. We compare $\mathcal{V}_{m,k}^{\text{REP}}$ with $\mathcal{V}_{m,k}^{\text{RANDOM}}$, a random subspace of the same rank constructed from an independently sampled orthonormal basis, which we refer to as RANDOM.

Let $S_{i,m}(v) \in \mathbb{R}^C$ denote the scores for C candidate labels, obtained by adding an edit v at the selected answer position and running the remaining frozen layers, following prior hidden-state intervention and activation-patching analyses (Vig et al., 2020; Zhang & Nanda, 2024). To remove score shifts shared by all classes, we use an orthonormal contrast basis $Q \in \mathbb{R}^{C\times(C-1)}$ satisfying $Q^\top Q = I_{C-1}$ and $Q^\top \mathbf{1} = \mathbf{0}$, and define the effect of an edit as

$$f_{i,m}(v) = Q^\top \left[S_{i,m}(v) - S_{i,m}(\mathbf{0}) \right]. \tag{3}$$

The edit is constructed from the input states before candidate-specific scoring and reused across all candidate labels under the same input.

Both interventions start from the same missing-input state. The full displacement produces the effect $f_{i,m}(\delta h_{i,m})$, while REP uses only its projection $f_{i,m}(P_{m,k}\delta h_{i,m})$. On samples $\mathcal{H}$ not used to construct the subspace, we measure the fraction of hidden-displacement energy retained by the projection:

$$\rho_{m,k} = \frac{\sum_{i\in\mathcal{H}} \|P_{m,k}\delta h_{i,m}\|_2^2}{\sum_{i\in\mathcal{H}} \|\delta h_{i,m}\|_2^2}, \tag{4}$$

and the normalized mean squared error (NMSE) between the projected and full-displacement effects:

$$\epsilon_{m,k} = \frac{\sum_{i\in\mathcal{H}} \|f_{i,m}(P_{m,k}\delta h_{i,m}) - f_{i,m}(\delta h_{i,m})\|_2^2}{\sum_{i\in\mathcal{H}} \|f_{i,m}(\delta h_{i,m})\|_2^2}. \tag{5}$$

An NMSE of zero indicates exact reproduction, while a value below one indicates a smaller error than applying no edit.

Keeping the REP basis fixed, we subtract the training mean from each hidden displacement and recompute Equation 4, denoting the resulting retention by $\rho_{m,k}^{\text{c}}$. Across all twelve model–dataset settings, rank-six REP achieves lower mean effect NMSE than RANDOM (Table 1). REP errors range from 0.052 to 0.302, while RANDOM errors remain close to one. Thus, directions extracted from training displacements preserve the reference score changes on validation samples more accurately than random spaces of the same rank.

After subtracting the training mean, REP retains 23.73–51.28% of the remaining hidden-displacement energy, compared with 0.143–0.168% for RANDOM. The unchanged space therefore captures variation between individual samples beyond the shared training displacement. Together with the replay results, this supports using REP as a compact space for edits that preserve downstream score effects.

Table 1: Rank-six diagnostics averaged over Missing-V/A/T. ρ^c is hidden-displacement energy retention after subtracting the training mean, reported in percentages. Retention and effect NMSE entries are REP/RANDOM, with RANDOM averaged over ten seeds. ΔW-F1 is REP minus RANDOM in percentage points.

	IEMOCAP-4			CMU-MOSI			CMU-MOSEI		
Model	$\rho^c \uparrow$	Effect NMSE↓	ΔW-F1	$\rho^c \uparrow$	Effect NMSE↓	ΔW-F1	$\rho^c \uparrow$	Effect NMSE↓	ΔW-F1
Emotion-LLaMA-v2	40.95/0.146	.132/.996	+8.20	37.58/0.144	.052/.998	+10.91	33.59/0.143	.069/.998	+9.67
Qwen2.5-Omni	32.75/0.166	.234/.995	+4.74	23.73/0.164	.239/1.003	+0.76	25.58/0.164	.302/1.000	+1.22
AffectGPT	35.73/0.168	.121/.999	+8.40	25.83/0.164	.118/.997	+12.10	26.49/0.168	.150/.998	+14.59
MiniCPM-o	51.28/0.147	.205/.988	+3.53	46.13/0.143	.231/.996	+1.03	38.73/0.148	.210/1.000	+0.71

Should Edits Within This Space Depend on the Sample? REP provides one space shared across samples within each missing condition. We next examine whether the edit applied inside this space must depend on the individual input. Because REP is constructed from uncentered displacements, part of its effect could be reproduced by an average correction shared across samples.

Let

$$b_m = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \delta h_{i,m} \tag{6}$$

be the mean displacement on the training split. We compare each sample's own projected edit with two controls:

$$v_{i,m}^{\mathrm{MEAN}} = P_{m,k} b_m, \tag{7}$$

$$v_{i,m}^{\mathrm{SHUFFLED}} = P_{m,k} \delta h_{\pi(i),m}. \tag{8}$$

The edit $v_{i,m}^{\mathrm{MEAN}}$ applies the same training-derived correction to every sample under a missing condition. The edits $v_{i,m}^{\mathrm{SHUFFLED}}$ preserve the collection of projected edits but assign each edit to another sample. The permutation π has no self-matches and is restricted to the same missing condition and construction fold.

On CMU-MOSI and CMU-MOSEI, matched REP edits achieve lower mean effect NMSE than both $v_{i,m}^{\mathrm{MEAN}}$ and $v_{i,m}^{\mathrm{SHUFFLED}}$ in all eight model–dataset settings (Table 4). A shared average correction misses differences in score effects between samples. Shuffling preserves the collection of edits but increases the error by breaking their correspondence with inputs. The ability to reproduce individual score effects therefore depends on assigning each sample its corresponding edit.

These results motivate sharing the REP space within each missing condition while predicting edit coefficients from the individual incomplete input. The classification benefits of sample matching vary across models, so closer effect reproduction should also be evaluated alongside task performance.

Does Better Effect Fidelity Improve Classification? An intervention can reproduce the full-displacement effect accurately without moving the prediction toward the correct label. Let $s_{i,m}^{\mathrm{miss}} = S_{i,m}(\mathbf{0})$ denote the unedited candidate scores. We evaluate classification by adding the intervention's contrast-space effect to these scores:

$$\widetilde{s}_{i,m}(v) = s_{i,m}^{\mathrm{miss}} + Q f_{i,m}(v). \tag{9}$$

Rank-six REP improves Mean-Missing W-F1 over same-rank RANDOM in all twelve model–dataset settings (Table 1). However, these average gains do not imply improvements under every missing condition. MiniCPM-o gains 1.03 and 0.71 points on CMU-MOSI and CMU-MOSEI, respectively, but loses 0.95 points under Missing-V on CMU-MOSI and 0.51 points under Missing-A on CMU-MOSEI. Thus, preserving the full-displacement effect more accurately does not guarantee a classification benefit in every condition.

The rank analysis on IEMOCAP-4, CMU-MOSI, and CMU-MOSEI examines this relationship within each model. Increasing rank generally lowers effect NMSE (Figure 1(a)), while classification

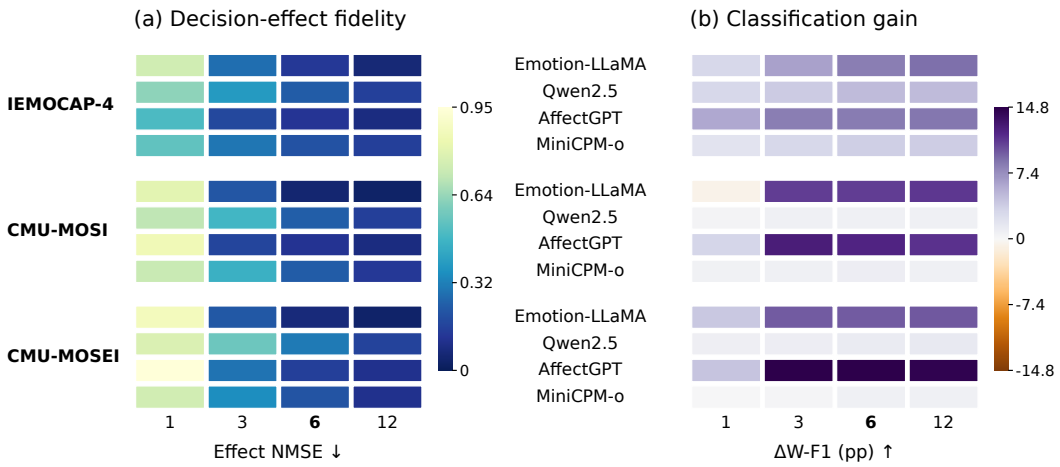


Figure 1: REP across subspace ranks: (a) effect NMSE relative to applying the entire representation difference at the same site (↓); (b) W-F1 gain over the unedited baseline in percentage points (↑). Results average V/A/T for every dataset. Better effect fidelity does not consistently yield larger classification gains.

gains can saturate or decline (Figure 1(b)). For example, AffectGPT on CMU-MOSI achieves lower effect NMSE at higher ranks, yet its W-F1 decreases from 84.26 at rank three to 83.17 at rank twelve. These results motivate learning corrections through both effect matching and task supervision.

4 EASE: Effect-Aware Subspace Editing

EASE learns sample-specific internal edits within a fixed representation-derived carrier. Each missing condition uses a REP basis constructed from paired training inputs and an independently trained lightweight editor. Given an incomplete input, the editor predicts coefficients in this basis and edits the frozen MLLM's intermediate state. Training combines task supervision with selective effect matching. Figure 2 summarizes the overall pipeline.

Constructing the REP Carrier For each missing condition m, we extract the complete-input and missing-input states $h_{i,\text{full}}^{\ell_m}$ and $h_{i,m}^{\ell_m} \in \mathbb{R}^d$ at the answer position before block ℓ_m, and define their displacement as

$$\delta h_{i,m} = h_{i,\text{full}}^{\ell_m} - h_{i,m}^{\ell_m}. \tag{10}$$

Following Section 3, we stack the training displacements and derive a rank-k basis $U_{m,k}$ at the intervention layer ℓ_m, whose columns span the REP carrier $\mathcal{V}_{m,k}^{\text{REP}}$. Each missing condition has its own fixed carrier and intervention layer, while the same basis is shared across all samples under that condition.

Predicting and Applying Sample-Specific Edits Given an incomplete input, we run the frozen MLLM to the input of block ℓ_m. Let $H_{i,m}^{\ell_m}$ be the hidden-state sequence and $h_{i,m}^{\ell_m} = H_{i,m,p_i}^{\ell_m}$ the state at the selected answer position p_i. A separate editor for condition m uses this state and a learned condition embedding e_m to predict coefficients $\widehat{c}_{i,m} \in \mathbb{R}^k$. Its trainable parameters, including the embedding, are denoted by θ_m. Writing $U_m = U_{m,k}$, we obtain

$$\widehat{c}_{i,m} = g_{\theta_m}\left(h_{i,m}^{\ell_m}, e_m\right), \qquad v_{i,m} = U_m \widehat{c}_{i,m}. \tag{11}$$

Thus, U_m specifies the condition-level carrier, whereas $\widehat{c}_{i,m}$ determines the edit for the individual sample.

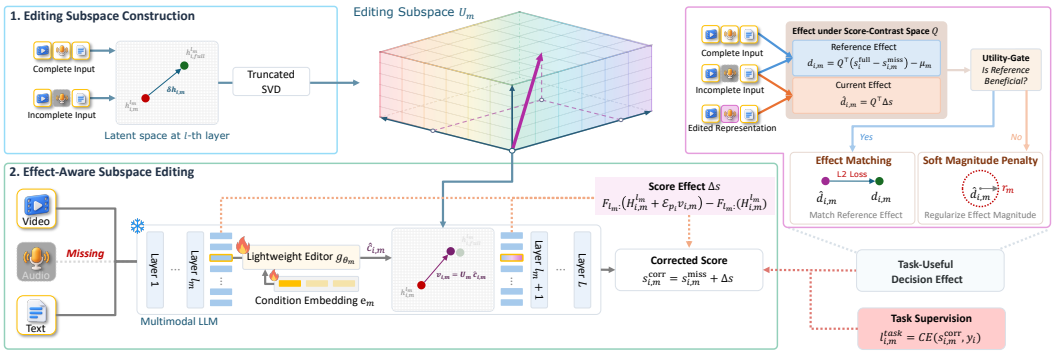


Figure 2: Overview of EASE. For each missing condition, truncated SVD of paired complete–incomplete representation differences defines a fixed low-dimensional editing subspace. Each condition's independently trained editor predicts input-dependent coefficients to modify the frozen MLLM's intermediate hidden state. Training applies classification loss to every sample, together with effect matching when the reference correction has lower classification loss than the current edit, or a soft magnitude penalty otherwise.

We inject $v_{i,m}$ only at the answer position:

$$\widetilde{H}_{i,m,t}^{\ell_m} = \begin{cases} H_{i,m,t}^{\ell_m} + v_{i,m}, & t = p_i, \\ H_{i,m,t}^{\ell_m}, & t \neq p_i. \end{cases} \tag{12}$$

Let $F_{\ell_m:}$ denote the remaining frozen blocks and candidate-score computation. The induced score effect and corrected scores are

$$\Delta s_{i,m} = F_{\ell_m:}\left(\widetilde{H}_{i,m}^{\ell_m}\right) - F_{\ell_m:}\left(H_{i,m}^{\ell_m}\right), \tag{13}$$

$$s_{i,m}^{\text{corr}} = s_{i,m}^{\text{miss}} + \Delta s_{i,m}. \tag{14}$$

Both branches use the same input context and scoring procedure, so the difference isolates the effect of the edit. All candidate labels share the same edit. Complete inputs are required only during training; inference uses the incomplete input and missing-condition identity.

Learning Task-Useful Decision Effects Section 3 shows that better effect fidelity does not always yield better classification. We therefore apply task supervision to every sample and use effect matching as a selective auxiliary objective.

Effect targets. Let $s_i^{\text{full}} \in \mathbb{R}^C$ denote the frozen model's candidate-label scores under complete input. Using the contrast basis Q defined in Equation 3, we define the predicted effect and complete–missing reference effect as

$$\widehat{d}_{i,m} = Q^\top \Delta s_{i,m}, \qquad D_{i,m} = Q^\top\left(s_i^{\text{full}} - s_{i,m}^{\text{miss}}\right). \tag{15}$$

The contrast projection removes score offsets shared by all classes. Section 3 evaluates effects relative to full-displacement replay at one site, whereas $D_{i,m}$ is computed from separate complete-input and incomplete-input forward passes. Their discrepancy is quantified in Appendix A.2.

With $N = |\mathcal{T}|$ training examples, we center the reference effects using training statistics:

$$\mu_m = \frac{1}{N}\sum_{i\in\mathcal{T}} D_{i,m}, \qquad d_{i,m} = D_{i,m} - \mu_m. \tag{16}$$

Centering expresses each reference effect as a deviation from its training mean. The centered effects determine the matching targets, utility gate, and magnitude statistics, while task supervision also permits shared corrections. We use $E_m = N^{-1}\sum_{i\in\mathcal{T}} \|d_{i,m}\|_2^2$ to normalize both auxiliary losses. The statistics μ_m and E_m are computed from training data and held fixed. The training mean μ_m is not added back at inference.

Task supervision. We use cross-entropy (CE) to increase the probability of the ground-truth class y_i under the corrected scores $s_{i,m}^{\text{corr}}$:

$$\ell_{i,m}^{\text{task}} = \text{CE}(s_{i,m}^{\text{corr}}, y_i). \tag{17}$$

This term directly optimizes classification on every sample, regardless of whether its reference effect is useful.

Utility-gated effect matching and regularization. Reference supervision can vary in usefulness across samples, and closer agreement with an imperfect reference can be harmful (He et al., 2025; Tiapkin et al., 2025). We therefore select the auxiliary loss according to the classification utility of the reference correction.

We form $s_{i,m}^{\text{target}} = s_{i,m}^{\text{miss}} + Qd_{i,m}$ by adding the centered reference correction to the unedited scores. We compare its classification loss with that of the current editor's corrected scores $s_{i,m}^{\text{corr}}$:

$$w_{i,m} = \mathbf{1}\left[\text{CE}(s_{i,m}^{\text{target}}, y_i) < \text{CE}(s_{i,m}^{\text{corr}}, y_i)\right]. \tag{18}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function and y_i is the ground-truth label. The gate activates effect matching when the reference correction yields a lower classification loss than the current edit; otherwise, it activates the magnitude penalty. The gate is recomputed during training and detached from gradients.

The matching loss measures the difference between the predicted effect $\widehat{d}_{i,m}$ and the centered reference $d_{i,m}$:

$$\ell_{i,m}^{\text{match}} = \frac{\|\widehat{d}_{i,m} - d_{i,m}\|_2^2}{E_m + \epsilon}, \tag{19}$$

where E_m is the training target energy defined above and $\epsilon > 0$ ensures numerical stability.

For magnitude regularization, let r_m be a fixed radius set to a chosen quantile of the training target norms $\{\|d_{j,m}\|_2 : j \in \mathcal{T}\}$. We define

$$\ell_{i,m}^{\text{mag}} = \frac{[\|\widehat{d}_{i,m}\|_2 - r_m]_+^2}{E_m + \epsilon}, \qquad [a]_+ = \max(a, 0). \tag{20}$$

This penalty is zero within the radius and grows quadratically with excess magnitude, leaving the task loss to guide the correction direction.

For each missing condition m, we independently optimize θ_m using

$$\mathcal{L}_m = \frac{1}{N} \sum_{i \in \mathcal{T}} \left[\ell_{i,m}^{\text{task}} + \lambda \left(w_{i,m}\ell_{i,m}^{\text{match}} + (1 - w_{i,m})\ell_{i,m}^{\text{mag}}\right)\right], \tag{21}$$

where $\lambda \geq 0$ weights the auxiliary loss. Task supervision applies to every sample, while the gate determines which auxiliary term contributes.

5 Experiments

Experimental setup. We evaluate EASE with a frozen Emotion-LLaMA-v2 backbone on IEMOCAP-4 for four-class emotion recognition and CMU-MOSI and CMU-MOSEI for binary sentiment classification. IEMOCAP-4 uses five grouped outer folds; the sentiment datasets use their official test splits with zero-valued labels excluded. Missing-V/A/T removes vision, audio, or text while retaining the other modalities. We report accuracy and weighted F1 (W-F1). The main EASE experiments use rank $k = 3$ on IEMOCAP-4 and $k = 6$ on CMU-MOSI and CMU-MOSEI. Supplementary rank-six IEMOCAP-4 results, including an additional training-seed evaluation, are reported in Appendix A.3.

We compare with six missing-modality methods. MMIN (Zhao et al., 2021) uses cascaded residual autoencoders and cycle consistency to predict missing-modality representations from available inputs. RedCore (Sun et al., 2024) learns cross-modal representations through a variational information bottleneck and adaptively balances supervision across modalities. MoMKE (Xu et al., 2024) combines knowledge from pretrained unimodal experts with joint representations to support prediction under incomplete inputs. ComP (He et al., 2026) combines cross-modal prompting, gradient

modulation, and adaptive output weighting to improve and balance modality-specific predictions. GCNet (Lian et al., 2022) models speaker and temporal dependencies with graph neural networks, jointly learning classification and missing-feature reconstruction. MissRAG (Pipoli et al., 2025) retrieves substitute modality tokens and uses prompts indicating missing inputs to support MLLM predictions without additional instruction tuning. These methods retain their respective features and prediction pipelines. ReFT-CE and LoReFT-CE follow representation fine-tuning (Wu et al., 2024). They use the same frozen backbone, data splits, scoring procedure, training budget, and checkpoint-selection rule as EASE.

Table 2: Results (%) under Missing-V/A/T. Acc denotes four-class accuracy on IEMOCAP-4 and Acc-2 on CMU-MOSI/MOSEI. Bold marks the best score; underlining marks the strongest external baseline.

	IEMOCAP-4						CMU-MOSI						CMU-MOSEI					
	Missing-V		Missing-A		Missing-T		Missing-V		Missing-A		Missing-T		Missing-V		Missing-A		Missing-T	
Method	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1
MMIN	70.93	70.84	70.26	70.11	60.13	59.76	79.12	79.17	79.27	79.35	55.34	54.60	84.67	84.51	84.65	84.36	63.46	58.11
RedCore	67.85	67.93	67.73	67.91	60.84	60.78	80.03	80.12	78.51	78.60	50.61	46.14	83.60	83.42	84.18	83.97	62.93	49.57
MoMKE	76.88	76.70	72.75	72.75	62.57	62.29	86.13	86.05	86.43	86.41	48.02	41.97	86.71	86.68	87.18	87.03	71.90	70.61
ComP	77.20	76.81	**73.39**	**72.95**	66.21	66.46	79.88	79.32	78.81	78.21	53.05	48.99	86.68	86.58	87.18	87.02	66.84	59.15
GCNet	65.41	65.42	68.20	67.01	50.89	49.41	66.77	66.93	65.70	65.44	53.96	54.23	87.26	87.15	86.93	86.72	62.85	48.86
MissRAG	68.22	68.18	49.34	50.09	55.49	54.53	74.39	72.45	77.59	76.96	53.66	52.38	85.94	85.62	86.90	86.64	63.73	55.61
Frozen backbone	65.63	64.81	60.50	60.20	49.47	45.89	88.87	88.91	85.98	85.99	57.16	54.81	87.59	87.39	86.85	86.51	64.39	56.54
EASE (ours)	**77.21**	**77.14**	69.87	69.91	**71.67**	**71.71**	**89.48**	**89.48**	**87.65**	**87.62**	**62.20**	**62.31**	**88.77**	**88.66**	**87.62**	**87.40**	**75.34**	**74.51**

Main Results Table 2 shows that EASE improves both metrics over the frozen backbone in all nine settings and ranks first in eight. These external comparisons evaluate complete systems with different backbones and feature pipelines. The largest W-F1 margins over external baselines occur under Missing-T on all three datasets. On CMU-MOSI/MOSEI, removing text weakens the backbone more than removing vision or audio. EASE provides substantial compensation from the remaining modalities, improving Missing-T W-F1 by 7.50 and 17.97 points, respectively. Under Missing-V/A, editing further improves an already strong backbone. These gains include supervised adaptation. The following same-backbone controls examine the additional benefit of effect-aware training over CE-only editing. Complete-input references and retention ratios appear in Appendix A.3.

On IEMOCAP-4, with the REP editing space and editor architecture fixed, CE-only training achieves 70.33 mean W-F1, while the full objective reaches 72.92 (Table 3(a)). CE-only training already provides substantial gains, particularly under Missing-T. The full objective adds 3.83, 2.38, and 1.57 points under Missing-V/A/T, respectively. These increments assess the additional benefit of the auxiliary objective. EASE also achieves higher W-F1 than ReFT-CE and LoReFT-CE in all three conditions; those comparisons assess different editing designs on the same backbone.

The centered reference effect $d_{i,m}$ in Equation 16 removes the condition-specific training mean from the complete–missing score difference. On IEMOCAP-4, matching this reference on every sample improves all three conditions. Utility gating further improves Missing-V/T but reduces Missing-A, while adding the magnitude penalty improves all three conditions. This last addition yields the largest mean increment in the ablation sequence. On CMU-MOSI, the full objective improves mean W-F1 over task-only training by 1.02 points (95% CI: $[0.16, 1.91]$). The separate increments from all-sample matching and utility gating have intervals that include zero (Appendix A.3). These results support the combined objective. Isolating the benefit of matching when magnitude regularization is already present requires an additional control.

LearnedCarrier-CE jointly learns orthonormal editing directions and the coefficient predictor using CE alone, with rank three on IEMOCAP-4 and rank six on CMU-MOSI. EASE improves mean W-F1 over this control by 2.28 points on IEMOCAP-4 and 1.25 points on CMU-MOSI (Table 3(b)). On CMU-MOSI, the mean improvement has a 95% confidence interval of $[0.33, 2.31]$; intervals also exclude zero under Missing-A/T, while Missing-V shows no clear difference (Appendix A.4). These comparisons support the combined choice of a representation-derived space and effect-aware supervision. The fixed-REP ablations separately examine the training objective.

Table 3: W-F1 (%) under Missing-V/A/T. (a) Baselines and objective ablations; the first row uses ReFT-CE on IEMOCAP-4 and Residual-MLP-CE on CMU-MOSI. (b) Learned-carrier comparisons. EASE results match the main table.

(a) Editing baselines and objectives

	IEMOCAP-4			CMU-MOSI		
Method	Missing-V	Missing-A	Missing-T	Missing-V	Missing-A	Missing-T
ReFT-CE / MLP-CE	76.57	68.33	69.59	88.35	86.56	61.16
LoReFT-CE	75.53	68.42	70.87	88.71	86.40	53.90
Task only	73.31	67.53	70.14	88.69	86.76	60.89
All-sample matching	73.48	68.76	70.41	**89.64**	86.29	61.14
Utility-gated matching	73.74	67.98	71.50	89.02	86.86	61.62
Full EASE	**77.14**	**69.91**	**71.71**	89.48	**87.62**	**62.31**
(b) Learned-carrier comparison						
LearnedCarrier-CE	73.84	67.08	71.01	89.47	86.25	59.93
EASE	**77.14**	**69.91**	**71.71**	**89.48**	**87.62**	**62.31**

Understanding the Learned Corrections On IEMOCAP-4, replacing predicted coefficients with their training mean or shuffling them across samples reduces W-F1 by 6.15 and 13.21 points, respectively; both paired 95% confidence intervals exclude zero. On CMU-MOSI Missing-T, the original coefficients achieve slightly higher W-F1 than the training-mean and shuffled controls, by 0.16 and 0.12 points, respectively. These small differences provide limited evidence for a classification benefit from sample-specific coefficient assignment in this setting (Appendix A.4).

To examine whether reference alignment matters beyond effect magnitude, we remove the positively aligned component of the learned score effect and compare this with rescaling the original effect to the same remaining norm. The rescaled effect achieves higher W-F1 by 3.50 points on IEMOCAP-4 and 6.01 points on CMU-MOSI Missing-T. These interventions support useful directional information in reference alignment beyond effect magnitude. Details appear in Appendix A.4.

Limitations. Additional mechanistic control experiments should also be included to determine the effects attributable to the specific factor and the magnitude of control. Future work will examine additional backbones, simultaneous removal of multiple modalities, and learning with fewer paired training examples.

6 Conclusion

We present EASE, a lightweight framework for compensating for missing modalities through representation editing in a frozen MLLM. Across four models, controlled replay shows that compact representation-derived spaces retain sample variation beyond the average hidden displacement and preserve downstream score effects more accurately than same-rank random spaces. Our diagnostics further show that reference-aligned score components provide useful task information beyond effect magnitude, while closer effect reproduction does not consistently improve classification. Building on these findings, EASE predicts input-dependent edits within a fixed representation-derived space. Training combines task supervision with utility-gated effect matching and a soft magnitude penalty, using reference score changes to guide corrections according to their classification utility. Experiments on IEMOCAP-4, CMU-MOSI, and CMU-MOSEI demonstrate improvements over the frozen backbone in all nine missing-modality settings. Controlled comparisons on IEMOCAP-4 and CMU-MOSI further show gains over task-only training with the same editor, supporting the combined use of task supervision, selective effect matching, and magnitude regularization for missing-modality compensation.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the replay analysis, editing framework, and training objective. Section 5 and Appendix A.1 describe the evaluation protocols and implementation settings. Additional replay results appear in Appendix A.2, while supplementary classification results and correction diagnostics are provided in Appendices A.3 and A.4. Reported uncertainty estimates specify their evaluation protocols and whether checkpoints are held fixed.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure, and the remaining required disclosure tasks are not applicable to this work. We used generative AI tools to edit and polish the manuscript for improved readability, clarity, and grammatical correctness, as well as to assist with retrieval and discovery of potentially relevant literature. All AI-assisted edits were carefully reviewed by the authors to ensure that they accurately reflected the authors' intended meaning and did not alter the scientific content, claims, or conclusions. Literature identified with the assistance of generative AI tools was independently reviewed and verified by the authors before being cited or incorporated into the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335–359, 2008. doi: 10.1007/s10579-008-9076-6. URL https://doi.org/10.1007/s10579-008-9076-6.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c7f43ada17acc234f568dc66da527418-Paper-Conference.pdf.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.94. URL https://aclanthology.org/2024.acl-long.94/.

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. DA-KD: Difficulty-aware knowledge distillation for efficient large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22379–22391. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/he25c.html.

Wen-Jue He, Xiaofeng Zhu, and Zheng Zhang. Cross-modal prompting for balanced incomplete multi-modal emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17463–17471, 2026. doi: 10.1609/aaai.v40i21.38800. URL https://ojs.aaai.org/index.php/AAAI/article/view/38800.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge Bridger: Towards Training-Free Missing Modality Completion . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02409. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02409.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i17.33984. URL https://doi.org/10.1609/aaai.v39i17.33984.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. Aoept: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. In *International Conference on Machine Learning (ICML)*, 2026.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuai Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12458–12468, 2024. URL https://api.semanticscholar.org/CorpusID:269362788.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24068–24077, October 2025.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. GCNet: Graph completion network for incomplete multimodal learning in conversation, 2022. URL https://arxiv.org/abs/2203.02177.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3612836. URL https://doi.org/10.1145/3581783.3612836.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lian25a.html.

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. Cyin: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz,

M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17605–17642. Curran Associates, Inc., 2025. doi: 10.52202/085713-0594. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1994798981a8467d7d50d8603c6caecb-Paper-Conference.pdf.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3215–3224, October 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i13.29440. URL https://doi.org/10.1609/aaai.v38i13.29440.

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Rame, and Mathieu Blondel. On teacher hacking in language model distillation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59552–59569. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tiapkin25a.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962. Curran Associates, Inc., 2024. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/75008a0fba53bf13b0bb3b7bff986e0e-Paper-Conference.pdf.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 438–446, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681683. URL https://doi.org/10.1145/3664647.3681683.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016. URL https://arxiv.org/abs/1606.06259.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.203. URL https://aclanthology.org/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 *The Fourteenth International Conference on Learning Representations*, 2026.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Cmad: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636, October 2025.

APPENDIX

A.1 IMPLEMENTATION AND EVALUATION

Editor and Training. We train a separate editor for each missing condition. Each editor consists of layer normalization, a $4096 \to 128$ projection, and a $128 \to 64 \to k$ prediction network. A 128-dimensional condition embedding is added after the first projection, and GELU follows both hidden layers. The final projection is initialized to zero. Including the three-entry embedding table, each editor has $541{,}443$ trainable parameters for $k = 3$ and $541{,}638$ for $k = 6$. Each editor uses only its corresponding condition embedding, which acts as a learned bias.

The main experiments use $k = 3$ on IEMOCAP-4 and $k = 6$ on CMU-MOSI and CMU-MOSEI. The main IEMOCAP-4 runs use seed 2061. The supplementary rank-six IEMOCAP-4 runs and the CMU-MOSI/MOSEI runs use Adam with learning rate 10^{-3} and effective batch size 128, with seeds 2062 and 2063 used across these runs. Training allows up to 100 epochs with an early-stopping patience of 10.

Dataset-Specific Settings. The main experiments use rank $k = 3$ on IEMOCAP-4 and $k = 6$ on CMU-MOSI and CMU-MOSEI. The sentiment experiments use the frozen Emotion-LLaMA-v2 stage-2 model and edit a single answer position. The scoring is based on the log-probabilities of the next token corresponding to the single tokens for the neutral, negative, and positive categories; thus, $C = 3$ and $Q \in \mathbb{R}^{3\times 2}$. The Adam optimizer employs default beta coefficients and uses neither weight decay nor a learning-rate scheduler. For the supplementary default rank-six experiment on the IEMOCAP-4 dataset, the following settings were used: the answer-position block-input layer index was set to 23, $\lambda = 0.25$, the radius quantiles q_m for the Missing-V/A/T modalities were set to $0.95/0.50/0.95$ respectively, and $\epsilon = 10^{-12}$. Here, q_m is used to determine the radius r_m based on the norm of the training targets.

IEMOCAP-4 splits. We use five leave-one-session-out outer folds. Each fold reserves one complete session for testing. The remaining four sessions are divided using `StratifiedGroupKFold` with eight partitions, shuffling enabled, split seed 12061, and dialogue IDs as groups. The first partition provides validation examples, and the remaining partitions provide training examples. Dialogues do not overlap across training, validation, and test within an outer fold.

For each outer fold and missing condition, the REP basis is constructed from uncentered training displacements. The reference mean μ_m, normalization energy E_m, and magnitude threshold r_m are computed from the corresponding training split. The threshold r_m is the q_m-quantile of training target norms. Complete-input reference effects on validation and test examples are used only for diagnostics. Checkpoint selection uses validation W-F1.

Evaluation. The non-zero label test subsets for the CMU-MOSI and CMU-MOSEI datasets contain 656 and 3,634 samples, respectively. In our evaluation, we distinguish between negative and non-negative sentiment: samples with a ground-truth label of zero are excluded, and predicted neutral and positive labels are mapped to the non-negative class. The W-F1 metric calculates a weighted average of F1 scores based on the number of samples (support) for each binary classification category in the evaluation set. The replay diagnostics process utilized the same binary mapping scheme on the official validation set and involved neither training the editor nor accessing the test set.

A.2 ADDITIONAL REPLAY RESULTS

Local replay versus complete-input effects. On IEMOCAP-4 Session2, we compare $f_{i,m}(\delta h_{i,m})$ with the complete–missing effect $D_{i,m}$ using 549 validation examples. The answer-position block-input sites are $23/15/23$ for Missing-V/A/T. The errors

$$\frac{\sum_i \|f_{i,m}(\delta h_{i,m}) - D_{i,m}\|_2^2}{\sum_i \|D_{i,m}\|_2^2}$$

are $0.093/0.776/0.047$, averaging 0.305. Thus, replaying the natural displacement at one site does not fully reproduce the complete-input effect. This diagnostic uses $D_{i,m}$, whereas training matches

the centered target $d_{i,m}$. The measured discrepancy concerns this particular edit and does not establish a lower bound on the error achievable by other edits.

Protocol and sample correspondence. CMU-MOSI uses $1{,}284/229$ training/validation examples from $52/10$ videos; MOSEI uses $16{,}326/1{,}871$ examples from $2{,}249/300$ videos. Answer-position block-input sites are $23/20/20/26$ for Emotion-LLaMA-v2, Qwen2.5-Omni, AffectGPT, and MiniCPM-o, respectively. Table 4 supplements the main rank-six results with the mean and shuffled edits defined there, and the remaining displacement $v^{\perp}_{i,m} = (I - P_{m,6})\delta h_{i,m}$. Matched edits have lower effect error than every control in all eight settings.

Table 4: Rank-six effect NMSE on validation data, averaged over Missing-V/A/T. The reference is full-displacement replay at the same site.

Dataset	Model	Matched	v^{MEAN}	v^{SHUFFLED}	$v^{\perp}$
CMU-MOSI	Emotion-LLaMA-v2	.052	.846	1.634	.897
	Qwen2.5-Omni	.239	.739	1.224	.746
	AffectGPT	.118	.946	1.357	.807
	MiniCPM-o	.231	.795	1.533	.809
CMU-MOSEI	Emotion-LLaMA-v2	.069	.907	1.775	.883
	Qwen2.5-Omni	.302	.796	1.262	.667
	AffectGPT	.150	1.036	1.560	.838
	MiniCPM-o	.210	.794	1.354	.746

Variation across random spaces. Table 5 reports variation across ten random spaces. For each seed, results are first averaged over Missing-V/A/T.

Table 5: Rank-six RANDOM results on validation data. Entries are mean $\pm$ sample standard deviation across ten seeds. W-F1 is reported in percentages.

Dataset	Model	Effect NMSE	W-F1
CMU-MOSI	Emotion-LLaMA-v2	$.998 \pm .002$	77.72 ± 0.05
	Qwen2.5-Omni	$1.003 \pm .005$	88.33 ± 0.08
	AffectGPT	$.997 \pm .001$	71.69 ± 0.09
	MiniCPM-o	$.996 \pm .007$	81.37 ± 0.16
CMU-MOSEI	Emotion-LLaMA-v2	$.998 \pm .001$	77.16 ± 0.05
	Qwen2.5-Omni	$1.000 \pm .005$	86.09 ± 0.05
	AffectGPT	$.998 \pm .002$	69.14 ± 0.05
	MiniCPM-o	$1.000 \pm .004$	81.72 ± 0.05

MELD and MER2023. Table 6 evaluates samples not used to construct the spaces. REP improves effect fidelity throughout, but Qwen2.5-Omni has slightly lower W-F1 than RANDOM on both datasets, reinforcing that effect fidelity and classification utility need not improve together.

Table 6: Rank-six replay on MELD (Poria et al., 2019) and MER2023 (Lian et al., 2023), averaging Missing-V/A/T and Missing-V/A, respectively. NMSE entries are REP/RANDOM; ΔW-F1 is REP minus RANDOM in percentage points. RANDOM averages ten seeds after averaging missing conditions.

	MELD		MER2023	
Model	Effect NMSE	ΔW-F1	Effect NMSE	ΔW-F1
Emotion-LLaMA-v2	.172/.997	+6.78	.083/.997	+7.56
Qwen2.5-Omni	.246/1.004	-0.21	.178/.996	-0.92
AffectGPT	.202/.997	+1.97	.144/1.002	+7.91
MiniCPM-o	.168/.995	+2.05	.268/1.062	+5.00

A.3 Supplementary Classification Results

Table 7 adds complete-input baseline results and accuracy for the main controlled comparisons. MissRAG uses Emotion-LLaMA on IEMOCAP-4 and CMU-MOSEI, and OneLLM on CMU-MOSI. Edited incomplete-input predictions can exceed the unedited complete-input backbone because the editor receives task supervision. Visual inputs may also contain ambiguous or misaligned cues, so removing them can occasionally help. Effect matching does not require EASE to reproduce every complete-input score change. These explanations are plausible, but the reported comparisons do not isolate their contributions.

Table 7: Supplementary classification results (%). Panel (a) reports complete-input W-F1 / Accuracy.

(a) Complete-input external baselines			
Method	IEMOCAP-4	CMU-MOSI	CMU-MOSEI
MMIN	75.09 / 75.25	79.60 / 79.57	84.54 / 84.76
RedCore	72.24 / 72.14	78.92 / 78.81	84.08 / 84.31
MoMKE	76.23 / 76.30	86.97 / 87.04	87.10 / 87.23
ComP	77.82 / 78.25	78.21 / 78.81	86.24 / 86.41
GCNet	64.39 / 64.56	82.37 / 82.32	86.83 / 86.90
MissRAG	75.98 / 75.99	82.03 / 82.16	87.24 / 87.51
(b) Accuracy for controlled comparisons			
Method	Missing-V	Missing-A	Missing-T
Task only	73.25	67.61	70.21
All-sample matching	73.32	68.54	70.43
Utility-gated matching	73.54	67.85	71.43

Across the five outer folds, W-F1 standard deviations for Missing-V/A/T are $1.76/3.50/1.46$ for ReFT-CE and $3.76/1.99/0.32$ for LoReFT-CE.

CMU-MOSI objective comparisons. LearnedCarrier-CE uses rank $k = 6$ at the answer-position input of block 23, with the same coefficient predictor as EASE. Its trainable carrier $B_{\text{raw}} \in \mathbb{R}^{4096\times 6}$ is initialized from a Gaussian matrix and orthonormalized by QR decomposition. The predictor's final layer is initialized to zero. The carrier adds $24{,}576$ parameters, giving $566{,}214$ trainable parameters in total, compared with $541{,}638$ for EASE.

Each missing condition is trained independently on the official training split using seed 2061, Adam with learning rate 10^{-3}, effective batch size 128, and replay batch size 16. Training allows up to 100 epochs with early-stopping patience 10; checkpoints are selected by validation W-F1. Training uses three-class CE over neutral, negative, and positive labels, including neutral training examples. Binary test metrics exclude zero-valued gold labels. Reference effects are used only for post-training diagnostics.

Table 8 supports the improvement of the full objective over task-only training. The separate increments from all-sample matching and utility gating remain inconclusive.

Table 8: CMU-MOSI outer-test differences in Mean-Missing W-F1 (percentage points). Mean-Missing averages Missing-V/A/T. Intervals use $10{,}000$ paired video-bootstrap replicates with checkpoints held fixed. EASE uses the main-table checkpoints.

Comparison	ΔW-F1	95% CI
All-sample matching − Task only	$+0.24$	$[-0.76, 1.29]$
Utility-gated matching − All-sample matching	$+0.14$	$[-0.29, 0.60]$
EASE − Frozen backbone	$+3.23$	$[0.95, 5.53]$
EASE − Task only	$+1.02$	$[0.16, 1.91]$

Recovery under Missing-T. For CMU-MOSI Missing-T, the fraction of the frozen backbone's performance gap closed by EASE is

$$R_{\text{gap}} = \frac{M_{\text{EASE}} - M_{\text{miss}}}{M_{\text{full}} - M_{\text{miss}}} \times 100\%, \tag{22}$$

Here, M_{full} and M_{miss} are the same frozen backbone's complete-input and Missing-T scores, and M_{EASE} is the edited score. Using the main-table scores, EASE closes 23.49% of the W-F1 gap and 17.04% of the Acc-2 gap. These gains include supervised adaptation.

Performance retention under Missing-T. We compare the fraction of complete-input performance retained under Missing-T on CMU-MOSI:

$$R_{\text{ret}} = \frac{M_{\text{Missing-T}}}{M_{\text{full}}} \times 100\%, \tag{23}$$

Here, M_{full} is each method's complete-input reference, and $M_{\text{Missing-T}}$ is its Missing-T performance. EASE uses the unedited backbone as its reference, whereas external baselines use their trained complete-input models. The ratios therefore describe systems with different adaptation states.

Table 9: CMU-MOSI Missing-T performance and descriptive retention ratios (%). Full reference denotes each baseline's complete-input model and the unedited backbone for EASE. Ratios use unrounded scores where available. Bold marks the best Missing-T scores.

	Full reference		Missing-T		Retention	
Method	Acc-2	W-F1	Acc-2	W-F1	Acc-2	W-F1
MMIN	79.57	79.60	55.34	54.60	69.54	68.60
RedCore	78.81	78.92	50.61	46.14	64.22	58.46
MoMKE	87.04	86.97	48.02	41.97	55.17	48.26
ComP	78.81	78.21	53.05	48.99	67.31	62.63
GCNet	82.32	82.37	53.96	54.23	65.55	65.84
MissRAG	82.16	82.03	53.66	52.38	65.31	63.86
Frozen backbone	86.74	86.74	57.16	54.81	65.91	63.19
EASE	86.74	86.74	**62.20**	**62.31**	**71.71**	**71.84**

Table 10: Rank-six IEMOCAP-4 outer-test W-F1 (%). Session rows report the tuned configuration. Summary rows give the five-fold mean ± sample standard deviation. Default uses the settings(seed 2062) in Appendix A.1.

Test fold	Missing-V	Missing-A	Missing-T
Session1	76.08	70.46	72.96
Session2	79.72	72.64	73.81
Session3	75.34	68.44	70.65
Session4	79.36	67.58	71.30
Session5	78.77	65.35	68.71
Tuned mean ± SD	77.85 ± 2.00	68.89 ± 2.78	71.49 ± 2.00
Default mean ± SD	77.14 ± 2.02	69.91 ± 1.45	71.08 ± 1.97
Tuned − default	$+0.71$	-1.02	$+0.41$

To complement the main rank-three evaluation, we conducted a supplementary rank-six EASE run on IEMOCAP-4 with training seed 2063 (Table 11). The five-fold mean accuracies under Missing-V/A/T are 76.25%, 69.87%, and 71.05%, respectively, with corresponding W-F1 scores of 76.24%, 69.91%, and 71.08%. The Mean-Missing W-F1 is 72.41%. These results provide supplementary evidence from an additional training seed in the rank-six setting.

Table 11: IEMOCAP-4 outer-test results (%) for a supplementary rank-six EASE run with training seed 2063. Each session is held out for testing in one outer fold. The final row reports the arithmetic mean across the five folds.

	Missing-V		Missing-A		Missing-T	
Test fold	Acc	W-F1	Acc	W-F1	Acc	W-F1
Session1	75.48	75.15	69.68	69.65	71.71	71.94
Session2	78.40	78.30	72.34	72.35	73.80	73.72
Session3	74.46	74.56	70.03	69.90	70.81	71.14
Session4	77.79	77.91	68.57	68.95	70.51	70.19
Session5	75.10	75.26	68.73	68.70	68.41	68.42
Five-fold mean	76.25	76.24	69.87	69.91	71.05	71.08

A.4 Correction Diagnostics and Carrier Choice

Fixed-checkpoint interventions. The IEMOCAP-4 diagnostics modify corrections on the validation splits of the five outer folds without retraining; confidence intervals use paired dialogue resampling. Coefficient means come from training data, and shuffling excludes self-matches within each missing condition and construction fold. For a nonzero centered reference d, let $u = d/\|d\|_2$. The positively aligned component, remaining effect, and same-norm control are

$$a = [\langle \widehat{d}, u \rangle]_+ u, \qquad b = \widehat{d} - a, \qquad z = \frac{\|b\|_2}{\|\widehat{d}\|_2} \widehat{d}.$$

Set $a = \mathbf{0}$ when $d = \mathbf{0}$ and $z = \mathbf{0}$ when $\widehat{d} = \mathbf{0}$. Predictions use $s^{\text{miss}} + Qa$, $s^{\text{miss}} + Qb$, and $s^{\text{miss}} + Qz$. The norm-controlled contrast is W-F1(z)−W-F1(b); reference-direction shuffles recompute it on common eligible samples, globally or within the same ground-truth class. Table 12 reports paired differences for these interventions. The intervals support input-dependent coefficients and useful reference alignment beyond effect magnitude. The within-class shuffle difference remains inconclusive.

Table 12: Fixed-checkpoint IEMOCAP-4 validation diagnostics. Differences are W-F1 percentage points, with 95% paired dialogue-bootstrap confidence intervals. The last two rows compare the norm-controlled contrast W-F1(z) − W-F1(b) under true and shuffled reference directions.

Comparison	Difference	95% CI
Predicted − mean coefficients	+6.15	[3.24, 8.80]
Predicted − shuffled coefficients	+13.21	[9.70, 16.23]
Same-norm control z − aligned removed b	+3.50	[2.88, 4.12]
True − globally shuffled directions	+2.37	[1.82, 2.94]
True − within-class shuffled directions	−0.01	[−0.50, 0.50]

Table 13: CMU-MOSI Missing-T outer-test diagnostics (%), using the main-table checkpoint and a common candidate-score readout in both panels. Shuffled coefficients report mean ± standard deviation across five permutations, differences are computed before rounding.

Variant	Acc-2	W-F1
(a) Coefficient interventions		
Original coefficients	62.20	62.31
Training-mean coefficients	61.89	61.99
Shuffled coefficients	61.92 ± 0.13	62.02 ± 0.13
(b) Score-effect components		
ZERO-CANDIDATE	56.71	54.36
ORIGINAL-CANDIDATE	62.04	62.15
REMOVE-ALIGNED	53.51	53.66
ALIGNED-ONLY	65.55	65.09
NORM-MATCHED	59.45	59.67

The original coefficients slightly outperform the mean and shuffled controls, but the small differences and absence of paired confidence intervals limit conclusions about sample correspondence. Removing the reference-aligned component lowers W-F1 from 62.15 to 53.66. The norm-matched control achieves 59.67, exceeding the removal variant by 6.01 points. The aligned-only variant achieves 65.09 W-F1. Together, these results support useful directional information in reference alignment beyond effect magnitude.

Trainable carrier on IEMOCAP-4. LearnedCarrier-CE initializes $B_{\text{raw}} \in \mathbb{R}^{4096\times 3}$ from a Gaussian matrix and obtains an orthonormal B by reduced QR decomposition at each forward pass. It learns $v = B\widehat{c}$ using CE alone, with $553{,}731$ parameters versus $541{,}443$ for the EASE editor; both use $k = 3$. The comparison and the following sample-level analysis use the same EASE checkpoints as the main results. The alternative-correction analysis selects samples where this control is correct and has CE no higher than EASE. For Missing-V/A/T, the subsets contain $2{,}993/2{,}656/2{,}157$ samples; $90.6/86.1/82.0\%$ have larger squared reference-effect error $\|\widehat{d} - d\|_2^2$ than EASE.

CMU-MOSI carrier comparison. Table 14 reports EASE's improvements over LearnedCarrier-CE with checkpoints held fixed. The intervals exclude zero under Missing-A/T and for the condition average, while Missing-V remains inconclusive.

Table 14: CMU-MOSI outer-test differences: EASE minus LearnedCarrier-CE in percentage points, with 95% confidence intervals. Mean averages Missing-V/A/T. Differences are computed before rounding.

Condition	ΔAcc-2 [95% CI]	ΔW-F1 [95% CI]
Missing-V	$0.00\ [-0.91, 1.00]$	$+0.02\ [-0.90, 1.03]$
Missing-A	$+1.37\ [0.16, 2.57]$	$+1.37\ [0.17, 2.56]$
Missing-T	$+2.13\ [0.31, 4.44]$	$+2.38\ [0.43, 4.84]$
Mean	$+1.17\ [0.27, 2.19]$	$+1.25\ [0.33, 2.31]$

REP and JAC. JAC constructs editing directions from the dominant right-singular vectors of stacked training class-score Jacobians at missing-input hidden states. We compare rank-three REP and JAC carriers on IEMOCAP-4 using five-fold outer-test classification results. The separate oracle replay diagnostic gives REP/JAC NMSE of $0.2770/0.1287$ on the Session2-fold validation split.

Table 15: Five-fold IEMOCAP-4 outer-test W-F1 (%). Both carriers use rank three. REP-EASE reproduces the main-table results. Mean averages Missing-V/A/T; bold marks the higher score.

	Oracle replay	Learned-editing W-F1			
Carrier	NMSE↓	Missing-V	Missing-A	Missing-T	Mean
REP	.2770	**77.14**	**69.91**	**71.71**	**72.92**
JAC	**.1287**	76.80	68.79	70.54	72.05

REP-EASE achieves higher five-fold W-F1 under all three missing conditions, with a mean advantage of 0.87 points. Together with JAC's lower oracle replay error, these results show that better reproduction of an oracle effect does not necessarily yield better learned corrections.