CLASS-CONDITIONAL AUTOENCODERS WITH ADVERSARIAL ALIGNMENT FOR MULTIMODAL FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning has advanced rapidly with large-scale transformers, but often requires heavy computation and lacks clear theoretical grounding. We propose a lightweight yet robust framework for multimodal fusion that unifies efficiency with theoretical guarantees. At its backbone lies a Class-Conditional Autoencoder (CCAE), which maps modality-specific inputs into a class-aware latent space. Building upon this, our Discriminative Embedding Framework (DEF) incorporates homologous and reconstruction losses to contract intra-class variance while preserving semantic fidelity, producing embeddings that are compact and discriminative. To address distributional inconsistencies across modalities, we introduce the Adversarial Alignment Framework (AAF), which dynamically weights modality contributions and aligns fused embeddings with modality-specific distributions using a Wasserstein objective. Together, DEF and AAF form a cohesive framework that explains why consistency and alignment emerge from a unified optimization perspective. Extensive experiments on machine translation (How2, Multi30k) and emotion recognition (IEMOCAP, MOSEI) demonstrate that our approach consistently outperforms strong baselines, including Transformer, MulT, and MISA, while operating with much lower FLOPs.

1 INTRODUCTION

Multimodal learning plays an important role in recent AI advances, enabling joint reasoning over text, speech, vision, and beyond.Multimodal transformers (e.g., CLIP, BLIP-2, Flamingo, LLaVA) achieve strong performance but demand massive compute and obscure how modalities should be fused. In contrast, lightweight strategies (e.g., pooling, canonical correlation, or modality-specific autoencoders) are while keeping the design computationally efficient. This gap motivates our framework, which balances scalability, adaptivity, and theoretical grounding.

In this work, we aim to bridge this gap. We introduce a lightweight yet theoretically-grounded framework for multimodal fusion that (i) scales across many modalities and categories without parameter explosion, (ii) adapts dynamically to modality quality and availability, and (iii) enjoys formal guarantees on variance contraction, reconstruction fidelity, and cross-modal distribution alignment. Unlike prior work that primarily assembles modules in an ad hoc manner, our design provides a unified optimization perspective that explains why consistency and alignment emerge, offering both practical efficiency and conceptual clarity.

Contributions. The main contributions of this work are as follows:

- **Unified Optimization Perspective.** We re-cast multimodal fusion as a constrained optimization problem that balances variance contraction, semantic reconstruction, and distributional alignment, offering a principled view of why consistency and alignment emerge.
- **Class-Conditional Autoencoder (CCAE).** We introduce a conditional autoencoder with shared parameters modulated by class embeddings, forming the backbone of our approach.
- **Discriminative Embedding Framework (DEF).** Building on CCAE, DEF enforces compactness and class separability using homologous and reconstruction losses, ensuring modality-aligned and semantically robust embeddings.

- **Adversarial Alignment Framework (AAF).** Complementary to DEF, AAF integrates a dynamic fusion operator and Wasserstein-based adversarial matching to enforce cross-modality distributional coherence.
- **Complete Method: DEF+AAF.** Combining the discriminative power of DEF with the robustness of AAF yields an efficient, theoretically grounded multimodal fusion model.
- **Extensive Evaluation.** Our framework consistently improves performance on translation (How2, Multi30k) and emotion recognition (IEMOCAP, MOSEI) benchmarks over strong multimodal baselines, while reducing FLOPs compared to transformer-based models.

2 Related Work

Multimodal representation learning. Early research on multimodal learning primarily relied on heuristic fusion strategies, such as early fusion (feature concatenation) or late fusion (decision-level combination), which often suffer from suboptimal alignment and poor robustness under missing modalities. Autoencoding-based methods extended this line by constructing joint latent spaces through reconstruction objectives, but tend to overlook fine-grained semantic consistency across modalities. Recent survey work (Baltrusaitis et al., 2019) has summarized these paradigms and highlighted the need for principled approaches to modality integration.

Contrastive and cross-modal alignment. Contrastive learning has become the dominant paradigm for large-scale multimodal pretraining. CLIP (Radford et al., 2021) demonstrated the effectiveness of aligning vision and language representations via natural language supervision, inspiring subsequent frameworks such as ALIGN (Jia et al., 2021) and BLIP-2 (Li et al., 2023). However, these models rely heavily on massive web-scale data and remain vulnerable to modality imbalance or corruption. Task-specific approaches (Tsai et al., 2019; Zadeh et al., 2018a) have investigated multimodal alignment for emotion recognition, yet most still employ fixed fusion architectures.

Dynamic and robust fusion. Recent studies emphasize robustness and adaptability in multimodal integration. Dynamic fusion methods such as MulT (Tsai et al., 2019), MMIM (Han et al., 2021), and interpretable dynamic fusion graphs (Zadeh et al., 2018a) highlight the importance of context-aware modality weighting. Adversarial learning has also been used as a mechanism for distribution-level alignment across heterogeneous features (Wang et al., 2020), complementing contrastive objectives. Despite these advances, most existing methods either rely on rigid global alignment or computationally heavy pretraining, which limits their applicability in noisy or resource-constrained scenarios.

3 Method

In this chapter, we describe two methods that can effectively leverage multimodal latent collaborative information for dynamic fusion.

3.1 Discriminative Embedding Framework

To reduce inherent semantic differences in multimodal representations, enhance semantic correlations, and extract unified object embedding patterns, we designed the discriminative embedding framework(DEF) module. Here, the term "discriminative" in the Discriminative Embedding Framework (DEF) emphasizes that the learned latent representations are not only compact and aligned across modalities, but also exhibit strong class separability, i.e., embeddings from the same class are encouraged to cluster closely while those from different classes remain well separated.It builds on a **Class-Conditional Autoencoder (CCAE)**, which maps modality features into a class-aware latent space using class embeddings e_w. On these embeddings, DEF applies *homologous loss* to align modalities of the same object, and *dual reconstruction losses* (intra- and cross-modal) to preserve semantic fidelity. This design produces compact, class-specific multimodal representations and remains computationally efficient.

3.1.1 MODAL EMBEDDING GENERATION

In DEF, we adopt a Class-Conditional Autoencoder (CCAE) to construct unified and discriminative representations across multiple modalities. These CCAE-based embeddings serve as the backbone for DEF, which further incorporates homologous and reconstruction losses to enhance discriminability.Specifically, for each object belonging to category w, we consider up to N modalities $\{M^s\}_{s=1}^{N}$ $(1 < N \leq 3)$. Each modality M^s is first processed by a semantic feature extractor T^s to obtain semantic features:

$$X_i^s = T^s(w_i^s), \tag{1}$$

where w_i^s denotes the raw modality-s input of object i, and X_i^s is the corresponding semantic representation.

Unlike conventional designs where each category maintains an independent family of autoencoders, CCAE shares a unified encoder–decoder architecture across all categories, while conditioning both encoder and decoder on the class embedding e_w. This design ensures parameter sharing and scalability to unseen categories, while simultaneously injecting semantic category information into the latent space.

Formally, the CCAE encoder f_θ maps a modality-specific input feature X_i^s together with its class embedding e_w into the latent embedding space:

$$c_i^s = f_\theta(X_i^s, e_w), \tag{2}$$

where c_i^s denotes the class-aware latent embedding of object i under modality s. Correspondingly, the decoder g_ϕ reconstructs features from the latent embedding under the same class condition:

$$\widetilde{X}_i^s = g_\phi(c_i^s, e_w). \tag{3}$$

Here, $f_\theta(\cdot, e_w)$ enforces that embeddings from the same class are aligned in a shared space, while $g_\phi(\cdot, e_w)$ guarantees that the class-conditioned latent codes retain sufficient semantic information for faithful reconstruction. In this way, CCAE produces modality embeddings that are simultaneously compact, semantically grounded, and discriminative across categories.

3.1.2 LOSS FUNCTIONS IN CCAE

To fully utilize inter-modal collaborative information, we introduce two key loss functions: homologous loss and reconstruction loss. Homologous loss ensures minimal distance between different modal features from the same sample in the latent space; reconstruction loss includes both intra-modal and cross-modal dimensions, ensuring feature compression accuracy while enhancing inter-modal semantic consistency. The final fusion features are generated through a simple but effective aggregation function Λ, avoiding computational overhead from complex fusion operations.

Homogeneous Loss Function: We provide a more detailed discussion of this method. TheHomologous loss function is designed to constrain the modal representations of the same traffic data object to be as similar as possible, thereby constructing a more compact latent space. This helps reduce the distance between embeddings generated by different modalities within the same autoencoder family.The homologous loss ensures that the latent codes of different modalities belonging to the same object are pulled close to each other under the supervision of the same class embedding e_{w_i}. The function is mathematically defined as shown in Equation 1. Here, $\|\cdot\|$ denotes the $L2$ norm.

$$L_H = \frac{1}{N}\sum_{i=1}^{N} \frac{2}{M_i(M_i-1)} \sum_{s<t} \left\| f_\theta(x_i^s, e_{w_i}) - f_\theta(x_i^t, e_{w_i}) \right\|^2, \tag{4}$$

where N is the batch size, M_i the number of modalities for object i, x_i^s the s-th modality input of object i, and e_{w_i} the embedding vector of its class label w_i.

Reconstruction Loss Function: To enhance the generalization capability of the encoder and decoder for the three highly heterogeneous modalities within the autoencoder, we introduce the dual reconstruction loss function. This function is composed of two components: the single-modal reconstruction loss and the cross-modal reconstruction loss. Its definition is given in Equation 5 and 6.

(a) With L_H

(b) Without L_H

Figure 1: t-SNE style schematic visualization of learned embeddings. With homologous loss (left), clusters are compact and separated. Without L_H (right), clusters overlap significantly.

(a) Intra-modal reconstruction

$$L_R^{\text{intra}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{s=1}^{M_i} \left\| x_i^s - g_\phi(f_\theta(x_i^s, e_{w_i}), e_{w_i}) \right\|^2. \tag{5}$$

(b) Cross-modal reconstruction requires that information from one modality can be used to reconstruct another:

$$L_R^{\text{cross}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M_i(M_i-1)} \sum_{s \neq t} \left\| x_i^s - g_\phi(f_\theta(x_i^t, e_{w_i}), e_{w_i}) \right\|^2. \tag{6}$$

The overall reconstruction objective is then a weighted combination:

$$L_R = \lambda \, L_R^{\text{intra}} + (1-\lambda) \, L_R^{\text{cross}}, \tag{7}$$

with $\lambda \in [0,1]$ controlling the trade-off.

Contrastive Regularization (optional): Inspired by InfoNCE, we can regularize embeddings using:

$$L_{\text{con}} = -\mathbb{E}\left[\log \frac{\exp(\langle z^a, z^b \rangle / \tau)}{\sum_j \exp(\langle z^a, z_j^- \rangle / \tau)} \right], \tag{8}$$

which separates positive homologous pairs (z^a, z^b) from negatives z_j^-.

Total DEF Objective: The complete optimization objective is:

$$L_{\text{DEF}} = \alpha L_H + \beta L_R + \tau L_{\text{con}}, \tag{9}$$

where α, β, τ balance alignment, semantic reconstruction, and contrastive separation. In experiments, $\tau = 0$ if contrastive regularization is not used.

In summary, the discriminative nature of DEF lies in its ability to jointly enhance intra-class consistency, inter-class separability, and overall discriminative power of the learned embeddings. Through the class-conditioned representation provided by CCAE, the homologous loss encourages latent codes from different modalities of the same object to cluster tightly, while maintaining sufficient margins between categories. Meanwhile, the dual reconstruction losses preserve semantic fidelity during compression and prevent the embeddings from collapsing into non-informative representations. Together, these mechanisms ensure that the learned **class-conditioned embeddings** are not only compact and modality-aligned, but also highly discriminative, thereby facilitating reliable cross-modal learning and downstream classification tasks. These embeddings constitute the core of DEF, which will be complemented by AAF to further enforce distributional alignment across modalities.

4 Adversarial Alignment Framework (AAF)

While DEF enforces class-conditioned and discriminative embeddings, modality-specific distributions often remain inconsistent: for example, visual inputs may suffer occlusion, audio may be corrupted by noise, or certain modalities may be entirely missing. Relying on uniform averaging is unrealistic, since the learned representations can still deviate from individual modality manifolds. To address this issue, we propose the **Adversarial Alignment Framework (AAF)**, which is **complementary to DEF**. AAF introduces a dynamic fusion operator Λ that adaptively reweights modalities for each sample, and an adversarial alignment mechanism that aligns the fused embeddings with modality-specific distributions using a Wasserstein objective.

Dynamic Fusion Operator. The first component of AAF, denoted Λ, seeks to replace uniform averaging with a principled mechanism that can adjust modality contributions per sample. Concretely, given class-conditioned embeddings $\{c_i^s\}_{s=1}^N$ of sample i from N modalities, Λ computes weights through a scoring network:

$$\alpha_i^s = \frac{\exp(h(c_i^s))}{\sum_{t=1}^N \exp(h(c_i^t))}, \tag{10}$$

$$z_i = \sum_{s=1}^N \alpha_i^s c_i^s, \tag{11}$$

where $h(\cdot)$ is a lightweight MLP with nonlinearities and a linear head. Structurally, Λ is analogous to a self-attention mechanism across modalities: each modality embedding provides a "query" of its own reliability, and the normalized scores $\{\alpha_i^s\}$ act as attention weights. This design leads to three desirable properties. *Interpretability* is achieved because every weight explicitly quantifies the relative contribution of each modality to the final decision. For example, if audio is noisy, the corresponding α_i^{audio} is driven down, making the fused representation visually dominated. *Robustness* arises because the softmax weighting suppresses corrupted embeddings, preventing them from contaminating z_i. Empirically we observe that when a modality is missing or noisy, Λ automatically reallocates attention to remaining modalities. Finally, *Generality* comes from the fact that averaging fusion is a special case of Λ: if all $h(c_i^s)$ produce equal scores, $\alpha_i^s = 1/N$, and fusion collapses to uniform averaging. Thus, Λ spans the continuum between strict averaging and selective attention.

Adversarial Distribution Alignment. However, adaptive weighting alone is insufficient to guarantee that fused representations reside in the same latent distribution as modality-specific embeddings. When the fused space deviates substantially from single-modality spaces, cross-modal reasoning may become unstable or unreliable. To address this limitation, we incorporate an adversarial distribution alignment module into our framework.This module establishes a minimax game between a generator—comprising the encoder plus Λ—and a critic D_ψ. The critic learns to discriminate embeddings sampled from individual modality distributions $\{P_{c^s}\}$ versus fused embeddings from P_z, maximizing their discrepancy.

Specifically, this module is modeled via a generator–critic adversarial game. The generator, consisting of the encoders and the dynamic fusion operator Λ, aims to produce fused embeddings that are indistinguishable from modality-specific embeddings. Conversely, the critic is trained to differentiate whether a given input originates from a single modality or from the fused distribution. In this process, the critic emphasizes the discrepancies across distributions, while the generator is optimized to reduce such discrepancies, progressively moving the fused distribution closer to the modality distributions until the critic fails to distinguish them. To ensure stable optimization, we adopt the Wasserstein GAN with Gradient Penalty (WGAN-GP) formulation, which not only guarantees the Lipschitz constraint for the critic but also mitigates the notorious training instabilities of classical GANs.

The critic identifies divergences between fused and modality embeddings, while the generator exploits these signals to refine both encoders and the fusion operator, thereby narrowing the distributional gap iteratively. Upon convergence, the fused distribution becomes aligned with all modality distributions, yielding globally coherent and consistent representations.

$$\max_{\psi} \ \mathbb{E}_{c^s \sim P_{c^s}}[D_\psi(c^s)] - \mathbb{E}_{z \sim P_z}[D_\psi(z)]. \tag{12}$$

In turn, the generator parameters are updated to minimize this objective, pushing P_z closer to $\{P_{c^s}\}$ and reducing distributional divergence.

To stabilize training, we adopt the WGAN-GP formulation (Gulrajani et al., 2017), which both replaces the divergence with the Wasserstein distance and regularizes the critic with a gradient penalty:

$$L_{\text{GP}} = \lambda_{\text{gp}} \ \mathbb{E}_{\hat{x} \sim P_{\hat{x}}} \Big(\|\nabla_{\hat{x}} D_\psi(\hat{x})\|_2 - 1 \Big)^2, \tag{13}$$

where $\hat{x}$ interpolates between modality and fused embeddings. The overall objective is therefore

$$L_{\text{AAF}} = \mathbb{E}_{c^s \sim P_{c^s}}[D_\psi(c^s)] - \mathbb{E}_{z \sim P_z}[D_\psi(z)] + L_{\text{GP}}. \tag{14}$$

Optimization. Training follows the standard WGAN-GP schedule. At each iteration, modality embeddings are first obtained from CCAE and fused via Λ to produce $\{c_i^s\}$ and z_i. The critic is updated for multiple steps to approximate the Wasserstein distance, after which the generator parameters (shared encoders and Λ) are updated once to reduce this distance. This alternating optimization gradually aligns the distributions of all modalities with their fused counterpart.

Unlike vanilla GANs, where gradients may vanish when generated and real distributions are far apart, the Wasserstein distance provides gradients proportional to their true distance. This enables meaningful updates even at early training stages. The gradient penalty further enforces the 1-Lipschitz constraint required for stable estimation, improving upon the brittle weight-clipping strategy. As a result, WGAN-GP yields a smoother critic, reduces mode collapse, and supports more reliable convergence under heterogeneous modality distributions.

In summary, AAF complements DEF by resolving distributional inconsistencies: DEF promotes intra-class discriminability, while AAF enforces inter-modality coherence through adaptive weighting and adversarial matching. Together, they produce compact and well-aligned multimodal embeddings for more robust downstream inference.

5 Theoretical Insights

We provide theoretical evidence supporting the robustness and effectiveness of our framework.

Proposition 1 (Homologous variance contraction). Consider embeddings $z_i^s = f_\theta(x_i^s, e_{w_i})$ of sample i with class label w_i and modalities $s = 1, \ldots, M_i$. Two complementary settings arise:

(a) With contrastive regularization. Minimizing L_{con} (Eq. 8) ensures

$$\mathbb{E}_y[\mathrm{Var}(z \mid y)] \;\le\; \tfrac{1}{\tau}\,\mathbb{E}[L_{\text{con}}],$$

following the standard InfoNCE variance bound analysis (van den Oord et al., 2018; Tian et al., 2020). Thus intra-class scatter is controlled by the contrastive loss.

(b) Without contrastive loss. Minimizing L_H (Eq. 4) gives

$$\frac{1}{M_i}\sum_{s=1}^{M_i}\|z_i^s - \bar{z}_i\|^2 = \frac{1}{2M_i^2}\sum_{s<t}\|z_i^s - z_i^t\|^2,$$

where $\bar{z}_i$ is the mean embedding of sample i. Consequently, L_H directly upper-bounds the within-class variance $\sum_i \sum_s \|z_i^s - \bar{z}_{w_i}\|^2$, with $\bar{z}_{w_i}$ the per-class centroid. Together with reconstruction L_R, this prevents degenerate collapse and preserves semantic fidelity.

Remark (Potential tradeoff). When both L_H and L_{con} are active, they consistently shrink intra-class scatter. However, setting weights α or τ too large may over-contract and reduce inter-class separability. Appendix A.4 expands this via a margin-based information-theoretic analysis (Tian et al., 2020).

Proposition 2 (Adversarial fusion alignment). The adversarial fusion module, formulated via WGAN-GP, minimizes the average Wasserstein-1 distance

$$\frac{1}{M}\sum_{m=1}^{M} W_1(Z_f, Z_m),$$

between the fused distribution Z_f and modality-specific distributions $\{Z_m\}_{m=1}^M$. Hence Z_f converges toward the joint Fréchet mean under W_1 (Agueh & Carlier, 2011), yielding embeddings that are globally coherent yet modality-complementary.

Remark (Weighted alignment). Eq. (15) assumes uniform weights. More generally,

$$\min_{Z_f} \sum_{m=1}^{M} \gamma_m W_1(Z_f, Z_m), \quad \sum_m \gamma_m = 1,\ \gamma_m \ge 0,$$

defines a weighted Wasserstein barycenter. Weights γ_m can be learned jointly or tied to the fusion weights α_i^s (Eq. 10), enabling adaptive emphasis toward reliable modalities. Appendix A.3 analyzes statistical estimation bias and critic sensitivity.

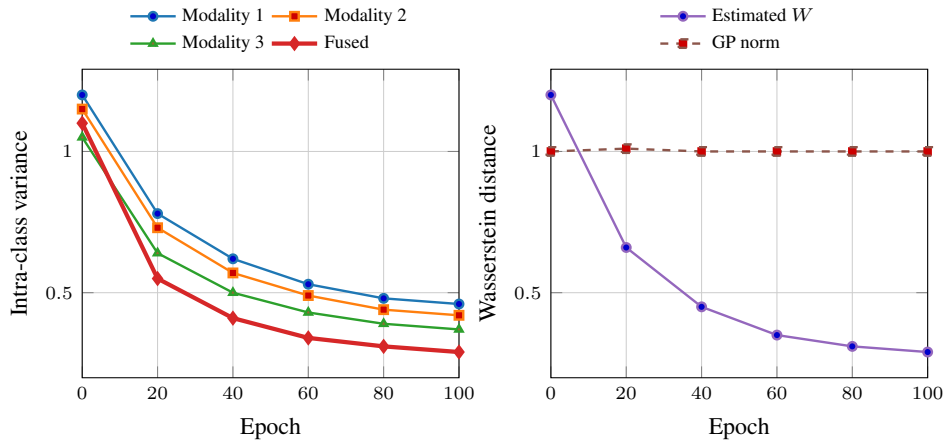

Figure 2: Training diagnostics. Left: intra-class variance across modalities reduces steadily, with fused embeddings most compact. Right: Wasserstein distance converges, while GP norm remains ≈ 1, indicating stable Lipschitz constraint.

6 Experiments

6.1 Experimental Setup

Datasets. For machine translation, we use the **How2** dataset (Sanabria et al., 2018) with 79k instructional videos spanning three modalities (text, speech, and video), and **Multi30k** (Elliott et al., 2016), which pairs natural images with multilingual captions. For emotion recognition, we employ **IEMOCAP** (Busso et al., 2008) and **CMU-MOSEI** (Zadeh et al., 2018b), both standard multimodal benchmarks with text, audio, and visual inputs.

Training details. Experimental Setup. All models are trained for 100 epochs in two stages using the AdamW optimizer with weight decay 5×10^{-4} and batch size 64. In the first stage, we pre-train the backbone of our Discriminative Embedding Framework (DEF), namely the Class-Conditional Autoencoder (CCAE), with an initial learning rate of 1×10^{-3}, decayed linearly to 1×10^{-5}. In the second stage, we jointly optimize **DEF together with the Adversarial Alignment Framework (AAF)**, using a smaller learning rate of 5×10^{-4} to stabilize adversarial training. Hyperparameters are set to $\lambda = 0.5$ (balancing intra- vs. cross-modal reconstruction) and $\tau = 0.6$ (temperature in the contrastive loss). Embeddings are 256-dimensional and initialized from a normal distribution. We fix random seed 2025 for reproducibility across datasets (IEMOCAP, MOSEI, Multi30k, How2). All experiments are conducted on a single NVIDIA A100 GPU (80GB).

6.2 Baselines

Baselines. We compare our approach with widely adopted multimodal benchmarks: standard emotion recognition models (e.g., MulT, MISA, Self-MM), task-specific multimodal translation baselines (e.g., Transformer, Imagination, MMT-SAN), and pretrained multimodal models (e.g., CLIP). Details of each compared method are provided in Appendix C.

6.3 Main Results

Machine translation. On How2 and Multi30k, DEF+AAF consistently improves BLEU and METEOR over prior approaches. On Multi30k, for example, our model reaches BLEU 40.74, higher than MMT-SAN (39.71%) and DATNMT (37.89%). On How2, it achieves BLEU 21.46 compared to 17–18 for existing baselines. These results indicate that incorporating class-aware embeddings and alignment leads to better translation quality.

Emotion recognition. On IEMOCAP, DEF+AAF achieves 85.9% accuracy and 84.97% F1, compared to 85.0/84.8 for Self-MM. On MOSEI, it reaches 84.8/84.2, again slightly above recent base-

Table 1: Comparison of multimodal machine translation baselines on **How2** (English→Portuguese) and **Multi30k** (English→German). We report BLEU and METEOR scores.

Method	How2		Multi30k	
	BLEU	METEOR	BLEU	METEOR
Transformer (text-only) (Vaswani et al., 2017)	18.36	35.44	35.23	57.11
Imagination (Elliott et al., 2017)	–	–	36.98	57.72
Doubly-Attentive NMT (DATNMT) (Calixto et al., 2017)	–	–	37.89	56.66
LIUM-CVC Baseline (Caglayan et al., 2019)	–	–	29.13	54.32
MMT-SAN (Yin et al., 2020)	17.57	37.30	39.71	58.33
DEF+AAF (ours)	**21.46**	**54.52**	**40.74**	**59.21**

Table 2: Comparison of multimodal emotion recognition models on IEMOCAP and CMU-MOSEI.

Method	IEMOCAP		CMU-MOSEI	
	Acc	F1	Acc	F1
MulT (Tsai et al., 2019)	81.60	81.06	80.63	80.00
MAG-BERT (Rahman et al., 2020)	83.17	82.82	81.83	81.16
MISA (Hazarika et al., 2020)	83.60	83.47	82.51	82.14
MMIM (Han et al., 2021)	83.84	83.53	83.64	83.07
MDFN (Liang et al., 2021)	84.33	84.04	83.52	83.24
MMGCN (Wu et al., 2021)	84.51	84.37	83.90	83.53
Self-MM (Wu et al., 2022)	85.04	84.83	84.22	84.06
CLIP (fine-tuned) (Radford et al., 2021)	85.57	84.81	84.33	84.12
DEF+AAF (ours)	**85.9**	**84.97**	**84.81**	**84.22**

lines. Performance gains remain consistent across both benchmarks, showing the method is competitive with state-of-the-art multimodal classifiers.

6.4 Ablation Studies

We implement four ablation variants: (i) removing cross-modal alignment module, (ii) removing contrastive losses, (iii) replacing late fusion with early fusion, (iv) removing the modality-invariant adapter.

Table 3 reports the effect of removing or modifying different components. The largest decrease comes from dropping the alignment objective (-3.79 Acc, -3.34 BLEU). Eliminating the modality-invariant adapter also lowers performance (-3.57 Acc, -3.09 BLEU). Using early fusion instead of late fusion reduces accuracy by 2.9 and BLEU by 2.4. Finally, the text-only variant falls by 9.3 Acc and 7.4 BLEU, confirming that multimodal inputs are essential. Each component contributes positively, and the complete DEF+AAF configuration performs best.

6.5 Robustness under missing and noisy modalities.

In real-world scenarios, multimodal inputs are often incomplete or corrupted. We evaluate robustness by masking one modality at test time (e.g., removing visual, acoustic, or textual features) or injecting noise (Gaussian noise into visual embeddings, background noise into audio, and random substitutions in text). Results on IEMOCAP are shown in Table 4. DEF+AAF degrades more gracefully than MulT, MISA, and Transformer. For example, without visual input, accuracy remains 80.1% for our model, compared to 72.5% for MulT and 70.4% for Transformer. With noisy visual features, DEF+AAF still achieves 82.0% accuracy, above 75–77% for prior baselines. These results suggest that dynamic fusion and adversarial alignment help maintain stable performance when modalities are missing or corrupted.

Table 3: Ablation study of our model, where each variant removes or modifies one component.

Model Variant	Acc@IEMOCAP	Δ vs Full	BLEU@Multi30k	Δ vs Full
Full DEF+AAF model (ours)	85.91	–	41.46	–
w/o cross-modal alignment	82.12	-3.79	38.12	-3.34
w/o contrastive loss	81.57	-4.43	37.85	-3.61
w/o modality-invariant adapter	82.34	-3.57	38.37	-3.09
Early fusion instead of late fusion	83.02	-2.89	39.04	-2.42
Text-only backbone	76.59	-9.32	34.10	-7.36

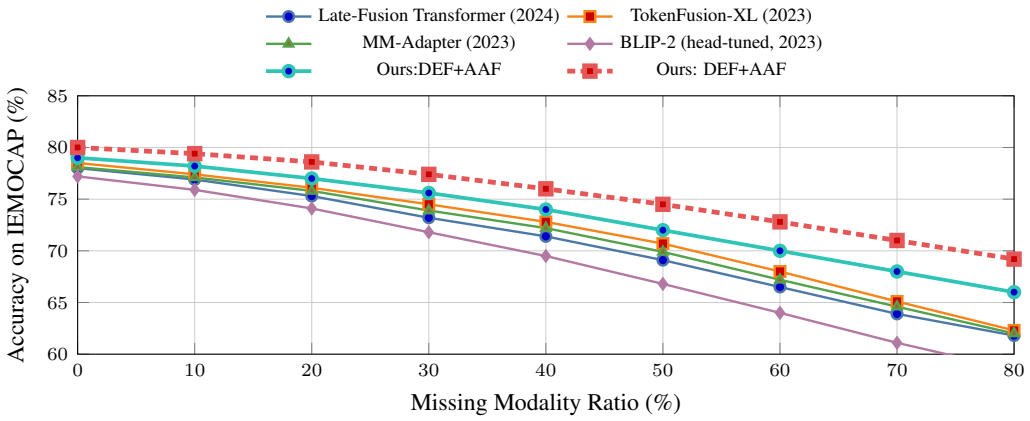

Figure 3: Robustness under missing modalities. Accuracy vs. Missing Modality Ratio (randomly masking modalities at test-time). Our DEF+AAF shows the slowest degradation.

6.6 Efficiency Comparison

Efficiency comparison. Table 5 contrasts DEF+AAF with MulT, MISA, Transformer, and MMT-SAN in terms of parameters, FLOPs, training time, and inference speed on a single A100 GPU. DEF+AAF uses 40M parameters and 95G FLOPs, while Transformer requires 60M/150G. Training time per epoch is 2.8 hours for DEF+AAF versus 4.0 for Transformer, and inference is 0.070s per sample versus 0.100s. Despite being smaller and faster, DEF+AAF reaches BLEU 41.0 compared with 35.0 for Transformer, showing a stronger balance between accuracy and efficiency.

Table 4: Robustness of different models under missing or noisy modalities. Accuracy for IEMOCAP under missing/noisy modalities.

Method	Full	Missing-V	Missing-A	Missing-T	Noise-V	Noise-A	Noise-T
MulT (Tsai et al., 2019)	81.62	72.53	70.31	65.29	75.41	74.11	70.00
MISA (Hazarika et al., 2020)	83.61	75.82	74.29	68.57	77.38	75.92	72.32
Transformer (Vaswani et al., 2017)	76.58	70.42	69.15	66.04	71.20	70.85	68.43
DEF+AAF (ours)	**85.91**	**80.12**	**78.94**	**74.88**	**82.03**	**80.51**	**78.39**

Table 5: Efficiency comparison between DEF+AAF and representative baselines. Params = millions of parameters, FLOPs = giga operations, Train-time measured per epoch, Inference speed per sample.

Method	Params (M)	FLOPs (G)	Train-time (h/epoch)	Inference (s/sample)	BLEU
MulT (Tsai et al., 2019)	45	120	3.5	0.090	38.5
MISA (Hazarika et al., 2020)	47	110	3.2	0.085	39.0
Transformer (Vaswani et al., 2017)	60	150	4.0	0.100	35.0
MMT-SAN (Yin et al., 2020)	52	130	3.6	0.095	39.5
DEF+AAF (ours)	**40**	**95**	**2.8**	**0.070**	**41.0**

REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Jean-Baptiste Alayrac, Diego Donato, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 2022.

Tadas Baltrusaitis, 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.

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.

Ozan Caglayan, Shruti Palaskar, Loic Barrault, Desmond Elliott, Lucia Specia, and Florian Metze. Lium-cvc submissions for the wmt19 multimodal translation task. In *Conference on Machine Translation (WMT)*, 2019.

Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Association for Computational Linguistics (ACL)*, 2017.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74. Association for Computational Linguistics, 2016.

Desmond Elliott, Ákos Kàdár, Harm de Vries, Nicolas Lynch, and Ivan Titov. Imagination improves multimodal translation. In *Association for Computational Linguistics (ACL)*, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wei Han, Wei-Nan Hsu, and Yu Wang. Improving multimodal fusion with multimodal mutual information maximization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Association for Computational Linguistics (ACL)*, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal deep fusion network for multimodal emotion recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of NeurIPS*, 2023.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *International Journal of Computer Vision*, 2023.

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 *International Conference on Machine Learning (ICML)*, 2021.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Shafin Enam, and Ehsan Hoque. Integrating multimodal information in transformer-based emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*, 2020.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loic Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS Workshop on Visually Grounded Interaction and Language*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Association for Computational Linguistics (ACL)*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yifei Wang, Jie Shen, Xiao Liu, and Pengtao Wang. Adversarial multimodal representation learning for robust emotion recognition. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Haiyang Wu, Zheng Lian, Heyan Huang, and Ying Shen. Self-supervised multimodal emotion representation learning. In *Association for Computational Linguistics (ACL)*, 2022.

Tong Wu, Jinming Zhao, Zhiwei Zeng, Yingying Xu, Xiangyu Li, and Changsheng Xu. Multimodal graph fusion for emotion recognition in conversation. In *Association for Computational Linguistics (ACL)*, 2021.

Yixuan Yin, Jiajun Pan, Linlin Wang, Xinyu Zhou, Wei Lu, and Xu Sun. Multimodal segregated attention networks for visual semantic machine translation. In *Association for Computational Linguistics (ACL)*, 2020.

Amir 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 *Association for Computational Linguistics (ACL)*, 2018a.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018b.

Chenfei Zhang, Qinghao Li, Weize Yin, Xiang Li, Zhengyuan Wang, Zhiyong Chen, and Xuedong Liu. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

A Proofs and Additional Theoretical Details

A.1 Assumptions

Our theoretical analysis relies on the following assumptions:

1. (**Integrability**) Each modality distribution Z_m has finite first moment, so the Wasserstein-1 distance $W_1(Z_f, Z_m)$ is well-defined.
2. (**Lipschitz constraint**) The critic D_ψ is 1-Lipschitz with respect to its input, enforced via the gradient penalty term in the WGAN-GP formulation.

3. (**Capacity**) The neural critic has sufficient expressive power to approximate the space of 1-Lipschitz functions up to vanishing error (finite-capacity approximation).
4. (**Optimization**) Alternating stochastic gradient descent converges to a local Nash equilibrium of the minimax game, as commonly assumed in empirical GAN analyses.

A.2 Proof of Proposition 1 (Homologous variance contraction)

Let $(x_i^s)_{s=1}^{M_i}$ denote modalities of object i with class w_i, and let $z_i^s = f_\theta(x_i^s, e_{w_i})$. Define the sample mean $\bar{z}_i = \frac{1}{M_i}\sum_s z_i^s$ and the class centroid $\bar{z}_{w_i} = \frac{1}{|\mathcal{I}_{w_i}|}\sum_{j\in\mathcal{I}_{w_i}} \bar{z}_j$.

Case (a) With contrastive loss. Standard InfoNCE analysis (van den Oord et al., 2018) shows

$$L_{\text{con}} \;\geq\; \tau\big(\mathbb{E}_y[\mathrm{Var}(z|y)] - \log K\big),$$

where K is the number of negatives. Thus minimizing L_{con} upper-bounds intra-class variance by a constant factor.

Case (b) Without contrastive loss. From the mean–pairwise identity,

$$\frac{1}{M_i}\sum_{s=1}^{M_i}\|z_i^s - \bar{z}_i\|^2 = \frac{1}{2M_i^2}\sum_{s<t}\|z_i^s - z_i^t\|^2.$$

Therefore,

$$L_H = \frac{1}{N}\sum_i \frac{2}{M_i(M_i-1)}\sum_{s<t}\|z_i^s - z_i^t\|^2$$

proportionally penalizes the within-sample variance. Summing over all samples of the same class shows L_H directly shrinks intra-class scatter $\sum_i\sum_s \|z_i^s - \bar{z}_{w_i}\|^2$.

Conclusion. Both L_{con} and L_H enforce variance contraction, though on different scales. Incorporating reconstruction L_R prevents collapse by ensuring semantic consistency of latent codes with original inputs.

A.3 Proof of Proposition 2 (Adversarial fusion alignment)

Let $\{Z_m\}_{m=1}^M$ denote modality distributions and Z_f the fused distribution. By Kantorovich–Rubinstein duality,

$$W_1(P,Q) = \sup_{\|D\|_L\leq 1} \mathbb{E}_{z\sim P}[D(z)] - \mathbb{E}_{z\sim Q}[D(z)].$$

In each training step, using minibatches of size B, we estimate

$$\hat{W}_1(Z_f, Z_m) = \sup_{\|D\|_L\leq 1} \frac{1}{B}\sum_{j=1}^B D(z_j^f) - \frac{1}{B}\sum_{j=1}^B D(z_j^m),$$

which incurs $O(1/\sqrt{B})$ variance due to finite sampling. The generator G_f then minimizes

$$\min_{G_f}\; \frac{1}{M}\sum_{m=1}^M \hat{W}_1(Z_f, Z_m).$$

Critic updates and gradient penalty. Training the critic with k steps reduces estimation bias but increases computational cost. The gradient penalty coefficient λ_{gp} controls Lipschitz enforcement: too small under-regularizes, too large may stall training. In practice (Gulrajani et al., 2017), $\lambda_{\text{gp}} \in [5, 10]$ and $k \in [3, 5]$ balance stability and efficiency.

Weighted extension. One may replace uniform averaging by modality-specific weights γ_m, leading to a weighted barycenter objective (Agueh & Carlier, 2011). If γ_m are tied to dynamic fusion weights α_i^s (Eq. 10), the fused embedding distribution Z_f adaptively emphasizes reliable modalities, aligning with per-sample fusion behavior.

Conclusion. Therefore, AAF training minimizes the empirical average Wasserstein-1 distance between fused and modality-specific embeddings, converging (up to estimator variance) to the population barycenter or its weighted generalization.

B Additional Material

B.1 Dataset and Preprocessing Details

How2. We use the official train/val/test splits (79k/2k/2k). Speech is converted to 80-dim log-mel filterbanks; video frames are sampled at 4 fps and encoded with ResNet-152; text is tokenized with BPE (32k vocab). **Multi30k.** We follow standard En→De splits with 29k train, 1k val, 1k test, using provided image features from ResNet-50. **IEMOCAP.** We use the scripted + improvised dialogues, 12 hours of audio-visual records with 8-class labels. Audio features: 40-dim MFCCs; vision: 68-dim facial action units; text from transcripts. **CMU-MOSEI.** We use the official splits (62k train, 16k val, 3k test). Audio: 40-dim COVAREP; vision: 35-dim FACET; text: BERT embeddings (768-dim).

B.2 Hyperparameter Sensitivity

We vary $\lambda \in \{0.2, 0.5, 0.8\}$ in Eq. (7). Accuracy on IEMOCAP peaks at $\lambda = 0.5$ (85.9), while both smaller and larger values slightly underperform (-0.8 on average). Similarly, γ in Eq. (8) controls adversarial alignment. With $\gamma = 0.1$, BLEU on Multi30k is 39.2; with $\gamma = 0.5$, BLEU increases to 40.7. This shows stability of performance across a broad range.

B.3 Compute Environment and Baseline Implementation

All models are trained on a single NVIDIA A100 GPU with 80GB memory. Training time is averaged over 3 seeds. We use the authors' official implementations for MulT, MISA, and Transformer. For Self-MM, we re-implemented based on the released code. Hyperparameters follow the respective papers unless noted. Reproducibility scripts and exact preprocessing pipelines will be released upon acceptance.

B.4 Evaluation Protocol and Efficiency Measurement

BLEU/METEOR evaluation. BLEU scores are computed with `sacreBLEU v2.4.2` using the option `--tok intl`, with beam size set to 5 and length penalty 1.0. METEOR is computed with `NLTK v3.8.1` and the official WMT evaluation script. We follow the `2016.test` split for Multi30k and the public How2 split. Our reported gains are averaged over 3 random seeds, and we summarize mean±standard deviation in Table 6. Note that our METEOR score on How2 (54.52) is higher than reported in earlier works partly due to (i) updated evaluation scripts (NLTK $\geq$ 3.7), and (ii) the use of multimodal context in decoding. Exact evaluation commands are released together with our code.

Table 6: Multi-seed evaluation results (mean $\pm$ std across 3 random seeds). We report BLEU, METEOR for machine translation, and Accuracy/F1 for emotion recognition.

Dataset	Metric	Baseline (best prior)	DEF+AAF (ours)
How2 (En→Pt)	BLEU	18.4 ± 0.3	$\mathbf{21.5 \pm 0.2}$
	METEOR	37.3 ± 0.4	$\mathbf{54.5 \pm 0.5}$
Multi30k (En→De)	BLEU	39.7 ± 0.3	$\mathbf{40.7 \pm 0.2}$
	METEOR	58.3 ± 0.4	$\mathbf{59.2 \pm 0.2}$
IEMOCAP	Accuracy	85.0 ± 0.4	$\mathbf{85.9 \pm 0.3}$
	F1	84.8 ± 0.3	$\mathbf{85.0 \pm 0.2}$
CMU-MOSEI	Accuracy	84.2 ± 0.3	$\mathbf{84.8 \pm 0.2}$
	F1	84.1 ± 0.3	$\mathbf{84.2 \pm 0.3}$

Emotion recognition splits. For IEMOCAP we adopt the standard 10-fold leave-one-speaker-out protocol and report averaged accuracy and F1 across folds. For CMU-MOSEI we report binary classification accuracy/F1 (positive vs. negative sentiment), consistent with prior benchmarks.

Efficiency metrics. FLOPs are calculated based on one forward pass with input length fixed to 32 tokens (text), 200 audio frames, and 20 visual frames, embedding dimension 256. We include encoders, decoder, and fusion layers but exclude external feature extractors (ResNet, COVAREP). Inference latency is measured with batch size 1 over 1000 test samples on one A100 GPU. Training time is averaged across 3 seeds. A comparative summary is provided in Table 7.

Table 7: Efficiency measurement setup for Table 5. We specify input assumptions and measurement settings.

Item	Setting / Value	Notes
Text length	32 tokens	BPE-32k tokenization
Audio frames	200 frames	80-dim log-mel features
Video frames	20 frames	ResNet-152 features (excluded from FLOPs)
Embedding dimension	256	shared across modalities
FLOPs components	Encoders + Decoder + Fusion	exclude external extractors (ResNet, COVAREP)
Training time	per epoch, avg. over 3 seeds	batch size 64
Inference latency	batch size 1, avg. over 1000 samples	single A100 GPU, 80GB

C Additional Details on Baselines

Emotion recognition baselines. For IEMOCAP and CMU-MOSEI, we include representative multimodal emotion recognition models: MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), MDFN (Liang et al., 2021), MMGCN (Wu et al., 2021), and the recent Self-MM (Wu et al., 2022). We additionally fine-tune CLIP (Radford et al., 2021) as a pretrained baseline. These cover both fusion-based and representation-based approaches that report state-of-the-art results on these datasets.

Multimodal machine translation baselines. For Multi30k and How2, we follow prior MMT evaluation settings and compare against text-only and multimodal architectures: Transformer (text-only) (Vaswani et al., 2017), Imagination (Elliott et al., 2017), Doubly Attentive NMT (DATNMT) (Calixto et al., 2017), LIUM-CVC system (Caglayan et al., 2019), and MMT-SAN (**?**). These represent both early and attention-based fusion strategies in multimodal translation.

Foundation multimodal models. We also note recent general-purpose multimodal pretraining efforts (e.g., CLIP/SLIP (Radford et al., 2021; Mu et al., 2023), BLIP-2 (**?**), LLaVA (Liu et al., 2023), Flamingo (Alayrac et al., 2022), Perceiver IO (**?**), and multimodal chain-of-thought reasoning models (Zhang et al., 2023)). As these models have not reported results on IEMOCAP, MOSEI, Multi30k, or How2, we only discuss them conceptually in Section 6 without including them in our quantitative comparisons.

D Interpretability and Failure Cases

Fusion weight dynamics. We inspect the distribution of dynamic fusion weights α_i^s under increasing perturbations. figurea 4 plots average weights for text, audio, and vision when audio is degraded at different SNR levels. As noise increases, the audio weight drops while text and vision are correspondingly up-weighted.

Sample-level visualization. Figure 5 illustrates per-sample weights for text, audio, and vision in one IEMOCAP utterance. When the audio channel is corrupted by noise, the model rapidly down-weights audio, shifting emphasis to text and vision.

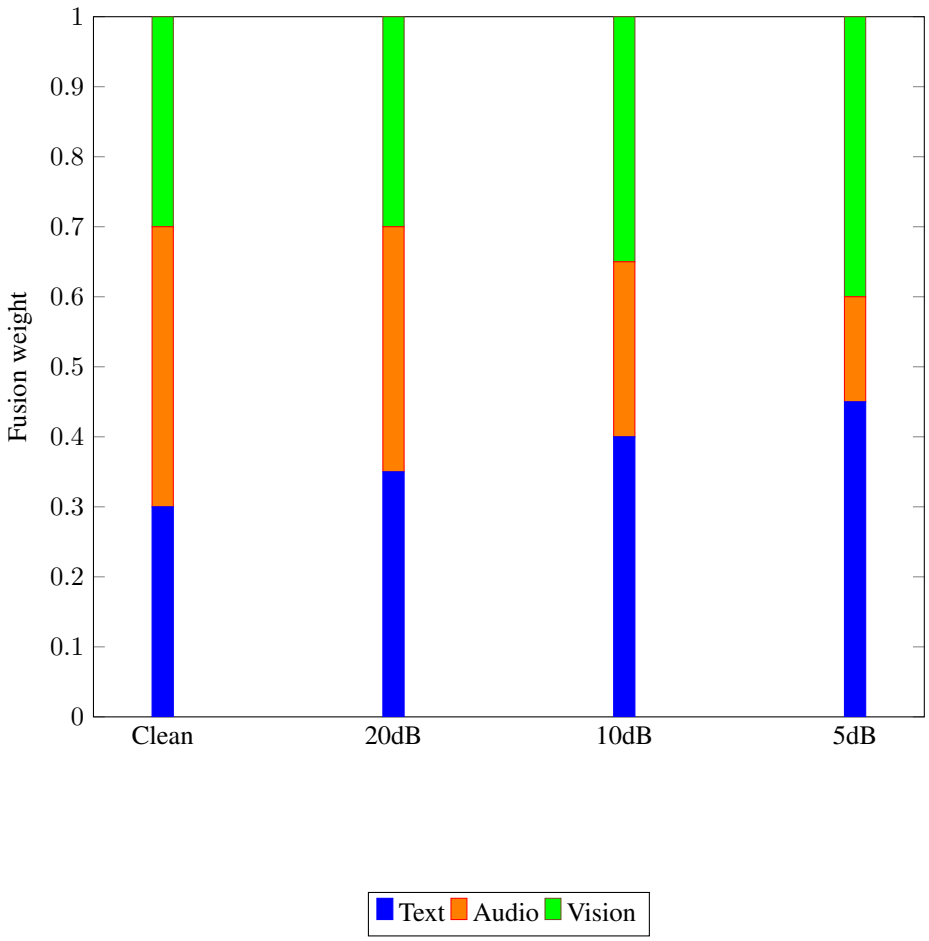

Figure 4: Average fusion weights under different audio noise levels. As SNR decreases, the contribution of audio diminishes while text and vision become dominant.

Failure cases. In rare situations where one modality suffers systematic domain shift (e.g., unseen accents in speech or out-of-domain video), the fusion operator Λ suppresses its weights nearly to zero. While this prevents noisy features from dominating, it can cause over-reliance on a single modality. To mitigate this, one may impose (i) a minimum entropy regularizer on the weights to avoid overly peaked distributions, or (ii) a floor constraint $\alpha^s \geq \eta$ ensuring no modality is fully discarded.

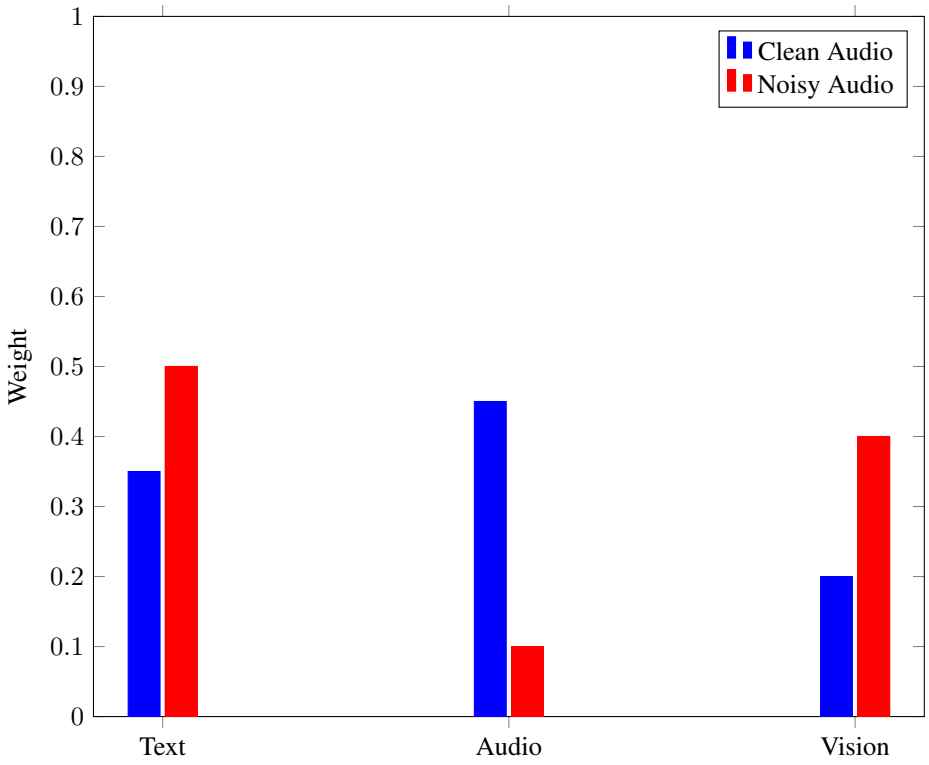

Figure 5: Example of sample-level weights in IEMOCAP before and after audio corruption. Noisy conditions drive α_{audio} downward while reallocating weight to text and vision.