Anchoring the Affective Manifold: Learning Canonical and Disentangled Representations via Generative Cross-Modal Alignment

Weibin Li[1,2,†] **Jintao Cheng**[3,†] **Xiaoyu Tang**[1,†,*] **Chi Man VONG**[2,*]
[1]South China Normal University [2]University of Macau
[3]Hong Kong University of Science and Technology
tangxy@scnu.edu.cn cmvong@um.edu.mo

Abstract

Dominant multimodal emotion recognition paradigms often neglect the intrinsic geometric structure of affect, resulting in representations heavily entangled with non-affective factors. To address this, we propose a Canonical Disentangled Multimodal Generative Framework aimed at recovering the canonical affective manifold from raw data. We explicitly decompose the latent space into a canonical Shared Affective Subspace (z_{vad}) and a Private Modality Subspace (z_{priv}). We facilitate this factorization through Supervised Manifold Anchoring and Cross-Modal Manifold Alignment. Experiments demonstrate that our model effectively disentangles affect from private attributes (e.g., identity), achieving superior robustness in zero-shot cross-domain transfer compared to fully supervised baselines, while enabling controllable emotion generation.

1 Introduction

Human emotional expression is intrinsically a multimodal cognitive process, orchestrating complex interactions across linguistic semantics, acoustic prosody, and facial dynamics (Picard, 2000; Lai et al., 2023). With the proliferation of social media and human-machine interaction systems, Multimodal Affective Computing has emerged as a pivotal field (Baltrušaitis et al., 2018). Unlike unimodal analysis, multimodal learning aims to leverage the complementarity among heterogeneous data sources to resolve ambiguity inherent in isolated signals (D'mello and Kory, 2015). However, recovering robust, latent affective representations from high-dimensional, asynchronous, and noisy multimodal streams remains a fundamental scientific challenge (Bengio et al., 2013).

The research on multimodal continuous emotion computation has undergone an evolution from pure prediction to structure-awareness. The earliest and most mature line of work is dedicated to Continuous Affect Recognition. Represented by the AVEC Challenge series, these studies utilize LSTMs or Transformers to directly map multimodal inputs to Valence-Arousal values (Ringeval et al., 2015; Kollias et al., 2019). Although these methods achieve continuous value output, they essentially belong to Supervised Prediction rather than representation learning. The models focus solely on the precision of the $X \rightarrow Y$ mapping, while ignoring the intrinsic structure of the latent space Z. To remedy this defect, transitional works in recent years have started to explore VAD-aware Representation. Specifically, the VAD model characterizes affective states as points within a continuous three-dimensional space defined by Valence (pleasure), Arousal (intensity), and Dominance (control) (Mehrabian, 1996). For instance, some approaches attempt to introduce explicit VAD latent variables (Yang et al., 2023) or bridge discrete emotions with continuous spaces (Jia et al., 2025). However, these attempts mostly still rely on full supervision signals and follow discriminative paradigms, failing to construct a generative emotional manifold truly independent of the task.

In the other dimension of Multimodal Representation Learning, although generative models and disentanglement ideas have emerged, there remains a significant gap towards the ideal canonical continuous affective representation. Existing Multimodal VAE works (Wu and Goodman, 2018) introduce a shared latent variable z, but their latent space is typically an unstructured isotropic Gaussian distribution. This space lacks clear physical semantics and leads to serious Entanglement of emotional information with private attributes like speaker identity. On the other hand, disentanglement learning works represented by MISA (Hazarika et al., 2020) successfully separate Modality-Invariant and Modality-Specific features, but they are intrinsi-

[†]Equal contribution
[*]Corresponding Authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41605–41614

cally discriminative and lack explicit modeling of Continuous Affective Geometry. Recent advanced explorations have begun to address these issues, such as fuzzy VAD representations for EEG (Asif et al., 2024) and variational disentanglement for stance detection (Xu et al., 2025). In summary, a systematic solution for canonical, disentangled, and manifold-based continuous affective representation is still lacking. To date, the restoration of the affective manifold under a unified structurally anchored generative framework remains a largely unexplored challenge.

To transcend these limitations, we propose a Canonical Disentangled Multimodal Generative Framework that reconceptualizes affective modeling from discriminative label-fitting to generative manifold recovery. Specifically, we instantiate a Variational Autoencoder (VAE) (Kingma and Welling, 2013) to explicitly factorize the joint multimodal distribution into two orthogonal subspaces: a Shared Affective Subspace (z_{vad}) serving as a canonical coordinate system, and a Private Modality Subspace (z_{priv}) encapsulating non-affective residues. To bridge the modality gap within this disentangled structure, we implement a Cross-Modal Manifold Alignment mechanism, where mutual reconstruction objectives compel heterogeneous encoders to reach a consensus on the affective state, thereby filtering out modality-specific noise. Crucially, to refine the manifold geometry, we orchestrate a Dual-Anchoring Strategy that enforces strict geometric constraints using ground-truth labels for empirical precision, while simultaneously injecting Linguistic Priors (via the NRC-VAD lexicon (Mohammad, 2025)) as a semantic regularizer to enhance structural robustness against dataset-specific biases.

In summary, our main contributions are as follows:

- We propose a multimodal generative framework that integrates explicit disentanglement with canonical VAD anchoring. By encouraging an orthogonal decomposition between affective and private factors in the latent space, we establish a psychologically interpretable coordinate system that improves the structural semantics of conventional latent representations.
- We design a Cross-Modal Mutual Reconstruction mechanism that acts as a synchronization signal across modalities. It encourages semantic consistency among heterogeneous modalities and helps align the learned manifold toward a shared affective structure.
- Extensive experiments demonstrate the robustness and interpretability of the learned manifold under our evaluation settings. Our model shows strong performance in disentanglement verification and indicates that the learned representation can provide meaningful affective guidance for generation, beyond surface-level statistical correlations.

2 Related Work

2.1 Multimodal Emotion Modeling

Multimodal emotion recognition has evolved significantly with the advancement of deep learning strategies for fusing heterogeneous signals. Early influential works focused on capturing intermodal dynamics through tensor-based operations, such as the Tensor Fusion Network (TFN) (Zadeh et al., 2017) and Low-rank Multimodal Fusion (LMF) (Liu et al., 2018), which modeled the outer product of unimodal representations. With the advent of attention mechanisms, the focus shifted towards alignment and cross-modal interaction. The Multimodal Transformer (MulT) (Tsai et al., 2019) and MAG-BERT (Rahman et al., 2020) leveraged cross-attention to dynamically align unaligned multimodal sequences. Recent approaches have further refined representation learning. Self-MM (Yu et al., 2021) employs self-supervised multi-task learning to improve unimodal feature quality, while MMIM (Han et al., 2021) maximizes mutual information to preserve task-relevant content while reducing redundancy. Other architectures like CubeMLP (Sun et al., 2022) and hierarchical fusion networks (Lv et al., 2021) explore efficient fusion pathways. Despite their discriminative success, these methods typically map inputs directly to labels, neglecting the generative process and the intrinsic topological structure of the affective latent space.

2.2 Generative Representation Learning

Disentangled Representation Learning (DRL) seeks to partition data into independent latent factors (Bengio et al., 2013). Early unsupervised paradigms, such as β-VAE (Higgins et al., 2017) and FactorVAE (Kim and Mnih, 2018), utilized information bottlenecks and total correlation constraints to enforce independence, while InfoGAN (Chen et al., 2016) employed adversarial mutual information maximization. However, following Locatello et al.'s (Locatello et al., 2019) proof

that unsupervised DRL is fundamentally ill-posed without inductive biases, research has pivoted toward supervised (Reed et al., 2014) and weakly-supervised approaches (Bouchacourt et al., 2018; Locatello et al., 2020) that leverage external signals or grouped observations. Our work extends these principles to the multimodal affective domain. We extend DRL principles to recover a canonical VAD manifold, orchestrated via a dual-anchoring mechanism that synergizes empirical supervision with linguistic regularization.

3 Method

In this section, we elucidate our proposed Canonical Disentangled Multimodal Generative Framework. As illustrated in Figure 1, the overarching objective is to learn a structurally anchored, continuous affective manifold by explicitly factorizing the joint multimodal distribution into a canonical shared subspace and orthogonal modality-specific private subspaces.

3.1 Problem Formulation and Notation

Formally, let the multimodal dataset be denoted as $\mathcal{D} = \{(x_t^{(i)}, x_a^{(i)}, x_v^{(i)})\}_{i=1}^{N}$, where x_t, x_a, x_v represent the aligned input sequences for textual, acoustic, and visual modalities, respectively. Our primary objective is to learn a joint mapping function Φ that projects these heterogeneous modalities into a disentangled latent space $\mathcal{Z}$.

We explicitly decompose the latent space $\mathcal{Z}$ into two orthogonal subspaces:

- Shared Affective Subspace ($\mathcal{S}$): Represented by the latent variable $z_{vad} \in \mathbb{R}^3$. To endow this space with interpretability, we constrain its three dimensions to correspond to the psychological primitives of Valence, Arousal, and Dominance. This subspace captures the affective state invariant across modalities.

- Private Modality Subspace ($\mathcal{P}_m$): Represented by the latent variable $z_{priv}^m \in \mathbb{R}^{d_m}$, where $m \in \{t, a, v\}$. This subspace captures modality-specific variations unrelated to emotion, such as textual syntax, acoustic timbre, or visual identity features.

3.2 Unimodal Backbones and Preprocessing

To extract robust feature representations from heterogeneous data sources, we employ specific pre-trained backbone networks for each modality. For the textual modality, we encode input utterances using RoBERTa (Liu et al., 2019) and select the embedding of the start token '<s>' from the final layer as the global semantic feature h_t. For the acoustic modality, raw audio waveforms are processed by the data2vec (Baevski et al., 2022) model. We apply a mean-pooling operation over the temporal dimension of the output sequence to obtain a fixed-length acoustic embedding h_a. Finally, for the visual modality, we utilize TimeSformer (Bertasius et al., 2021) to model the spatiotemporal dynamics of facial video clips, taking the output of the classification token '[CLS]' as the resulting visual representation h_v.

3.3 Disentangled Variational Encoding and Generation

To recover the underlying emotional manifold from heterogeneous multimodal data, we have developed a generative model based on variational inference. Our central hypothesis is that the data generation process of any modality is governed by two statistically independent latent factors: one is the cross-modal shared emotional state z_{vad}, and the other is the modality-specific private attribute z_{priv}^m. We factorize the latent space into two statistically independent distributions:

$$q_\phi(z_{vad}^m, z_{priv}^m | h_m) = q_\phi(z_{vad}^m | h_m) \cdot q_\phi(z_{priv}^m | h_m) \quad (1)$$

In implementation, the encoder E_m maps the input h_m to the parameters of two Gaussian distributions. We define two independent projection heads (MLPs):

$$[\mu_{vad}^m;\ \log(\sigma_{vad}^m)^2] = \mathcal{F}_{vad}(h_m; \phi_{vad}) \quad (2)$$

$$[\mu_{priv}^m;\ \log(\sigma_{priv}^m)^2] = \mathcal{F}_{priv}(h_m; \phi_{priv}) \quad (3)$$

where $\mathcal{F}$ denotes a Multi-Layer Perceptron, and $[\cdot;\cdot]$ denotes vector concatenation. Utilizing the Reparameterization Trick, the sampling process for the latent variables is formalized as:

$$z_k^m = \mu_k^m + \sigma_k^m \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad k \in \{vad, priv\} \quad (4)$$

The decoder D_θ is designed to reconstruct the original high-level features based on these latent representations. In the Self-Reconstruction phase, the decoder receives the pair of latent variables from the same source modality. The generation process is denoted as $\hat{h}_{m \to m} = D_m(z_{vad}^m, z_{priv}^m; \theta_m)$. Consequently, the intra-modal reconstruction loss is defined as:

$$\mathcal{L}_{self} = \sum_{m \in \mathcal{M}} \mathbb{E}_{q_\phi(z|h_m)} \left[||h_m - D_m(z_{vad}^m, z_{priv}^m)||_2^2 \right] \quad (5)$$

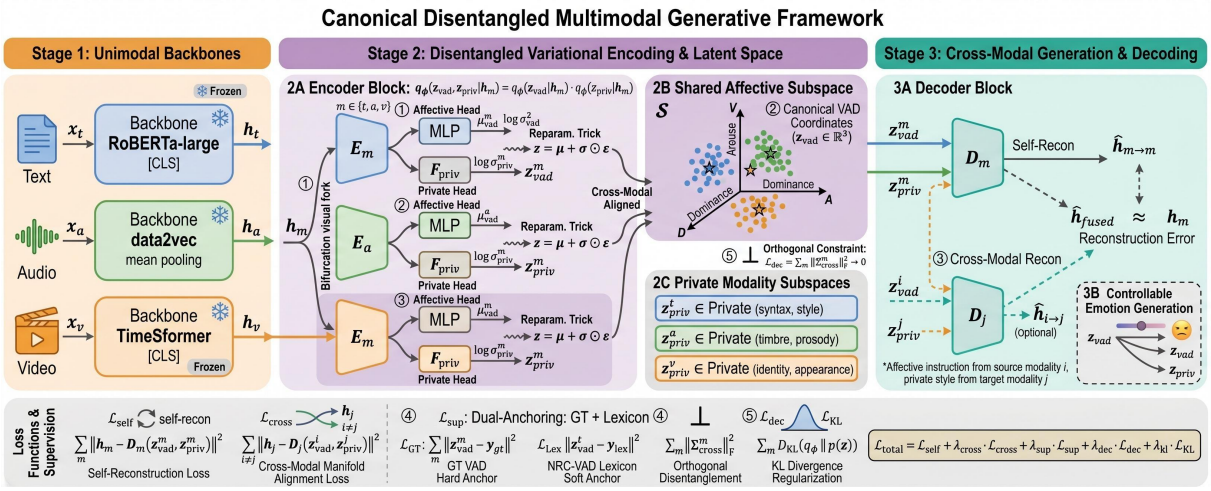

Figure 1: Overall framework of the proposed Canonical Disentangled Multimodal Generative Framework.

3.4 Cross-Modal Manifold Alignment

To map heterogeneous modality features onto a unified affective manifold $\mathcal{S}$, we define a cross-modal generation function $\mathcal{G}(i, j)$, where the source modality i provides the affective instruction z^i_{vad}, and the target modality j provides the private content z^j_{priv}. The core of this alignment mechanism lies in minimizing the discrepancy between the generated features and the ground-truth features of the target modality.

For all non-identical modality pairs $(i, j) \in \mathcal{M} \times \mathcal{M}, i \neq j$, the cross-modal reconstructed feature is calculated as:

$$\hat{h}_{i\to j} = D_j(z^i_{vad}, z^j_{priv}; \theta_j) \tag{6}$$

We define the Manifold Alignment Loss $\mathcal{L}_{cross}$ as the expected reconstruction error over all permutations:

$$\mathcal{L}_{cross} = \sum_{i\in\mathcal{M}} \sum_{j\in\mathcal{M}, j\neq i} \mathbb{E}_{q_\phi}\left[||h_j - \hat{h}_{i\to j}||_2^2\right] \tag{7}$$

This objective function not only enforces cross-modal semantic consistency within the z_{vad} space but also acts as a self-supervised regularizer, suppressing the encoding of modality-specific noise into the shared affective variables.

3.5 Dual-Anchored Manifold Supervision

To construct a latent manifold that is both empirically precise and semantically robust, we employ a Dual-Anchoring Strategy. First, we utilize dataset-provided VAD labels y_{gt} as a hard geometric constraint, strictly anchoring the shared latent variable z_{vad} to physical psychological coordinates across all modalities via $\mathcal{L}_{GT} = \sum_{m\in\{t,a,v\}} ||z^m_{vad} - y_{gt}||_2^2$. Simultaneously, to mitigate overfitting to dataset biases, we leverage the NRC-VAD lexicon as a task-agnostic semantic regularizer. We compute sentence-level pseudo-labels y_{lex} via EmotionDynamics (aggregating word-level scores with context adjustments) and apply them to the textual embedding z^t_{vad} through $\mathcal{L}_{Lex} = ||z^t_{vad} - y_{lex}||_2^2$. The final supervision objective is a weighted sum: $\mathcal{L}_{sup} = \mathcal{L}_{GT} + \lambda_{lex}\mathcal{L}_{Lex}$.

3.6 Orthogonal Disentanglement and Regularization

To encourage statistical independence between the shared space $\mathcal{S}$ and the private space $\mathcal{P}_m$, we employ an orthogonality constraint to approximate the minimization of Mutual Information. For a batch containing B samples, let $Z^m_{vad} \in \mathbb{R}^{B\times 3}$ and $Z^m_{priv} \in \mathbb{R}^{B\times d_m}$ denote the de-meaned latent variable matrices, respectively. We compute the cross-covariance matrix between them:

$$\Sigma^m_{cross} = \frac{1}{B-1}(Z^m_{vad})^\top Z^m_{priv} \tag{8}$$

The disentanglement regularization term $\mathcal{L}_{dec}$ is defined as the squared Frobenius norm of this matrix:

$$\mathcal{L}_{dec} = \sum_{m\in\mathcal{M}} ||\Sigma^m_{cross}||_F^2 \tag{9}$$

Furthermore, to constrain the topology of the latent space and support generative tasks, we introduce the standard KL divergence loss, forcing the posterior distribution to approximate the standard

normal prior $p(z) = \mathcal{N}(0, I)$:

$$\mathcal{L}_{KL}=\sum_{m\in\mathcal{M}} \beta\cdot D_{KL}\big(q_\phi(z^m_{vad},z^m_{priv}|h_m) \;||\; p(z_{vad})p(z_{priv})\big) \tag{10}$$

3.7 Optimization Objective

To strike a balance between feature reconstruction fidelity, cross-modal manifold alignment, and the degree of latent disentanglement, we formulate the overall optimization objective as a weighted summation of the constituent losses:

$$\begin{aligned}\min_{\phi,\theta} \mathcal{L}_{total} = \mathcal{L}_{self} + \lambda_{cross}\mathcal{L}_{cross} \\ + \lambda_{sup}\mathcal{L}_{sup} + \lambda_{dec}\mathcal{L}_{dec} \\ + \lambda_{kl}\mathcal{L}_{KL}\end{aligned} \tag{11}$$

where $\lambda_{cross}, \lambda_{sup}, \lambda_{dec}$, and λ_{kl} are hyperparameters that govern the relative importance of each regularization term. By minimizing this objective, the model learns a structured and disentangled shared affective manifold aligned with the NRC-VAD linguistic priors, while simultaneously preserving modality-specific private characteristics.

4 Experiment

4.1 Datasets

To comprehensively evaluate the quality, disentangling capability, and generalization performance of the learned representations, we selected three widely used multimodal emotion benchmark datasets.

The IEMOCAP(Busso et al., 2008) dataset consists of approximately 12 hours of dyadic conversation videos, performed by 10 actors in both improvised and scripted scenarios. It provides not only discrete emotional category labels (such as Happy, Sad, Angry, Neutral) but also detailed continuous ratings for the Valence, Arousal, and Dominance (VAD) dimensions, with a scale ranging from 1 to 5. Additionally, the explicit speaker identity annotations (Speaker ID) make it an ideal experimental platform for validating the disentangling performance, i.e., the separation of emotion from speaker identity.

The CMU-MOSI(Zadeh et al., 2016) and CMU-MOSEI(Zadeh et al., 2018) datasets primarily consist of single-speaker movie review videos from YouTube, characterized by more natural scenes and noise. They provide fine-grained sentiment intensity ratings, ranging from -3 to +3. In this paper, we treat Sentiment as a proxy for the Valence dimension. To evaluate the model's zero-shot transfer ability, we train the model on IEMOCAP and directly assess the effectiveness of its representations on MOSI/MOSEI.

4.2 Semantic Interpretability of the Affective Manifold

A central premise of our proposed framework is that the learned shared latent variable, z_{vad}, constitutes a canonical orthogonal basis corresponding to the psychological primitives of Valence, Arousal, and Dominance. To validate this hypothesis, we investigate the statistical correlations between the specific dimensions of z_{vad} and the ground-truth affective attributes.

Table 1 reports the Pearson correlations on the IEMOCAP test set. The prominent diagonal coefficients ($r = +0.533, +0.630, +0.503$) shows that the latent axes align well with the intended psychological primitives. Specifically, $z_{vad}[0]$ shows high specificity to Valence with minimal leakage. The moderate coupling observed between Arousal ($z_{vad}[1]$) and Dominance ($z_{vad}[2]$) reflects established psychological realities where emotional intensity and perceived control often co-vary in naturalistic speech (Russell and Mehrabian, 1977). Notably, our model matches or even surpasses the Inter-rater Agreement in the A and D dimensions, suggesting that the learned manifold captures stable multimodal affective regularities beyond individual annotation variability.

Table 1: **Semantic Alignment Analysis on IEMOCAP.** We evaluate whether the learned latent dimensions (z_V, z_A, z_D) statistically correlate with ground-truth affective attributes. High diagonal values indicate successful semantic grounding.

Learned Axis	GT Valence	GT Arousal	GT Dominance
Shared $z_{vad}[0]$	**+0.533**	-0.120	-0.190
Shared $z_{vad}[1]$	-0.017	**+0.630**	+0.493
Shared $z_{vad}[2]$	-0.133	+0.523	**+0.503**
Inter-rater Agreement	+0.645	+0.511	+0.383

To provide visual insight into the topological structure of the learned manifold, Figure 2 visualizes the distribution of z_{vad} across different input sources. Notably, sample points from heterogeneous modalities—text, audio, and video—are thoroughly intermixed within the same latent space. This overlap provides qualitative evidence that the cross-modal reconstruction mechanism helps reduce the modality gap and improves manifold align-

ment. The resulting distribution is consistent with a more unified, modality-aligned geometry, suggesting that the framework distills shared affective structure across modalities.

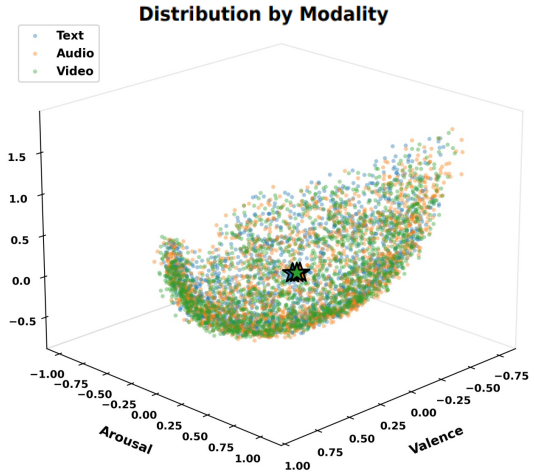

Figure 2: Visualization of the shared affective manifold z_{vad} across different modalities. Sample points from text, audio, and video are thoroughly intermixed within the unified latent space, indicating that the "Modality Gap" has been effectively bridged. Notably, the centroids of each modality, represented by stars, nearly perfectly overlap.

4.3 Disentangled Latent Space Analysis

One of fundamental objective of our framework is to ensure that the Shared Affective Subspace (z_{vad}) is statistically independent of the Private Modality Subspace (z_{priv}). We seek to verify whether the learned representations achieve a functional dichotomy, where z_{vad} captures core affective states while z_{priv} successfully filters out modality-specific nuisances (e.g., speaker identity, background noise). To this end, we evaluate the disentanglement quality through statistical independence metrics and functional ablation tasks.

We first quantify the factorization quality using three complementary metrics: Mean Absolute Correlation ($|r|$), Centered Kernel Alignment (CKA), and Mean Mutual Information (MI). As reported in Table 2, our model achieves near-zero values across all criteria (e.g., MI ≈ 0.005) for text, audio, and video modalities. This statistical evidence is further visually corroborated by the correlation heatmaps in Figure 3. The vertical and horizontal low-correlation bands indicate that the canonical affective dimensions (V, A, D) are largely orthogonal to the private modal residuals. Notably, within the affective subspace, the moderate coupling between Arousal and Dominance reflects psychological reality, yet their relationship with the private space remains minimal.

Modality	**Mean $\lvert r\rvert$ ↓**	**CKA ↓**	**Mean MI ↓**
Text	0.0255	0.0126	0.0058
Audio	0.0297	0.0164	0.0056
Video	0.0235	0.0109	0.0055

Table 2: **Statistical Independence Analysis.** We measure the factorization quality between the shared subspace z_{vad} and private subspaces z_{priv} using Mean Absolute Correlation ($|r|$), Centered Kernel Alignment (CKA), and Mutual Information (MI). Near-zero values across all modalities provide strong evidence of effective factorization between affective and modality-specific information.

To verify if the information is effectively partitioned, we conduct an ablation study by training linear probes for emotion classification on different latent components. As shown in Table 3, restricting the input to z_{vad} only yields high accuracy across all modalities, even outperforming the "Full" ($z_{vad} + z_{priv}$) configuration in Text (+6.9%) and Audio (+2.7%). This phenomenon suggests a denoising effect, where the affective manifold distills core emotional cues from modality-specific residuals. Conversely, using z_{priv} only results in a drastic performance drop (Δ up to -16.4%), suggesting that much of the affective information has been removed from the private subspaces. This double dissociation suggests that our framework learns a more canonical and disentangled representation without requiring explicit instance-level annotations.

Table 3: **Ablation Study on Latent Factorization.** Comparison of emotion recognition performance across modalities when restricting information to specific latent subspaces.

Modality	**Configuration**	**Acc (%)**	**Δ (vs. Full)**
Text	Full ($z_{vad} + z_{priv}$)	50.0	–
	Affective (z_{vad})	**56.9**	+6.9
	Private (z_{priv})	37.2	-12.8↓
Audio	Full ($z_{vad} + z_{priv}$)	54.0	–
	Affective (z_{vad})	**56.7**	+2.7
	Private (z_{priv})	38.3	-15.7↓
Video	Full ($z_{vad} + z_{priv}$)	**55.1**	–
	Affective (z_{vad})	54.0	-1.1
	Private (z_{priv})	38.7	-16.4↓

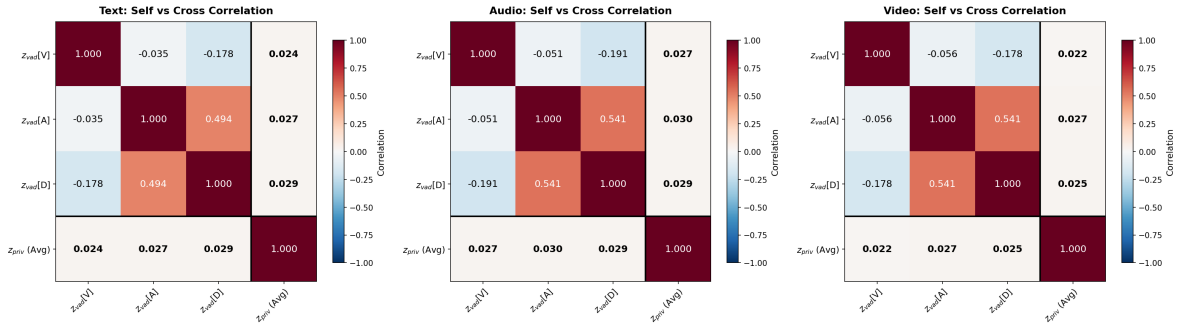

Figure 3: Pearson correlation matrices between the shared affective subspace (z_{vad}) and the private modality subspace (z_{priv}) for text, audio, and video modalities. Each heatmap displays the intra-modal correlations among the three affective dimensions (V, A, D) and their cross-subspace correlations with the average private modality representation.

4.4 Zero-Shot Cross-Domain Transfer

To assess the generalizability of our learned manifold, we conduct zero-shot transfer from IEMOCAP to MOSI and MOSEI. Unlike probing, we perform direct inference using the frozen IEMOCAP encoder and correlate the resulting $z_{vad}[0]$ with ground-truth sentiment. This label-free protocol provides a direct evaluation of how well the learned VAD space transfers across unseen domains.

We select MVAE and MISA as baselines because their architectures are specifically designed to learn structured latent representations, which—unlike purely discriminative models—allow for a direct and rigorous evaluation of whether the learned affective 'coordinates' remain valid under domain shifts. While these paradigms were originally proposed for joint inference or invariant feature extraction, they serve as the most structurally comparable benchmarks for assessing the intrinsic generalizability of our proposed affective manifold.

Table 4: **Zero-shot Domain Transfer (IEMOCAP → MOSI/MOSEI).** We evaluate the generalizability of the learned affective manifold by performing direct inference on unseen datasets. Metrics reported are Pearson Correlation (r) for Valence. All models are frozen after training on IEMOCAP without any fine-tuning on the target domains.

Method	Latent Structure	Target Performance (r) ↑	
		CMU-MOSI	CMU-MOSEI
MVAE	Entangled	0.13	0.15
MISA	Invariant	0.26	0.28
Ours	**Canonical VAD**	**0.40**	**0.43**

Table 4 presents the zero-shot transfer performance from IEMOCAP to the MOSI and MOSEI datasets. Our model outperforms all baselines, achieving a Pearson correlation (r) of 0.40 on MOSI and 0.43 on MOSEI. The low performance of MVAE (0.13–0.15) suggests that entangled representations are highly susceptible to domain noise and speaker-specific variances. While MISA improves results through invariant feature alignment, it lacks the explicit semantic grounding provided by our canonical VAD basis. These results suggest that our model learns a more transferable affective manifold that remains better aligned across datasets, rather than relying as heavily on dataset-specific cues.

4.5 validation of factorization

To demonstrate that the learned affective manifold shows useful generative steerability, we integrate our framework with the Qwen2.5-3B large language model (LLM) for controllable text generation. In this setup, the LLM remains frozen to preserve its pre-trained linguistic knowledge, while a lightweight trainable projector is optimized to map the concatenated latent variables (z_{vad}, z_{priv}) into the LLM's input embedding space. We adopt a controlled perturbation strategy: we first utilize our encoders to extract the private content vector z_{priv} and the original affective state $\mathbf{v}_{orig}$ from seed utterances in the test set. By fixing z_{priv} to lock in the core thematic information and speaker-specific style, we perform linear interpolation along the Shared Valence axis ($z_{vad}[0]$) from -1.0 to $+1.0$. This allows us to observe how the projector effectively steers the frozen LLM to re-interpret the same underlying content across a continuous spectrum of affective intensities.

As illustrated in Table 5, the framework exhibits a generally smooth semantic evolution across diverse themes. In Theme 2, for example, the gen-

Table 5: **Qualitative Demonstration of Controllable Emotion Manipulation via Latent Traversal.** For each theme, we extract the private content vector z_{priv} and the original affective state $\mathbf{v}_{orig}$ using our encoders. We then fix z_{priv} and traverse the shared valence axis $z_{vad}[0]$ from -1.0 to $+1.0$. The model generates sentences that strictly preserve the core semantic topic while exhibiting a smooth tonal progression. Affective keywords are highlighted in **bold**.

Valence	Affective State	Generated Text (Fixed z_{priv} content)
Theme 1: "My whole entire, bloody life time and time again." Original Model VAD: $[V: -0.614, A: 0.121, D: 0.469]$		
-1.0	Extreme Negative	Feeling utterly **hopeless**, my life repeats endlessly in **agony**.
-0.5	Negative	I've been stuck in this endless cycle of **worry and anxiety** for years.
0.0	Neutral	In various instances throughout my life, I have faced this repeatedly.
$+0.5$	Positive	I am filled with endless **optimism** for every moment ahead.
$+1.0$	Extreme Positive	I'm **bursting with excitement** every single time; to experience this joy anew is **thrilling**!
Theme 2: "The business does not inspire me." Original Model VAD: $[V: -0.534, A: -0.280, D: 0.511]$		
-1.0	Extreme Negative	Despite the **bleak** prospects, my spirit **sinks lower** than ever before in this **failing** enterprise.
-0.5	Negative	Despite my concerns, the business still **fails** to excite or motivate me.
0.0	Neutral	Despite my lack of enthusiasm, I do not find the business inspiring.
$+0.5$	Positive	Instead of feeling uninspired, I **look forward** to what the business can achieve.
$+1.0$	Extreme Positive	I am **thrilled and energized** by the business potential!
Theme 3: "Yes. On this I would." Original Model VAD: $[V: 0.580, A: 0.141, D: -0.168]$		
-1.0	Extreme Negative	Feeling utterly **defeated**, I say simply "Yes. On this I would not."
-0.5	Negative	I feel **worry and anxiety** saying yes, on this I would.
0.0	Neutral	Yes. As for this, I agree.
$+0.5$	Positive	With **great positivity**, I agree on this point; to my **confidence**, it holds true.
$+1.0$	Extreme Positive	Yes! On this I am **extremely excited** to agree!

erated output transitions logically from describing "bleak prospects" to a neutral "not inspiring" statement, and finally to being "thrilled and energized" by business potential. Crucially, the core thematic anchors—such as "life repeats," "business," or "agreement"—remain largely stable during affective changes. This thematic constancy provides evidence that z_{priv} retains non-affective semantic information, while the projector maps z_{vad} coordinates into usable affective guidance for the LLM. Furthermore, the observation that the model's original VAD estimates $\mathbf{v}_{orig}$ align closely with human intuition suggests that the learned manifold supports a coherent perception-to-manipulation pipeline, making affective variation more interpretable in the learned latent space.

5 Conclusion

In this work, we addressed the challenge of learning a structurally interpretable affective manifold from heterogeneous multimodal data by proposing a Canonical Disentangled Multimodal Generative Framework. Our comprehensive experimental evaluation provides several empirical findings. First, the distinct "double dissociation" observed in our linear probing tasks suggests that affective states and modality-private attributes function as orthogonal latent factors, which can be explicitly disentangled via generative constraints. Second, the smooth semantic transitions observed in the latent traversal experiments provide support for the effectiveness of our Dual-Anchoring Strategy. These results suggest that abstract affective semantics have been successfully grounded into a controllable and psychologically interpretable coordinate system, distinguishing our approach from unstructured latent representations. Future work will explore extending this manifold to capture complex non-linear dynamics, such as mixed emotions.

6 Limitation

While our framework establishes a useful canonical manifold, several directions remain open for further study. First, anchoring the shared affective space to a fixed three-dimensional VAD coordinate system offers a practical and psychologically grounded

prior, though richer emotional phenomena such as mixed states, culturally dependent expressions, and context-sensitive transitions may benefit from more adaptive or non-linear representations. Second, our affective/private decomposition is best understood as a modeling abstraction that improves analysis and control, rather than as a claim that these factors are fully independent in natural multimodal behavior, where lexical content, speaker identity, prosody, and facial style often remain partially coupled. Third, our generative analyses mainly focus on controllable text generation and latent traversal as diagnostic tools for representation quality; extending this validation to raw audio or video generation, perceptual quality assessment, and broader human evaluation would further strengthen the picture. Finally, our experiments primarily emphasize cross-domain transfer and disentanglement behavior, and future work can broaden this evaluation with additional in-domain analyses and comparisons to newer multimodal systems.

Acknowledgments

This work was supported by the Guangdong Philosophy and Social Science Foundation (No. GD25CJY17). We also thank the anonymous reviewers for their constructive feedback on our submission.

References

Mohammad Asif, Noman Ali, Sudhakar Mishra, Anushka Dandawate, and Uma Shanker Tiwary. 2024. Deep fuzzy framework for emotion recognition using eeg signals and emotion representation in type-2 fuzzy vad space. *arXiv preprint arXiv:2401.07892*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Jiehui Jia, Huan Zhang, and Jinhua Liang. 2025. Bridging discrete and continuous: A multimodal strategy for complex emotion detection. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929.

Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal sentiment analysis: A survey. *Displays*, 80:102563.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. 2020. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2554–2562.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4):261–292.

Saif M Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2359–2369.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. 2014. Learning to disentangle factors of variation with manifold interaction. In *International conference on machine learning*, pages 1431–1439. PMLR.

Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3722–3729.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.

Beiyu Xu, Zhiwei Liu, and Sophia Ananiadou. 2025. Disentangled vad representations via a variational framework for political stance detection. *arXiv preprint arXiv:2502.19276*.

Kailai Yang, Tianlin Zhang, and Sophia Ananiadou. 2023. Disentangled variational autoencoder for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 15(2):508–518.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246.