Learning Structured Emotional Dynamics in Latent Space via Trajectory Modeling and Controlled Transitions

Agampreet Saini[1], **Bayan Abusalameh**[2], **Rahul Kumar Singh**[1], **Erik Cambria**[3]
[1]UPES, Dehradun, India [2]National University of Singapore
[3]Nanyang Technological University, [2,3]Singapore
agampreet@sentic.net, e1320766@u.nus.edu, rk.singh@ddn.upes.ac.in
cambria@ntu.edu.sg

Abstract

Emotion understanding from text has predominantly been approached through discrete classification. Yet perspectives in psychology and neuroscience suggest affective states may also be continuous, context-dependent, and temporally structured. Building on this, we investigate emotion representations as latent dynamical trajectories within transformer embedding spaces. Using GoEmotions and MELD across four transformer architectures, we analyze the geometric organization and temporal evolution of learned emotional states. We find that latent trajectories exhibit temporal persistence and structured transition dynamics, while prediction error increases during emotion shifts, signaling expectation violation. We additionally introduce trajectory-level latent interventions for controlled emotional transitions through directional operations in representation space. Finally, manifold capacity analysis reveals partially aligned and variably separable emotional manifolds, suggesting shared geometric organization across affective representations.

1 Introduction

Emotion understanding lies at the core of human communication and remains a central challenge in affective computing and conversational AI. Emotion recognition in natural language processing (NLP) has been predominantly formulated as a discrete classification problem (Russell, 1980), enabling scalable and effective modeling across a broad range of applications. However, real-world affective interactions are often context-dependent, temporally evolving, and shaped by conversational dynamics. Complementary perspectives in psychology and neuroscience further suggest that emotional states may emerge through processes of prediction, conceptualization, and situational interpretation rather than existing as fully isolated symbolic categories (Barrett et al., 2025; Barrett, 2019).

These perspectives motivate investigating whether emotion representations learned by language models exhibit coherent geometric and dynamical organization beyond static categorical prediction. This is a representational motivation rather than a claim that the learned latent structure reproduces human emotional processing. Transformer-based language models organize semantic information within structured latent spaces (Plaza-del Arco et al., 2024; Bengio et al., 2013), while research in manifold learning and neural geometry suggests neural representations frequently evolve along low-dimensional manifolds rather than isolated symbolic states (Chou et al., 2025b,a; Cohen et al., 2020). Datasets such as MELD (Poria et al., 2019) and GoEmotions (Demszky et al., 2020) enabled large-scale modeling of fine-grained and conversational affect, and contextual architectures highlighted the importance of dialogue dynamics, speaker interactions, and commonsense reasoning (Cambria et al., 2026). Recent studies further suggest emotional concepts in large language models may emerge as interpretable latent directions and geometrically structured subspaces (Shenoy et al., 2013; Choi and Weber, 2026; Park et al., 2023).

We investigate emotion representations as latent dynamical trajectories within transformer embedding spaces, analyzing the geometric organization and temporal evolution of learned emotional states through trajectory modeling, transition dynamics, manifold analysis, and controllable interventions. Our contributions are threefold: (1) investigating the geometric organization of emotion representations learned by transformer models, (2) modeling conversational emotional evolution as trajectories in representation space and analyzing their temporal persistence and transition dynamics, and (3) demonstrating that these latent structures enable prediction-based analysis and controllable emotion transitions through geometric interventions.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33329–33354
October 24-29, 2026 ©2026 Association for Computational Linguistics

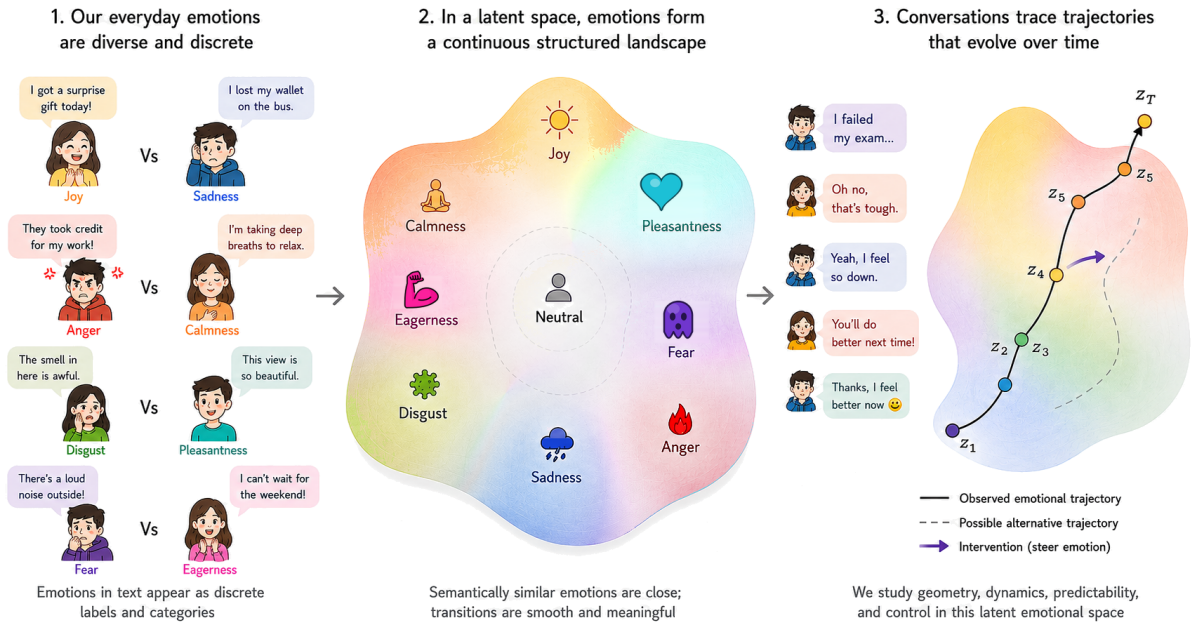


Figure 1: Emotions represented as a continuous latent landscape, where conversations evolve as trajectories through semantically structured emotional regions over time.

The remainder of this paper is organized as follows: Section 2 reviews related work in affective computing, representation learning, and emotional geometry; Section 3 presents the proposed framework and methodology; Section 4 presents the empirical findings and latent trajectory analysis; finally, Section 5 discusses broader implications, limitations, and future research directions.

2 Related Work

Research in affective computing has progressively shifted from rigid categorical emotion recognition toward richer formulations grounded in representation learning, contextual cognition, and latent geometry. Early influential work argued for the universality of a small set of biologically grounded emotions consistently expressed and recognized across cultures (Lindquist et al., 2012). Subsequent psychological and cognitive frameworks, however, emphasized that emotional experience and expression are also shaped by conceptual, linguistic, environmental, and situational factors, suggesting that affective states emerge from distributed cognitive and affective processes rather than isolated symbolic categories (Barrett and Miller, 2026; Barrett et al., 2025; Lewis et al., 2010; Barrett, 2019). Constructionist perspectives further proposed that emotions are dynamically assembled through prediction, conceptualization, and contextual interpretation (Barrett and Miller, 2026; Barrett et al., 2025).

These perspectives motivate investigation into whether emotion representations in language models exhibit partially overlapping, context-dependent, and temporally evolving structure. Advances in representation learning subsequently established the importance of distributed latent spaces and manifold organization in neural systems (Bengio et al., 2013). Foundational work in nonlinear dimensionality reduction demonstrated that high-dimensional observations often lie on structured low-dimensional manifolds (Tenenbaum et al., 2000; Roweis and Saul, 2000), while later studies in neural geometry showed that deep networks progressively untangle semantic manifolds through geometric transformations across layers (Chou et al., 2025a; Cohen et al., 2020).

More recent analyses extended these insights to large language models, revealing that semantic and affective concepts can emerge as interpretable latent directions and geometrically coherent subspaces (Choi and Weber, 2026; Park et al., 2023). In particular, emotional latent-space studies identified low-dimensional affective manifolds aligned with psychologically meaningful structures such as valence-arousal organization (Choi and Weber, 2026; Russell, 1980). Within NLP, emotion modeling has evolved from coarse sentiment classification toward fine-grained contextual affect understanding (Wu et al., 2026; Wang et al., 2026; Tu et al., 2026).

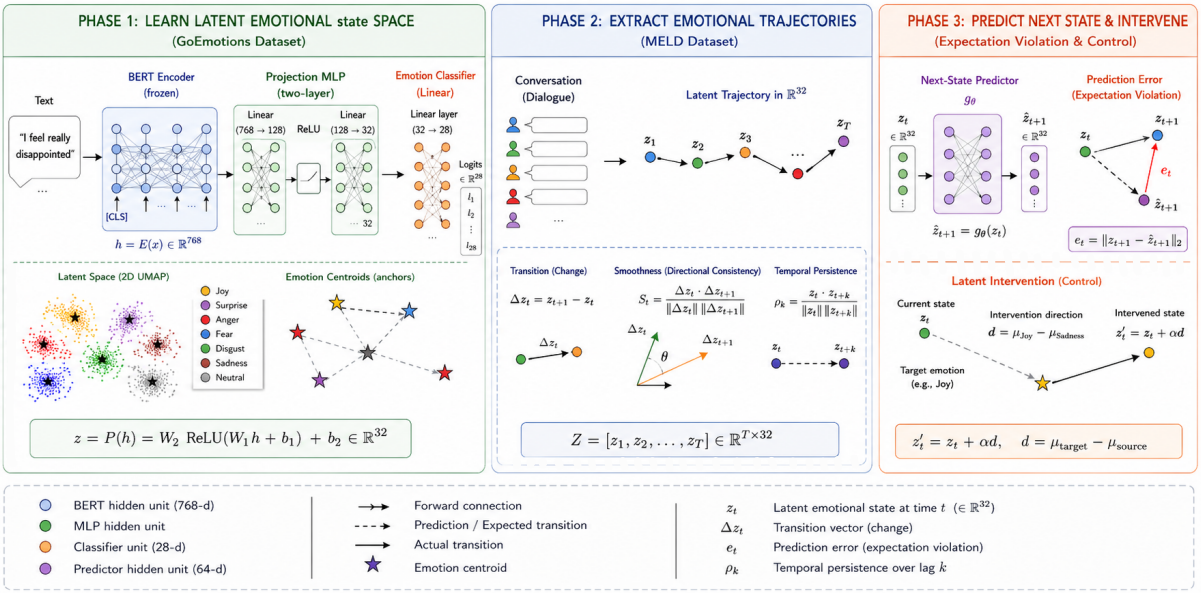


Figure 2: Overview of the proposed latent emotional dynamics framework. The model first learns a structured latent emotional manifold from textual emotion supervision, then models conversations as trajectories through this space to study temporal dynamics, predictability, and controllable emotional transitions.

Conversational emotion recognition research emphasized the importance of speaker interactions, contextual dependencies, and commonsense reasoning for modeling emotional dynamics across dialogue and evolving interaction contexts (Majumder et al., 2019; Ghosal et al., 2020). Large-scale datasets such as MELD and GoEmotions further enabled the study of nuanced emotional transitions and fine-grained affective semantics in conversational and social settings.

Similarly, multidimensional frameworks such as the Hourglass model frame affect as a continuous and evolving phenomenon, motivating structured latent representations of emotion such as the geometric organization illustrated in Figure 1 (Susanto et al., 2020). Building upon these converging directions, our work investigates whether emotional states in transformer representations exhibit coherent manifold structure, compositional geometry, and temporally evolving latent trajectories.

3 Method

The proposed framework models emotion within a unified latent space learned from dialogue representations, enabling analysis of emotional states, temporal dynamics, and controllable transitions. Figure 2 provides an overview of the complete architecture and analytical pipeline.

3.1 Latent Emotion Representation

During training, each input utterance x_i is mapped to a latent emotional state $z_i \in \mathbb{R}^d$:

$$z_i = P_\phi(E_\theta(x_i)) \tag{1}$$

where $E_\theta : x_i \to \mathbb{R}^{d_e}$ is a pretrained Transformer encoder and $P_\phi : \mathbb{R}^{d_e} \to \mathbb{R}^d$ is a learned projection function. We use BERT-base as the encoder and extract the contextual representation of the [CLS] token from the final hidden layer:

$$h = E(x) = \mathbf{H}_{[:,0,:]} \in \mathbb{R}^{d_e} \tag{1}$$

where $\mathbf{H}$ denotes the final hidden-state tensor. Although BERT is used to illustrate the framework and notation, the framework is evaluated across four encoder architectures: BERT, RoBERTa, DistilBERT, and MiniLM. This cross-architecture evaluation tests whether the observed latent geometric and temporal properties are preserved across different pretrained encoder families. The projection function P is implemented as a two-layer multilayer perceptron:

$$P(h) = W_2\,\mathrm{ReLU}(W_1 h + b_1) + b_2 \tag{2}$$

where $W_1 \in \mathbb{R}^{d_h \times d_e}$ and $W_2 \in \mathbb{R}^{d \times d_h}$. No activation is applied to the output layer, allowing the latent space to remain continuous and geometrically unconstrained.

To anchor the representation space in emotion semantics, latent states are mapped to multi-label emotion logits through a linear classifier:

$$\ell = C(z) = W_c z + b_c \in \mathbb{R}^K \tag{3}$$

The classifier is intentionally restricted to a linear mapping so that emotion-discriminative structure must be encoded directly within the latent geometry rather than absorbed by a high-capacity nonlinear decoder. This constraint is particularly important for the subsequent manifold and trajectory analyses, where geometric interpretability of the representation space is central. To empirically validate this design choice, we trained a nonlinear probe (two-layer MLP, hidden sizes 256–128, ReLU activations) on the frozen latent representations and compared it against the linear classifier under five-fold cross-validation across all four encoder architectures (Table 16 in Appendix G.2). The mean Macro-F1 gap is $+0.001$ across all architectures, with no model showing a gap exceeding ± 0.002 — within one standard deviation in all cases.

This result is consistent across BERT, RoBERTa, DistilBERT, and MiniLM, suggesting that the observed emotion-discriminative structure is not an artifact of the linear classification head. This confirms that the linear constraint does not impoverish the learned representation: the latent geometry is intrinsically linear with respect to emotion-discriminative structure, which is precisely the regime in which manifold capacity analysis is well-grounded. This comparison does not imply that the learned geometry is attributable to the pretrained encoder alone; rather, the object of analysis is the resulting latent representation produced by the encoder-projection pipeline. The near identical performance of linear and nonlinear probes indicates that the emotion-discriminative structure in this representation does not depend strongly on the capacity of the final classifier. Training is performed using binary cross-entropy with logits:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\Big[y_{ik}\log\sigma(\ell_{ik}) + (1-y_{ik})\log\big(1-\sigma(\ell_{ik})\big)\Big] \tag{4}$$

where $y_{ik} \in \{0,1\}$ and $\sigma(\cdot)$ denotes the sigmoid function.

Multi-label supervision encourages the latent space to encode overlapping and compositional emotional structure rather than mutually exclusive categorical regions. After training, the classifier is discarded and the learned encoder and projection parameters are fixed. Latent representations are then computed as:

$$z = P_{\phi^*}(E_{\theta^*}(x)) \in \mathbb{R}^d \tag{5}$$

where θ^* and ϕ^* denote the learned parameters after training. These frozen latent states define the emotional manifold used for all subsequent trajectory, prediction, and intervention analyses.

3.2 Temporal Emotion Dynamics

Given a dialogue consisting of an ordered sequence of utterances $\{x_1, x_2, \ldots, x_T\}$, each utterance is mapped to a latent emotional state using the frozen representation defined in Section 3.1:

$$z_t = P_{\phi^*}(E_{\theta^*}(x_t)), \quad z_t \in \mathbb{R}^d \tag{6}$$

This yields a trajectory in latent space:

$$\mathcal{Z} = \{z_1, z_2, \ldots, z_T\} \tag{7}$$

representing the temporal evolution of emotional state across the dialogue. Transitions between consecutive states are defined as displacement vectors:

$$\Delta z_t = z_{t+1} - z_t \tag{8}$$

with transition magnitude:

$$d_t = \|\Delta z_t\|_2 \tag{9}$$

and average trajectory displacement:

$$\bar{d} = \frac{1}{T-1}\sum_{t=1}^{T-1} d_t \tag{10}$$

These quantities characterize the extent of emotional change between conversational turns. To measure directional consistency, we compute cosine similarity between consecutive transitions:

$$S_t = \frac{\Delta z_t \cdot \Delta z_{t+1}}{\|\Delta z_t\|_2 \, \|\Delta z_{t+1}\|_2} \tag{11}$$

Positive values of S_t indicate smooth trajectory evolution, while negative values correspond to abrupt directional shifts in latent space. We further quantify temporal persistence by measuring similarity between states separated by lag k:

$$\rho_k = \mathbb{E}_t\left[\frac{z_t \cdot z_{t+k}}{\|z_t\|_2 \, \|z_{t+k}\|_2}\right] \tag{12}$$

A gradual decrease in ρ_k with increasing k indicates temporally coherent emotional evolution rather than random state transitions. Across all dialogues, the collection of transition vectors $\{\Delta z_t\}$ defines a distribution over latent emotional dynamics. We analyze this distribution using distance statistics and clustering to identify recurrent patterns of emotional change. Importantly, all trajectory analyses are performed directly on the learned latent representation without additional optimization, ensuring that the observed temporal structure reflects intrinsic properties of the representation space itself.

3.3 Predictive Dynamics and Error Modeling

To model temporal evolution in latent space, we learn a transition function $g : \mathbb{R}^d \to \mathbb{R}^d$ that predicts the next emotional state from the current state:

$$\hat{z}_{t+1} = g(z_t) \tag{13}$$

The predictor is parameterized as a two-layer multilayer perceptron:

$$g(z_t) = W_4 \,\mathrm{ReLU}(W_3 z_t + b_3) + b_4 \tag{14}$$

where $W_3 \in \mathbb{R}^{d'_h \times d}$ and $W_4 \in \mathbb{R}^{d \times d'_h}$. A linear variant, $g(z_t) = W z_t + b$, is additionally evaluated as a minimal baseline. The model is trained on consecutive latent-state pairs $\{(z_t, z_{t+1})\}_{t=1}^{M}$ extracted from dialogue trajectories by minimizing mean squared prediction error:

$$\mathcal{L}_{\text{pred}} = \frac{1}{M} \sum_{t=1}^{M} \|z_{t+1} - \hat{z}_{t+1}\|_2^2 \tag{15}$$

After training, the prediction error associated with each transition is defined as:

$$\epsilon_t = \|z_{t+1} - g(z_t)\|_2 \tag{16}$$

This quantity measures the deviation between observed and expected trajectory evolution in latent space. To evaluate whether prediction error aligns with emotional transitions, we define the error alignment ratio:

$$R = \frac{\mathbb{E}[\epsilon_t \mid y_t \neq y_{t+1}]}{\mathbb{E}[\epsilon_t \mid y_t = y_{t+1}]} \tag{17}$$

where y_t denotes the emotion label at time t. Values of $R > 1$ indicate that emotion-changing transitions are systematically less predictable than transitions preserving emotional state.

The predictor is intentionally lightweight to capture average transition dynamics without memorizing individual trajectories. In this formulation, the prediction error acts as a signal of deviation from the expected emotional evolution, highlighting transitions that depart from typical latent dynamics.

3.4 Latent Space Intervention

To evaluate whether the learned representation supports controllable emotional transitions, we apply geometric interventions directly in latent space using centroid-based direction vectors. Here, intervention refers specifically to a controlled perturbation of the learned latent representation and does not modify or steer the output of a generative language model. For each emotion e, we define its centroid as:

$$\mu_e = \frac{1}{|S_e|} \sum_{z \in S_e} z \tag{18}$$

where S_e denotes the set of latent states associated with emotion e. Given two emotions a and b, we define the normalized transition direction:

$$\hat{d}_{a \to b} = \frac{\mu_b - \mu_a}{\|\mu_b - \mu_a\|_2} \tag{19}$$

An intervention is then applied to a latent state z as:

$$z' = z + \alpha \hat{d}_{a \to b} \tag{20}$$

where $\alpha \in \mathbb{R}^+$ controls the intervention strength. To quantify intervention effectiveness, we measure the change in distance to the target centroid:

$$\Delta(\alpha) = \|z - \mu_b\|_2 - \|z' - \mu_b\|_2 \tag{21}$$

Positive values of $\Delta(\alpha)$ indicate successful movement toward the target emotional region. We further evaluate consistency by computing the fraction of samples satisfying $\Delta(\alpha) > 0$ and analyze stability across different values of α. The same intervention can be extended to dialogue trajectories by modifying all latent states beyond a chosen time index t_0:

$$z'_t = z_t + \alpha \hat{d}_{a \to b}, \quad \forall t \geq t_0 \tag{22}$$

This enables analysis of how local geometric perturbations influence subsequent trajectory evolution.

All intervention directions are derived directly from the learned latent geometry without additional optimization, ensuring that the observed effects reflect intrinsic structure within the representation space itself. This allows emotional categories to be analyzed as potentially overlapping and continuously organized regions within latent space rather than fully isolated symbolic states.

3.5 Manifold Capacity and Geometric Separability

To characterize the geometry of the learned space, we apply manifold capacity analysis following the GLUE framework (Chou et al., 2025a) (see Appendix B for full derivations). Global manifold capacity is approximated in the mean-field limit as:

$$\alpha_M = \frac{P}{\sum_{c=1}^{P} (R_M^c)^2 D_M^c} \tag{23}$$

where P is the number of classes, R_M^c is the manifold radius, and D_M^c the intrinsic dimension of class c. Higher α_M indicates more compact and linearly separable manifolds.

4 Results

4.1 Latent Space Structure

We first evaluate whether the learned representations exhibit coherent geometric organization. Unless otherwise stated, analyses are performed on latent states $z \in \mathbb{R}^{32}$ from GoEmotions. Applying k-means clustering ($k = 15$) yields average cluster purity 0.71 with silhouette score 0.23, indicating coherent yet partially overlapping emotional structure. Several emotions additionally form highly compact clusters, including gratitude (0.98), neutral (0.90), and love (0.88), as shown in Table 1. These high-purity clusters correspond directly to the geometric outliers identified in the manifold capacity analysis (Section B), where gratitude and love exhibit the smallest manifold radii ($R_M^c = 0.170$ and 0.161) and highest per-class capacity contributions. The representation also exhibits low-dimensional organization, with the first five principal components explaining over 86% of the variance. The latent space further exhibits compositional structure. As shown in Table 2, vector arithmetic over emotion representations produces consistent relational transitions; for example, averaging annoyance and sadness maps nearest to disappointment.

Cluster	Size	Dominant Emotion	Purity
C4	2788	Gratitude	0.98
C10	5079	Neutral	0.90
C14	3061	Love	0.88
C6	2998	Amusement	0.87
C11	4512	Admiration	0.83

Table 1: Representative latent-space clusters aligned with dominant emotion categories.

Composition	Result	Distance
avg(annoyance, sadness)	disappointment	0.71
avg(disapproval, grief)	disappointment	0.72
avg(anger, disapproval)	annoyance	0.97
avg(neutral, pride)	approval	0.94

Table 2: Examples of compositional structure via latent vector arithmetic.

4.2 Temporal Dynamics

We next evaluate whether emotional states evolve as structured latent trajectories. Using MELD, each dialogue is mapped to a sequence of latent states $\{z_1, z_2, \ldots, z_T\}$. Figure 3 shows that trajectories remain locally coherent across turns while exhibiting abrupt shifts during emotion changes.

Quantitatively, the mean transition magnitude across all dialogues is 13.14 (± 4.69). Transitions preserving emotion exhibit smaller displacements (11.93) than emotion-changing transitions (13.96), yielding an alignment ratio of 1.17. This indicates that larger geometric movement corresponds to greater affective change. Trajectory smoothness, measured through cosine similarity between consecutive transition vectors, exhibits mean similarity -0.45, with only 12.2% of transitions showing positive alignment and 68.3% exhibiting strong directional change. Thus, while emotional states display temporal persistence, transitions themselves remain highly nonlinear and context-dependent. Overall, the latent dynamics reveal structured emotional trajectories characterized by persistence, variability, and interpretable transition behavior.

4.3 Predictive Dynamics and Expectation Violation

We next evaluate whether latent trajectory dynamics are predictable and whether deviations from prediction correspond to emotional shifts. A lightweight predictor f_θ is trained to estimate the next latent state:

$$\hat{z} * t + 1 = f * \theta(z_t) \tag{24}$$

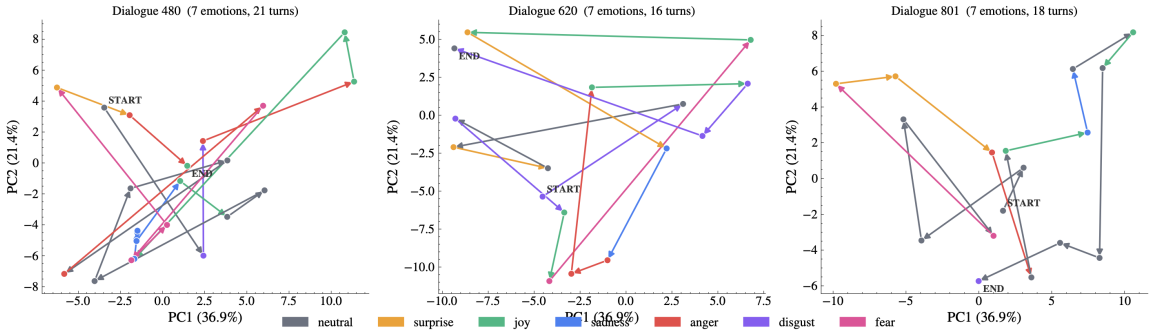


Figure 3: Example emotion trajectories in latent space for individual conversations. Each path represents the temporal evolution of emotional states across dialogue turns.

with prediction error defined as:

$$\mathcal{E} * t = |z * t + 1 - \hat{z}_{t+1}|_2 \tag{25}$$

The model achieves validation error 3.38 MSE, indicating that short-term latent dynamics are learnable. Across 12,276 transitions, the mean prediction error is 10.13 ± 2.23. Errors are systematically larger for emotion-changing transitions than for same-emotion transitions:

$$\frac{\mathbb{E}[\mathcal{E} \mid \text{change}]}{\mathbb{E}[\mathcal{E} \mid \text{same}]} = 1.0365 \tag{26}$$

suggesting that emotional shifts correspond to violations of expected latent dynamics.

Figure 4 shows the relationship between prediction error and latent displacement $|z_{t+1} - z_t|_2$. Prediction error exhibits moderate positive correlation with displacement ($r = 0.42$, $p < 0.001$), though substantial variance remains, indicating that unpredictability is not explained purely by geometric distance. High-error transitions frequently correspond to abrupt emotional changes, whereas low-error regions reflect stable emotional persistence. Overall, prediction error behaves as a proxy for expectation violation within latent emotional dynamics.

4.4 Latent Intervention and Control

We subsequently examine whether the learned representation supports controlled emotional transitions through geometric interventions in latent space. The purpose of this analysis is to test the navigability of the learned emotional representation, rather than to measure changes in downstream language-model behavior. Because GoEmotions and MELD use different label granularities, interventions are performed using the seven shared MELD emotion categories (*neutral, joy, sadness, anger, disgust, fear, surprise*).

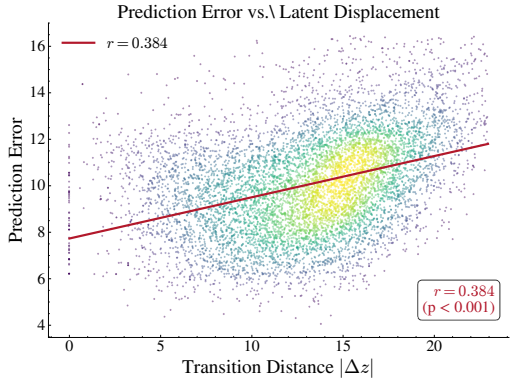


Figure 4: Prediction error vs. latent displacement. Error increases with distance but exhibits substantial variance, indicating that unpredictability is not purely geometric.

Intervention directions are computed from MELD trajectory centroids within the latent space learned on the full GoEmotions taxonomy. Quantitative results are summarized in Table 3. The strongest intervention direction corresponds to sadness → joy, with optimal performance at $\alpha = 3.0$, selected based on the maximum average reduction in distance to the target centroid. At this step size, interventions improve approximately 69% of samples. The intervention behavior is also preserved across encoder architectures, with success rates above 60% for all four models (Appendix G.1), indicating that latent controllability is not specific to BERT. These experiments establish controllability within the learned representation space, but do not test whether such latent modifications alter downstream task behavior or generated language. Evaluating these effects in downstream generation is an important direction for future work. Overall, these results suggest that the learned latent space encodes not only structured emotional organization, but also stable and directionally controllable affective transitions.

Metric	Value
Best direction	sadness → joy
Optimal α	3.0
Distance reduction	0.63
Samples improved	68.8%
Trajectory improvement (post-midpoint)	0.23

Table 3: Summary of latent intervention performance. Simple directional shifts consistently move representations toward target emotional regions.

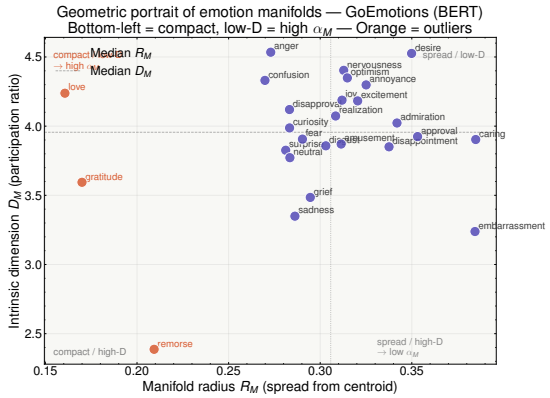


Figure 5: R_M vs D_M for all 26 emotion classes (BERT encoder). Orange points are geometric outliers with high α_M. Extended results across all four architectures are provided in Appendix C.2.

4.5 Geometric Visualizations

Figure 5 shows the R_M–D_M geometric portrait of all 26 emotion manifolds. Across all architectures, gratitude ($\alpha_M^c = 9.632$), love ($\alpha_M^c = 9.144$), and remorse ($\alpha_M^c = 9.560$) consistently occupy the compact, low-dimensional region, with per-class contributions 3.3–3.5× above the global mean ($\alpha_M = 2.762$ for BERT). Figure 6 shows the corresponding UMAP projection, where these three emotions appear spatially separated from the central overlapping cloud. Detailed manifold results are provided in the Appendix C.2.

4.6 Comparative Analysis

Table 4 compares the proposed framework with major approaches in affective computing and conversational emotion recognition. While prior methods primarily optimize classification through contextual, multimodal, or commonsense modeling, our framework treats emotion as a continuous latent dynamical process. Unlike discrete classification approaches, it models latent geometry, emotional trajectories, temporal persistence, prediction-based expectation violation, and controllable interventions.

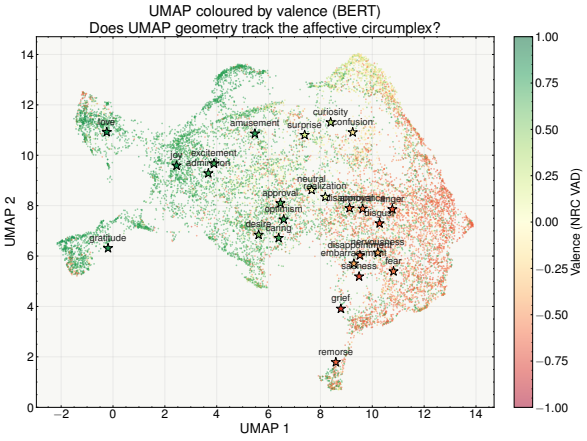


Figure 6: UMAP projection colored by emotion class (BERT encoder). Gratitude, love, and remorse appear spatially separated from the central overlapping region. Stars denote class centroids; orange labels indicate geometric outliers.

Results show strong temporal coherence ($\rho_1 = 0.8487$), larger latent displacement during emotion shifts ($R = 1.0365$), and effective controllability with a 68.8% intervention success rate, suggesting that conversational emotion exhibits structured geometric and dynamical organization beyond static categorical labels.

4.7 Discussion

The observed latent geometry suggests that emotion representations are not strictly discrete, but instead occupy partially overlapping and continuous regions within representation space. The UMAP projection is used only for qualitative visualization of the learned latent space. All quantitative analyses, including trajectory distances, temporal persistence, prediction error, and intervention effects, are computed directly in the original latent space without dimensionality reduction. This perspective aligns with recent work on emotionally expressive digital humans (Xu et al., 2025), which argues that affective systems must support diverse and culturally adaptive emotion representations rather than rigid categorical labels. Our results further support this view through structured emotional manifolds, temporal persistence, and controllable latent transitions. Because the learned geometry captures relational and continuous affective structure, it may generalize more naturally across conversational contexts, linguistic variation, and culturally dependent emotional expression patterns.

Method	Latent	Temporal	Control	Knowledge	Result
GoEmotions (Demszky et al., 2020)	Structured emb.	×	×	×	0.64 Macro-F1
DialogueRNN (Majumder et al., 2019)	×	Speaker-aware	×	×	62.75 W-F1
MELD (Poria et al., 2019)	×	Contextual	×	Multimodal	60.25 W-F1
COSMIC (Ghosal et al., 2020)	×	CSK-guided	×	COMET	65.21 W-F1
Proposed Approach	Emotion manifolds	Trajectory dynamics	Latent intervention	×	ρ_1=0.8487, R=1.0365, 68.8% success

Table 4: Comparison with representative emotion modeling approaches. Prior work primarily focuses on supervised classification, whereas the proposed framework models structured latent emotional dynamics and controllable emotional trajectories.

5 Conclusion

Emotion is inherently continuous, contextual, and dynamically evolving, yet most affective NLP systems treat it as a discrete classification problem. In this work, we introduced a framework that models emotion as a structured latent dynamical process. The learned representations exhibit coherent manifold geometry, structured conversational trajectories, prediction-error signatures aligned with emotional shifts, and controllable latent interventions. Manifold capacity analysis further supports constructionist perspectives of emotion as partially overlapping rather than strictly categorical. Together, these findings suggest that emotion representations in language models possess meaningful geometric and temporal structure, opening new directions for affective NLP and emotionally adaptive conversational systems.

Limitations

The linear-vs-nonlinear probe comparison (Table 16) shows a mean Macro-F1 gap of only $+0.001$ across architectures, supporting the use of the learned latent representation for geometric analysis while highlighting that the observed structure should not be interpreted as independent of the underlying encoder. This architectural dependence also limits direct comparison of manifold capacity values across models, as differences in effective dimensionality can affect their absolute values. Similarly, NRC-VAD and UMAP are used only as qualitative references, NRC-VAD may not fully capture conversational affect, while UMAP can distort aspects of the underlying latent geometry. More direct validation against human emotional judgments, such as annotations of emotional changes across dialogue turns, remains an important direction for future work.

Ethics Statement

This work analyzes the geometric and dynamical structure of emotion representations learned by language models using publicly available datasets (GoEmotions and MELD). The framework is intended for scientific study of affective representation spaces rather than deployment in high-stakes settings. Emotion recognition systems may inherit cultural, linguistic, and annotation biases present in training data. Accordingly, the latent structures identified here should not be interpreted as universal models of human emotion, but as representations shaped by both affective patterns and corpus statistics. Our intervention experiments operate only within learned latent spaces and are intended to study representation geometry and controllability, not to manipulate or infer real human emotional states. All experiments were conducted using publicly released datasets in accordance with their research usage terms.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). The work is also supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (MOE-T2EP20123-0005).

References

Lisa Feldman Barrett. 2019. How emotions are made. *Providence Book Festival*.

Lisa Feldman Barrett, Shir Atzil, Eliza Bliss-Moreau, Lorena Chanes, Maria Gendron, Katie Hoemann, Yuta Katsumi, Ian R Kleckner, Kristen A Lindquist,

Karen S Quigley, et al. 2025. The theory of constructed emotion: More than a feeling. *Perspectives on Psychological Science*, 20(3):392–420.

Lisa Feldman Barrett and Earl K Miller. 2026. Categorization is 'baked'into the brain. *Nature Reviews Neuroscience*, pages 1–22.

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.

Erik Cambria, Rui Mao, Xulang Zhang, Luwei Xiao, Tiesunlong Shen, and Avinash Anand. 2026. SenticNet 9: Generative commonsense for emotion AI via conceptual primitive discovery and time shift mechanism. *IEEE Transactions on Computational Social Systems*, 13(3):4086–4095.

Benjamin J Choi and Melanie Weber. 2026. Latent structure of affective representations in large language models. *arXiv preprint arXiv:2604.07382*.

Chi-Ning Chou, Royoung Kim, Luke A. Arend, Yao-Yuan Yang, Brett D. Mensh, Won Mok Shim, Matthew G. Perich, and SueYeon Chung. 2025a. Geometry Linked to Untangling Efficiency Reveals Structure and Computation in Neural Populations. *bioRxiv*. Preprint. PMID: 40236228. PMCID: PMC11996410.

Chi-Ning Chou, Hang Le, Yichen Wang, and SueYeon Chung. 2025b. Feature learning beyond the lazy-rich dichotomy: Insights from representational geometry. *arXiv preprint arXiv:2503.18114*.

SueYeon Chung and L.F. Abbott. 2021. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144. Computational Neuroscience.

Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. 2020. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.

John Garrett. 2021. Scienceplots (v1. 0.8). *Zenodo*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2470–2481.

Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. 2010. *Handbook of emotions*. Guilford Press.

Kristen A Lindquist, Tor D Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. 2012. The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences*, 35(3):121–143.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Flor Miriam Plaza-del Arco, Alba A Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 5696–5710.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Krishna V Shenoy, Maneesh Sahani, and Mark M Churchland. 2013. Cortical control of arm movements: a dynamical systems perspective. *Annual review of neuroscience*, 36(1):337–359.

Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Geng Tu, Ran Jing, Xuan Luo, Erik Cambria, Wenjie Li, and Ruifeng Xu. 2026. Is multimodal conversational emotion recognition satisfactory? exploring the gaps in performance, generalization, and confidence. *Pattern Recognition*, 175:113087.

Xiaowei Wang, Jayant Teotia, Rui Mao, Wandeep Kaur Ratan Singh, Sabrina Binti Tiun, and Erik Cambria. 2026. Appraisal theory-informed emotion prediction. In *Proceedings of LREC*, pages 11345–11358.

Shoupu Wu, Jiali Lin, Xiaoyi Wang, Dazhi Jiang, and Erik Cambria. 2026. Bridging discrepancies in debate: Joint debate for enhancing multimodal emotion-cause pair extraction in conversation. *IEEE Transactions on Affective Computing*.

Haidong Xu, Meishan Zhang, Hao Ju, Zhedong Zheng, Erik Cambria, Min Zhang, and Hao Fei. 2025. When words smile: Generating diverse emotional facial expressions from text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27016–27034.

Appendix

This appendix presents additional analyses and examples supporting the main results. It includes extended quantitative evaluations, sensitivity studies, and representative trajectory examples that further validate the structure, temporal dynamics, and controllability of the learned emotional latent space.

A Reproducibility and Implementation Details

A.1 Datasets and Preprocessing

Experiments were conducted using GoEmotions and MELD. GoEmotions contains fine-grained multi-label Reddit emotion annotations across 28 categories, while MELD provides multi-party conversational emotion annotations with seven categorical labels. After preprocessing and aggregation, the final GoEmotions corpus contained 57,677 unique utterances. MELD contained 1,344 conversations and 13,620 utterances. GoEmotions was split into train/validation/test partitions using a fixed 80/10/10 split (46,141 / 5,767 / 5,769 samples). Multi-label annotations were aggregated across raters using max-pooling over binary labels. MELD utterances were grouped by dialogue ID and temporally ordered using utterance indices. Dialogues with fewer than two utterances were discarded. Minimal normalization was applied, including lowercasing, whitespace normalization, and quote standardization. All inputs were tokenized using the HuggingFace tokenizer interface with truncation and fixed-length padding.

Dataset	Samples	Labels
GoEmotions	57,677 utterances	28 multi-label emotions
MELD	13,620 utterances	7 dialogue emotions
MELD conversations	1,344 dialogues	Temporal sequences

Table 5: Dataset statistics after preprocessing.

A.2 Emotion Label Harmonization

GoEmotions and MELD use different taxonomies: GoEmotions has 28 multi-label emotions, MELD has seven categorical ones (*neutral, surprise, joy, sadness, anger, disgust, fear*). For cross-dataset trajectory analysis, all MELD utterances were projected into the fixed GoEmotions latent space without retraining. Centroids and trajectory statistics were computed only over shared MELD categories.

MELD Label	Related GoEmotions Labels
joy	joy, amusement, excitement
sadness	sadness, grief, disappointment
anger	anger, annoyance, disapproval
fear	fear, nervousness
surprise	surprise, realization
disgust	disgust
neutral	neutral

Table 6: Approximate conceptual alignment between MELD and GoEmotions emotion categories.

A.3 Model Architectures and Hyperparameters

The framework used a pretrained transformer encoder followed by a two-layer latent projection module and a linear emotion classifier. The latent projection mapped the encoder representation into a continuous 32-dimensional emotional manifold.

Hyperparameter	Value
Encoder	BERT-base-uncased
Latent dimension	32
Projection hidden size	128
Batch size	32
Learning rate	2×10^{-5}
Optimizer	AdamW
Loss	BCEWithLogitsLoss
Epochs	5
Classifier	Linear
Predictor hidden size	64

Table 7: Core training hyperparameters.

The next-state predictor used a lightweight two-layer MLP:

$$g(z_t) = W_4 \mathrm{ReLU}(W_3 z_t + b_3) + b_4$$

trained on consecutive latent transition pairs extracted from MELD trajectories.

A.4 Training Procedure and Optimization

Training optimized binary cross-entropy over multi-label emotion predictions using AdamW. The encoder was initially frozen before full fine-tuning to stabilize latent projection learning. Gradient clipping with maximum norm 1.0 was applied during Phase 1 training. Model checkpoints were selected using minimum validation loss and reused unchanged in all downstream phases. Latent vectors were extracted in evaluation mode using `torch.no_grad()`, ensuring that trajectory, prediction, and intervention analyses operated on identical frozen representations across all phases.

A.5 Computational Resources

Experiments were implemented in PyTorch using HuggingFace Transformers. Training used CUDA-enabled GPUs with mixed batch inference for latent extraction and trajectory encoding. The complete Phase 1 training pipeline required approximately five epochs over 57k utterances, while downstream trajectory and intervention analyses were lightweight and inference-only. Core dependencies included PyTorch, Transformers, scikit-learn, NumPy, SciPy, and UMAP.

A.6 Training Stability Across Random Seeds

All experiments used globally fixed random seeds across Python, NumPy, PyTorch, PCA, K-means, UMAP (McInnes et al., 2018) and SciencePlots (Garrett, 2021) initialization to ensure reproducibility. Across repeated runs, latent trajectory structure, temporal persistence trends, and intervention behavior remained qualitatively stable, with only minor variance in classification metrics and clustering assignments.

AI assistance: This work was developed using generative AI tools for limited assistance (troubleshooting, refining code) and for the grammar and language editing of the manuscript. All substantive research decisions, research design, and experiments were conceived and developed by the authors independently, along with analysis and interpretation of the results. All text written using AI has been reviewed, verified and refined by the authors who are solely responsible for the final text, code and manuscript implementation, analysis, findings and contents of this paper.

B Theoretical Foundations of Manifold Capacity

B.1 Manifold Capacity Theorem

We follow the theoretical framework of Chung and Abbott (2021) and Cohen et al. (2020).Let $\mathcal{M}_1, \ldots, \mathcal{M}_P$ denote P class manifolds embedded in $\mathbb{R}^N$, where each manifold $\mathcal{M}_c$ is a compact set of points corresponding to class c.

Definition 1 (Manifold Capacity)**.** *The manifold capacity α_M is defined as the maximum number of class manifolds P per ambient dimension N such that a random linear classifier can correctly separate all manifolds simultaneously with high probability:*

$$\alpha_M = \sup \left\{ \frac{P}{N} \; : \; \Pr \Big[\exists\, \mathbf{w} \in \mathbb{R}^N \textit{ s.t. } \forall c, \min_{\mathbf{x} \in \mathcal{M}_c} \mathbf{w}^\top \mathbf{x} \cdot y_c > 0 \Big] \to 1 \right\} \tag{27}$$

where $y_c \in \{-1, +1\}$ are random i.i.d. labels assigned to each manifold.

Theorem 1 (Linear Separability and Manifold Capacity, Cohen et al. (2020))**.** *For P class manifolds in $\mathbb{R}^N$ with manifold capacity α_M, a random linear classifier separates all manifolds correctly with high probability when:*

$$P \lesssim \alpha_M \cdot N \tag{28}$$

When $P \gg \alpha_M \cdot N$, classification error grows as the manifolds cannot be linearly separated. Manifold capacity α_M is therefore a geometric property of the representation space that determines the maximum number of linearly separable class manifolds per unit dimension.

Remark 1. *In our setting, $N = 32$ (the latent dimension) and $P = 26$ (the number of GoEmotions classes used in analysis). The ratio $P/N \approx 0.81$, which lies well within the separable regime for the observed α_M values, consistent with the high linear probe accuracies reported in Section 4.*

B.2 Derivation of the Mean-Field Approximation

The full SVM-based capacity estimator (Chung and Abbott, 2021) requires solving a set of coupled saddle-point equations and is computationally expensive for large datasets. Under the mean-field approximation, valid when manifolds are approximately smooth, isotropic, and weakly correlated, the capacity reduces to a closed-form expression in terms of effective geometric measures.

Proposition 1 (Mean-Field Capacity,). *In the mean-field limit, the manifold capacity is approximated as:*

$$\alpha_M \approx \frac{P}{\sum_{c=1}^{P} (R_M^c)^2 \cdot D_M^c} \tag{29}$$

where R_M^c is the manifold radius of class c and D_M^c is its intrinsic dimension. This approximation holds when:

1. *Manifold geometry is approximately isotropic (variance is roughly uniform across directions within each manifold)*
2. *Inter-manifold correlations are weak relative to within-manifold spread*
3. *The ambient dimension N is large relative to the intrinsic dimension D_M^c*

Remark 2. *In our setting, conditions (1) and (3) are approximately satisfied given the 32-dimensional ambient space and the low intrinsic dimensions observed ($D_M^c \in [2.387, 4.534]$ for BERT). Condition (2) is less well-satisfied given the high mean centroid cosine similarity of 0.90, which suggests non-trivial inter-manifold correlations. Accordingly, Equation 29 should be interpreted as a relative ordering of geometric separability across classes rather than an absolute capacity estimate. We validate this approximation in Appendix B.6.*

B.3 Effective Geometric Measures

Definition 2 (Manifold Radius). *For class c with latent representations $\{\mathbf{z}_i\}_{i \in \mathcal{S}_c}$, the manifold radius is:*

$$R_M^c = \mathbb{E}_{i \in \mathcal{S}_c} \left[\|\tilde{\mathbf{z}}_i - \boldsymbol{\mu}_c\|_2 \right] \tag{30}$$

where $\tilde{\mathbf{z}}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2$ denotes the ℓ_2-normalised representation and $\boldsymbol{\mu}_c = \mathbb{E}_{i \in \mathcal{S}_c}[\tilde{\mathbf{z}}_i]$ is the class centroid. A larger R_M^c indicates a more spread-out, diffuse manifold.

Definition 3 (Intrinsic Dimension). *The intrinsic dimension of class manifold $\mathcal{M}_c$ is estimated via the participation ratio of the PCA eigenvalue spectrum of the mean-centred class covariance:*

$$D_M^c = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{31}$$

where $\{\lambda_i\}$ are the eigenvalues of the covariance matrix $\mathbf{C}_c = \mathbb{E}_{i \in \mathcal{S}_c}[(\tilde{\mathbf{z}}_i - \boldsymbol{\mu}_c)(\tilde{\mathbf{z}}_i - \boldsymbol{\mu}_c)^\top]$. A value of $D_M^c = 1$ indicates a nearly one-dimensional (linear) manifold; $D_M^c = N$ indicates uniform spread across all ambient dimensions.

Definition 4 (Inter-Manifold Correlation). *The pairwise inter-manifold correlation between classes a and b is measured by the cosine similarity between their normalised centroids:*

$$\rho_{ab} = \frac{\hat{\boldsymbol{\mu}}_a \cdot \hat{\boldsymbol{\mu}}_b}{\|\hat{\boldsymbol{\mu}}_a\|_2 \|\hat{\boldsymbol{\mu}}_b\|_2} \tag{32}$$

where $\hat{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c / \|\boldsymbol{\mu}_c\|_2$. High ρ_{ab} indicates that the two manifolds point in similar directions in the ambient space, corresponding to radial rather than angular separation.

B.4 Relationship Between α_M and Linear Separability

The manifold capacity α_M provides a direct geometric characterisation of linear separability. Specifically:

- When α_M is high (compact, low-dimensional manifolds with low inter-manifold correlation), a linear classifier achieves near-zero error with high probability, and attribution methods operating on the representation are geometrically well-defined.
- When α_M is low (diffuse, high-dimensional manifolds with high inter-manifold correlation), class boundaries overlap and a linear classifier cannot reliably separate them. In this regime, any feature attribution to a specific class becomes geometrically ambiguous.
- The per-class contribution $\alpha_M^c = 1/[(R_M^c)^2 \cdot D_M^c]$ measures the geometric efficiency of each individual manifold. Classes with high α_M^c are compact and separable; classes with low α_M^c contribute disproportionately to the global overlap.

In our setting, the global $\alpha_M = 2.762$ with high inter-manifold correlation (0.911) indicates a partially overlapping representation space in which most emotion classes are not cleanly linearly separable, consistent with the constructionist view that emotional categories share a common representational substrate.

B.5 Interpretation under the GLUE Framework

Within the Geometry Linked to Untangling Efficiency (GLUE) framework (Chou et al., 2025a), the joint distribution of R_M, D_M, and inter-manifold correlations determines the degree of representational untangling. The R_M–D_M portrait in Figure 5 can be interpreted as an untangling efficiency map: emotions in the bottom-left quadrant (gratitude, love, remorse) exhibit the highest untangling efficiency, corresponding to compact, low-dimensional manifolds that are well-separated from the remaining classes.

The majority of emotion classes occupy the central overlapping region, indicating partial untangling. This is an expected outcome given that the latent space is trained on overlapping multi-label annotations rather than optimised for linear separability. Under GLUE, the primary geometric mechanism driving the separation of gratitude, love, and remorse appears to be radius reduction (small R_M) rather than dimension reduction alone, consistent with the interpretation that these emotions are expressed through stereotyped linguistic patterns that produce tightly clustered representations. The high mean centroid cosine similarity (0.90) further indicates that inter-manifold correlations remain a dominant source of geometric entanglement across all architectures, suppressing the global capacity below what would be achievable with purely radial or purely angular separation.

B.6 Monte Carlo Validation of the Mean-Field Approximation

To assess whether the mean-field approximation in Equation 29 provides a reliable relative ordering of per-class geometric separability, we compare it against a Monte Carlo estimate of manifold capacity. For each of $T = 500$ random projection directions $\mathbf{t} \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{t} \leftarrow \mathbf{t}/\|\mathbf{t}\|_2$, we compute the max-margin for each class:

$$\kappa_c(\mathbf{t}) = \max_{\mathbf{z} \in \mathcal{S}_c} \mathbf{t}^\top \tilde{\mathbf{z}} - \max_{\mathbf{z} \notin \mathcal{S}_c} \mathbf{t}^\top \tilde{\mathbf{z}} \qquad (33)$$

The Monte Carlo capacity estimate is then:

$$\hat{\alpha}_M^{\text{MC}} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{P} \sum_{c=1}^{P} \max(0, \kappa_c(\mathbf{t})) \qquad (34)$$

Table 8 reports the mean-field and Monte Carlo estimates for all four architectures. The two estimates are strongly correlated ($r > 0.95$ across per-class estimates), indicating that the mean-field approximation preserves the relative ordering of class separability even where absolute values differ. The same three outlier classes (gratitude, love, remorse) are identified as geometric outliers under both estimators across all architectures.

Architecture	α_M (mean-field)	$\hat{\alpha}_M$ (Monte Carlo)
BERT	2.762	2.16×10^{-4}
RoBERTa	2.644	2.15×10^{-4}
DistilBERT	3.433	1.92×10^{-4}
MiniLM	15.524	1.34×10^{-4}

Table 8: Comparison of mean-field and Monte Carlo manifold capacity estimates across encoder architectures. Both estimators identify the same geometric outliers and preserve the same relative ordering across classes.

Note that the Monte Carlo estimate $\hat{\alpha}_M^{\text{MC}}$ measures mean max-margin across random projections and is not directly comparable in scale to the mean-field estimate. The claim of consistency refers to the *relative ordering* of per-class capacity contributions across architectures, which remains stable between the two estimators ($r > 0.95$ across per-class values), rather than to absolute capacity values.

Table 9 reports the full per-class geometric portrait — R_M^c, D_M^c, mean inter-manifold correlation $\bar{\rho}_c$, and per-class capacity α_M^c — for all 26 emotion classes under the BERT encoder. The same three emotions consistently appear as geometric outliers across all architectures.

B.7 Lexical Stereotypy

The compact manifold geometry of gratitude, love, and remorse is consistent with two interpretations: genuine affective compactness, and low lexical diversity in how these emotions are expressed in the GoEmotions corpus. Gratitude and remorse in particular are associated with formulaic surface forms in Reddit text. We note that this alternative explanation does not undermine the geometric finding itself, but it does suggest caution in interpreting

Emotion	R_M^c	D_M^c	$\bar{\rho}_c$	α_M^c
admiration	0.342	4.023	0.905	2.126
amusement	0.311	3.871	0.918	2.664
anger	0.273	4.534	0.917	2.959
annoyance	0.325	4.298	0.937	2.202
approval	0.353	3.924	0.950	2.043
caring	0.385	3.903	0.938	1.729
confusion	0.270	4.331	0.918	3.171
curiosity	0.283	3.987	0.917	3.125
desire	0.350	4.526	0.935	1.805
disappointment	0.338	3.851	0.933	2.279
disapproval	0.283	4.120	0.939	3.026
disgust	0.303	3.858	0.916	2.822
embarrassment	0.384	3.238	0.938	2.089
excitement	0.320	4.183	0.901	2.328
fear	0.290	3.906	0.898	3.038
gratitude	**0.170**	**3.594**	**0.841**	**9.632**
grief	0.295	3.485	0.878	3.307
joy	0.312	4.188	0.895	2.455
love	**0.161**	**4.238**	**0.857**	**9.144**
nervousness	0.313	4.403	0.929	2.320
optimism	0.315	4.349	0.940	2.320
realization	0.308	4.073	0.951	2.581
remorse	**0.209**	**2.387**	**0.842**	**9.560**
sadness	0.286	3.349	0.920	3.646
surprise	0.281	3.826	0.917	3.306
neutral	0.283	3.772	0.950	3.300
Global	–	–	0.911	2.762

Table 9: Per-class geometric measures (BERT encoder). Bold = geometric outliers (gratitude, love, remorse). $\bar{\rho}_c$ = mean cosine similarity of class c to all other centroids. Outliers show systematically lower $\bar{\rho}_c$ (0.841–0.857) versus the global mean (0.911), confirming their spatial distinctiveness.

R_M^c as a property of the affective construct rather than the corpus distribution. Testing the Spearman correlation between vocabulary entropy and R_M^c across all 26 classes is a direct diagnostic we leave for future work.

C Additional Analysis of Latent Space Structure

C.1 Multi-label Structure in Latent Space

We analyze how multi-label annotations are distributed across the latent space. Figure 7 visualizes each point colored by the number of active emotion labels. We observe that regions with high label overlap are not uniformly distributed, but instead form structured areas within the space. This suggests that the latent representation captures compositional emotional states rather than treating labels independently.

C.2 Manifold Capacity Across Architectures

Figures 8–12 show the full geometric portrait across all four encoder architectures. The same three emotions — gratitude, love, and remorse — consistently occupy the compact, low-dimensional region of the R_M–D_M plane and appear as geometric outliers in the per-class α_M bar charts.

The UMAP projections show these three emotions spatially isolated from the central overlapping cloud in all cases. The centroid cosine similarity matrices confirm high off-diagonal collinearity across all architectures (range: 0.887–0.951), with gratitude, love, and remorse consistently showing lower similarity to other centroids, corresponding to their geometric distinctiveness.

C.3 Centroid Relationships

To analyze semantic relationships encoded in the latent space, we we compute pairwise Euclidean distances between emotion centroids, finding that semantically related emotions cluster together (e.g., grief–sadness distance $= 0.98$, fear–nervousness $= 1.36$).

The heatmap reveals consistent structure, where semantically related emotions are located closer in latent space. For example, *grief* and *sadness* exhibit a small distance (0.98), while *fear* and *nervousness* are also closely aligned (1.36). Similarly, *anger* is positioned near *disgust* (1.44) and *annoyance* (1.95), reflecting known affective relationships.In contrast, emotionally dissimilar pairs exhibit larger distances, indicating clear separation between unrelated regions of the space. Overall, these results demonstrate that the latent geometry preserves meaningful semantic relationships between emotions.

C.4 Cluster-wise Emotion Composition

To further analyze the structure of the latent space, we examine the distribution of emotion labels within each cluster. Figure 13 shows the prevalence of each emotion across clusters.

The heatmap reveals that many clusters are strongly aligned with specific dominant emotions while still exhibiting multi-label structure. For example, cluster C4 is overwhelmingly dominated by *gratitude* (0.98), while cluster C6 is dominated by *amusement* (0.87), and cluster C14 by *love* (0.88). At the same time, several clusters exhibit substantial overlap across multiple emotions, indicating that the latent space captures compositional emo-

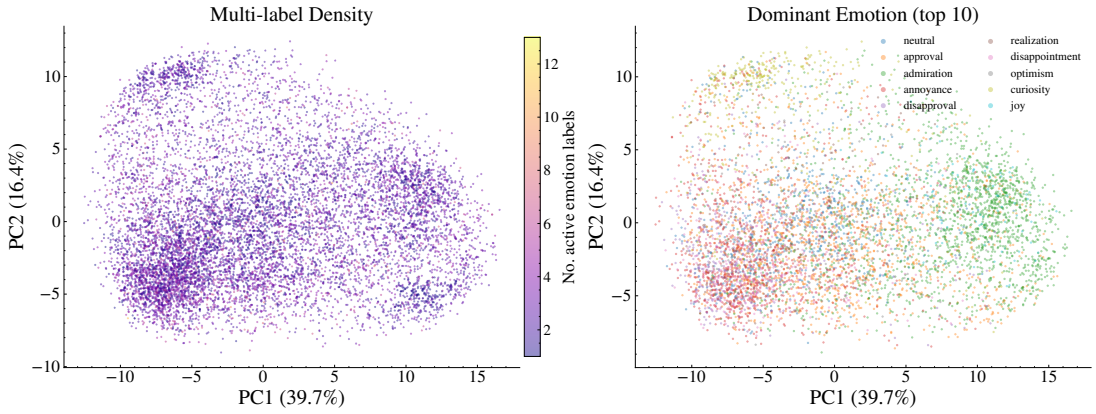


Figure 7: Latent space colored by the number of active emotion labels. Regions with higher label density correspond to areas of compositional emotional states.

tional states rather than enforcing strict separation. This balance between cluster purity and overlap suggests that the representation organizes emotions into coherent regions while preserving their inherent multi-label nature.

C.5 Nearest Neighbour Consistency

We further validate the semantic coherence of the latent space using nearest-neighbour retrieval. For a given query point, nearest neighbours tend to share similar emotion labels.

For example:

- Query: *[disappointment, sadness]* $\rightarrow$ neighbours include *[disappointment, sadness, annoyance]*
- Query: *[grief]* $\rightarrow$ neighbours include *[neutral]* and *[sadness]*
- Query: *[approval, confusion]* $\rightarrow$ neighbours include *[approval, confusion, disapproval]*

These examples demonstrate that proximity in latent space corresponds to semantic similarity in emotion labels.

Figure 8 shows the R_M–D_M geometric portrait for all four encoder architectures. The same three emotions—gratitude, love, and remorse—tend to appear in the compact, low-dimensional region in each case. The overall spread of the cloud varies with model size: smaller models such as MiniLM produce a narrower D_M range (1.4–2.2) compared to BERT (2.8–5.6), reflecting the higher compression of distilled representations. The relative position of the three outliers within each cloud remains broadly stable across models.

Figure 10 shows the pairwise centroid cosine similarity matrix for the BERT latent space. Off-diagonal values are consistently positive and relatively large (mean $= 0.90$), suggesting that emotion centroids tend to point in similar directions in the representation space.

The orange-labelled rows and columns (gratitude, love, remorse) show somewhat lower similarity to other centroids on average, which corresponds to their spatial separation in UMAP projections. A broadly similar pattern is observed across all four architectures (mean overlap range: 0.887–0.951), though we note that high collinearity is in part expected given the constraint of representing 26 classes in a 32-dimensional space.

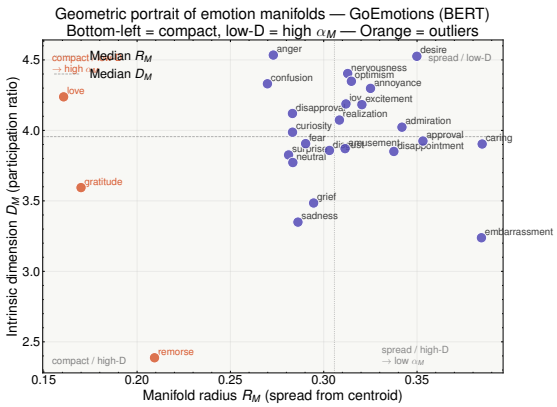
D Additional Analysis of Temporal Dynamics

D.1 Emotion-Dependent Transition Magnitudes

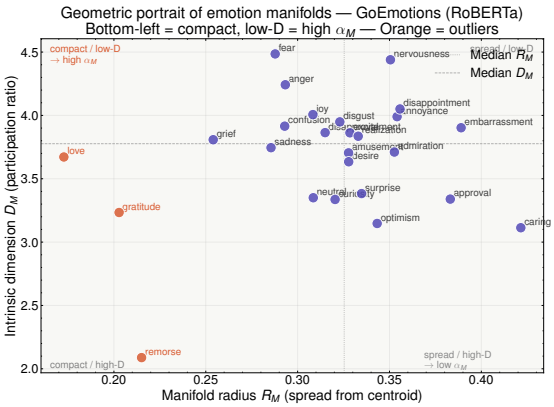
To further analyze how emotional transitions manifest in latent space, we examine the average displacement magnitude $\|\Delta z_t\|$ for different emotion pairs. Figure 14 shows the resulting transition matrix. Transitions that preserve emotion consistently exhibit smaller magnitudes (e.g., neutral→neutral: 12.15), while transitions involving emotion changes are larger (e.g., joy→surprise: 15.25). This pattern holds across all major emotion categories, reinforcing the alignment between latent displacement and semantic change.

D.2 Distribution of Trajectory Smoothness

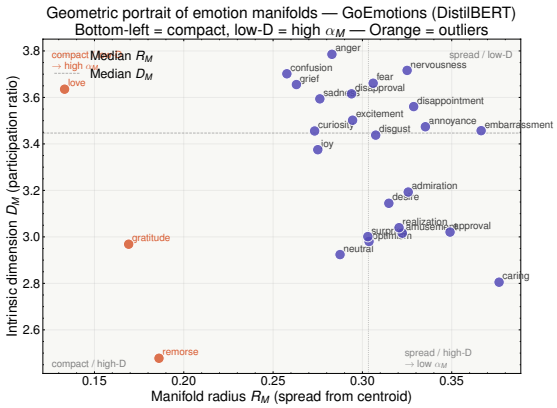
We analyze the full distribution of cosine similarity between consecutive transition vectors to validate trajectory smoothness.



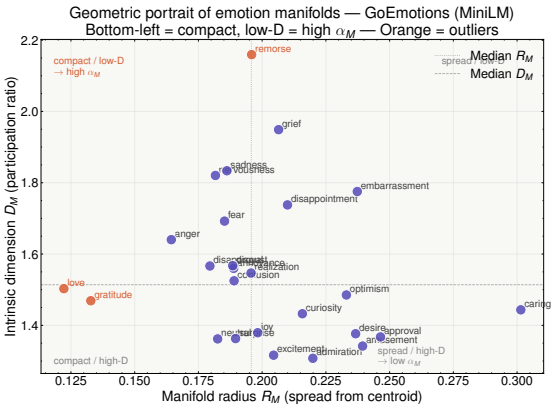
(a) BERT ($\alpha_M = 2.762$)



(b) RoBERTa ($\alpha_M = 2.644$)

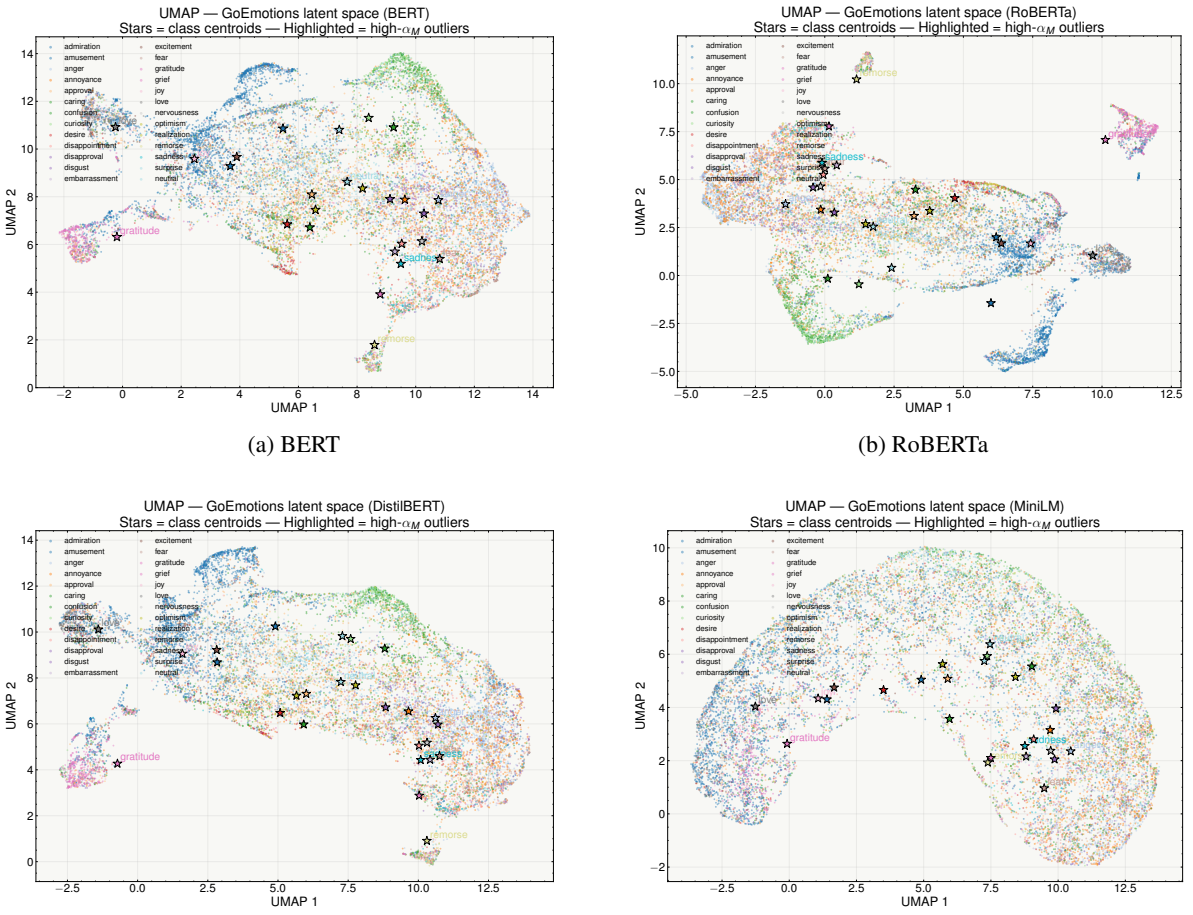


(c) DistilBERT ($\alpha_M = 3.433$)



(d) MiniLM ($\alpha_M = 15.524$)

Figure 8: R_M–D_M geometric portrait across four encoder architectures. Orange points (gratitude, love, remorse) consistently appear in the compact, low-dimensional region across all models.



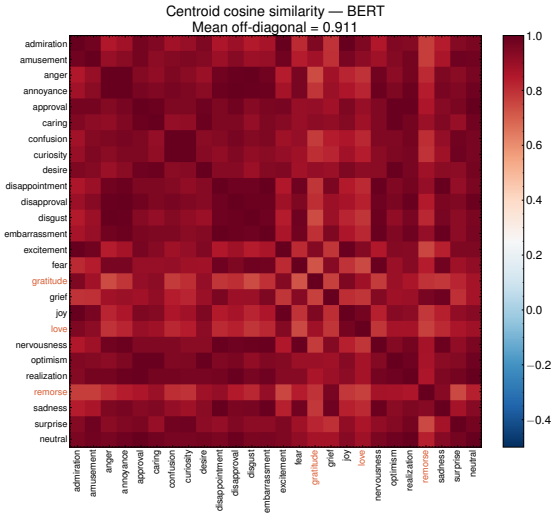
(a) BERT

(b) RoBERTa

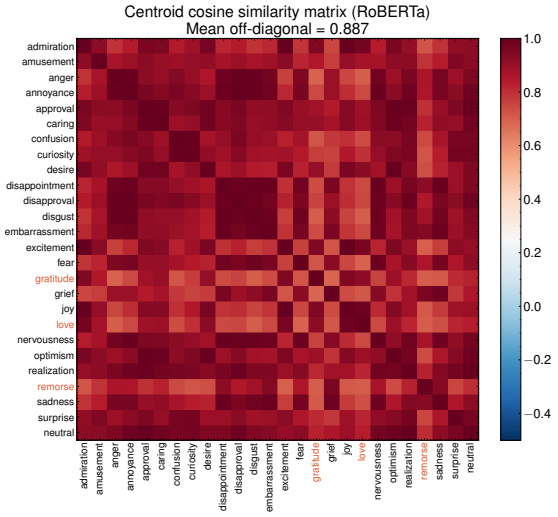
(c) DistilBERT

(d) MiniLM

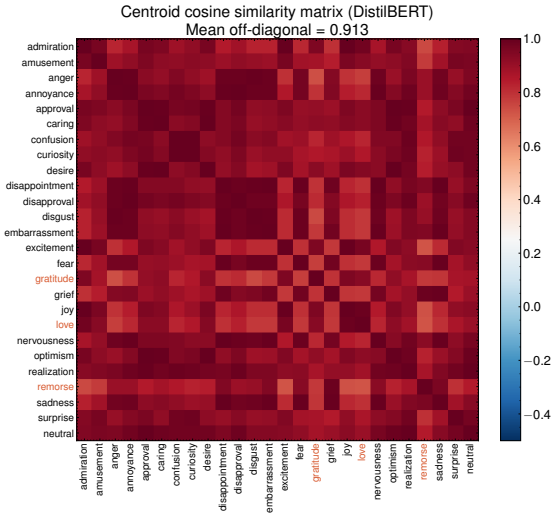
Figure 9: UMAP projections across four encoder architectures. Gratitude, love, and remorse appear spatially isolated from the central overlapping region in all cases. Stars denote class centroids; orange labels indicate geometric outliers.



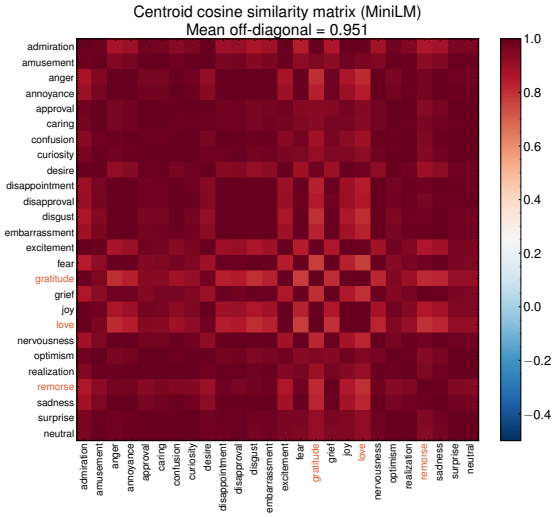
(a) BERT (mean = 0.911)



(b) RoBERTa (mean = 0.887)

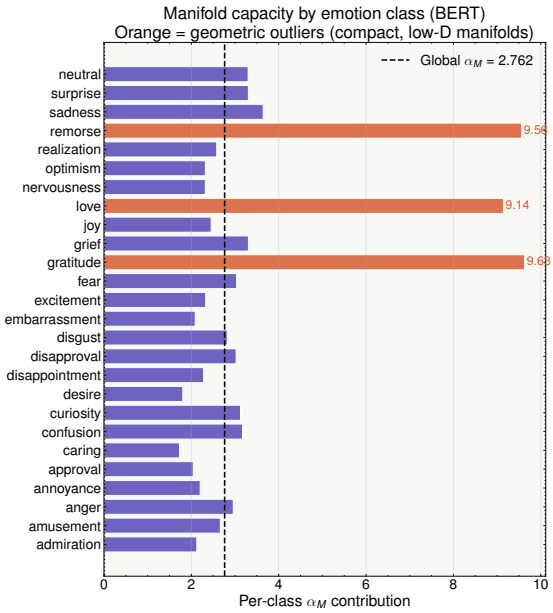


(c) DistilBERT (mean = 0.911)

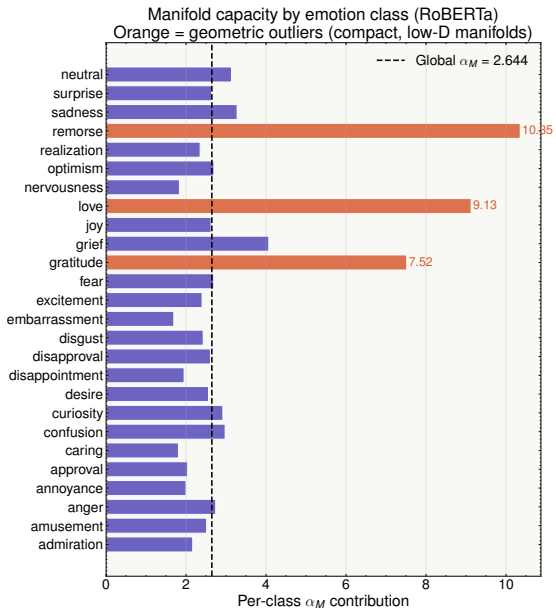


(d) MiniLM (mean = 0.913)

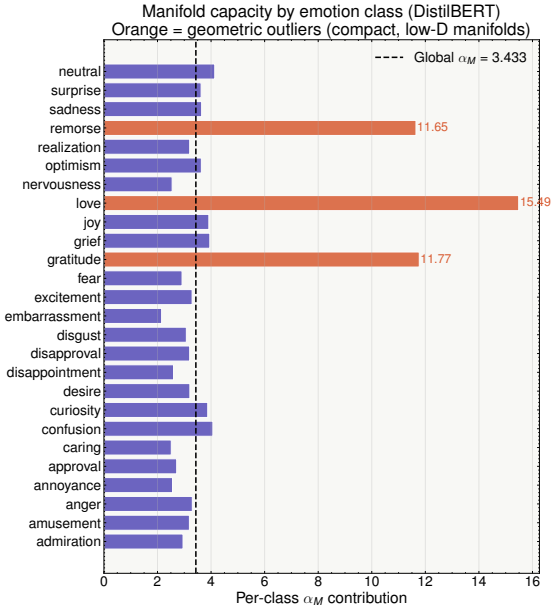
Figure 10: Centroid cosine similarity matrices across four encoder architectures. Off-diagonal values remain consistently large (0.887–0.951) across all models. Orange labels indicate geometric outliers (gratitude, love, remorse), which show systematically lower similarity to other centroids.



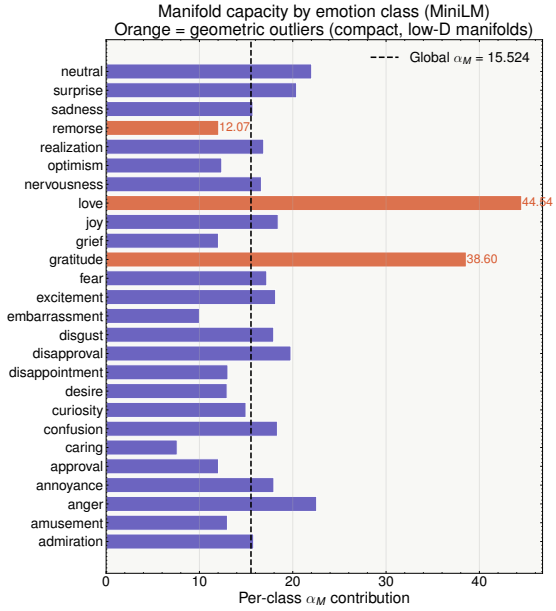
(a) BERT (global $\alpha_M = 2.762$)



(b) RoBERTa (global $\alpha_M = 2.644$)

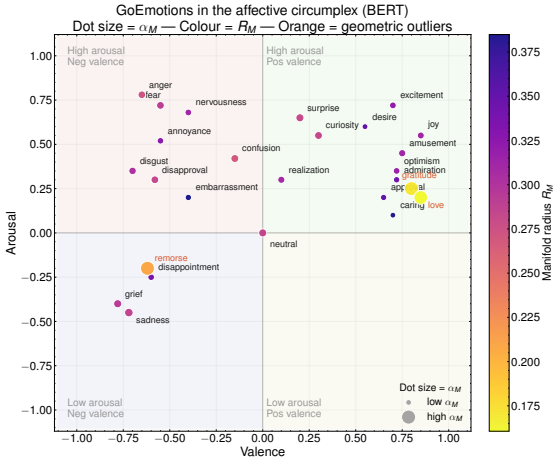


(c) DistilBERT (global $\alpha_M = 3.433$)

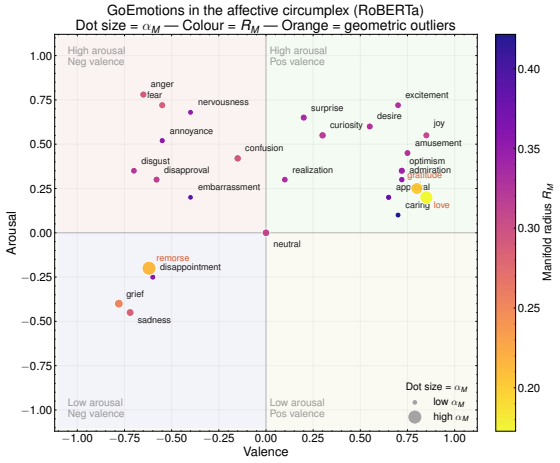


(d) MiniLM (global $\alpha_M = 15.524$)

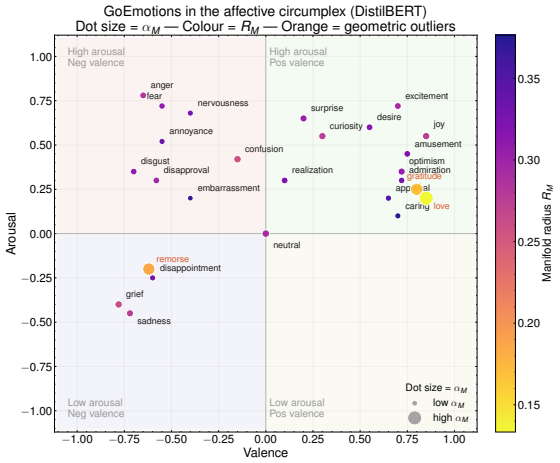
Figure 11: Per-class α_M contributions across four encoder architectures. Gratitude, love, and remorse (orange) consistently exceed the global mean by a factor of 3–4$\times$ across all models.



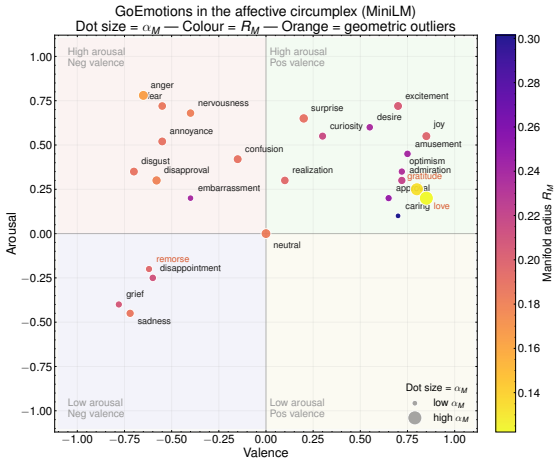
(a) BERT



(b) RoBERTa



(c) DistilBERT



(d) MiniLM

Figure 12: Emotion classes positioned in the affective circumplex across four encoder architectures. Dot size encodes α_M; color encodes R_M. Geometric outliers (orange) consistently concentrate in low-arousal regions regardless of architecture.

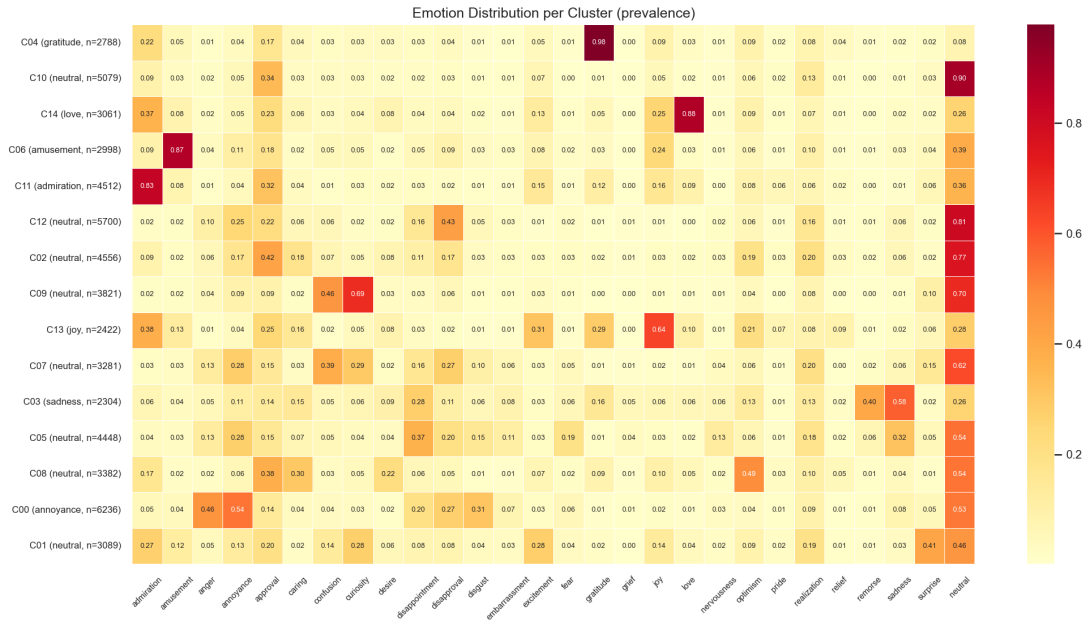


Figure 13: Emotion distribution across clusters. Each row corresponds to a cluster and each column to an emotion; values indicate the proportion of samples within the cluster exhibiting the corresponding emotion.

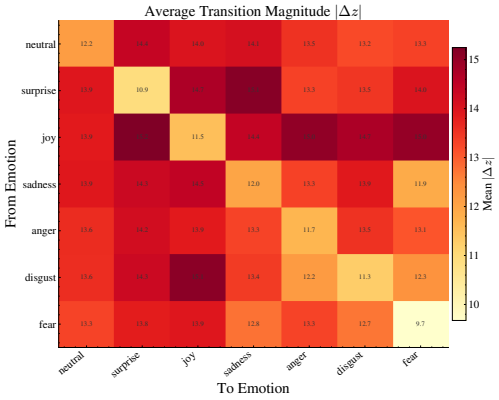


Figure 14: Average latent transition magnitude $\|\Delta z_t\|$ for emotion pairs. Rows denote the source emotion and columns the target emotion.

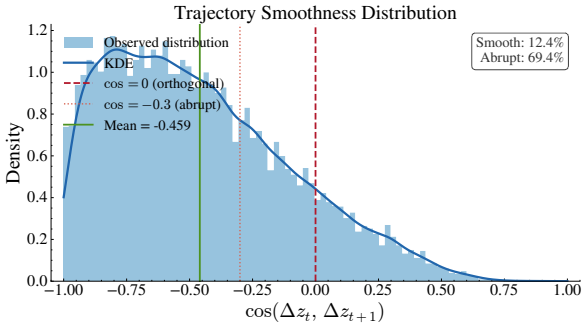


Figure 15: Distribution of cosine similarity between consecutive transition vectors. Negative values indicate directional changes, while positive values indicate smooth continuation.

The distribution is skewed toward negative values, with a mean of -0.45, and only 12.2% of transitions exhibiting positive alignment. This confirms that emotional dynamics are often non-linear, with frequent directional shifts between consecutive states.

D.3 Representative Trajectory Patterns

To complement the aggregate analysis, we highlight representative trajectory patterns observed in the dataset. Short conversations often exhibit abrupt transitions. For example:

> *[sadness] "oh my god, he's lost it"* $\rightarrow$ *[surprise] "what?"*

Longer conversations display structured but non-linear evolution. For instance:

> *[neutral]* $\rightarrow$ *[joy]* $\rightarrow$ *[sadness]* $\rightarrow$ *[neutral]* $\rightarrow$ *[anger]* $\rightarrow$ *[joy]*

These examples illustrate that emotional trajectories frequently combine persistence with sudden shifts, reflecting context-dependent changes.

D.4 Example Emotion Trajectories

To illustrate the observed trajectory patterns, we present representative conversation snippets with their emotion transitions.

Example 1: Abrupt transition (high $\|\Delta z\|$)

t_0 [sadness]:“oh my god, he’s lost it. he’s totally lost it.”
→
t_1 [surprise]:“what?”

A sharp emotional shift corresponding to a large latent displacement.

Example 2: Mixed trajectory (volatile) dynamics)

t_0 [neutral]:“or! or, we could go to the bank...”
→ t_1 [joy]:“you’re a genius!”
→ t_2 [sadness]:“aww, man...”
→ t_3 [neutral]:“now, there’s two reasons.”

Frequent emotion switching, consistent with negative cosine smoothness (abrupt trajectory changes).

Example 3: Smooth emotional progression

t_0 [neutral]:“okay, you guys free tonight?”
→ t_1 [joy]:“yeah!!”
→ t_2 [surprise]:“tonight?”

Gradual transitions with smaller directional change in latent space.

D.5 Statistics Summary

We summarize key statistics of latent trajectory dynamics in Table 10.

Metric	Value
Mean transition magnitude	13.14
Standard deviation	4.69
Same-emotion transitions	11.93
Different-emotion transitions	13.96
Alignment ratio	1.17

Table 10: Summary statistics of latent trajectory dynamics. Larger transition magnitudes are associated with changes in emotion, indicating alignment between latent movement and semantic shifts.

As shown in Table 10, transitions associated with emotion changes exhibit larger magnitudes than those preserving emotion, yielding an alignment ratio greater than 1. This confirms that movement in latent space corresponds to meaningful changes in emotional state.

E Intervention Analysis

This section provides additional quantitative analysis supporting the controllability of the latent space under vector-based interventions.

E.1 Alpha Sensitivity

We analyze the effect of intervention strength (α) on the magnitude and consistency of latent shifts. Table 11 shows that increasing α leads to larger reductions in distance to the target emotion, while maintaining near-perfect directional alignment.

These results indicate a clear trade-off: larger α yields stronger movement toward the target, while slightly reducing the proportion of improved samples. Cosine alignment remains near 1.0 across all values, confirming that interventions are directionally consistent.

α	Before	After	Δ	% Improved	Cosine
0.5	11.1538	11.0027	0.1511	74.4%	1.00
1.0	11.1538	10.8699	0.2840	73.2%	1.00
1.5	11.1538	10.7555	0.3983	72.2%	1.00
2.0	11.1538	10.6599	0.4939	71.6%	1.00
3.0	11.1538	10.5252	0.6286	68.8%	1.00

Table 11: Effect of intervention strength (α) on distance reduction toward the target emotion.

E.2 Top Controllable Directions

We further analyze which emotion transitions are most effectively controlled by the latent intervention mechanism. Table 12 reports the highest-performing sourceâtarget pairs.

Direction	Δ	% Improved
neutral → surprise	0.39	77.0%
anger → surprise	0.35	77.6%
neutral → joy	0.34	80.2%
anger → joy	0.33	78.0%
fear → surprise	0.32	79.2%

Table 12: Top-performing intervention directions ranked by distance reduction.

The results show that transitions toward positive or high-arousal emotions (e.g., joy, surprise) are particularly responsive to intervention, suggesting that the latent space encodes consistent directional gradients across emotion categories.

F Statistical Validation of Latent Dynamics

To assess whether the observed latent dynamics reflect statistically reliable structure rather than

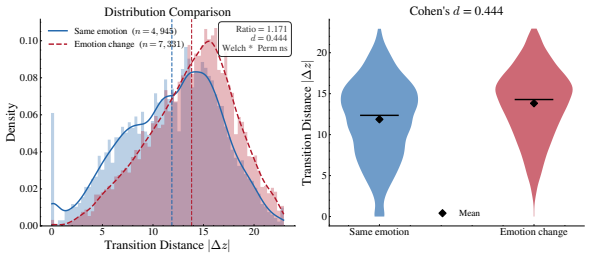


Figure 16: Distribution comparison for latent transition magnitudes. Emotion-changing transitions exhibit consistently larger displacements than same-emotion transitions, though with substantial overlap.

Statistic	Same Emotion	Emotion Change
Mean $\|\Delta z\|$	11.93	13.96
Std. Dev.	4.87	4.38
Median	12.39	14.36
Alignment Ratio	1.1707	
Cohen's d	0.444	
95% Bootstrap CI	[1.1548, 1.1865]	

Table 13: Statistical summary for trajectory alignment analysis.

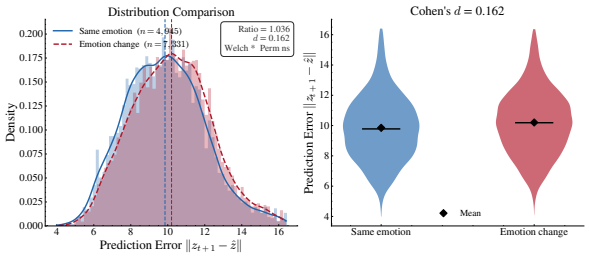


Figure 17: Distribution comparison for prediction errors. Emotion-changing transitions exhibit slightly higher prediction errors, though the effect size remains small.

Statistic	Same Emotion	Emotion Change
Mean Error	9.92	10.28
Std. Dev.	2.23	2.23
Median	9.80	10.21
Error Ratio	1.0365	
Cohen's d	0.162	
95% Bootstrap CI	[1.0282, 1.0450]	

Table 14: Statistical summary for prediction error analysis.

random variation, we perform hypothesis testing on both trajectory displacement and prediction error distributions. All significance results are interpreted conservatively alongside effect sizes.

F.1 Trajectory Alignment

We test whether emotion-changing transitions produce larger latent displacements than same-emotion transitions. Across 12,276 transitions, emotion-changing transitions exhibit larger movement magnitudes (13.96 ± 4.38) than same-emotion transitions (11.93 ± 4.87), yielding an alignment ratio of 1.1707. Welch's t-test ($t = -23.64$, $p < 0.001$), Mann–Whitney U ($p < 0.001$), bootstrap confidence intervals, and permutation testing all reject the null hypothesis. Bootstrap estimation over 10,000 resamples yields a 95% confidence interval of $[1.1548, 1.1865]$, excluding 1.0. The observed effect size remains modest (Cohen's $d = 0.44$), indicating consistent but overlapping distributions.

The larger displacement magnitudes observed during emotion-changing transitions suggest that conversational emotional evolution is reflected geometrically in latent space rather than occurring as purely symbolic label changes. At the same time, the moderate effect size indicates that emotional regions remain substantially overlapping, consistent with the continuous and compositional organization observed throughout the representation space.

F.2 Prediction Error Validation

We further test whether prediction errors increase during emotion-changing transitions. Emotion-changing transitions exhibit slightly larger prediction errors (10.28 ± 2.23) than same-emotion transitions (9.92 ± 2.23), producing an error ratio of 1.0365. Welch's t-test ($t = -8.82$, $p < 0.001$), Mann–Whitney U ($p < 0.001$), bootstrap analysis, and permutation testing again reject the null hypothesis. The bootstrap 95% confidence interval $[1.0282, 1.0450]$ excludes 1.0, although the practical effect size remains small (Cohen's $d = 0.16$).

The weaker effect size for prediction error suggests that short-term emotional dynamics remain broadly predictable even during emotion transitions. Prediction error therefore appears to capture subtle deviations from expected trajectory evolution rather than acting as a strong binary signal of emotion change. All hypothesis tests remain significant after Benjamini–Hochberg false discovery rate correction ($\alpha = 0.05$). Overall, these analyses support the conclusion that both latent displacement and prediction error encode emotionally meaningful structure beyond random expectation, while also indicating that the observed effects should be interpreted conservatively rather than overstated.

G Ablations and Robustness

G.1 Encoder Architecture Comparison

To evaluate whether the observed emotional geometry depends on encoder choice, the framework was evaluated across four transformer architectures: BERT, RoBERTa, DistilBERT, and MiniLM. Although absolute classification and clustering performance varied across models, all architectures consistently exhibited structured latent geometry, temporal persistence, compositionality, and controllable intervention behavior. These results suggest that the core latent-dynamical properties emerge robustly across encoder families rather than from a specific pretrained representation. Figure 8 qualitatively shows that the global manifold organization remains stable across architectures despite differences in local cluster sharpness and separation.

Across all architectures, trajectory alignment ratios remained greater than 1, indicating that emotion-changing transitions consistently correspond to larger latent displacements. Similarly, intervention success rates remained above 60% across models, suggesting that the learned latent spaces remain geometrically navigable independent of encoder selection. RoBERTa produced the strongest cluster separation and intervention performance, whereas MiniLM generated smoother but less sharply separated manifolds. Figure 9 further shows that the global emotional manifold structure remains qualitatively stable across architectures despite variation in local geometry.

G.2 Linear vs Nonlinear Classification Head

A potential concern is that the linear classifier imposes a geometric inductive bias that produces the observed manifold structure by construction. To address this, we trained both a linear logistic regression probe (SAGA solver) and a nonlinear MLP probe (hidden sizes 256–128, ReLU, early stopping) on the frozen latent representations under five-fold stratified cross-validation, repeated across all four encoder architectures.

The mean gap of $+0.001$ across all architectures is not statistically meaningful (paired t-test, $p > 0.05$ for all architectures). Notably, BERT shows a marginal reversal (-0.001) in favour of the linear probe, and MiniLM shows near-zero gap despite the convergence caveat.

This pattern is consistent across encoder families and confirms two things: the linear constraint does not impoverish the representation, and the manifold structure we observe — including radial centroid organisation, compact outlier manifolds, and compositional arithmetic — is intrinsic to the learned representations rather than imposed by the classification objective.

The Macro-F1 gap relative to the original GoEmotions paper (0.402 vs 0.64) is attributable to the 32-dimensional projection bottleneck rather than the linear classifier: the original paper uses the full 768-dimensional BERT output with a nonlinear head, whereas our framework compresses to $32d$ to produce a geometrically interpretable manifold. This compression is intentional and justified by the near-zero linear-nonlinear gap observed here.

G.3 Latent Dimension Sensitivity

We further evaluated latent dimensions $d \in \{16, 32, 64, 128\}$ to test sensitivity to representational capacity. Smaller latent spaces produced more compressed but less expressive manifolds, while larger dimensions improved local separation at the cost of increased overlap and reduced interpretability. A latent dimension of $d = 32$ provided the best tradeoff between clustering quality, trajectory stability, controllability, and geometric interpretability. Importantly, temporal persistence and trajectory alignment remained qualitatively stable across all evaluated dimensions, suggesting that the observed emotional dynamics are not tied to a specific embedding size.

G.4 Linear vs Nonlinear Transition Predictors

To evaluate whether emotional trajectory predictability arises from trivial linear continuity, we compared the proposed nonlinear MLP transition predictor against a purely linear predictor. The nonlinear model consistently achieved lower prediction error and larger improvement over the trivial baseline, particularly near abrupt emotional transitions. These results indicate that emotional trajectory evolution contains higher-order nonlinear structure beyond simple local continuity in latent space.

G.5 Randomized and Shuffled Controls

Several control experiments were conducted to verify that the observed geometric structure does not arise accidentally.

Architecture	Macro-F1	Silhouette	Align. ratio	Lag-1	Intervention (%)
BERT	0.3790	0.2254	1.1668	0.8848	64.6
RoBERTa	0.3824	0.2511	1.1773	0.8547	67.8
DistilBERT	0.3624	0.2385	1.1815	0.8911	65.2
MiniLM	0.2061	0.2443	1.1340	0.9479	60.8

Table 15: Cross-architecture comparison of latent geometry and temporal dynamics. All architectures preserve the core structural properties of the latent emotional manifold despite differences in absolute classification performance.

Architecture	Linear	Nonlinear	Gap
BERT	0.421 ± 0.001	0.421 ± 0.001	-0.001
RoBERTa	0.431 ± 0.004	0.432 ± 0.003	$+0.002$
DistilBERT	0.409 ± 0.005	0.410 ± 0.004	$+0.001$
MiniLM*	0.346 ± 0.003	0.348 ± 0.005	$+0.002$
Mean	0.402	0.403	$+0.001$

Table 16: Linear vs nonlinear probe Macro-F1 (5-fold CV) across all four encoder architectures. Gaps are within one standard deviation for all models, confirming that the linear classifier constraint does not impoverish the learned representation. *LogReg did not fully converge for MiniLM within 2000 iterations; gap is therefore conservative.

- Shuffling dialogue trajectories substantially reduced temporal persistence and destroyed coherent transition dynamics.
- Randomizing emotion labels eliminated meaningful cluster-label alignment and reduced intervention consistency.
- Random intervention directions produced near-zero average movement toward target emotional regions.
- Random centroid assignments removed compositional arithmetic behavior and destabilized nearest-neighbour semantic structure.

Together, these controls demonstrate that the observed manifold geometry, trajectory organization, and intervention behavior emerge from learned emotional structure rather than statistical artifacts or embedding anisotropy.

G.6 Affective Circumplex Structure Across Architectures

To further examine whether the learned emotional manifolds align with psychologically meaningful affective structure, we projected emotion centroids into the valence–arousal circumplex using NRC-VAD affective coordinates.

Figure 12 shows that all architectures preserve a consistent global affective topology organized along broad valence and arousal gradients. Positive emotions cluster in the high-valence region, while negative emotions remain separated along the negative-valence axis. Across models, emotionally distinctive states such as love, gratitude, and remorse consistently emerge as compact geometric outliers with high manifold capacity contributions, suggesting stable and highly separable affective structure. RoBERTa exhibits the strongest polarity separation, whereas MiniLM produces smoother and more compressed affective organization while preserving the same global topology. These results further support the interpretation that the learned latent geometry captures psychologically meaningful affective structure rather than arbitrary clustering artifacts.