CHARACTERIZING THE GEOMETRIC AND TEMPORAL ORGANIZATION OF EMOTION IN NEURAL REPRESENTATION SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion is encoded in the internal representations of language models, yet the principles governing its organization remain poorly understood. This raises a fundamental question: do emotional concepts form a structured geometry in representation space, and how does that structure evolve as emotional states unfold through conversation? We address these questions by modeling categorical, continuous, and temporal aspects of emotion through three representation forms, State, Continuous, and Dynamic, using BERT, DistilBERT, MiniLM, and RoBERTa. This choice of encoder-only Transformers provides contextualized representations while enabling comparison of emotional organization across architectures. Applying these representations to emotion classification and conversational datasets, we find that emotional concepts are broadly accessible through linear decoding. Yet this accessibility does not imply an unstructured representation. Emotional concepts concentrate in compact, structured subspaces with lower intrinsic dimensionality than the ambient representation space. The resulting organization varies with the representation objective, exposing different computational properties while preserving core geometric patterns across architectures. This organization also extends beyond individual emotional states into their evolution through dialogue. Across conversational trajectories, emotional representations exhibit reproducible temporal organization, with neighboring states showing predictive persistence and systematic changes in representation dynamics around emotional transitions. Together, these results establish a unified framework for characterizing emotion, linking the geometry of emotional states to representation learning and their evolution over time within conversational context.

1 INTRODUCTION

Emotion can be modeled as discrete affective states (Ekman, 1992; Demszky et al., 2020), represented through continuous variation in intensity and quality (Russell, 1980; Buechel & Hahn, 2017; Cowen & Keltner, 2017), or understood as evolving with contextual and conversational dynamics (Poria et al., 2019; Barrett, 2017) as illustrated in Figure 1. Understanding how these properties are organized within language model representations is therefore central to determining what those representations encode about emotion and how that information is structured (Zhang & Zhong, 2025; Sofroniew et al., 2026). However current evaluation methods focus on downstream performance, including emotion recognition, conversational modeling, and benchmark prediction (Fu et al., 2025; Sabour et al., 2024). Such evaluations establish what emotional concepts can be recovered from a representation, but they provide limited insight into how that information are organized within the underlying latent space. This distinction is important because representations with similar task performance may encode emotional concepts through fundamentally different geometric and temporal structures. Recent work on representation geometry has shown that concepts can be recovered along linear directions (Park et al., 2024), organized into structured and hierarchical subspaces (Park et al., 2025), and expressed through manifolds whose separability changes across model depth (Cohen et al., 2020; Mamou et al., 2020). Intrinsic dimensionality has likewise been shown to vary systematically across layers and representations (Valeriani et al., 2023; Cheng et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

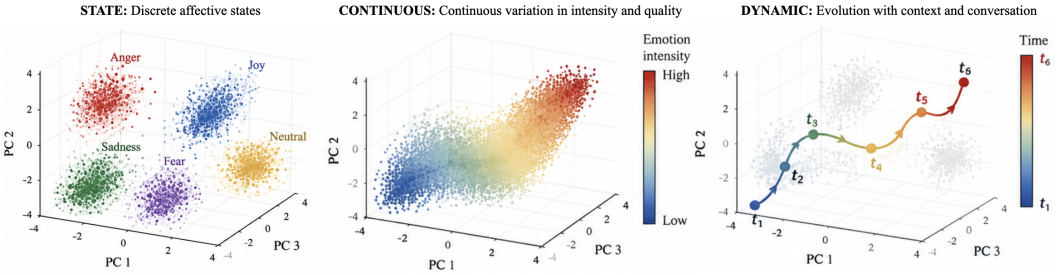


Figure 1: **Three complementary views of emotion in representation space.** The figure illustrates discrete states, continuous variation, and context-dependent trajectories in a 3D latent space.

In parallel, affective representation studies have identified low dimensional and steerable emotional subspaces (Reichman et al., 2026; Sun et al., 2026; Saini et al., 2026; Zhang et al., 2024), supporting broader accounts of emotion as a continuous and structured space rather than a collection of isolated categories (Russell, 1980; Susanto et al., 2020). Despite these advances, the two perspectives remain largely disconnected. Geometric studies rarely examine emotional representations specifically, while affective studies often focus on individual geometric properties, leaving unclear how accessible, compact, structured, and temporally organized emotional concepts are across representation objectives and encoder architectures. Further, the extent to which such structure is genuinely linear remains contested (Engels et al., 2025), and existing work provides limited characterization of how emotional representations evolve across conversational time (Fumero et al., 2026). We address these gaps by characterizing emotion representations jointly as latent spaces and temporal trajectories. We compare State, Continuous, and Dynamic representations across multiple affective tasks, datasets, and encoder backbones (Huh et al., 2024), examining information accessibility, latent geometry, state and process organization, and temporal dynamics. Our analyses show that emotional concepts are broadly linearly accessible while remaining concentrated in compact and structured subspaces, and that conversational trajectories exhibit reproducible temporal structure without evidence for a universal directional transition mechanism (Chun et al., 2026).

Our contributions are as follows: (1) We introduce a unified framework for characterizing emotion as State, Continuous, and Dynamic representations across four Transformer encoders and multiple affective datasets. (2) We systematically assess information accessibility through complementary tasks and probe complexities, showing when emotional information is linearly recoverable. (3) We characterize the geometry of emotional representations, revealing compact, low-dimensional and locally structured organization beyond task performance. (4) We model dialogue representations as temporal trajectories, quantifying persistence, directional structure, drift, and transition organization under matched temporal controls. The remainder of the paper covers: Section 2 reviews prior work on representation geometry and affective representation learning. Section 3 describes the representation objectives, datasets, experimental tasks, and evaluation metrics. Section 4 presents the empirical results across information accessibility, geometric structure, state and process organization, and temporal dynamics. Section 5 concludes with the implications and limitations of the findings.

2 Related Work

Recent studies of neural representations increasingly examine hidden spaces through properties such as dimensionality, local geometry, and organization across layers. Intrinsic dimensionality varies non-monotonically with depth and can peak in early to middle layers (Viswanathan et al., 2026; Ruppik et al., 2026), while intermediate representations can retain more usable structure than final layers (Skean et al., 2024).

This structure is also shaped by how representations are learned: next-token optimization can induce semantic organization (Zhao & Thrampoulidis, 2025), while analyses of representation manifolds suggest that their geometry has identifiable origins rather than arising solely from projection artifacts (Modell et al., 2025). Such work builds on classical manifold learning, where geodesic distances recover nonlinear low-dimensional structure (Tenenbaum et al., 2000), providing a basis

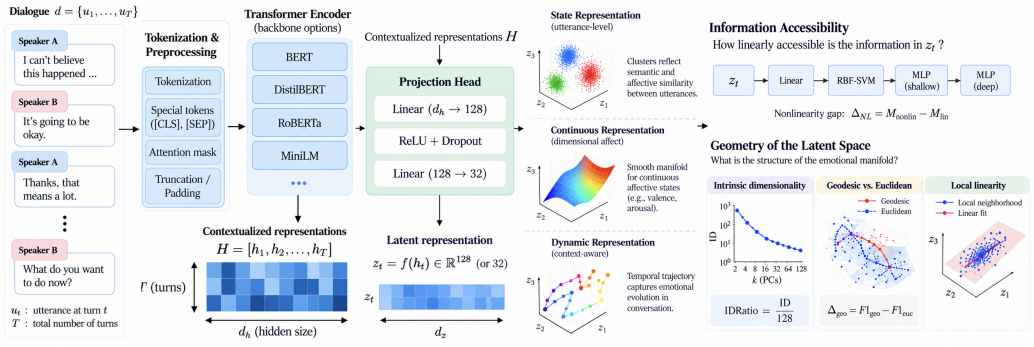


Figure 2: **Overview of the proposed framework.** Dialogue utterances are tokenized and encoded using four frozen transformer backbones, then projected into a shared 128-dimensional latent space to construct State, Continuous, and Dynamic representations. These representations are evaluated for information accessibility, latent geometry, and conversational organization.

for examining neural spaces beyond downstream performance. This perspective has more recently been extended to emotion. Emotional concepts exhibit a defined internal geometry that changes with model scale and depth (Zhang & Zhong, 2025), while emotion categories display hierarchical rather than purely flat organization (Okawa et al., 2025). Other work suggests that emotion concepts participate functionally in model computation (Sofroniew et al., 2026), with causal analyses tracing emotion inference to identifiable internal mechanisms (Tak et al., 2025; Cambria et al., 2026). These findings align with broader evidence that high-level model behavior can be mediated by individual directions in representation space (Arditi et al., 2024). Prediction-trained language models have also been shown to recover aspects of human-like conceptual organization (Xu et al., 2025), while surveys of affective language modeling identify rigid categorical treatment of emotion as an important limitation (Plaza-del Arco et al., 2024). Despite this progress, these lines of work remain fragmented across individual geometric or functional properties. Limited attention has been given to how emotional structure varies across representation objectives and encoder architectures, or to how it is organized over conversational time. Our work brings these perspectives together by jointly examining information accessibility, geometric structure, state and process organization, and temporal dynamics across multiple representation objectives and encoder backbones (Liu et al., 2025). We therefore characterize not only whether Emotional concepts are present, but how it is organized and evolves within the latent space.

3 Methodology

We study emotion representations through a controlled comparison of three representation objectives across four frozen transformer encoders. The analysis separates four questions: what affective information is accessible, how it is geometrically organized, whether it exhibits state or process structure, and whether it evolves systematically over conversational time. All representations are mapped to a common 128-dimensional space, enabling direct comparison of task performance and geometric properties. This framework is illustrated in Figure 2, providing a common basis for comparing how different representation objectives organize affective information.

3.1 Representation Construction

Let u_i denote an utterance and E_θ a frozen encoder. We first obtain its contextual representation

$$h_i = E_\theta(u_i), \qquad h_i \in \mathbb{R}^D, \tag{1}$$

and project it into a common latent space,

$$z_i = P_\phi(h_i), \qquad z_i \in \mathbb{R}^{128}. \tag{2}$$

All geometric analyses are performed in the projected space z_i, unless otherwise stated.

To assess whether the observed compactness is introduced by the projection, we additionally evaluate intrinsic dimensionality and linear probe performance in the encoder's native representation h_i under the same protocol. We construct three representations that impose different assumptions about affect. The *State* representation models emotion as a discrete property of the current utterance,

$$z_i^S = f_S(h_i). \tag{3}$$

The *Continuous* representation instead organizes affect along continuous dimensions,

$$z_i^C = f_C(h_i), \qquad a_i \approx g_C(z_i^C), \tag{4}$$

where a_i denotes the available dimensional affective targets. This formulation follows Russell's circumplex model (Russell, 1980; Cowen & Keltner, 2021; 2017) and prior work on dimensional affective representations (Sun et al., 2026). Finally, the *Dynamic* representation conditions the current state on preceding dialogue context,

$$z_t^D = f_D(h_t, \mathcal{H}_{<t}), \qquad \mathcal{H}_{<t} = \{h_1, \ldots, h_{t-1}\}. \tag{5}$$

This follows contextual emotion modeling in dialogue systems (Ghosal et al., 2020; Fu et al., 2025; Jiang et al., 2024). The three objectives therefore provide controlled state, continuous, and context-dependent views of affect. We evaluate these objectives on GoEmotions, MELD, DailyDialog, IEMOCAP, and EmoBank using BERT, DistilBERT, MiniLM, and RoBERTa. Dialogue identity and turn order are retained for conversational analyses, while splits are performed at the dialogue level where applicable. Primary BERT experiments are evaluated across three random seeds, while all cross-backbone comparisons use seed 42 for a controlled like-for-like comparison.

3.2 Information Accessibility

We assess the computational affordances of each representation across complementary tasks: classification (macro-F1), dimensional affect prediction (R^2), next-turn forecasting (macro-F1), editing (target attainment), explainability (probe faithfulness), and ambiguity (entropy on ambiguous instances). Editing is motivated by evidence that high-level attributes can be mediated by directions in activation space (Arditi et al., 2024; Park et al., 2024), while explainability complements causal analyses of emotion representations (Tak et al., 2025). To separate absent information from information that is difficult to decode, we evaluate a probe ladder comprising a linear probe, RBF-SVM, shallow MLP, and deep MLP. For task score M, we define the nonlinearity gap as:

$$\Delta_{\mathrm{NL}} = M_{\mathrm{nonlin}} - M_{\mathrm{lin}}, \tag{6}$$

where M_{nonlin} is the strongest nonlinear probe score. A near-zero gap indicates that linear decoding captures most accessible information, whereas a positive gap indicates additional nonlinear accessibility (Park et al., 2024; Engels et al., 2025).

3.3 Geometry and Temporal Structure

We characterize the latent spaces independently of probe complexity, first measuring their intrinsic geometry and then, for conversational data, their organization over time. For conversational data, Figure 3 illustrates this temporal view through latent-state transitions and emotion trajectories in the latent space. Intrinsic dimensionality is estimated for each representation and normalized by the ambient dimensionality of the corresponding representation:

$$\mathrm{IDRatio} = \frac{\mathrm{ID}}{D}, \tag{7}$$

where $D = 768$ for the encoder representation h and $D = 128$ for the projected representation z. We further compare Euclidean and graph-based geodesic kNN classification, with

$$\Delta_{\mathrm{geo}} = F1_{\mathrm{geo}} - F1_{\mathrm{euc}}, \tag{8}$$

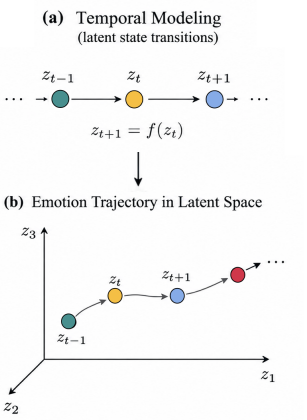


Figure 3: Successive latent states and their resulting conversational trajectories.

measuring whether manifold-aware neighborhoods recover structure missed by Euclidean distance (Tenenbaum et al., 2000; Modell et al., 2025). Local linearity provides a complementary measure of neighborhood structure (Cohen et al., 2020; Mamou et al., 2020; Chun et al., 2026). We also retain the leading principal components for $k \in \{2, 4, 8, 16, 32, 64, 128\}$ and track intrinsic dimensionality, task performance, and explained variance to assess how rapidly useful structure saturates. For conversational data, we treat each dialogue as an ordered latent trajectory (Fumero et al., 2026). For consecutive states, the displacement is:

$$\Delta z_t = z_{t+1} - z_t. \tag{9}$$

We evaluate chronological direction using a forward-versus-backward probe and a within-dialogue shuffled null that preserves utterances and dialogue identity while removing temporal order. Directional coherence measures the alignment of successive displacements,

$$C_{\text{dir}} = \frac{1}{T-2} \sum_{t=1}^{T-2} \cos(\Delta z_t, \Delta z_{t+1}), \tag{10}$$

while latent drift measures displacement as a function of intervening turns,

$$D(\delta) = \mathbb{E}_t \left[\|z_{t+\delta} - z_t\|_2 \right]. \tag{11}$$

Because consecutive displacements share z_{t+1}, directional coherence can have a negative expectation under the shuffled null. We therefore interpret it relative to this matched null rather than against zero. To separate temporal structure from general dialogue similarity, we compare adjacent pairs with non-adjacent pairs from the same dialogue and with cross-dialogue pairs. The non-adjacent condition controls for dialogue, speaker, and broader context while removing immediate adjacency. Forward and backward displacement probes test directional asymmetry, while antecedent–emotion analysis measures the predictive gain from the preceding emotional state relative to a matched control (Cowen & Keltner, 2017). Together, these analyses characterize latent geometry, temporal organization, and directed emotion transitions.

3.4 Representation-Level Indices

We complement task-specific scores with representation-level indices that capture properties not exposed by any single downstream task. The Information Preservation Index (IPI) quantifies how much affective information remains linearly accessible across the evaluated affective facets, the Representation Agreement Score (RAS) measures agreement between representation views on the same instances, and the Task Affordance Score (TAS) normalizes performance within each task to compare the relative computational strengths of the representations. The State–Dynamics Index (SDI) further characterizes whether emotion-specific information is more strongly recoverable from the current State representation or the Dynamic representation. Full definitions and derivations of all representation-level indices are provided in Appendix B. All indices are computed independently for BERT, DistilBERT, MiniLM, and RoBERTa using the same preprocessing, splits, latent dimensionality, and evaluation procedures, enabling representation-level comparisons without changing the experimental conditions across architectures.

4 Experiments and Results

We evaluate the three representation objectives from complementary perspectives, moving from downstream task affordances to information accessibility, latent geometry, and temporal organization. Across four encoder backbones, this analysis reveals both what Emotional concepts are accessible and how it is structured within the underlying representation space.

4.1 Task Affordances and Information Accessibility

We first examine whether the three representation objectives support the same downstream computations. Figure 4 shows a clear task- and dataset-dependent pattern rather than a universal winner. State is strongest for GoEmotions classification (0.4693 macro-F1) and editing (0.8120). Continuous is strongest for classification on DailyDialog and IEMOCAP, both regression tasks, and

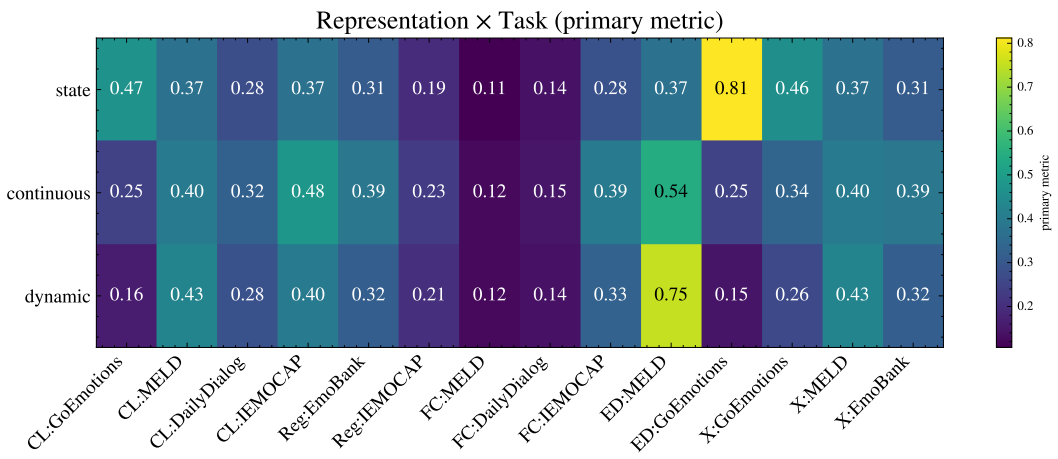


Figure 4: **Representation–task matrix.** Primary performance across State, Continuous, and Dynamic representations for BERT, averaged over three seeds. Task abbreviations: CL - classification, Reg - regression, FC -forecasting, ED - editing, X - explainability.

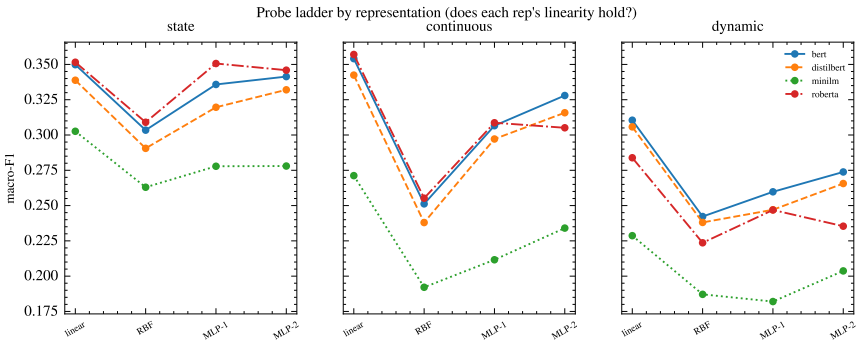


Figure 5: **Probe complexity and information accessibility.** Performance of linear, RBF-SVM, shallow-MLP, and deep-MLP probes across representations and encoder backbones. Increasing probe complexity provides little systematic improvement over linear decoding.

forecasting on DailyDialog and IEMOCAP. Dynamic is strongest for classification (0.4261), forecasting (0.1242), editing (0.7487), and explainability (0.4262) on MELD. Explainability on GoEmotions and EmoBank instead favors State (0.4580) and Continuous (0.3862), respectively. The winning representation therefore depends on the task and dataset rather than following a single performance hierarchy. Detailed performance values for all representation–task pairs are reported in Appendix F.4. We next ask whether the information exposed by these representations requires nonlinear decoding. Figure 5 compares linear, RBF-SVM, shallow-MLP, and deep-MLP probes. Increasing probe complexity provides little systematic improvement over linear decoding. For example, for State on GoEmotions, the linear probe achieves 0.4477 macro-F1, compared with 0.4339 for the deep MLP. On MELD, the corresponding scores are 0.3673 and 0.3662. Across encoder backbones, the mean nonlinearity gap is -0.0200 for BERT, -0.0192 for DistilBERT, -0.0270 for MiniLM, and -0.0204 for RoBERTa, with near-zero gaps in 83.3%, 75.0%, 91.7%, and 66.7% of settings, respectively. Thus, under the tested probe families, task-relevant information is largely accessible through linear decoding, with little systematic benefit from increasing probe complexity.

4.2 Low-Dimensional and Structured Geometry

Linear accessibility does not imply geometrically trivial representations. We therefore characterize the latent spaces independently of probe performance. Although all representations occupy a 128-dimensional ambient space, their corrected intrinsic dimensionality remains substantially lower

Table 1: **Cross-backbone geometric and accessibility statistics.** Seed-42 results are reported for all four encoders. Mean nonlinearity gap, maximum absolute gap, fraction of settings with near-zero gaps, intrinsic dimensionality, and IEMOCAP geodesic-edge statistic are reported for each encoder. Bold values indicate the most favorable value for each statistic: Δ_{NL} closest to zero, smallest maximum absolute gap, highest near-zero fraction, lowest intrinsic dimensionality, and highest IEMOCAP geodesic-edge statistic.

Backbone	Δ_{NL}	$\max\lvert\Delta_{\mathrm{NL}}\rvert$	Near-zero	Mean ID	IEMOCAP Geo.
BERT	−0.0200	0.0928	0.833	9.691	0.0122
DistilBERT	**−0.0192**	0.0805	0.750	9.388	0.0198
MiniLM	−0.0270	0.0866	**0.917**	**6.187**	**0.0505**
RoBERTa	−0.0204	**0.0703**	0.667	8.745	0.0300

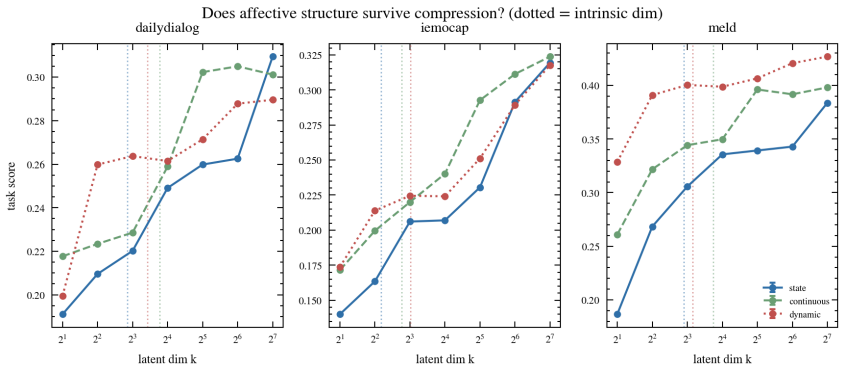


Figure 6: **Task performance across retained dimensions.** Task performance as the dimensionality of the retained representation increases.

across encoder backbones, with mean ID values of 9.69 for BERT, 9.39 for DistilBERT, 6.19 for MiniLM, and 8.75 for RoBERTa. The largest observed ID is 15.76, only 12.3% of the ambient dimensionality, indicating a consistently compact effective space. Local structure is likewise not fully captured by Euclidean neighborhoods: on IEMOCAP, the geodesic-edge statistic is positive for all four backbones, ranging from 0.0122 to 0.0505. These results indicate that emotional concepts are both readily accessible and geometrically concentrated within structured latent spaces. To test whether compactness is introduced by the projector, we additionally evaluate intrinsic dimensionality in the encoder's native space h under the same protocol. Intrinsic dimensionality is already low before projection, with a mean of 14.2 out of 768 ambient dimensions (1.8%), compared with 9.4 in the projected space (1.2%). On held-out IEMOCAP, dimensionality is preserved or slightly increased after projection, while linear probe accuracy is consistently higher in h than in z, indicating that the objective-specific projection can discard information that transfers across domains. We further examine how much of this structure is retained as the representation dimensionality is reduced. Across datasets and representation objectives, task performance exhibits diminishing returns as additional dimensions are retained (Fig. 6). The corresponding intrinsic-dimensionality and variance analyses, reported in Appendix E.3, show that geometric complexity and explained variance largely stabilize well before the full 128 dimensions are used. Thus, the compactness of the representations is not simply a consequence of evaluating them at a fixed low dimension.

4.3 Latent Density Structure

To complement the quantitative geometric measures, we visualize the concentration of latent samples using density surfaces. Figure 7 shows the BERT representations on DailyDialog after projection onto the first two principal components. The three objectives exhibit visibly different spatial organizations: State forms comparatively compact regions, Continuous produces smoother localized structures, and Dynamic shows more extended configurations. These differences provide an intuitive view of the distinct geometries induced by the representation objectives, while the cor-

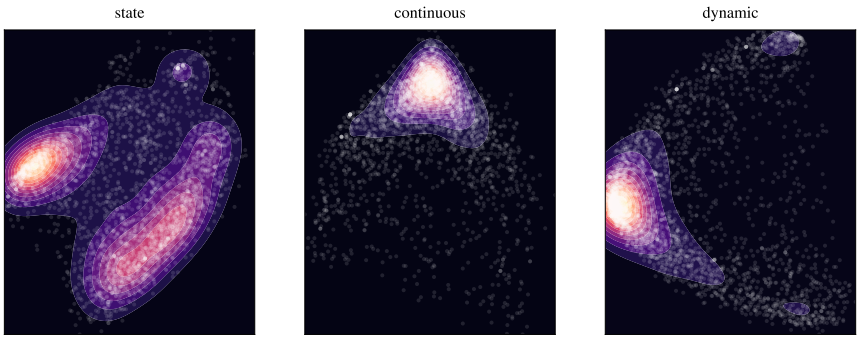


Figure 7: **Latent density surfaces for BERT on DailyDialog.** PCA-2D density visualizations of the State, Continuous, and Dynamic representations. Surfaces indicate regions of higher latent-sample concentration. The visualization is qualitative and complements the quantitative geometric analyses.

responding density surfaces and quantitative manifold analyses across datasets and backbones are reported in Appendix G.

4.4 Robustness Across Encoder Backbones

We next assess whether the representation–task specializations persist across encoder architectures. Table 2 compares the winning representation for each of the 14 task–dataset pairs across BERT, DistilBERT, MiniLM, and RoBERTa. The overall pattern is stable: DistilBERT matches BERT on all 14 pairs, MiniLM on 12 (85.7%), and RoBERTa on 10 (71.4%). The disagreements are concentrated on MELD, while DailyDialog and IEMOCAP classification, both regression settings, DailyDialog and IEMOCAP forecasting, and GoEmotions editing and explainability retain the same winner across all four backbones. This consistency indicates that the observed representation specializations are not artifacts of a particular encoder. At the same time, the MELD-specific deviations show that these preferences are not architecture-invariant, supporting a view of representation specialization as a robust tendency rather than a fixed specialization pattern.

Table 2: **Representation–task winners across encoder backbones.** Each entry denotes the representation achieving the highest primary score for the corresponding task–dataset pair. S, C, and D denote State, Continuous, and Dynamic, respectively.

Task	Dataset	BERT	DistilBERT	MiniLM	RoBERTa
Classification	GoEmotions	S	S	S	S
	MELD	D	D	D	C
	DailyDialog	C	C	C	C
	IEMOCAP	C	C	C	C
Regression	EmoBank	C	C	C	C
	IEMOCAP	C	C	C	C
Forecasting	MELD	D	D	C	C
	DailyDialog	C	C	C	C
	IEMOCAP	C	C	C	C
Editing	MELD	D	D	C	D
	GoEmotions	S	S	S	S
Explainability	GoEmotions	S	S	S	S
	MELD	D	D	D	C
	EmoBank	C	C	C	C

4.5 Temporal Structure in Dialogue Trajectories

We examine whether latent representations exhibit systematic conversational drift by treating each dialogue as a trajectory through latent space. Figure 8 illustrates representative trajectories, while Table 3 summarizes the effect on DailyDialog. Temporal direction is recoverable above the within-dialogue-shuffled control on DailyDialog across all four backbones, with Δ ranging from $+0.077$ to $+0.086$. In contrast, MELD and IEMOCAP show substantially weaker effects. The complete dataset-level results are reported in Appendix E.4. The DailyDialog effect is comparable for State ($+0.079$) and Dynamic ($+0.078$), despite State having no recurrent component, indicating that the observed drift does not originate from the GRU. We therefore interpret the result as systematic latent position drift within structured dialogue rather than evidence that affect itself is temporally organized.

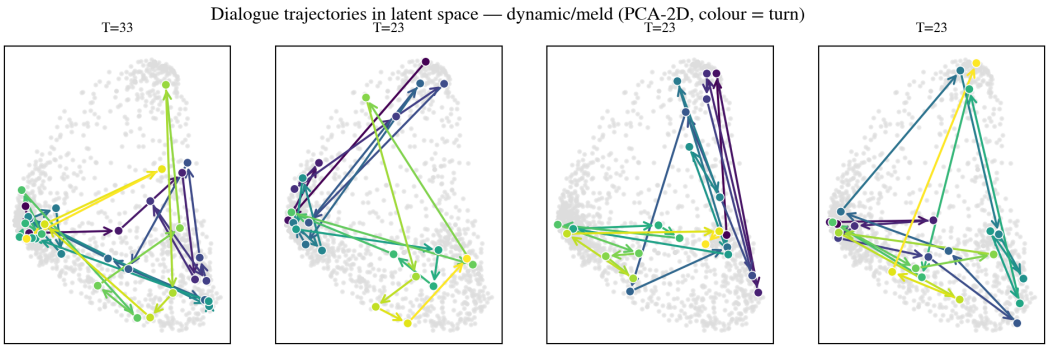


Figure 8: **Latent dialogue trajectories.** Representative Dynamic trajectories from BERT on MELD, projected onto the first two principal components, with arrows indicating conversational progression.

4.6 Representation-Level Metrics

We quantify representation-level properties that are not captured by downstream task performance alone. For BERT at seed 42, the Information Preservation Index (IPI) varies across representation objectives and datasets: State is highest on GoEmotions (0.461), Continuous on EmoBank (0.321) and IEMOCAP (0.318), and Dynamic on MELD (0.390). The State–Dynamics Index (SDI) reveals emotion-specific organization rather than a uniform state–process structure, with the direction and magnitude of the effect varying across emotions, datasets, and encoder backbones. In contrast, the capacity-corrected Representation Complementarity Score (RCS^*) indicates additional linearly accessible information from combining the three representation views beyond the strongest individual representation. Across the 16 backbone–dataset settings, RCS^* is positive in 15 settings, with the only negative value occurring for DistilBERT on GoEmotions (-0.002). The largest corrected gains are observed on MELD for DistilBERT (0.062), DailyDialog for RoBERTa (0.060), and IEMOCAP for RoBERTa (0.058). All values reported here use seed 42 for the controlled cross-backbone comparison. Complete results are provided in Appendix F.

Table 3: **Temporal drift on DailyDialog.** Δ denotes chronological minus within-dialogue-shuffled accuracy.

	BERT	DistilBERT	MiniLM	RoBERTa
Δ	$+0.086$	$+0.084$	$+0.079$	$+0.077$

Limitations and Future Work: Our work covers emotional representations across multiple encoder architectures and emotion and dialogue datasets. However, settings involving other model families, languages, and data modalities are an important direction for future work. Temporal effect in particular may partly reflect dataset-specific conversational structure, motivating further evaluation in broader conversational settings.

5 Conclusion

Our results show that emotional concepts in neural representations are structured across accessibility, geometry, and conversational dynamics. Across three representation objectives, four encoder backbones, and multiple affective datasets, emotional concepts are broadly accessible through simple linear decoding while occupying lower-dimensional and locallt structured spaces. Representation objectives nevertheless induce distinct computational affordances, with categorical, continuous, and context-dependent representations showing different strengths across recognition, regression, forecasting, editing, and explainability. Representation-level analyses reveal differentiated information preservation and emotion-specific state and process organization, while dialogue trajectories exhibit weak but reproducible temporal structure, particularly in DailyDialog. Those patterns hold across encoder architectures, although their strength varies by dataset and task. Overall, emotion in neural representations is neither a purely categorical label nor an unstructured latent property, but structured phenomenon shaped by geometry, representation objective and conversational context, providing foundation for studying emotional representations as organized latent spaces.

Ethics Statement

This work uses publicly available datasets for emotion and dialogue analysis and does not involve the collection of new data from human participants, human-subject experimentation, or interaction with individuals. We use the datasets solely for research purposes and follow the usage conditions and licenses specified by their respective sources. The pretrained transformer encoders used in our experiments are applied in a frozen setting, and no additional personal or sensitive information is collected or inferred beyond the information contained in the datasets. Because the study analyzes emotional representations, we acknowledge potential concerns regarding bias and the interpretation of emotion labels. Emotion categories and annotations may reflect cultural, linguistic, and annotator-specific biases and should not be interpreted as definitive measurements of an individual's internal emotional state. Our analyses concern patterns in model representations and do not make clinical or psychological diagnoses. We report the experimental methodology and limitations to support reproducibility and responsible interpretation of the findings. The authors take responsibility for the ethical use and reporting of the data and methods presented in this work.

Reproducibility Statement

We provide detailed descriptions of the representation construction, experimental protocols, evaluation metrics, and temporal analyses in the main paper and Appendix A. The datasets, preprocessing procedures, encoder architectures, representation objectives, training settings, and evaluation protocols are documented in the corresponding sections and appendices. Additional implementation details, complete experimental results, ablation analyses, and statistical procedures are provided in the appendices. The theoretical formulation and derivation of the Representation Complementarity Score are provided in Appendix B. Source code and supplementary materials will be made available upon acceptance to facilitate reproduction of the reported results.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, 2017.

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

Erik Cambria, Rui Mao, Xulang Zhang, Luwei Xiao, Tiesunlong Shen, and Avinash Anand. 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, 2026. doi: 10.1109/TCSS.2026.3677113.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations*, volume 2025, pp. 35389–35414, 2025.

Chanwoo Chun, Alexandre Polo, and SueYeon Chung. Emergent manifold separability during reasoning in large language models, 2026. URL https://arxiv.org/abs/2602.20338.

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

Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

Alan S Cowen and Dacher Keltner. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 25(2):124–136, 2021.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Li. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics, 2025. URL https://arxiv.org/abs/2403.07260.

Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. In *International Conference on Learning Representations*, volume 2026, pp. 24765–24793, 2026.

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

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Dazhi Jiang, Hao Liu, Geng Tu, Runguo Wei, and Erik Cambria. Self-supervised utterance order prediction for emotion recognition in conversations. *Neurocomput.*, 577(C), April 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.127370. URL https://doi.org/10.1016/j.neucom.2024.127370.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, 2017.

Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Dazhi Jiang, and Erik Cambria. Knowing what and why: Causal emotion entailment for emotion recognition in conversations. *Expert Syst. Appl.*, 274(C), May 2025. ISSN 0957-4174. doi: 10.1016/j.eswa.2025.126924. URL https://doi.org/10.1016/j.eswa.2025.126924.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. Emergence of separable manifolds in deep language representations, 2020. URL https://arxiv.org/abs/2006.01095.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

Maya Okawa, Bo Zhao, Eric J Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models. *arXiv preprint arXiv:2507.10599*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Flor Miriam Plaza-del Arco, Alba A Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 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)*, pp. 5696–5710, 2024.

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

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In *International Conference on Learning Representations*, volume 2026, pp. 49924–49947, 2026.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, et al. Less is more: Local intrinsic dimensions of contextual language models. *Advances in Neural Information Processing Systems*, 38:67795–67828, 2026.

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

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, 2024.

Agampreet Saini, Bayan Abusalameh, Rahul Kumar Singh, and Erik Cambria. Learning structured emotional dynamics in latent space via trajectory modeling and controlled transitions. In *The 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://openreview.net/forum?id=UMjSN84jUu.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*, 2026.

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

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

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

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The intrinsic dimension of prompts in internal representations of large language models, 2026. URL https://arxiv.org/abs/2501.10573.

Ningyu Xu, Qi Zhang, Chao Du, Qiang Luo, Xipeng Qiu, Xuanjing Huang, and Menghan Zhang. Revealing emergent human-like conceptual representations from language prediction. *Proceedings of the National Academy of Sciences*, 122(44):e2512514122, 2025.

Jingxiang Zhang and Lujia Zhong. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion, 2025. URL https://arxiv.org/abs/2510.04064.

Xulang Zhang, Rui Mao, and Erik Cambria. Senticvec: Toward robust and human-centric neurosymbolic sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4851–4863, 2024.

Yize Zhao and Christos Thrampoulidis. Geometry of semantics in next-token prediction: How optimization implicitly organizes linguistic representations. *arXiv preprint arXiv:2505.08348*, 2025.

APPENDIX

AI USE STATEMENT

In this work, we used generative AI tools for tasks under the recommended disclosure category, including editing the manuscript for readability and grammar correction, assisting with code debugging and identifying implementation errors in author-developed code, and minor visual cleanup of selected scientific figures. For the figures, the underlying content, architecture, and scientific design were created by the authors. Generative AI tools were not used for tasks requiring disclosure under the ICLR 2027 AI Policy for Authors, including developing the research ideas or conceptual framework, formulating hypotheses, designing the methodology or experiments, formulating mathematical claims, interpreting experimental results, or developing the scientific conclusions and claims. The research methodology, experimental design, analyses, and scientific decisions were developed by the authors. All AI-assisted text, code corrections, and figure modifications were reviewed by the authors. Code changes were tested and verified for correctness, and AI-assisted manuscript edits were checked against the underlying experiments and results. The authors take responsibility for the final content of this work, including the text, code, figures, claims, analyses, and scientific conclusions.

A REPRODUCIBILITY DETAILS

Datasets: We evaluate the representations across five emotion datasets, chosen to cover discrete emotion recognition, continuous affect, and conversational dynamics.

- **DailyDialog (Li et al., 2017):** A multi-turn conversational dataset with utterance-level emotion annotations. We use its predefined train, validation, and test splits for emotion classification and temporal forecasting.
- **EmoBank (Buechel & Hahn, 2017):** A sentence-level affective dataset annotated along three continuous dimensions: valence, arousal, and dominance (VAD). We retain its original train/dev/test split for continuous affect prediction.
- **GoEmotions (Demszky et al., 2020):** A fine-grained emotion dataset containing 27 emotion categories plus neutral. We remove examples marked as "very unclear" and retain an emotion label when selected by at least two annotators. The resulting data are split 80/10/10 into train, validation, and test sets.
- **MELD (Poria et al., 2019):** A multi-party conversational dataset with utterance-level emotion and speaker annotations. Its dialogue structure is used for classification and temporal analyses, with utterances ordered by dialogue and turn index.

- **IEMOCAP (Busso et al., 2008) :** A multi-session emotional dialogue dataset providing categorical emotion labels together with continuous VAD annotations. It is used as an external validation set, with the canonical mapping used for cross-dataset comparisons and ambiguous labels excluded from the categorical analysis.

Architectures: The primary experiments use BERT, with additional pretrained backbones used to assess whether the observed properties generalize across architectures.

- **BERT:** We use `bert-base-uncased` and extract the final-layer `[CLS]` representation (768 dimensions). A shared projection head maps this representation to a 128-dimensional latent space using a $768 \to 256 \to 128$ architecture with GELU activation and 0.1 dropout.
- **DistilBERT:** A compact distilled Transformer backbone evaluated under the same representation-learning and downstream evaluation framework.
- **MiniLM:** A lightweight Transformer backbone included to test the consistency of the representation-level analyses under a smaller pretrained architecture.
- **RoBERTa:** A pretrained Transformer backbone evaluated using the same latent dimensionality and downstream evaluation protocol as the other architectures.

For all representation types, the latent dimensionality is fixed to 128 to ensure that differences in downstream behavior are not attributable to dimensionality. The State, Continuous, and Dynamic representations therefore share the same backbone and projection configuration, while the Dynamic representation additionally processes sequential utterance representations using a single-layer, unidirectional GRU.

Training and preprocessing: Models are trained using AdamW with a learning rate of 2×10^{-5}, weight decay of 0.01, and six training epochs. The encoder is frozen for the first two epochs and jointly fine-tuned thereafter. We use linear warm-up with a 0.1 warm-up ratio and gradient clipping at 1.0. The Dynamic representation uses a batch size of 2 dialogues, while the other training procedures use a batch size of 16.

Seeds and evaluation protocol: The primary BERT experiments use seeds 42, 256, and 512. Analyses for which three runs are available report the corresponding three-seed results. For uniformity and fairness in cross-architecture comparison tables, results are reported using the common seed-42 evaluation protocol, since seed 42 is available for all evaluated architectures. The fixed seeds are also used for downstream evaluation and statistical procedures as specified by each experiment.

Downstream evaluation: All downstream probes use fixed hyperparameters across representations. Classification uses logistic regression with $C = 1.0$, the LBFGS solver, and a maximum of 2000 iterations, while regression uses ridge regression with $\alpha = 1.0$. Latent representations are standardized using statistics fitted on the training set only. Classification is evaluated primarily using macro-F1, with micro-F1 and accuracy as auxiliary measures. Regression uses R^2, with MAE and RMSE as auxiliary measures. Forecasting uses macro-F1 and accuracy.

Statistical analysis: Confidence intervals are estimated using 1,000 bootstrap iterations with a 95% confidence level and bootstrap seed 42. Geometry and nonlinearity analyses use fixed subsampling limits and neighborhood sizes across representations to maintain comparable computational budgets. In particular, geometry measures use $k = 10$ nearest neighbors and at most 4,000 points, while the nonlinearity analyses use fixed training and manifold subsampling limits.

Each analysis is run on the datasets whose annotations support it i.e. dataset–analysis combinations that are not applicable are omitted rather than reporting them as failures.

B Derivations

B.1 Derivation of the Representation Complementarity Score

Let $z^{(1)}, \ldots, z^{(K)}$ denote K representation views of the same instances, and let $\Phi(S)$ denote the best population macro-F1 attainable by the fixed linear probe family using the representations indexed by S. We seek a scalar measure of the label-relevant information that becomes linearly accessible only when multiple views are considered jointly.

Table 4: **Abbreviations used in the paper.**

Abbreviation	Meaning
S	State
C	Continuous
D	Dynamic
GoEmo	GoEmotions
MELD	Multimodal Emotion Lines Dataset
DailyD	Daily Dialog
IEMO	Interactive Emotional Dyadic Motion Capture

Joint representation: We form the joint representation by concatenation,

$$z^{(1)} \oplus \cdots \oplus z^{(K)}. \tag{1}$$

A linear probe on the concatenated representation can reproduce any single-view linear probe. In particular,

$$w^\top \left(z^{(1)} \oplus \cdots \oplus z^{(K)} \right) = \sum_{i=1}^{K} w_i^\top z^{(i)}. \tag{2}$$

Setting the weights associated with all but one view to zero recovers the corresponding single-view probe. Hence,

$$\Phi\left(z^{(1)} \oplus \cdots \oplus z^{(K)} \right) \geq \max_i \Phi\left(z^{(i)} \right). \tag{3}$$

Concatenation therefore provides the lossless linear union of the available representation views without imposing additional pooling or weight-sharing constraints.

Redundancy baseline: The joint representation is compared against the strongest individual representation,

$$\Phi_{\max} = \max_i \Phi\left(z^{(i)} \right). \tag{4}$$

This baseline is zero whenever the joint representation provides no additional linearly recoverable information beyond the best individual view. In particular, if all representation views are linearly redundant, concatenation cannot improve upon the strongest view. Using the mean of the individual scores would instead produce a positive gain whenever one representation is stronger than another, even if the weaker representation contributes no complementary information. The maximum therefore provides the appropriate redundancy baseline.

Gain operator: Let

$$u = \Phi\left(z^{(1)} \oplus \cdots \oplus z^{(K)} \right), \qquad a = \max_i \Phi\left(z^{(i)} \right). \tag{5}$$

The complementarity measure should quantify the additional performance of the joint representation relative to the strongest individual view. We therefore use the additive gain $u - a$. This difference is invariant to a common additive shift in both scores,

$$(u + \delta) - (a + \delta) = u - a, \tag{6}$$

whereas ratio-based alternatives depend on the absolute score level and can become unstable when a is small.

The resulting population definition is

$$\boxed{\mathrm{RCS} = \Phi\left(z^{(1)} \oplus \cdots \oplus z^{(K)} \right) - \max_i \Phi\left(z^{(i)} \right)}. \tag{7}$$

Because the joint probe can reproduce the strongest individual probe, the population quantity satisfies

$$\mathrm{RCS} \geq 0. \tag{8}$$

Consequently, negative empirical values should be interpreted as finite-sample or regularization effects rather than as evidence of negative complementarity.

Empirical estimator: In experiments, RCS is estimated using the same train/test split and identical linear-probe protocol for the individual and joint representations:

$$\widehat{\mathrm{RCS}} = F_1^M\left(z^{(1)} \oplus \cdots \oplus z^{(K)}\right) - \max_i F_1^M\left(z^{(i)}\right). \tag{9}$$

All representation views are aligned by instance before concatenation so that the comparison measures complementary information rather than differences in sample composition.

Finite-sample capacity control: The population capacity-neutrality property does not hold exactly for a finite-sample regularized probe. Concatenating multiple representation views increases the number of probe parameters and can therefore produce a spurious performance gain even when the additional dimensions are label-independent. We therefore introduce a dimension-matched control to estimate the gain attributable solely to the increased probe width. Let i^* denote the strongest individual representation and let η denote a label-independent control block with dimensionality matched to the additional representation dimensions introduced by concatenation. The control block is paired with the same instances and evaluated using the same train/test split, probe architecture, regularization, and optimization procedure as the corresponding joint representation. Because η carries no label-dependent information, any improvement obtained from $z^{(i^*)} \oplus \eta$ over $z^{(i^*)}$ estimates the finite-sample capacity effect associated with increasing the probe width. The resulting corrected complementarity score is

$$\mathrm{RCS}^* = \left[F_1^M\left(z^{(1)} \oplus \cdots \oplus z^{(K)}\right) - F_1^M\left(z^{(i^*)}\right)\right] - \left[F_1^M\left(z^{(i^*)} \oplus \eta\right) - F_1^M\left(z^{(i^*)}\right)\right]. \tag{10}$$

The second bracket therefore serves as a width-matched null estimate, separating performance gains attributable to finite-sample probe capacity from gains attributable to complementary information in the additional representation views.

B.2 Derivation of the State–Dynamics Index

For each emotion class c, we seek a scalar measure of whether that emotion is more strongly recoverable from the Dynamic representation or from the State representation. The index provides a per-emotion coordinate on a state–process spectrum: positive values indicate greater recoverability from the Dynamic representation, while negative values indicate greater recoverability from the State representation.

Design requirements: Let R_c^{dyn} and R_c^{stat} denote the per-class recall obtained from the Dynamic and State representations, respectively. The index should (i) be defined separately for each emotion, (ii) measure recoverability of the target class directly, (iii) be independent of class prevalence for a fixed decision rule, (iv) be bounded and sign-interpretable, and (v) reverse sign when the two representations are exchanged.

Choice of per-class functional: For a fixed class c, standard per-class candidates include precision, recall, and F_1. Writing the corresponding confusion counts as (TP_c, FN_c, FP_c),

$$R_c = \frac{TP_c}{TP_c + FN_c}, \qquad P_c = \frac{TP_c}{TP_c + FP_c}, \qquad F_{1,c} = \frac{2P_cR_c}{P_c + R_c}. \tag{12}$$

Recall is used because it depends only on instances belonging to the target class:

$$R_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbb{1}[\hat{y}_i = c]. \tag{13}$$

Unlike precision and F_1, recall does not depend on false positives from other emotion classes. It therefore provides a direct measure of how reliably instances of emotion c are recovered. In population form,

$$R_c = \Pr(\hat{Y} = c \mid Y = c), \tag{14}$$

which is independent of the class prior for a fixed decision rule.

Representation-specific recalls: Let z_t^{stat} and z_t^{dyn} denote the State and Dynamic representations associated with an utterance at turn t. Their corresponding class-specific recalls are

$$R_c^{\text{stat}} = \Pr(\hat{Y}_c^{\text{stat}} = c \mid Y = c), \qquad R_c^{\text{dyn}} = \Pr(\hat{Y}_c^{\text{dyn}} = c \mid Y = c). \tag{15}$$

The State representation provides the state-oriented reference, whereas the Dynamic representation incorporates sequential information and therefore provides the process-oriented reference. SDI compares these two representations directly rather than attempting to isolate a causal contribution of conversational context.

Choice of contrast: Consider the candidate contrasts

$$D = u - a, \qquad Q = \frac{u}{a}, \qquad L = \log \frac{u}{a}, \qquad G = \frac{u - a}{1 - a}, \tag{16}$$

where u and a denote the recalls under the two representations. Ratio-based forms are undefined when $a = 0$ and do not provide a bounded signed difference. The ceiling-normalized gain G can also become unstable as a approaches one. In contrast, subtraction is defined for all recall values, is antisymmetric,

$$D(a, u) = -D(u, a), \tag{17}$$

and, since $u, a \in [0, 1]$,

$$-1 \leq D(u, a) \leq 1. \tag{18}$$

Subtraction therefore provides a bounded and directly interpretable contrast between the Dynamic and State representations.

The resulting State–Dynamics Index is

$$\boxed{\text{SDI}(c) = R_c^{\text{dyn}} - R_c^{\text{stat}}}. \tag{19}$$

A positive value indicates that emotion c is more reliably recovered from the Dynamic representation, whereas a negative value indicates greater recoverability from the State representation. A value near zero indicates similar class-specific recall in the two representations.

The extreme values have direct interpretations:

$$\text{SDI}(c) = 1 \iff \left(R_c^{\text{dyn}}, R_c^{\text{stat}}\right) = (1, 0), \tag{20}$$

and

$$\text{SDI}(c) = -1 \iff \left(R_c^{\text{dyn}}, R_c^{\text{stat}}\right) = (0, 1). \tag{21}$$

Empirical estimator: For each emotion c, the empirical per-class recall is

$$\widehat{R}_c = \frac{1}{N_c} \sum_{i: y_i = c} \mathbb{1}[\hat{y}_i = c]. \tag{22}$$

The empirical State–Dynamics Index is consequently

$$\boxed{\widehat{\text{SDI}}(c) = \widehat{R}_c^{\text{dyn}} - \widehat{R}_c^{\text{stat}}}. \tag{23}$$

We report the index separately for each emotion to preserve heterogeneity in state–process organization across affective categories and datasets. Because recall estimates for rare emotions can have substantial sampling variance, confidence intervals are obtained by bootstrap resampling within each emotion class.

B.3 Derivation of the Information Preservation Index

The Information Preservation Index (IPI) summarizes how much affective information remains linearly accessible from a representation across the affective facets supported by each dataset. Let $\mathcal{F}_D$ denote the set of evaluated affective facets available for dataset D. For each facet $f \in \mathcal{F}_D$, we define a bounded evaluation score $f(r, D)$ and average the scores across the available facets.

Design requirements: The facet-level scores should (i) lie in a common bounded range, (ii) equal one under perfect prediction, and (iii) preserve the relative ordering of representations within a fixed dataset and facet. Because categorical and continuous affect are evaluated using different native metrics, their scores are mapped to a common $[0, 1]$ range before aggregation.

Categorical affective information: For categorical emotion prediction, let $F_1^M(r, D)$ denote the macro-F1 obtained from the linear probe. The implementation uses the raw macro-F1 directly, without chance correction or baseline subtraction:

$$f_{\text{cat}}(r, D) = F_1^M(r, D). \tag{24}$$

Since macro-F1 lies in $[0, 1]$, the categorical score is already bounded and satisfies

$$F_1^M(r, D) = 1 \quad \Longrightarrow \quad f_{\text{cat}}(r, D) = 1. \tag{25}$$

Thus, higher f_{cat} indicates greater categorical affective information accessible through the linear probe. Because the score is not chance-corrected, it should be interpreted as a task-level accessibility measure rather than as an absolute estimate of information above chance.

Continuous affective information: For continuous affect prediction, let $R^2(r, D)$ denote the coefficient of determination obtained from the linear probe. Because R^2 can take negative values, we truncate non-positive values at zero:

$$f_{\text{cont}}(r, D) = \max\left(0, R^2(r, D)\right). \tag{26}$$

This transformation places the continuous score on the same bounded scale as macro-F1. It also treats non-positive R^2 as zero accessible continuous affective information under the evaluated linear probe, while retaining positive R^2 values unchanged. Perfect prediction satisfies

$$R^2(r, D) = 1 \quad \Longrightarrow \quad f_{\text{cont}}(r, D) = 1. \tag{27}$$

Aggregation across affective facets: The IPI averages the facet-level scores available for each dataset, giving equal weight to each evaluated facet:

$$\boxed{\text{IPI}(r, D) = \frac{1}{|\mathcal{F}_D|} \sum_{f \in \mathcal{F}_D} f(r, D)}. \tag{28}$$

For datasets with both categorical and continuous affective evaluations, this becomes

$$\text{IPI}(r, D) = \frac{1}{2}\left[f_{\text{cat}}(r, D) + f_{\text{cont}}(r, D)\right]. \tag{29}$$

For datasets supporting only one affective facet, IPI reduces to that facet-level score. Thus, the index does not introduce a missing-facet value or impute an unevaluated modality.

Because every facet-level score lies in $[0, 1]$, the aggregate index is also bounded:

$$0 \leq \text{IPI}(r, D) \leq 1. \tag{30}$$

Empirical estimation: In the reported experiments, the facet-level scores are computed using the same representation splits, linear-probe settings, and evaluation protocols as the corresponding representation–task experiments. The empirical IPI is therefore

$$\widehat{\text{IPI}}(r, D) = \frac{1}{|\mathcal{F}_D|} \sum_{f \in \mathcal{F}_D} \widehat{f}(r, D), \tag{31}$$

where $\widehat{f}_{\text{cat}}$ is the observed raw macro-F1 and $\widehat{f}_{\text{cont}} = \max(0, \widehat{R}^2)$.

The facet-level results are retained in the complete representation–task evaluation so that the aggregate IPI can be interpreted together with the individual categorical and continuous affective scores.

C Complete Nonlinearity and Probe-Ladder Results

We evaluate four probe families: a linear probe, RBF-SVM, shallow MLP (MLP-S), and deep MLP (MLP-D). The nonlinearity gap Δ_{NL} is defined as the performance difference between the strongest nonlinear probe and the linear probe under the implemented evaluation. All encoder backbones are evaluated at seed 42 for a controlled cross-backbone comparison, as shown in Table 5. Across encoder backbones, the mean gap remains small or negative, with occasional positive deviations in individual settings. Overall, the probe ladder provides no systematic evidence that increasing decoder complexity improves access to affective information.

D Detailed Geometric Analysis

All representations occupy a 128-dimensional ambient space. Before intrinsic-dimensionality estimation, duplicate latent vectors are removed to mitigate bias from repeated points produced by the deterministic encoder. Across backbones, the resulting spaces remain substantially lower-dimensional: mean intrinsic dimensionality is 9.691 for BERT, 9.388 for DistilBERT, 6.186 for MiniLM, and 8.745 for RoBERTa, with maximum observed values of 15.756, 14.710, 8.983, and 14.441, respectively. The largest observed value therefore uses only 12.3% of the ambient dimensionality.

Table 6 reports the complete geometric analysis across representation objectives, datasets, and encoder backbones. The updated intrinsic-dimensionality estimates remain substantially below the 128-dimensional ambient space across all settings. Geodesic F1 is generally comparable to Euclidean kNN F1, with the IEMOCAP geodesic-edge statistic remaining positive across all four backbones, indicating a modest but consistent benefit from manifold-aware neighborhoods in this setting.

E Detailed Temporal Structure Across Encoder Backbones

We evaluate temporal structure using chronological-direction accuracy, within-dialogue shuffled controls, reversed trajectories, directional coherence, latent displacement–position association, and next-step forecasting. Across all four backbones, the clearest chronological signal occurs on DailyDialog, where accuracy remains approximately 0.56–0.60 across representation objectives while shuffled controls remain near chance. MELD and IEMOCAP show substantially weaker directional separation. Directional coherence is consistently negative, while displacement–position associations and forecasting gains vary across datasets and representations. The complete results are reported in Table 7.

Furthermore the reversed-trajectory condition is a diagnostic rather than an independent temporal null. Because trajectory reversal changes each displacement from Δz_t to $-\Delta z_t$, the directional probe and coherence statistic preserve their attainable structure under reversal. Thus, agreement between chronological and reversed conditions should not be interpreted as evidence for an absolute past-to-future orientation.

E.1 Transition Structure and Antecedent–Emotion Analysis

We compare adjacent, non-adjacent within-dialogue, and cross-dialogue latent pairs, together with forward and backward displacement predictability. Adjacent pairs are generally closer than non-adjacent and cross-dialogue pairs for Continuous and Dynamic representations, with the clearest local contraction on MELD and IEMOCAP. Directional predictability is weak on MELD and DailyDialog and strongest on IEMOCAP, while forward–backward asymmetries remain small (0.009–0.017). Antecedent–emotion gains over the matched control are likewise modest, providing evidence for local transition structure without a strong universal directional encoding.

E.2 Transition and Temporal Diagnostics

We use complementary diagnostics to distinguish local latent structure from directional and antecedent-specific effects. Figure 9 summarizes directional predictability and distance structure,

Table 5: **Complete probe-ladder results across encoder backbones.** Scores for linear, RBF-SVM, shallow MLP (MLP-S), and deep MLP (MLP-D) probes. Δ_{NL} denotes the difference between the strongest nonlinear probe and the linear probe. All backbones are evaluated at seed 42 for a controlled like-for-like comparison. Within each representation–dataset combination, the highest probe score is highlighted in light green. For Δ_{NL}, the value closest to zero is highlighted, corresponding to the smallest nonlinear advantage or disadvantage relative to the linear probe.

Backbone	Rep.	Dataset	Linear	RBF	MLP-S	MLP-D	Δ_{NL}
BERT	S	GoEmo	**0.444**	0.411	0.440	0.437	**-0.003**
	S	MELD	0.354	0.319	0.353	**0.375**	**+0.022**
	S	DailyD	**0.266**	0.202	0.251	0.212	-0.015
	S	IEMO	**0.352**	0.256	0.317	0.312	-0.035
	C	GoEmo	**0.292**	0.155	0.249	0.268	**-0.024**
	C	MELD	**0.393**	0.352	0.389	0.372	-0.004
	C	DailyD	**0.340**	0.212	0.263	0.259	-0.078
	C	IEMO	**0.412**	0.284	**0.412**	0.410	**-0.001**
	D	GoEmo	**0.216**	0.064	0.121	0.123	-0.093
	D	MELD	0.410	0.397	0.393	**0.409**	**-0.001**
	D	DailyD	**0.284**	0.266	0.253	0.274	-0.011
	D	IEMO	0.317	0.248	0.274	**0.320**	**+0.003**
DistilBERT	S	GoEmo	**0.457**	0.414	0.443	0.447	-0.010
	S	MELD	**0.363**	0.302	0.337	0.329	-0.026
	S	DailyD	**0.248**	0.200	0.218	0.212	-0.030
	S	IEMO	0.288	0.246	0.281	**0.339**	**+0.052**
	C	GoEmo	**0.313**	0.135	0.246	0.201	-0.067
	C	MELD	0.375	0.331	0.325	**0.380**	**+0.004**
	C	DailyD	**0.294**	0.205	0.260	0.273	-0.021
	C	IEMO	0.387	0.280	0.357	**0.411**	**+0.023**
	D	GoEmo	**0.225**	0.085	0.108	0.145	-0.081
	D	MELD	**0.409**	0.380	0.378	0.389	-0.020
	D	DailyD	**0.270**	0.265	0.251	0.259	-0.005
	D	IEMO	**0.319**	0.222	0.252	0.270	-0.049
MiniLM	S	GoEmo	**0.400**	0.371	0.363	0.373	-0.027
	S	MELD	**0.324**	0.282	0.305	0.319	-0.004
	S	DailyD	0.221	0.196	**0.223**	0.203	**+0.002**
	S	IEMO	**0.266**	0.204	0.220	0.217	-0.046
	C	GoEmo	**0.192**	0.068	0.099	0.106	-0.087
	C	MELD	**0.344**	0.314	0.323	0.334	-0.011
	C	DailyD	**0.260**	0.203	0.205	0.215	-0.045
	C	IEMO	**0.288**	0.183	0.219	0.282	**-0.006**
	D	GoEmo	**0.095**	0.032	0.033	0.053	-0.042
	D	MELD	**0.360**	0.337	0.318	0.345	-0.015
	D	DailyD	**0.243**	0.199	0.198	0.220	-0.023
	D	IEMO	**0.217**	0.181	0.179	0.196	-0.021
RoBERTa	S	GoEmo	**0.464**	0.428	0.456	0.443	-0.009
	S	MELD	0.362	0.341	0.357	**0.374**	**+0.011**
	S	DailyD	0.273	0.210	**0.282**	0.241	**+0.009**
	S	IEMO	0.307	0.258	0.308	**0.326**	**+0.020**
	C	GoEmo	**0.324**	0.157	0.253	0.219	-0.070
	C	MELD	**0.429**	0.358	0.395	0.354	-0.034
	C	DailyD	**0.325**	0.244	0.235	0.266	-0.059
	C	IEMO	0.351	0.262	0.351	**0.381**	**+0.030**
	D	GoEmo	**0.160**	0.040	0.097	0.054	-0.063
	D	MELD	**0.413**	0.393	0.401	0.397	-0.012
	D	DailyD	**0.269**	0.242	0.252	0.249	-0.017
	D	IEMO	**0.294**	0.220	0.238	0.242	-0.052

while Figure 10 shows antecedent gains for BERT and Figure 11 shows latent drift across all models.

Table 6: **Complete geometric analysis across encoder backbones.** Intrinsic dimensionality (ID), ID ratio, local linearity, Euclidean kNN F1, and geodesic F1 for each representation and dataset. ID Ratio = ID / 128. All backbones are evaluated at seed 42. Best values across encoder backbones for each representation–dataset combination are highlighted in light blue.

Backbone	Representation	Dataset	ID	ID Ratio	Local Lin.	kNN F1	Geo. F1
BERT	S	GoEmotions	7.147	0.0558	0.607	0.431	0.355
	S	MELD	7.453	0.0582	0.628	0.344	0.279
	S	DailyDialog	7.216	0.0564	0.605	0.210	0.282
	S	IEMOCAP	5.145	0.0402	0.638	0.324	0.382
	C	GoEmotions	15.756	0.1231	0.430	0.151	0.134
	C	MELD	13.407	0.1047	0.489	0.362	0.307
	C	DailyDialog	13.625	0.1064	0.437	0.235	0.260
	C	IEMOCAP	8.122	0.0635	0.486	0.314	0.390
	D	GoEmotions	10.448	0.0816	0.575	0.118	0.110
	D	MELD	8.919	0.0697	0.642	0.404	0.360
	D	DailyDialog	10.692	0.0835	0.545	0.272	0.272
	D	IEMOCAP	8.365	0.0654	0.617	0.272	0.346
DistilBERT	S	GoEmotions	7.678	0.0600	0.596	0.421	0.361
	S	MELD	7.762	0.0606	0.617	0.321	0.295
	S	DailyDialog	7.118	0.0556	0.595	0.205	0.289
	S	IEMOCAP	5.112	0.0399	0.634	0.285	0.365
	C	GoEmotions	14.710	0.1149	0.459	0.155	0.131
	C	MELD	12.185	0.0952	0.517	0.327	0.285
	C	DailyDialog	12.482	0.0975	0.463	0.221	0.278
	C	IEMOCAP	7.822	0.0611	0.513	0.263	0.367
	D	GoEmotions	10.190	0.0796	0.588	0.120	0.105
	D	MELD	8.684	0.0678	0.646	0.394	0.344
	D	DailyDialog	10.222	0.0799	0.573	0.265	0.271
	D	IEMOCAP	8.689	0.0679	0.619	0.261	0.321
MiniLM	S	GoEmotions	5.432	0.0424	**0.683**	0.376	0.324
	S	MELD	5.641	0.0441	**0.704**	0.317	0.264
	S	DailyDialog	5.267	0.0411	**0.684**	0.210	0.282
	S	IEMOCAP	4.805	0.0375	**0.719**	0.265	0.319
	C	GoEmotions	8.519	0.0666	**0.701**	0.095	0.098
	C	MELD	7.961	0.0622	**0.719**	0.312	0.269
	C	DailyDialog	8.983	0.0702	**0.637**	0.212	0.246
	C	IEMOCAP	7.123	0.0557	**0.688**	0.234	0.314
	D	GoEmotions	4.994	0.0390	**0.866**	0.071	0.074
	D	MELD	5.046	0.0394	**0.821**	0.342	0.299
	D	DailyDialog	5.645	0.0441	**0.761**	0.228	0.239
	D	IEMOCAP	4.821	0.0377	**0.846**	0.203	0.290
RoBERTa	S	GoEmotions	7.216	0.0564	0.629	**0.434**	**0.370**
	S	MELD	6.934	0.0542	0.636	**0.341**	**0.296**
	S	DailyDialog	6.495	0.0507	0.624	**0.228**	**0.288**
	S	IEMOCAP	4.856	0.0379	**0.644**	**0.266**	**0.342**
	C	GoEmotions	14.441	0.1128	0.481	**0.156**	**0.158**
	C	MELD	12.640	0.0988	0.536	**0.353**	**0.320**
	C	DailyDialog	12.370	0.0966	0.473	**0.237**	**0.271**
	C	IEMOCAP	7.595	0.0593	0.530	**0.341**	**0.372**
	D	GoEmotions	8.173	0.0639	0.651	**0.103**	**0.096**
	D	MELD	7.665	0.0599	0.698	**0.399**	**0.345**
	D	DailyDialog	9.246	0.0722	0.602	**0.252**	**0.271**
	D	IEMOCAP	7.310	0.0571	0.673	**0.288**	**0.328**

Directional predictability is strongest on IEMOCAP, while MELD and DailyDialog remain close to their within-setting baselines. Forward-backward asymmetries are small, indicating that predictive transition structure is largely symmetric rather than strongly directional. Adjacent utterances are generally closer than non-adjacent same-dialogue pairs, although this ordering is not universal.

Table 7: **Complete temporal structure analysis across encoder backbones.** Chron. denotes chronological-direction accuracy. Shuffle denotes the within-dialogue shuffled control. Rev. denotes the reversed-trajectory condition. Δ_{chance} denotes chronological accuracy above chance. Coherence measures local directional alignment. ρ measures the association between latent displacement and chronological position. Forecast denotes next-step forecasting performance, and Δ_{forecast} denotes the difference between real and shuffled forecasting performance. Best values across encoder backbones for each representation–dataset combination are highlighted in light pink. For Shuffle and Rev., values closest to chance are highlighted.

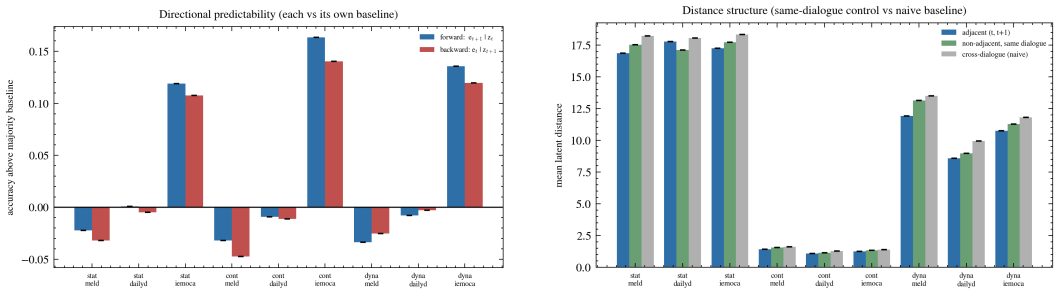
Backbone	Representation	Dataset	Chron.	Shuffle	Rev.	Δ_{chance}	Coherence	ρ	Forecast	Δ_{forecast}
	S	MELD	0.5022	**0.5023**	**0.5011**	0.0022	-0.4601	0.0540	0.1222	0.0098
	S	DailyDialog	0.5825	**0.5032**	0.5836	0.0825	-0.5440	0.0367	0.1426	0.0075
	S	IEMOCAP	0.5031	**0.5009**	**0.5044**	0.0031	-0.4934	0.0393	0.2423	0.0503
	C	MELD	0.5068	**0.4906**	**0.5048**	0.0068	-0.4109	0.0857	0.1524	0.0283
BERT	C	DailyDialog	**0.5992**	**0.4994**	0.5988	**0.0992**	-0.4411	**0.0834**	**0.1616**	0.0011
	C	IEMOCAP	**0.5160**	**0.4935**	0.5159	**0.0160**	-0.4345	0.0386	**0.3288**	**0.1039**
	D	MELD	0.5035	**0.5009**	**0.5002**	0.0035	-0.3987	0.0880	0.1346	0.0188
	D	DailyDialog	0.5856	0.5081	0.5848	0.0856	-0.4501	0.0685	0.1436	0.0118
	D	IEMOCAP	0.5104	**0.4982**	0.5099	0.0104	-0.4487	0.0308	0.2802	0.0735
	S	MELD	0.4881	0.5063	0.4853	-0.0119	-0.4603	0.0443	0.1329	0.0109
	S	DailyDialog	**0.5857**	0.4963	0.5862	**0.0857**	-0.5421	0.0410	0.1407	0.0048
	S	IEMOCAP	**0.5118**	0.4919	0.5122	**0.0118**	-0.4881	**0.0476**	0.2340	0.0430
	C	MELD	0.5119	0.5147	0.5133	0.0119	-0.3972	0.0699	0.1401	0.0122
DistilBERT	C	DailyDialog	0.5843	0.5007	0.5853	0.0843	-0.4309	0.0736	0.1580	0.0065
	C	IEMOCAP	0.5100	0.4979	**0.5070**	0.0100	-0.4228	0.0396	0.2881	0.0636
	D	MELD	**0.5077**	0.4839	0.5063	**0.0077**	-0.4070	0.0864	0.1387	**0.0344**
	D	DailyDialog	0.5833	**0.5047**	0.5823	0.0833	-0.4637	**0.0702**	0.1349	-0.0013
	D	IEMOCAP	**0.5115**	0.4946	0.5096	**0.0115**	-0.4474	0.0370	0.2462	0.0378
	S	MELD	0.5244	0.4895	0.5216	0.0244	**-0.4458**	0.0342	0.1047	0.0093
	S	DailyDialog	0.5823	0.5061	**0.5813**	0.0823	**-0.5314**	0.0350	0.1361	**-0.0000**
	S	IEMOCAP	0.5051	0.4961	0.5047	0.0051	**-0.4780**	0.0293	0.1871	0.0223
	C	MELD	0.5105	0.5105	0.5189	0.0105	**-0.3735**	0.0777	0.1095	0.0004
MiniLM	C	DailyDialog	0.5877	0.4993	**0.5848**	0.0877	**-0.3985**	0.0737	0.1511	0.0105
	C	IEMOCAP	0.5156	0.5036	0.5111	0.0156	**-0.3974**	0.0271	0.2270	0.0321
	D	MELD	0.4895	0.5077	0.4881	-0.0105	**-0.3723**	0.0865	0.0925	0.0013
	D	DailyDialog	0.5602	0.4865	**0.5607**	0.0602	-0.4101	0.0566	0.1306	-0.0003
	D	IEMOCAP	0.5017	0.5021	**0.5017**	0.0017	-0.4045	0.0096	0.1456	0.0109
	S	MELD	**0.5398**	0.4867	0.5356	**0.0398**	-0.4587	0.0458	0.1297	0.0002
	S	DailyDialog	0.5843	0.5012	0.5833	0.0843	-0.5409	**0.0484**	0.1397	0.0077
	S	IEMOCAP	0.5085	0.4961	0.5081	0.0085	-0.4988	0.0422	0.2177	0.0327
	C	MELD	**0.5230**	0.5091	0.5203	**0.0230**	-0.4048	0.0821	0.1616	**0.0286**
RoBERTa	C	DailyDialog	0.5848	0.5057	**0.5848**	**0.0848**	-0.4531	0.0726	0.1533	0.0067
	C	IEMOCAP	0.5107	0.4994	0.5133	0.0107	-0.4406	**0.0419**	**0.2897**	0.0684
	D	MELD	0.4979	0.4839	0.4937	-0.0021	-0.3765	**0.0904**	0.1163	-0.0055
	D	DailyDialog	0.5656	**0.4958**	0.5646	0.0656	-0.4298	0.0522	**0.1516**	0.0100
	D	IEMOCAP	0.4946	0.5032	0.4994	-0.0054	-0.4258	0.0200	0.2164	0.0254

Table 8: **Latent transition structure across representations and datasets.** Adj. and Non-adj. denote within-dialogue adjacent and non-adjacent latent distances. Cross denotes cross-dialogue distance. Fwd. and Bwd. measure directional predictability, and Asym. is their difference.

Rep.	Dataset	Adj.	Non-adj.	Cross	Fwd.	Bwd.	Asym.
S	MELD	16.674	17.355	17.897	-0.022	-0.032	0.010
S	DailyDialog	**17.613**	**16.992**	**17.931**	0.001	-0.002	**0.003**
S	IEMOCAP	**16.970**	17.457	18.038	**0.086**	**0.077**	0.009
C	MELD	**1.404**	**1.551**	**1.610**	-0.043	-0.039	-0.004
C	DailyDialog	1.061	1.140	1.265	-0.008	-0.007	**0.000**
C	IEMOCAP	1.239	1.319	1.373	**0.122**	**0.105**	0.017
D	MELD	12.206	13.442	13.781	-0.016	-0.016	**−0.000**
D	DailyDialog	**8.933**	**9.335**	**10.379**	-0.003	-0.002	-0.002
D	IEMOCAP	11.128	11.677	12.113	**0.097**	**0.080**	0.017

Cross-dialogue distances are consistently larger but provide a weaker control because they do not preserve dialogue context.

Antecedent information provides its clearest incremental gain on IEMOCAP, with gains of approximately $+0.043$, $+0.050$, and $+0.038$ for State, Continuous, and Dynamic, respectively. On MELD, negative raw gains largely disappear under the shuffled-antecedent control, indicating that they



(a) Directional predictability. (b) Distance structure.

Figure 9: **Transition diagnostics.** Directional predictability and latent-distance structure across representation objectives and dialogue datasets.

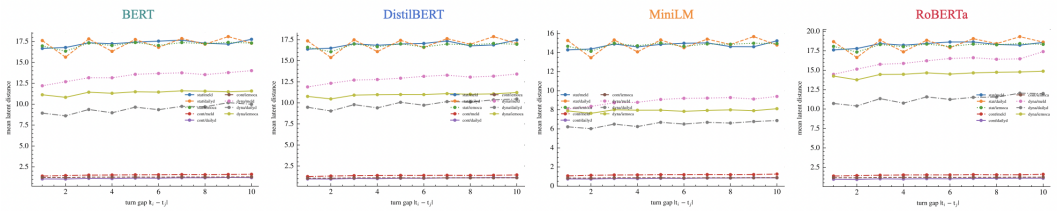


Figure 11: Latent drift across all models.

should not be interpreted as evidence against antecedent information. Latent distance also varies systematically with conversational separation, supporting structured temporal organization. However, raw distances are not directly comparable across representations because their latent scales differ. Together, these diagnostics support local and temporally structured latent organization, with modest antecedent contributions and limited evidence for directional asymmetry. The results therefore favor an interpretation in terms of persistence and structured temporal variation rather than a universal directed mechanism.

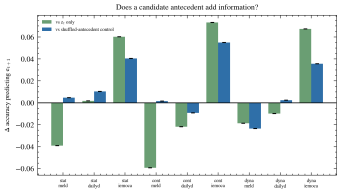


Figure 10: Transition antecedent gain for BERT.

E.3 Dimensionality and Variance Compression

We vary the retained dimensionality over $k \in \{2, 4, 8, 16, 32, 64, 128\}$ to characterize how intrinsic dimensionality and variance evolve with representation size. Figure 12 and Figure13 reports the corresponding trends for all models. Intrinsic dimensionality rises rapidly at low k and then saturates. At $k = 128$, State reaches approximately 7.2–7.5 dimensions on MELD and DailyDialog and 4.5 on IEMOCAP. Dynamic reaches 8.9–10.7 dimensions on MELD and DailyDialog, while Continuous reaches 13.4 and 13.6, respectively. Variance is similarly concentrated in a small number of dimensions. State exceeds 94% cumulative variance by $k = 8$ on MELD and DailyDialog, while Dynamic exceeds 98% on MELD by $k = 4$. Continuous is more distributed, reaching approximately 99% on MELD at $k = 32$ and 98.3% on DailyDialog. Thus, both intrinsic dimensionality and variance stabilize substantially before the full 128-dimensional representation is retained.

E.4 Detailed Temporal Direction Analysis

We evaluate temporal direction using chronological-direction accuracy relative to a within-dialogue-shuffled control. Δ denotes the paired difference between chronological and shuffled accuracy. Confidence intervals are computed over the evaluated representation types. For the controlled cross-backbone comparison, all results are reported using seed 42.

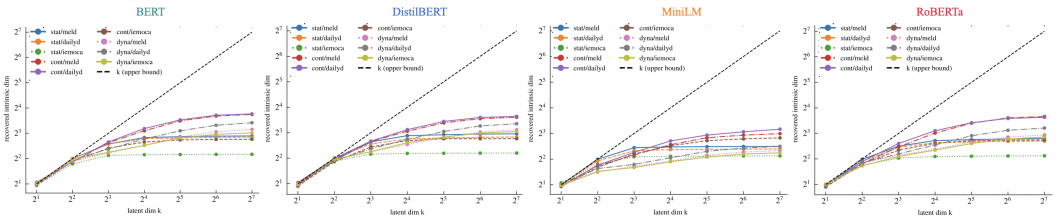


Figure 12: Intrinsic dimensionality for all models. Intrinsic dimension saturates once k exceeds it.

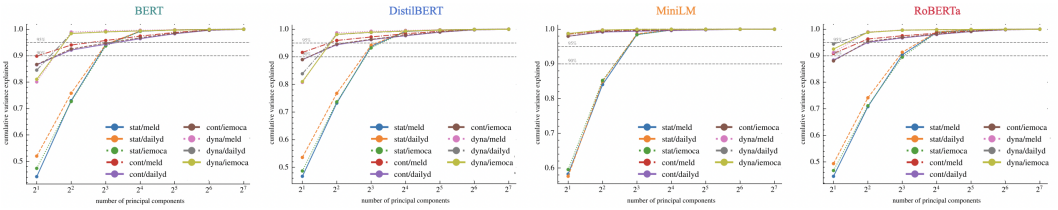


Figure 13: Cumulative variance across all models. This shows how many coordinates does the space actually use?

Across datasets, the temporal-direction effect is concentrated on DailyDialog. All four backbones show positive Δ values whose 95% CIs exclude zero, whereas MELD shows no significant effect and IEMOCAP shows only a single uncorrected significant cell (DistilBERT, $p = 0.019$), which does not survive correction across the 12 backbone–dataset tests. The consistency across four independently trained encoder architectures supports the reproducibility of the DailyDialog effect. We do not interpret this result as evidence that affect is temporally organized. DailyDialog contains scripted dialogues with strong positional conventions, such as greetings, exchanges, and closings, so recoverable direction may reflect discourse position or lexical cues rather than emotional dynamics. Its absence on MELD and IEMOCAP is consistent with this interpretation.

F Representation-Level Metrics

We further quantify properties of the learned representations that are not captured by downstream task performance or geometric measures alone. These metrics characterize state–process organization, information preservation, representational agreement, and complementarity across the three representation objectives.

F.1 Complete Representation–Task Evaluation

Table 10 reports the complete downstream evaluation across the four encoder backbones and three representation objectives. The evaluation covers five complementary tasks that probe different computational uses of the learned representations. *Classification* tests whether the representation preserves information needed to identify the affective category of an utterance. *Regression* evaluates whether continuous affective properties can be recovered from the representation, using R^2 as the primary score. *Forecasting* tests whether the representation contains information useful for predicting the affective state of a subsequent utterance. *Editing* evaluates whether the representation can support controlled changes to the affective content while preserving the desired target, measured by target attainment. Finally, *Explainability* evaluates whether the representation supports faithful recovery of the information used for its predictions through probing, measured by probe faithfulness. Macro-F1 is used for classification and forecasting, while R^2, target attainment, and probe faithfulness are used for regression, editing, and explainability, respectively.

Together, these tasks evaluate complementary properties of the representations: affective recognition, continuous affect recovery, temporal prediction, controllable manipulation, and interpretability. Table 10 reports the seed-42 results for all four encoder backbones to provide a controlled cross-backbone comparison. The results show that performance remains task- and dataset-dependent, with

Table 9: Recoverability of temporal direction across datasets and encoder backbones. Δ is the paired difference between chronological and within-dialogue-shuffled accuracy. Chance is 0.500 by construction. Intervals are 95% CIs over the three representations.

Backbone	Dataset	Δ	95% CI
BERT	MELD	$+0.006$	$[-0.016, +0.028]$
	DailyDialog	$+0.086$	$[+0.055, +0.116]$
	IEMOCAP	$+0.012$	$[-0.013, +0.038]$
DistilBERT	MELD	$+0.001$	$[-0.052, +0.054]$
	DailyDialog	$+0.084$	$[+0.070, +0.097]$
	IEMOCAP	$+0.016$	$[+0.006, +0.026]$
MiniLM	MELD	$+0.006$	$[-0.061, +0.073]$
	DailyDialog	$+0.079$	$[+0.060, +0.099]$
	IEMOCAP	$+0.007$	$[-0.009, +0.023]$
RoBERTa	MELD	$+0.027$	$[-0.029, +0.083]$
	DailyDialog	$+0.077$	$[+0.060, +0.094]$
	IEMOCAP	$+0.005$	$[-0.024, +0.034]$

Table 10: **Complete representation–task evaluation across encoder backbones.** Primary scores for State, Continuous, and Dynamic representations. Macro-F1 is used for classification and forecasting, R^2 for regression, target attainment for editing, probe faithfulness for explainability, and ambiguous entropy for ambiguity analysis.

Backbone	Task	Dataset	State	Cont.	Dyn.
BERT	Classification	GoEmotions	**0.473**	0.244	0.158
	Classification	MELD	0.362	0.395	0.421
	Classification	DailyDialog	0.270	0.290	0.278
	Classification	IEMOCAP	0.389	**0.489**	0.384
	Regression	EmoBank	0.316	**0.387**	0.319
	Regression	IEMOCAP	0.192	**0.231**	0.209
	Forecasting	MELD	0.105	0.125	**0.129**
	Forecasting	DailyDialog	0.143	**0.151**	0.140
	Forecasting	IEMOCAP	0.292	**0.408**	0.320
	Editing	MELD	0.362	0.528	**0.850**
	Editing	GoEmotions	**0.806**	0.280	0.164
	Explainability	GoEmotions	**0.457**	0.332	0.258
	Explainability	MELD	0.361	0.395	**0.422**
	Explainability	EmoBank	0.316	**0.387**	0.319
	Ambiguity	GoEmotions	1.132	1.719	1.922
DistilBERT	Classification	GoEmotions	0.462	0.256	0.178
	Classification	MELD	0.358	0.383	0.406
	Classification	DailyDialog	0.264	0.303	0.276
	Classification	IEMOCAP	0.349	**0.449**	0.372
	Regression	EmoBank	0.306	0.363	0.293
	Regression	IEMOCAP	0.190	0.216	0.191
	Forecasting	MELD	0.109	0.117	**0.124**
	Forecasting	DailyDialog	0.140	**0.147**	0.139
	Forecasting	IEMOCAP	0.270	**0.400**	0.300
	Editing	MELD	0.362	0.528	0.694
	Editing	GoEmotions	**0.806**	0.368	0.164
	Explainability	GoEmotions	**0.465**	0.342	0.262
	Explainability	MELD	0.358	0.387	**0.406**
	Explainability	EmoBank	0.306	**0.363**	0.293
	Ambiguity	GoEmotions	1.189	1.700	1.892

Backbone	Task	Dataset	State	Cont.	Dyn.
MiniLM	Classification	GoEmotions	0.364	0.147	0.071
	Classification	MELD	0.328	0.347	0.363
	Classification	DailyDialog	0.257	0.267	0.243
	Classification	IEMOCAP	0.306	**0.364**	0.270
	Regression	EmoBank	0.274	**0.360**	0.265
	Regression	IEMOCAP	0.173	**0.203**	0.160
	Forecasting	MELD	0.096	**0.104**	0.096
	Forecasting	DailyDialog	0.136	**0.143**	0.131
	Forecasting	IEMOCAP	0.221	**0.256**	0.172
	Editing	MELD	0.212	**0.528**	**0.528**
	Editing	GoEmotions	**0.648**	0.208	0.122
	Explainability	GoEmotions	**0.410**	0.245	0.130
	Explainability	MELD	0.324	0.346	**0.363**
	Explainability	EmoBank	0.274	**0.360**	0.265
	Ambiguity	GoEmotions	1.340	1.934	2.222
RoBERTa	Classification	GoEmotions	**0.474**	0.284	0.092
	Classification	MELD	0.363	**0.428**	0.417
	Classification	DailyDialog	0.281	**0.328**	0.270
	Classification	IEMOCAP	0.356	**0.429**	0.341
	Regression	EmoBank	0.328	**0.413**	0.318
	Regression	IEMOCAP	0.161	**0.230**	0.180
	Forecasting	MELD	0.107	0.121	0.116
	Forecasting	DailyDialog	0.137	**0.155**	0.140
	Forecasting	IEMOCAP	0.261	**0.348**	0.253
	Editing	MELD	0.362	0.528	**0.694**
	Editing	GoEmotions	**0.892**	0.340	0.130
	Explainability	GoEmotions	**0.470**	0.359	0.199
	Explainability	MELD	0.362	0.425	0.417
	Explainability	EmoBank	0.328	**0.413**	0.318
	Ambiguity	GoEmotions	**1.113**	1.661	2.042

the State, Continuous, and Dynamic representations exhibiting different computational affordances rather than a single representation consistently dominating across all tasks. Ambiguous entropy is reported separately as an ambiguity analysis rather than as one of the five primary downstream tasks.

Evaluation on IEMOCAP: IEMOCAP is used only for external evaluation and has no training split. Therefore, for every encoder and representation, classification and next-turn forecasting probes use stratified 5-fold cross-validation with shuffling seeded by the run seed, and the reported score is the

Table 11: **State–Dynamics Index (SDI) across encoder backbones.** SDI values for each emotion across MELD, DailyDialog, and IEMOCAP. Positive values indicate more process-like organization under the implemented definition. "–" denotes an unavailable emotion–dataset combination.

	BERT			**DistilBERT**			**MiniLM**			**RoBERTa**		
Emotion	MELD	DailyDialog	IEMOCAP	MELD	DailyDialog	IEMOCAP	MELD	DailyDialog	IEMOCAP	MELD	DailyDialog	IEMOCAP
Anger	0.163	**0.029**	-0.055	**0.183**	-0.114	**−0.012**	0.077	-0.114	**−0.012**	**0.183**	-0.143	-0.049
Disgust	-0.050	-0.071	**0.000**	**0.050**	-0.071	**0.000**	**0.000**	**0.000**	**0.000**	-0.100	-0.143	**0.000**
Fear	**0.000**	**0.000**	-0.083	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Happiness	–	-0.023	–	–	-0.026	–	–	**0.026**	–	–	-0.026	–
Joy	**0.190**	–	**0.000**	0.083	–	**0.012**	**0.256**	–	-0.084	0.182	–	-0.065
Neutral	**−0.008**	**−0.002**	**0.000**	-0.013	**0.001**	**0.004**	-0.019	-0.013	-0.012	-0.011	**−0.002**	**0.021**
Sadness	**0.065**	**0.032**	**0.095**	-0.145	0.000	0.049	0.048	0.000	-0.083	**0.081**	-0.129	-0.006
Surprise	0.012	**0.229**	0.062	**0.071**	0.171	**0.125**	0.036	0.143	-0.094	-0.012	0.171	-0.031

mean macro-F1 across folds. All other datasets are evaluated on fixed train/test splits and involve no cross-validation. Utterances whose IEMOCAP annotation is `xxx` (no annotator agreement) or `oth` (emotion outside the label set), 2,510 of 10,039, are excluded from categorical evaluation. Classes with fewer utterances than folds cannot be represented in every fold and are also excluded. This removes only *disgust* ($n = 2$), leaving six classes: anger ($n = 2,952$), neutral ($n = 1,708$), joy ($n = 1,636$), sadness ($n = 1,084$), surprise ($n = 107$), and fear ($n = 40$). The same exclusion is applied to next-turn targets in forecasting. VAD regression on IEMOCAP uses all 10,039 utterances with unstratified 5-fold cross-validation.

F.2 State–Dynamics Index

The State–Dynamics Index (SDI) measures whether an emotion is more strongly organized as a conversational process or as a relatively stable state. Rather than using overall classification performance, SDI compares the per-class recall of the Dynamic and State representations, allowing the comparison to be made separately for each emotion. A positive value indicates that the emotion is better recovered from the Dynamic representation, suggesting a more process-like organization, whereas a non-positive value indicates that the State representation provides at least as much class-specific information.

$$\mathrm{SDI}(c) = \mathrm{recall}_{\mathrm{Dynamic}}(c) - \mathrm{recall}_{\mathrm{State}}(c). \tag{32}$$

Here, recall is used because it measures how reliably instances of emotion c are recovered, without being affected by predictions made on other emotion classes. Thus, SDI directly captures the relative contribution of dynamic versus state-based representations to recognizing each emotion. Values near zero indicate similar recoverability in the two representations, while larger positive or negative values indicate a stronger association with the Dynamic or State representation, respectively. Table 11 reports the complete SDI results across MELD, DailyDialog, and IEMOCAP for all four encoder backbones. The variation across emotions and datasets shows that state–process organization is not uniform across affective categories or encoder architectures.

F.3 Information Preservation Across Representations

The Information Preservation Index (IPI) summarizes how much affective information remains linearly accessible from a representation across the affective facets supported by each dataset. It combines categorical emotion information, measured by raw macro-F1, and continuous affect information, measured by R^2, into a common bounded score. Higher IPI indicates that more of the evaluated affective information is recoverable from the representation, while lower values indicate less information accessible through the linear probe.

$$\mathrm{IPI}(r, D) = \frac{1}{|F_D|} \sum_{f \in F_D} f(r, D), \qquad f(r, D) \in \left\{ F1_{\mathrm{macro}}(r, D),\, \max(0, R^2(r, D)) \right\}. \tag{33}$$

Here, F_D denotes the affective facets available for dataset D. IPI therefore averages the affective information recoverable across the facets that can be evaluated for a given dataset, rather than relying on a single task. Table 12 reports the complete IPI results across encoder backbones and representation objectives. Bold values indicate the highest IPI across encoder backbones for each

Table 12: **Information Preservation Index (IPI) across encoder backbones.** IPI values for State, Continuous, and Dynamic representations. Bold values indicate the highest IPI across encoder backbones for each representation objective within each dataset.

	BERT			DistilBERT			MiniLM			RoBERTa		
Dataset	State	Cont.	Dyn.	State	Cont.	Dyn.	State	Cont.	Dyn.	State	Cont.	Dyn.
GoEmotions	0.461	0.243	0.174	0.437	**0.305**	**0.225**	0.392	0.173	0.085	**0.467**	0.261	0.150
MELD	0.345	0.360	**0.390**	0.348	0.380	0.375	0.301	0.341	0.344	**0.361**	**0.394**	0.391
EmoBank	0.305	0.321	0.234	0.280	0.343	0.255	0.222	**0.363**	0.235	**0.286**	0.352	**0.277**
IEMOCAP	**0.270**	**0.318**	**0.259**	0.238	0.295	0.253	0.219	0.241	0.181	0.234	0.291	0.237

representation objective within a given dataset. The results show that the representation retaining the most affective information varies across datasets and objectives, with no single representation consistently dominating across all settings.

Evaluation on IEMOCAP: For IEMOCAP, the categorical facet is computed only on utterances with a canonical emotion label (7,529 of 10,039; `xxx` and `oth` are excluded), while the VAD facet uses all 10,039 utterances. IPI is the mean of the two facets.

F.4 Representation Agreement and Task Affordance

The Representation Agreement Score (RAS) measures how consistently the three representations produce the same predictions for the same instances. It therefore captures agreement in their downstream decisions rather than similarity in their latent spaces. With $K = 3$ representations, RAS is the average pairwise prediction agreement across all instances.

$$\mathrm{RAS} = \frac{2}{K(K-1)} \sum_{i<j} \frac{1}{N} \sum_{n} \mathbb{I}\left[\hat{y}_n^{(i)} = \hat{y}_n^{(j)}\right], \qquad K = 3. \tag{34}$$

A higher RAS therefore indicates that different representations lead to more consistent decisions, while a lower value indicates greater disagreement in their downstream predictions. Table 13 reports the complete results across encoder backbones. RAS is highest on DailyDialog for all four backbones (0.922–0.942) and lowest on GoEmotions (0.521–0.571), indicating that the degree of representational agreement varies across datasets. For BERT, the three-seed RAS values are 0.544, 0.570, and 0.566 for GoEmotions, 0.728, 0.742, and 0.738 for MELD, 0.925, 0.936, and 0.941 for DailyDialog, and 0.673, 0.674, and 0.686 for IEMOCAP. These values are closely aligned with the seed-42 results, with only small differences across representations and datasets, indicating that the observed representation agreement is not strongly sensitive to the random seed.

The Task Affordance Score (TAS) measures how well each representation supports a particular computational task relative to the other representations. Rather than comparing raw scores across different tasks, TAS normalizes the primary task score by the strongest representation for that task, making the value relative to the best available representation within the same setting.

$$\mathrm{TAS}(r, \tau) = \frac{\bar{s}(r, \tau)}{\max_{r'} \bar{s}(r', \tau)}, \tag{35}$$

where $\bar{s}(r, \tau)$ is the mean primary score of representation r over the datasets associated with task τ. Thus, TAS equals 1 for the representation achieving the highest mean task score, with lower values indicating a relative performance gap within that task. Table 14 reports the complete seed-42 results across all four encoder backbones. The scores show clear task-specific structure: Continuous is consistently strongest for regression and forecasting, while State is strongest for classification and editing across all four backbones. For explainability, State is strongest for BERT, DistilBERT, and MiniLM, whereas Continuous is strongest for RoBERTa. The three-seed BERT analysis gives mean TAS values of 1.000, 0.973, and 0.849 for classification; 0.821, 1.000, and 0.847 for regression; 0.801, 1.000, and 0.887 for forecasting; 1.000, 0.671, and 0.758 for editing; and 1.000, 0.985, and 0.883 for explainability, respectively for State, Continuous, and Dynamic.

F.5 Representation Complementarity Score

The capacity-corrected Representation Complementarity Score (RCS^*) measures how much additional label-relevant information becomes linearly accessible when the three representation views

Table 13: **Representation Agreement Score (RAS) across encoder backbones.**

Dataset	BERT	DistilBERT	MiniLM	RoBERTa
GoEmotions	0.544	**0.571**	0.564	0.521
MELD	0.707	**0.743**	0.737	0.737
DailyDialog	0.927	0.937	**0.942**	0.922
IEMOCAP	0.669	0.673	**0.685**	0.657

Table 14: **Task Affordance Score (TAS) across encoder backbones.** Values are normalized within each task under the implemented definition.

	BERT			DistilBERT			MiniLM			RoBERTa		
Task	State	Cont.	Dyn.	State	Cont.	Dyn.	State	Cont.	Dyn.	State	Cont.	Dyn.
Classification	**1.000**	0.949	0.831	**1.000**	0.971	0.859	**1.000**	0.895	0.754	**1.000**	0.997	0.761
Regression	0.823	**1.000**	0.855	0.855	**1.000**	0.836	0.793	**1.000**	0.755	0.760	**1.000**	0.773
Forecasting	0.788	**1.000**	0.860	0.783	**1.000**	0.849	0.900	**1.000**	0.792	0.808	**1.000**	0.816
Editing	**1.000**	0.692	0.868	**1.000**	0.767	0.735	**1.000**	0.856	0.756	**1.000**	0.692	0.657
Explainability	**1.000**	0.983	0.881	**1.000**	0.967	0.851	**1.000**	0.944	0.752	0.969	**1.000**	0.780

are combined, beyond what is available from the strongest individual representation. The correction is important because concatenating the representations increases the probe width and can produce a performance gain even when the added dimensions contain no complementary information. RCS^* therefore subtracts a dimension-matched, label-independent control gain from the observed joint-representation gain.

$$RCS^* = \left[F1_{\text{macro}}\left(z^{(1)} \oplus z^{(2)} \oplus z^{(3)}\right) - F1_{\text{macro}}\left(z^{(i^*)}\right)\right] - \left[F1_{\text{macro}}\left(z^{(i^*)} \oplus \eta\right) - F1_{\text{macro}}\left(z^{(i^*)}\right)\right]. \tag{36}$$

Here, $z^{(i^*)}$ is the strongest individual representation and η is a label-independent control block with the same added dimensionality as the remaining representation views. The first bracket measures the gain from combining the representations, while the second estimates the gain that can be attributed to increased probe capacity alone. Thus, positive RCS^* indicates additional linearly accessible information beyond the width-matched null, while values near zero indicate little evidence for complementary information after this correction.

Table 15 reports the complete capacity-corrected results across the four encoder backbones and four datasets. RCS^* is positive in 15 of the 16 backbone–dataset settings; the only negative value occurs for DistilBERT on GoEmotions (-0.002). BERT was used as the primary encoder and was evaluated across three random seeds to assess seed sensitivity, while DistilBERT, MiniLM, and RoBERTa were included for the controlled cross-backbone comparison. The table reports seed-42 values for all four backbones, while the three-seed BERT mean $\pm$ standard deviation is reported in the main text.

G Qualitative Latent Density Structure

To complement the quantitative geometric analyses, we visualize the distribution of latent samples within the learned representation spaces. For each representation, the 128-dimensional vectors are projected onto their first two principal components, followed by density estimation in the resulting two-dimensional space. The plots therefore show sample concentration rather than individual latent points: brighter regions indicate higher density, darker regions indicate lower density, and nested contours trace density levels. This provides a qualitative view of how the State, Continuous, and Dynamic objectives organize the same underlying data.

The interpretation is structural. Compact contours indicate concentrated representations, broader contours indicate greater spatial dispersion, multiple local maxima indicate multimodality, and elongated or curved density ridges indicate anisotropic structure. The principal-component axes are data-derived and do not represent predefined affective dimensions. Apparent gaps between density regions should also not be interpreted as disconnected manifolds, since PCA can distort distances and

Table 15: **Capacity-corrected Representation Complementarity Score (RCS^*) across encoder backbones.** RCS^* measures the additional linearly accessible information obtained from the concatenated representations relative to the best single representation after correcting for the gain attributable to the dimension-matched shuffled null. Results use seed 42.

Dataset	BERT	DistilBERT	MiniLM	RoBERTa
GoEmotions	**0.025**	-0.002	0.012	0.023
MELD	0.004	**0.062**	0.016	0.033
DailyDialog	0.047	0.036	0.050	**0.060**
IEMOCAP	0.043	0.048	0.040	**0.058**

topology. Accordingly, these plots are used to characterize qualitative organization, while quantitative claims about dimensionality and local geometry are based on the analyses reported elsewhere.

Figure 14 shows DailyDialog across the four encoder backbones. A consistent objective-level pattern is visible: State representations occupy broader but relatively compact regions with concentrated density cores, Continuous representations are more localized, and Dynamic representations tend to form elongated or curved structures with greater spatial extension. Despite changes in orientation and contour shape across backbones, this relative organization remains visible.

A similar pattern appears in EmoBank (Fig. 15), although the density structures are more varied. State representations generally occupy broader regions, Continuous representations remain more localized, and Dynamic representations tend to exhibit elongated or curved structure. The precise locations and orientations of the density modes vary across encoders, while the broader objective-level organization remains apparent.

GoEmotions exhibits richer and more multimodal structure (Fig. 16). State representations frequently contain multiple high-density regions, Continuous representations are comparatively localized but can form curved or arch-like contours, and Dynamic representations show stronger elongated and curved configurations, including separated or partially enclosed high-density regions in several backbones. These differences illustrate that representations with similar overall extent can nevertheless differ substantially in how their density is organized.

MELD provides another clear separation among the objectives (Fig. 17). State representations generally form broad, high-density regions with multiple localized maxima, Continuous representations are more spatially concentrated, and Dynamic representations more often exhibit elongated or separated structures. Several Dynamic representations contain multiple density regions connected by lower-density areas, consistent with greater contextual variation. The detailed geometry of Dynamic representations also varies more across backbones than that of State or Continuous representations.

IEMOCAP shows a similarly multimodal organization (Fig. 18). State representations frequently contain several distinct density modes, Continuous representations are generally more compact and localized, and Dynamic representations again tend toward elongated or curved structures. Although the precise modes and orientations differ across encoders, the relative organization of the three objectives is recurrent.

Across datasets, the density surfaces therefore reveal a recurring qualitative organization rather than a single universal geometry. Continuous representations tend to be more localized, State representations tend to be broader or more multimodal, and Dynamic representations more often exhibit elongated, curved, or spatially separated structures. The strength of these patterns varies by dataset: Dynamic extensions are particularly visible in DailyDialog and MELD, while GoEmotions and IEMOCAP show stronger multimodality and EmoBank exhibits smoother but still distinct objective-dependent structure. Across backbones, orientations and local density patterns vary, but the broader objective-level organization is repeatedly preserved. These visualizations therefore provide qualitative evidence complementary to the quantitative analyses of intrinsic dimensionality, local linearity, and manifold-aware neighborhood structure.

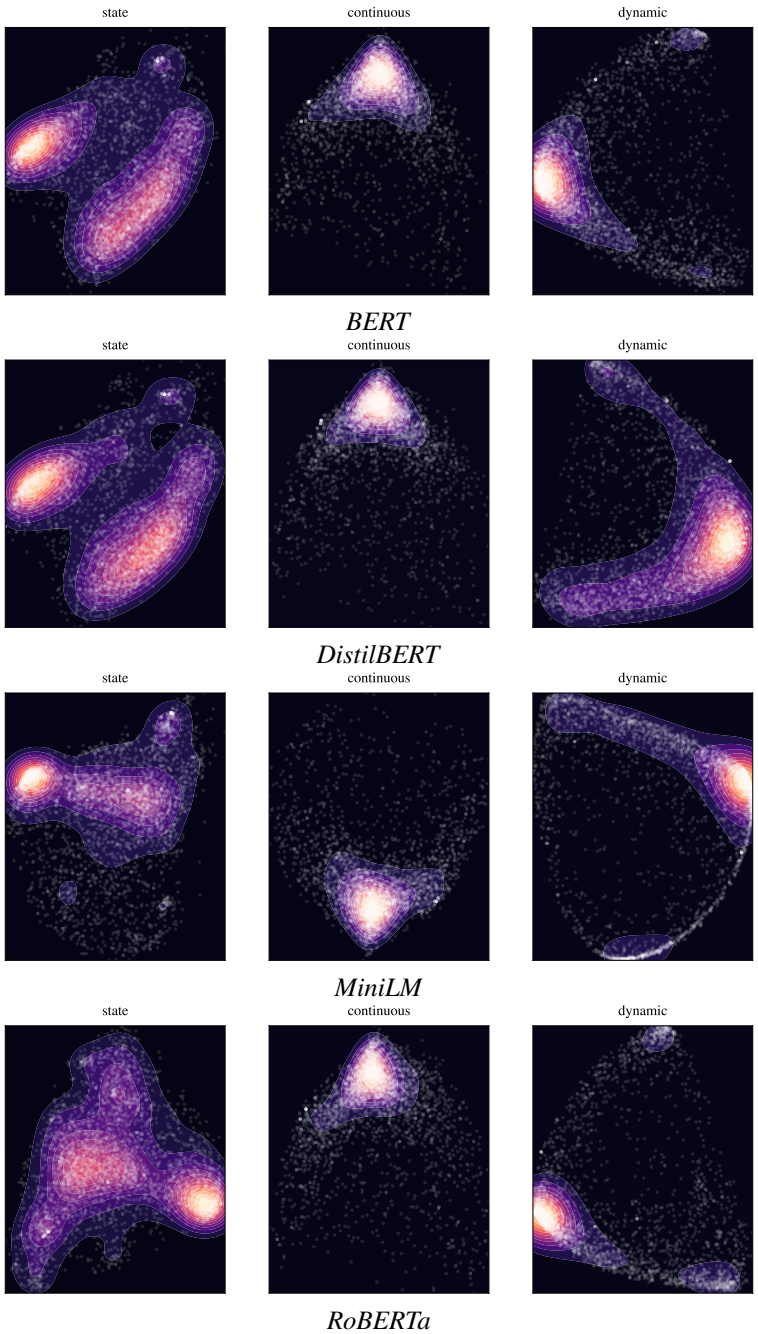


Figure 14: **Latent density surfaces for DailyDialog.** PCA-2D density visualizations of State, Continuous, and Dynamic representations across four encoder backbones.

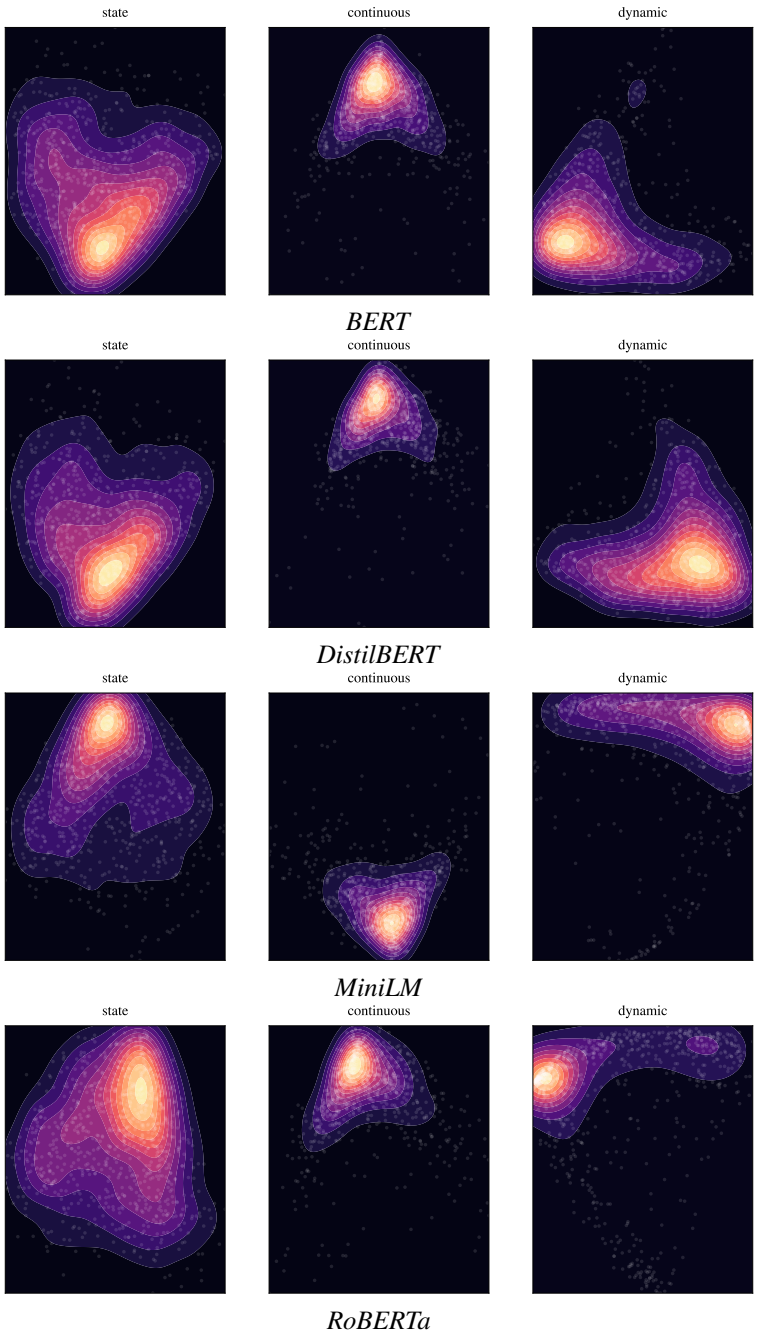


Figure 15: **Latent density surfaces for EmoBank.** PCA-2D density visualizations of State, Continuous, and Dynamic representations across four encoder backbones.

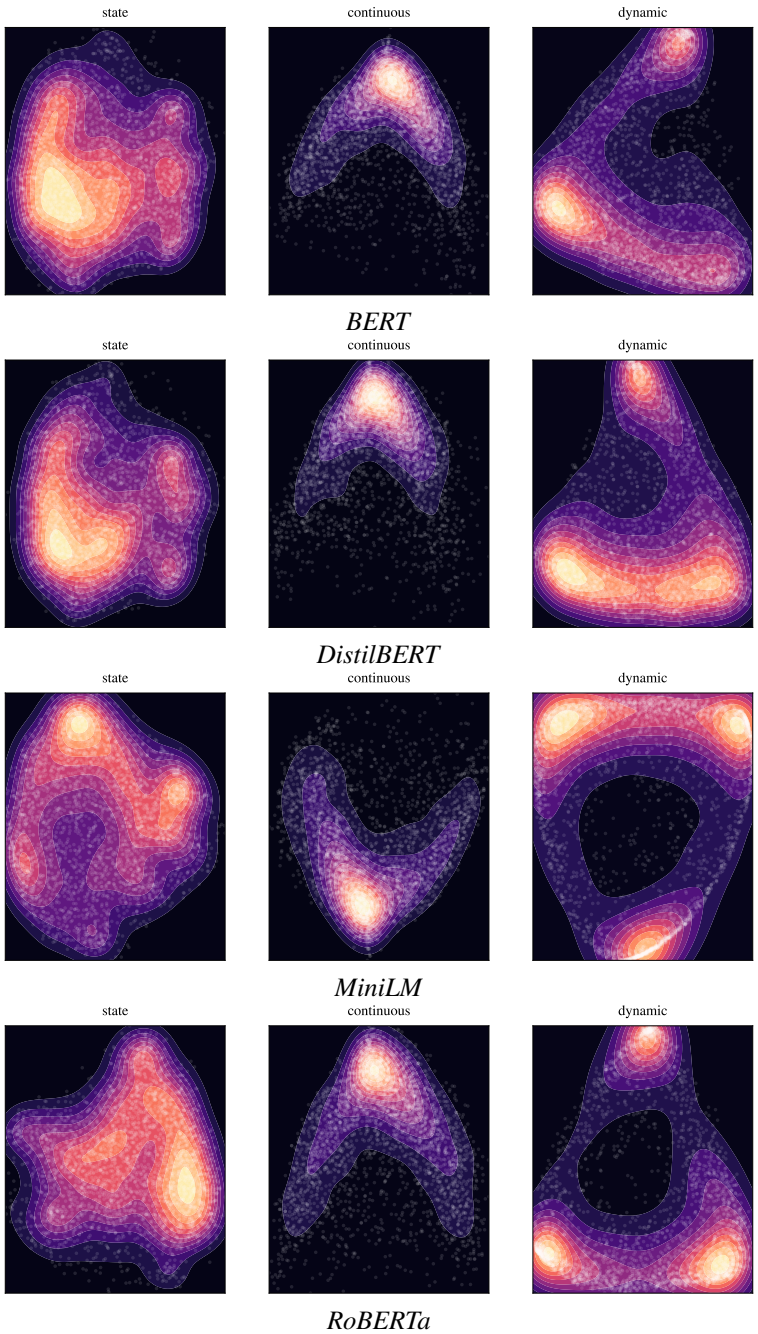


Figure 16: **Latent density surfaces for GoEmotions.** PCA-2D density visualizations of State, Continuous, and Dynamic representations across four encoder backbones.

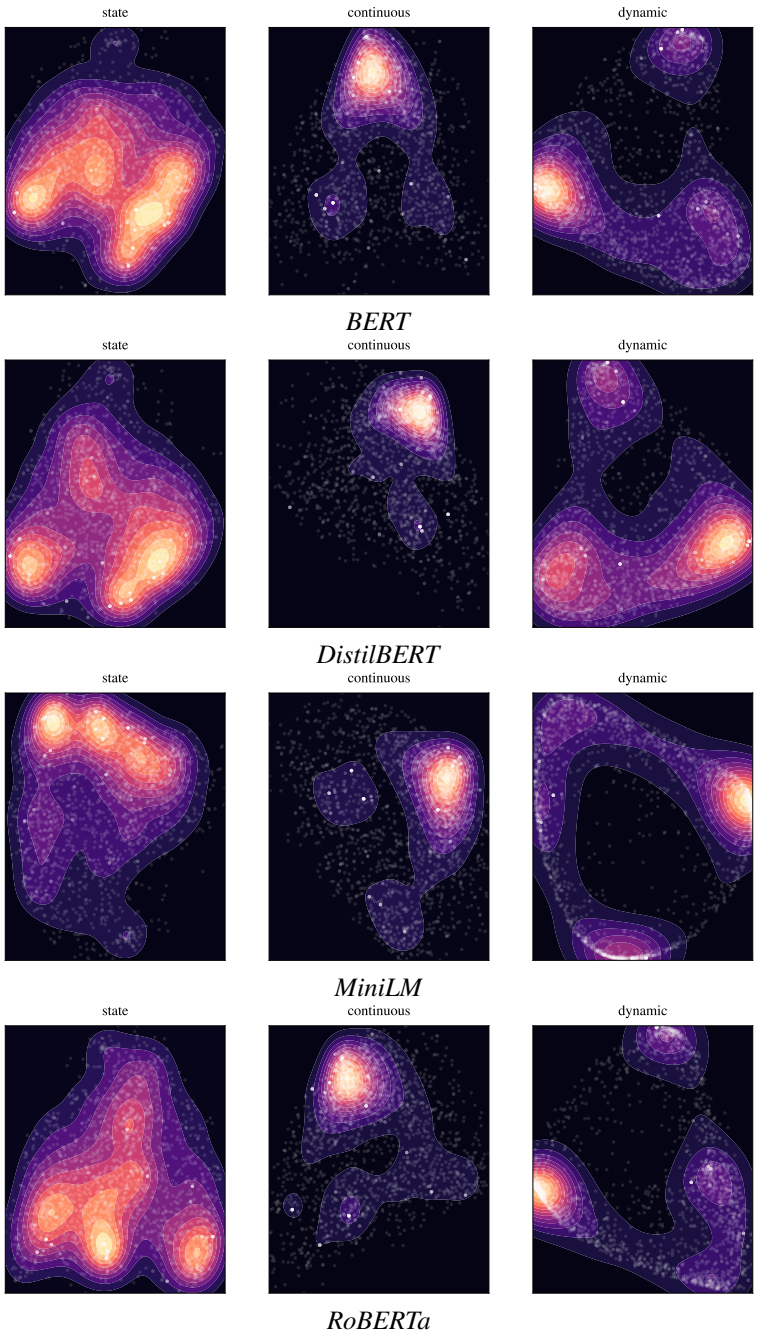


Figure 17: **Latent density surfaces for MELD.** PCA-2D density visualizations of State, Continuous, and Dynamic representations across four encoder backbones.

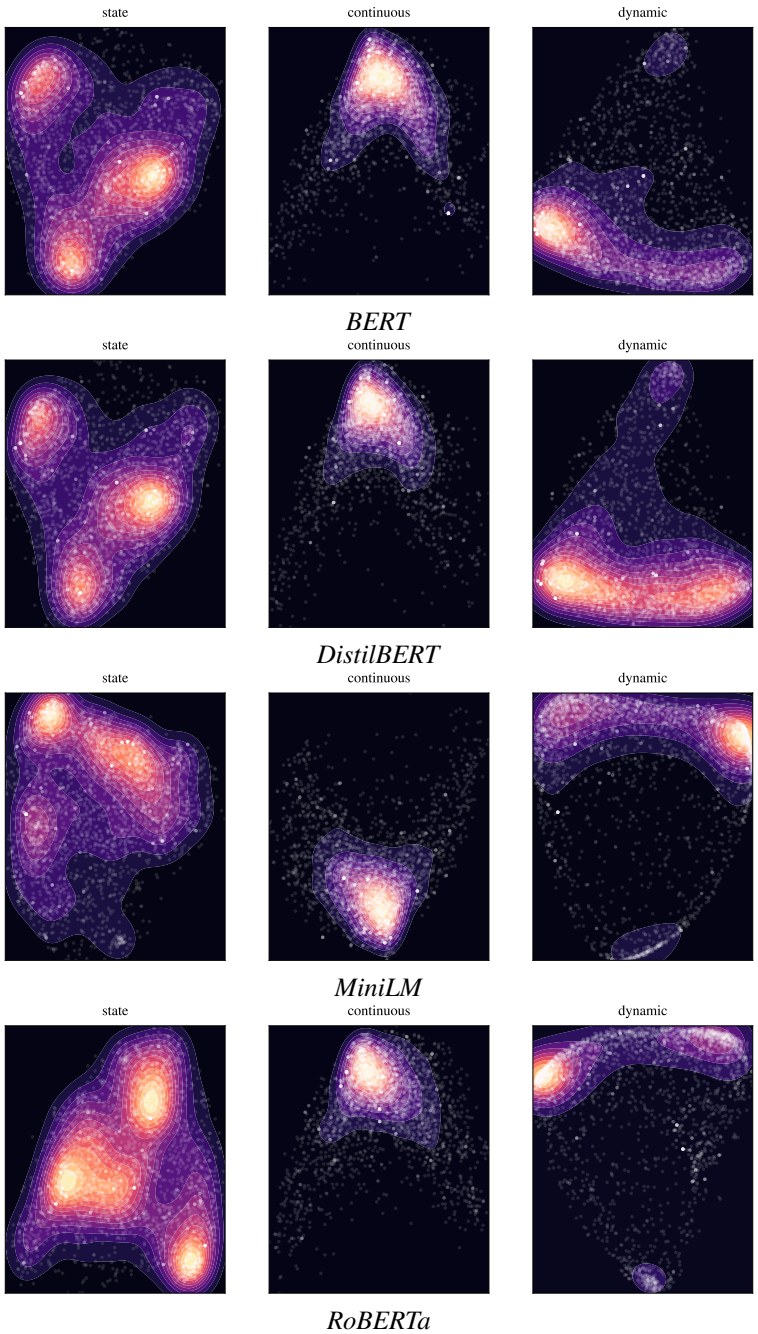


Figure 18: **Latent density surfaces for IEMOCAP.** PCA-2D density visualizations of State, Continuous, and Dynamic representations across four encoder backbones.