MULTIMODAL AFFECTIVE REPRESENTATION LEARNING: A TECHNICAL EVOLUTION FROM FEATURE DISENTANGLEMENT TO SEMANTIC ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal affective representation learning aims to learn discriminative, transferable, and robust affective representations from heterogeneous information, including text, speech, visual signals, physiological signals, and eye movements. This review systematically surveys the field from four perspectives: emotion theories, modality-specific information, representation-learning methods, and application-oriented evaluation. Existing approaches are organized into five families: cross-modal alignment and representation disentanglement, cross-modal interaction and dynamic fusion, context and dialogue-structure modeling, self-supervised and robust learning, and large-model-based and prompt-based learning. We further analyze these methods through six affective representation metrics: cross-modal comparability, complementarity, redundancy control, disentanglement of irrelevant factors, robustness to missing modalities, and sensitivity to dynamics, and examine their roles in affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization. Finally, we summarize current evaluation practices and their limitations, with the aim of moving multimodal affective computing beyond single-task performance toward a more systematic assessment of affective representation quality.

1 INTRODUCTION

Emotion can be represented in multiple forms, including discrete emotion categories, continuous dimensions such as valence and arousal, and natural-language descriptions that further express its causes, objects, and contextual conditions (Ekman, 1992; Russell, 1980; Barrett, 2017). These forms define the affective semantics that a model is expected to capture, but they are not equivalent to the internal representations actually learned by the model. An effective affective representation should not only support accurate prediction for a specific task, but also preserve transferable affective information while minimizing its dependence on emotion-irrelevant factors, such as speaker identity or recording conditions (Bengio et al., 2013; Schölkopf et al., 2021).

As illustrated in Fig. 1, raw multimodal observations must first be transformed into internal affective representations before they can reliably support downstream tasks. Textual, acoustic, visual, physiological, and contextual information is encoded and organized through **fusion**, **disentanglement**, **alignment**, and **robust learning**, thereby forming affective representations that can support different downstream tasks. The quality of this intermediate representation determines not only performance on the current task, but also the model's ability to generalize across new settings, modality configurations, and populations; consequently, even when a model achieves strong performance on a particular dataset, its transferability may remain limited if the learned representation also captures irrelevant cues such as speaker identity, recording devices, background conditions, or linguistic style.

A desirable affective representation should exhibit **generalizability**, **disentanglement of irrelevant factors**, **unification**, and **robustness**, such that it can transfer across new data distributions or modality configurations, distinguish affective information from non-affective factors such as identity or linguistic style, relate discrete emotion labels to continuous affective dimensions, and remain stable under missing modalities, noise, and individual variation (Lin & Hu, 2023; Wang et al., 2026; Hazarika et al., 2020). These requirements become particularly challenging in multimodal

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.

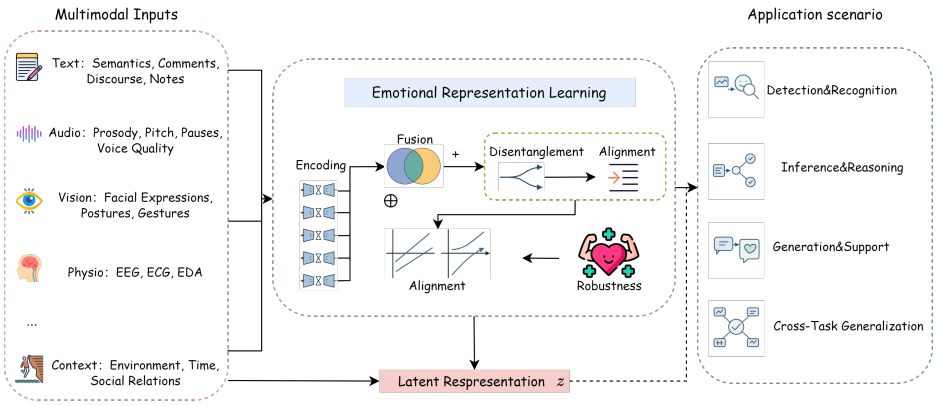

Figure 1: Schematic of the affective representation process.

settings, where text, speech, visual behavior, and physiological signals provide semantic, prosodic, behavioral, and bodily cues, respectively, while differing substantially in temporal scale, expression mechanism, and information density (Eyben et al., 2015; Baltrušaitis et al., 2016; Kosti et al., 2019; Zhong et al., 2020). Text may directly describe the cause or object of an emotion, speech conveys prosody and intensity, visual observations capture facial expressions and posture, and physiological signals reflect bodily responses; consequently, different modalities may provide complementary evidence, but may also exhibit redundancy, conflict, or temporal asynchrony. Multimodal affective representation learning must therefore determine which information should be integrated, which modality-specific distinctions should be preserved, which irrelevant factors should be disentangled, and how stable representations can be maintained when observations are noisy or incomplete.

Recent studies have addressed these **affective representation metrics** through different mechanisms. Tensor Fusion Networks explicitly model higher-order interactions among modalities (Zadeh et al., 2017), while multimodal Transformers employ cross-modal attention to process sequences that are not strictly aligned in time (Tsai et al., 2019); MISA separates modality-shared from modality-specific information through shared–private representation spaces (Hazarika et al., 2020), hierarchical alignment approaches establish semantic correspondences at different levels of granularity (Qian et al., 2026), and methods designed for missing modalities improve representation stability under incomplete observations (Lin & Hu, 2023; Li et al., 2024). Collectively, these studies have advanced **fusion**, **disentanglement of irrelevant factors**, **semantic alignment**, and **robustness**, although their effectiveness is still predominantly demonstrated through performance gains on individual downstream tasks. Existing surveys have examined multimodal learning, the understanding and expression of emotion by affective agents (Zall et al., 2025), emotion recognition, and affective computing with large models (Baltrušaitis et al., 2018; Hegde & Jayalath, 2025; Hu et al., 2024; Wu et al., 2025; Zhang et al., 2026); however, comparatively less attention has been devoted to using affective representation quality as a unified analytical perspective for explaining how different techniques influence generalizability, disentanglement of irrelevant factors, unification, and robustness, or how these properties can be assessed beyond task-level scores.

This review follows the theme **"from feature disentanglement to semantic alignment"** and organizes the literature around four interconnected elements: **information**, **representation metrics**, **methods**, and **evaluation**. It first examines the psychological foundations and semantic descriptions of emotion to clarify what affective representations are intended to encode; it then analyzes the affective cues provided by text, speech, vision, and physiology, and reviews methods for cross-modal alignment and disentanglement, interaction and fusion, dialogue-structure modeling, self-supervised and robust learning, and large-model-based learning. Finally, these representation mechanisms are connected to four application settings: affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization, thereby establishing a coherent analytical chain from affective information and representation mechanisms to downstream evaluation.

Rather than treating task performance as the sole criterion, this review examines multimodal methods in terms of the properties of the representations they learn. Its contributions are as follows:

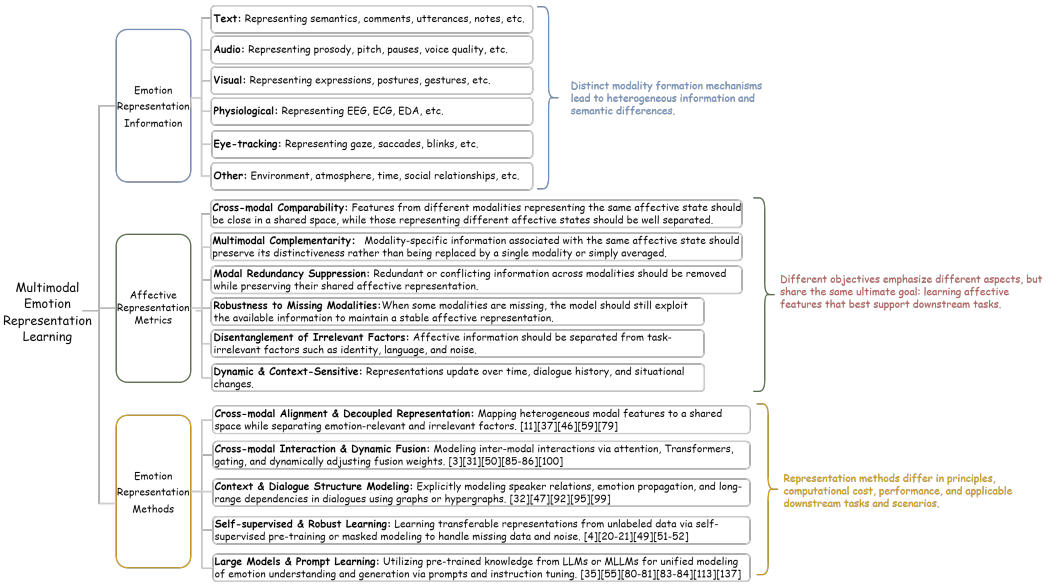

Figure 2: Framework for multimodal affective representation learning. The references shown inside the figure correspond to representative works on alignment and disentanglement (Hazarika et al., 2020; Yang et al., 2022; Qian et al., 2026; Hu et al., 2022; Meng et al., 2026), dynamic fusion (Tsai et al., 2019; Arevalo et al., 2017; Tan & Zhang, 2025; Lee et al., 2025; Hu et al., 2025a; Zadeh et al., 2017), context and dialogue structure modeling (Hu et al., 2021; Ai et al., 2025; Li et al., 2023; Ghosal et al., 2019; Hazarika et al., 2018), robust learning (Lin & Hu, 2023; Baevski et al., 2020; Hsu et al., 2021; Fan et al., 2024; Li et al., 2024; Zhuang et al., 2026), and large-model-based learning (Alayrac et al., 2022; Huang et al., 2024; 2026; Hu et al., 2025b; Yang et al., 2025; Song et al., 2025; Lian et al., 2025a; Luo et al., 2026).

- **we establish an affective-representation-centered framework.** We connect psychological theories and categorical, dimensional, and open-ended descriptions of emotion with textual, acoustic, visual, and physiological observations. We also specify six affective representation metrics: cross-modal comparability, complementarity, redundancy control, disentanglement of irrelevant factors, robustness to missing modalities, and sensitivity to dynamics.
- **we provide a systematic synthesis of existing methods and their scope.** We organize existing work into five families of methods and examine how they structure shared and modality-specific information, establish cross-modal semantic relationships, and address contextual changes and incomplete inputs.
- **we establish a connection between methods, applications, and evaluation.** Across affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization, we summarize the corresponding datasets, representative models, and evaluation measures.

The remainder of this paper is organized as follows. Section 2, *Review Methodology*, describes the literature search, screening, and inclusion procedures. Section 3, *Models of Emotion*, reviews discrete, dimensional, cognitive-appraisal, and open-ended formulations of emotion to clarify the semantic forms underlying affective representations. Section 4, *Affective Information*, examines affective cues and representational differences across textual, acoustic, visual, physiological, and eye-tracking modalities. Section 5, *Multimodal Affective Representation Learning*, systematically summarizes the major technical paradigms, including cross-modal alignment and disentanglement, dynamic fusion, context and dialogue-structure modeling, self-supervised and robust learning, and large-model- and prompt-based approaches. Section 6, *Applications of Affective Representations*, covers four application settings—affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization—and analyzes the roles of representative datasets, models, and evaluation metrics, followed by an integrated experimental analysis comparing their

representational requirements and evaluation characteristics. Finally, Section 7, *Conclusion*, summarizes the main findings, discusses current limitations, and outlines directions for future research in multimodal affective representation learning.

2 Review Methodology

We conducted a systematic literature search across IEEE Xplore, ScienceDirect, and Google Scholar using "multimodal affective representation learning" and "multimodal emotion recognition" as the primary search terms, complemented by additional keywords including "feature disentanglement", "cross-modal alignment", and "missing modality" to broaden the coverage of studies on representation disentanglement, semantic alignment, and robustness to incomplete modalities. To reduce potential omissions from database-based retrieval, we further supplemented the search through citation tracing of representative studies and large-language-model-assisted retrieval.

The collected records were subsequently subjected to deduplication, title and abstract screening, and full-text eligibility assessment, with particular attention to their relevance to affective representation, fusion, disentanglement, alignment, and robust learning. After screening, the final review included 146 references, including foundational studies identified through citation tracing and theoretical supplementation. Detailed search strategies, exclusion criteria, and the complete literature-selection procedure are provided in Appendix A.

3 Models of Emotion

Emotion models define the semantic space of affective representations. Existing accounts can be organized into four complementary perspectives—categorical, dimensional, appraisal-based, and open-ended descriptions (Ekman, 1992; Russell, 1980; Sander et al., 2005; Barrett, 2017)—which characterize affective states through discrete emotion classes, continuous valence–arousal spaces, evaluative relationships among events and goals, and context-dependent concepts expressed in natural language, respectively (Picard, 1997; Calvo & D'Mello, 2010). These perspectives impose different levels of supervision and representational constraints, thereby shaping how affective semantics are organized within the learned representation.

These perspectives involve distinct representational trade-offs. Categorical models provide explicit emotion labels that facilitate standardized annotation and classification, but their discrete formulation may obscure intensity, mixed emotional states, and fine-grained affective variation (Ekman, 1992; Ekman et al., 1999; Plutchik, 1980; Russell, 1980; Barrett, 2017); dimensional models preserve continuous affective variation and are therefore well suited to modeling intensity and temporal dynamics, although they provide comparatively limited information about the causes, objects, and contextual conditions of emotion (Russell, 1980; Sander et al., 2005; Lee et al., 2025; Cowen & Keltner, 2017). Appraisal-based models further represent structural relationships among events, goals, evaluations, and emotional responses, thereby supporting causal and contextual affective understanding, but the underlying appraisal processes are often difficult to observe and annotate directly (Sander et al., 2005; Schölkopf et al., 2021; Yeo & Ong, 2024; Poria et al., 2021). By contrast, open-ended descriptions retain richer contextual and fine-grained emotional semantics, while simultaneously introducing greater semantic ambiguity and evaluation difficulty (Barrett, 2017; Schölkopf et al., 2021; Huang et al., 2024; Havaldar et al., 2023; Belay et al., 2025; Song et al., 2025; Lian et al., 2025a). These four perspectives should therefore be viewed as complementary rather than interchangeable forms of supervision for affective representation learning. Detailed theoretical foundations, representative models, and their strengths and limitations are provided in Appendix B.

4 Affective Information

Affective cues are modality-specific. Multimodal affective representations draw on heterogeneous information from text,(Koelstra et al., 2011; Baltrušaitis et al., 2018; Wang et al., 2026; Eyben et al., 2015; Baltrušaitis et al., 2016; Zhong et al., 2020; Karasawa et al., 2023). speech, visual behavior, physiological signals, and eye movements These modalities characterize affect from different perspectives, including semantics, prosody, behavior, bodily responses, and attention, while differing

substantially in temporal scale, expression form, and observability. Text primarily conveys semantic and appraisal information (Devlin et al., 2019; Socher et al., 2013); speech provides prosodic and paralinguistic cues (Eyben et al., 2015; Baevski et al., 2020; Hsu et al., 2021); visual observations capture facial expressions, posture, and embodied behavior (Baltrušaitis et al., 2016; He et al., 2016; Dosovitskiy, 2020; Kosti et al., 2019); physiological signals reflect internal bodily responses (Koelstra et al., 2011; Zhong et al., 2020; Siegel et al., 2018); and eye movements provide information related to attention and cognitive load (Karasawa et al., 2023; Collins & Davies, 2023; Liu et al., 2025b).

Each modality has distinct limitations. Text may omit paralinguistic and nonverbal information (Eyben et al., 2015; Baevski et al., 2020; Hsu et al., 2021; Baltrušaitis et al., 2016; Zhong et al., 2020); speech often entangles affective variation with linguistic content and speaker-specific characteristics (Hazarika et al., 2020; Yang et al., 2022; Rizos et al., 2020); visual signals are influenced by context, identity, culture, occlusion, and viewpoint (Kosti et al., 2019; Chen et al., 2022; Havaldar et al., 2023; Belay et al., 2025; Lamme, 2006). Physiological responses exhibit substantial individual variability and do not map uniquely onto emotion semantics (Koelstra et al., 2011; Zhong et al., 2020; Siegel et al., 2018), while eye movements depend on task conditions and individual viewing patterns (Collins & Davies, 2023; Liu et al., 2025b). These differences motivate multimodal representations that exploit complementary information while controlling modality-specific noise, redundancy, and inconsistency (Baltrušaitis et al., 2018; Wang et al., 2026). A detailed modality-wise analysis is provided in Appendix C.

5 Multimodal Affective Representation Learning

Multimodal affective representation learning mainly includes cross-modal alignment and representation disentanglement, cross-modal interaction and dynamic fusion, context and dialogue-structure modeling, self-supervised and robust learning, and large-model-based and prompt-based learning (Baltrušaitis et al., 2018; Wang et al., 2026; Hu et al., 2024; Wu et al., 2025; Zhang et al., 2026). These families organize and optimize affective representations from different perspectives, including cross-modal semantic correspondence and information separation, adaptive multimodal interaction, contextual and dialogue dependency modeling, robustness to missing or noisy observations, and unified affective modeling with large pretrained models (Tsai et al., 2019; Lin & Hu, 2023; Hazarika et al., 2020; Ghosal et al., 2019; Lian et al., 2025a).

Table 1 summarizes the main objectives, representative mechanisms, and limitations of these method families. A more fine-grained taxonomy, together with extended characteristics and applicable tasks, is provided in Appendix D, Table 3. Representative architectures for disentanglement and dynamic fusion are further illustrated in Appendix D, Figs. 5 and 6.

6 Applications of Affective Representations

Affective representations support multi-level applications. Their role has expanded beyond conventional emotion classification to four major application areas: affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization. As task complexity increases, affective representations progressively extend from predicting emotional states to modeling emotion causes and contextual relations, generating context-appropriate responses, and transferring affective knowledge across domains and tasks, thereby forming a progression from perception and understanding to generation and transfer.

Table 2 provides a compact comparison of the four application areas in terms of representation requirements, representative datasets, and evaluation criteria. Their overall relationships are illustrated in Appendix E, Fig. 7. More fine-grained functional mappings between the five representation-learning method families and the four application areas are summarized in Tables 4–7, while the complete dataset comparison is provided in Table 8.

6.1 Affect Detection and Recognition

Affect detection and recognition identify emotion categories or polarities and estimate affective intensity from multimodal observations (Picard, 1997; Calvo & D'Mello, 2010). Depending on the

Table 1: Taxonomy of multimodal affective representation learning methods.

Method Family	Representation Operation	Core Mechanisms	Main Limitation
Alignment & Disentanglement (Hazarika et al., 2020; Yang et al., 2022; Qian et al., 2026; Hu et al., 2022; Meng et al., 2026)	Separate shared and modality-specific factors while establishing semantic correspondence	Contrastive learning; shared–private decomposition; hierarchical alignment	Over-alignment may suppress useful modality-specific cues
Dynamic Fusion (Tsai et al., 2019; Arevalo et al., 2017; Tan & Zhang, 2025; Lee et al., 2025; Hu et al., 2025a; Zadeh et al., 2017)	Adaptively integrate complementary information across modalities	Cross-modal attention; gating; tensor fusion; Transformers	Reliability estimation and computational cost
Context Modeling (Hu et al., 2021; Ai et al., 2025; Li et al., 2023; Ghosal et al., 2019; Hazarika et al., 2018)	Encode temporal, speaker, and interaction dependencies into the representation	Graph neural networks; hypergraphs; dialogue modeling	Dependence on contextual or structural information
Robust Learning (Lin & Hu, 2023; Baevski et al., 2020; Hsu et al., 2021; Fan et al., 2024; Li et al., 2024; Zhuang et al., 2026)	Preserve representation stability under missing, noisy, or incomplete observations	Masking; reconstruction; self-supervision; consistency learning	Information loss under severe missingness or noise
Large Models (Alayrac et al., 2022; Huang et al., 2024; 2026; Hu et al., 2025b; Yang et al., 2025; Song et al., 2025; Lian et al., 2025a; Luo et al., 2026)	Organize heterogeneous affective cues within unified semantic and generative spaces	Prompting; instruction tuning; multimodal foundation models	Limited representation-level interpretability and diagnostics

annotation scheme, the task can be formulated as discrete classification or continuous regression, with outputs ranging from emotion categories and polarity to valence–arousal–dominance scores and open-vocabulary emotion descriptions (Russell, 1980; Poria et al., 2019a;b; 2021; Baltrušaitis et al., 2018; Wang et al., 2026; Zhong et al., 2020). Representative benchmarks include CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), DEAP (Koelstra et al., 2011), CPED (Zhang et al., 2021), EMOTIC (Kosti et al., 2019), CASME II (Yan et al., 2014), MER-2024 (Lian et al., 2024), and MER-2025 (Lian et al., 2025b).

Methods in this area require both cross-modal integration and separation of affective information from modality-specific or non-affective factors. Existing approaches include shared–private representation learning such as MISA (Hazarika et al., 2020; Yang et al., 2022), contrastive alignment such as UniMSE (Hu et al., 2022), interaction and fusion models such as TFN (Zadeh et al., 2017) and MulT (Tsai et al., 2019), graph-based context modeling such as DialogueGCN (Ghosal et al., 2019), MMGCN (Hu et al., 2021), and GraphCFC (Li et al., 2023), and large-model approaches including AffectGPT (Lian et al., 2025a), FEALLM (Hu et al., 2025b), Omni-Emotion (Yang et al., 2025), and Emotion-o1 (Song et al., 2025). Classification commonly uses Accuracy, Precision, Recall, F1, Weighted-F1, and UAR (Poria et al., 2019a; Busso et al., 2008), whereas continuous affect prediction uses MAE, RMSE, Pearson correlation, and CCC (Lee et al., 2025; Zadeh et al., 2018; 2016).

A detailed functional mapping of the five representation-learning method families for affect detection and recognition is provided in Appendix E, Table 4.

6.2 Affective Reasoning

Affective reasoning extends emotion recognition toward explaining why an emotion arises, who or what it concerns, and how contextual events and mental states jointly contribute to its formation (Schölkopf et al., 2021; Poria et al., 2021; Luo et al., 2026). Unlike classification or intensity estimation, affective reasoning focuses on modeling the relationships among events, appraisals, causes, and emotional responses (Sander et al., 2005; Poria et al., 2021; Luo et al., 2026). Representative resources include RECCON (Poria et al., 2021), ECF (Wang et al., 2022), and the ECPE corpus (Xia & Ding, 2019).

Contextual information and external knowledge play particularly important roles in this setting. DialogueGCN (Ghosal et al., 2019) and dialogue memory networks (Hazarika et al., 2018) model speaker-level dependencies and historical conversational information, respectively, while ComFact (Gao et al., 2022) and COMET (Bosselut et al., 2019) introduce commonsense knowledge. ECR-Chain (Huang et al., 2024) adopts a stimulus–appraisal–emotion reasoning structure, HitEmotion (Luo et al., 2026) introduces mental-state-guided reasoning, and Emotion-o1 (Song et al., 2025) supports broader affective reasoning tasks. Since generated reasoning chains may not faithfully reflect the evidence underlying model predictions (Turpin et al., 2023), evaluation should consider not only task accuracy but also explanation quality. Existing studies commonly report Precision, Recall, F1, Macro-F1, Weighted-F1, Micro-F1, Exact Match F1, and semantic evaluation for open-ended outputs (Poria et al., 2021; Huang et al., 2024; Luo et al., 2026).

The functional roles of alignment, fusion, disentanglement, and robustness in affective reasoning are summarized in Appendix E, Table 5.

6.3 Affective Response and Support

Affective response and support extend emotion understanding toward the generation of contextually appropriate interaction. Affective response includes empathetic responses, emotional expressions, and multimodal affective descriptions (Rashkin et al., 2019; Lian et al., 2025a; Blume et al., 2024), while affective support further addresses counseling, companionship, and distress-management scenarios in which systems must also consider support strategies, user needs, and safety (Zhang et al., 2024; Liu et al., 2021). Representative datasets include EmpatheticDialogues (Rashkin et al., 2019), ESConv (Liu et al., 2021), MENTOS (Kim & Kim, 2026), PsyQA (Sun et al., 2021), CPsyCoun (Zhang et al., 2024), and MERCI (Althubyani et al., 2025).

These applications require multimodal representations to jointly incorporate the user's current affective state, conversational context, and prior interaction. Alignment and disentanglement help isolate and organize affect-relevant information, fusion models integrate multimodal and dialogue signals, context modeling captures speaker- and turn-level dependencies, self-supervised and robust learning address limited annotations and missing modalities, and large-model-based approaches further support response generation and strategy planning. Evaluation typically combines automatic generation metrics such as BLEU, ROUGE, and Distinct (Rashkin et al., 2019; Liu et al., 2021) with human judgments of empathy, contextual relevance, support-strategy appropriateness, and response consistency (Zhang et al., 2024; Liu et al., 2021).

An extended comparison of representation-learning methods for affective response and support is provided in Appendix E, Table 6.

6.4 Cross-task Generalization

Cross-task generalization examines whether affective representations remain effective when datasets, languages, cultures, modalities, or task objectives change (Baltrušaitis et al., 2018; Wang et al., 2026; Belay et al., 2025). Major challenges include distribution shift, cultural variation, incompatible annotation schemes, and missing modalities (Lin & Hu, 2023; Havaldar et al., 2023; Belay et al., 2025). Common benchmarks include CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2019a).

Representative approaches include UniMSE (Hu et al., 2022), which connects multimodal sentiment analysis and conversational emotion recognition through contrastive learning; MissModal (Lin & Hu, 2023), which constrains representations under both complete and missing inputs; XLM-EMO (Bianchi et al., 2022), which addresses multilingual emotion prediction; and AffectGPT (Lian et al., 2025a), which supports multimodal affect understanding. In this setting, alignment contributes to transferable affective semantics, while robustness becomes particularly important under distribution shifts and missing modalities. Evaluation depends on the target task and may include Accuracy, F1, Weighted-F1, MAE, RMSE, Pearson correlation, and the performance gap between in-domain and target-domain testing (Belay et al., 2025; Lian et al., 2026).

The functional profile of the five method families under cross-task generalization is provided in Appendix E, Table 7.

Table 2: Applications and evaluation of multimodal affective representations.

Application	Representation Requirement	Representative Datasets	Evaluation
Detection & Recognition	Discriminative affective semantics, multimodal complementarity, temporal/context modeling	MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), MER-2024 (Lian et al., 2024)	Accuracy, F1, UAR, MAE, RMSE, CCC
Affective Reasoning	Context, cause–emotion relations, mental-state and commonsense information	RECCON (Poria et al., 2021), ECF (Wang et al., 2022), ECPE (Xia & Ding, 2019)	Precision, Recall, F1, Macro-F1, semantic evaluation
Response & Support	Context-aware affect representation, strategy modeling, multimodal interaction	EmpatheticDialogues (Rashkin et al., 2019), ESConv (Liu et al., 2021), MENTOS (Kim & Kim, 2026), PsyQA (Sun et al., 2021)	BLEU, ROUGE, Distinct, human evaluation
Cross-task Generalization	Transferable semantics, robustness to domain and modality shifts	MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a)	Target-task metrics and transfer-performance gap

A complete comparison of the datasets used across these four application areas, including dataset scale, annotation scheme, and modality configuration, is provided in Appendix E, Table 8.

6.5 Analysis of Experimental Findings

Affect Detection and Recognition. This setting primarily examines whether a model can learn stable and discriminative multimodal affective representations. Typical tasks involve the prediction of discrete emotion categories, sentiment polarity, or continuous affective dimensions such as valence and arousal, making the effective integration of textual, acoustic, and visual information central to task performance (Russell, 1980; Poria et al., 2019a;b; Zadeh et al., 2016; 2018). Existing approaches improve multimodal complementarity and semantic consistency through mechanisms such as dynamic fusion, shared–private representation decomposition, and contrastive alignment, as exemplified by TFN, MulT, MISA, and UniMSE (Zadeh et al., 2017; Tsai et al., 2019; Hazarika et al., 2020; Hu et al., 2022). Accordingly, this setting primarily reflects the discriminability, complementarity, and cross-modal consistency of learned representations. However, task-level metrics such as Accuracy, F1, UAR, MAE, and RMSE remain insufficient to directly characterize the quality of the internal representations themselves (Poria et al., 2019a; Busso et al., 2008; Lee et al., 2025; Zadeh et al., 2018; 2016).

Affective Reasoning. This setting further examines whether a model can move beyond emotion recognition to organize and reason over relationships among events, emotion causes, mental states, and emotional outcomes (Sander et al., 2005; Poria et al., 2021; Luo et al., 2026). Rather than merely identifying what emotion is present, affective reasoning seeks to explain why the emotion arises and which contextual evidence supports the corresponding judgment. Datasets such as RECCON and ECF provide annotations of emotion causes or emotion–cause relations, while methods such as ECR-Chain and HitEmotion incorporate reasoning chains and mental-state modeling, respectively (Poria et al., 2021; Wang et al., 2022; Huang et al., 2024; Luo et al., 2026). Accordingly, this setting mainly reflects the ability of representations to organize contextual information, model relational structure, and support explanation, while also requiring careful assessment of whether generated explanations faithfully reflect the evidence underlying the model's predictions (Turpin et al., 2023).

Affective Response and Support. This setting primarily examines whether affective understanding can be translated into responses that are appropriate to the user's current state, dialogue history, and interaction goals (Rashkin et al., 2019; Liu et al., 2021). Representations in this setting must jointly preserve current affect, contextual information, and support strategies to enable empathetic response generation, emotional support, and mental-health-oriented interaction. EmpatheticDialogues, ESConv, PsyQA, CPsyCoun, MENTOS, and MERCI provide representative resources for empathy, support strategies, counseling, and personalized multimodal response generation (Rashkin et al., 2019; Liu et al., 2021; Sun et al., 2021; Zhang et al., 2024; Kim & Kim, 2026; Althubyani et al., 2025). Accordingly, this setting primarily reflects contextual sensitivity, dynamic adaptation, and generative utility. Its evaluation should therefore extend beyond surface-level generation metrics

such as BLEU, ROUGE, and Distinct to include empathy, relevance, support-strategy appropriateness, and response consistency (Rashkin et al., 2019; Liu et al., 2021; Zhang et al., 2024).

Cross-task Generalization. This setting primarily examines whether learned affective information remains useful when datasets, languages, cultures, modality configurations, or task objectives change (Baltrušaitis et al., 2018; Wang et al., 2026; Belay et al., 2025). Compared with fixed in-domain evaluation, cross-task generalization places greater emphasis on whether representations overfit dataset-specific distributions or annotation schemes. UniMSE, MissModal, XLM-EMO, and AffectGPT address generalization from the perspectives of task unification, missing-modality robustness, cross-lingual transfer, and multimodal affect understanding, respectively (Hu et al., 2022; Lin & Hu, 2023; Bianchi et al., 2022; Lian et al., 2025a). Accordingly, this setting primarily reflects representation transferability, robustness, and distributional stability, and can be more effectively assessed through performance changes between in-domain and out-of-domain settings, as well as between complete- and missing-modality conditions (Belay et al., 2025; Lian et al., 2026).

Overall, the four application settings evaluate affective representations from four complementary perspectives: discrimination, reasoning, generation, and transfer. This comparison suggests that representation quality should not be judged solely by performance on a single downstream task, but should instead be assessed jointly in terms of semantic consistency, contextual modeling, generative utility, and cross-domain robustness.

7 Conclusion

This review surveys multimodal affective representation learning from the perspectives of emotion theories, semantic descriptions, and multimodal observations, with emphasis on **fusion**, **disentanglement**, **alignment**, and **robust learning**. Existing studies show that different modalities contain both shared affective information and modality-specific cues, while differences in language, modality configuration, and annotation schemes limit direct cross-dataset comparison (Zadeh et al., 2018; Busso et al., 2008; Yu et al., 2020); cultural variation further complicates semantic alignment (Havaldar et al., 2023; Belay et al., 2025). Future research should place greater emphasis on representation consistency, factor separability, and robustness (Lin & Hu, 2023; Eastwood & Williams, 2018; Qian et al., 2026), while incorporating counterfactual fairness (Kusner et al., 2017), long-context analysis (Liu et al., 2025a), and auxiliary tools such as t-SNE and attention visualization (Wattenberg et al., 2016; Turpin et al., 2023; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Overall, evaluation should move beyond task performance toward a more direct assessment of representation quality.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11418–11426, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Amelia Aldao, Susan Nolen-Hoeksema, and Susanne Schweizer. Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical psychology review*, 30(2):217–237, 2010.

Mohammed Althubyani, Zhijin Meng, Shengyuan Xie, Francisco Cruz, Imran Razzak, Mukesh Prasad, Eduardo B Sandoval, and Baki Kocaballi. Merci: A multimodal dataset for personalised and emotionally-aware dialogues. In *2025 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–7. IEEE, 2025.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems*, volume 33, pp. 12449–12460, 2020.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016.

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

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.

Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. Culemo: Cultural lenses on emotion-benchmarking llms for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18894–18909, 2025.

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

Federico Bianchi, Debora Nozza, and Dirk Hovy. Xlm-emo: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pp. 195–203, 2022.

Florian Blume, Runfeng Qu, Pia Bideau, Martin Maier, Rasha Abdel Rahman, and Olaf Hellwich. How do you perceive my face? recognizing facial expressions in multi-modal context by modeling mental representations. In *DAGM German Conference on Pattern Recognition*, pp. 20–36. Springer, 2024.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4762–4779, 2019.

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

Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. Towards unbiased visual emotion recognition via causal intervention. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 60–69, 2022.

Mackenzie L Collins and T Claire Davies. Emotion differentiation through features of eye-tracking and pupil diameter for monitoring well-being. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4. IEEE, 2023.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

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

Paul Ekman, Tim Dalgleish, and M Power. Basic emotions. *San Francisco, USA*, 1, 1999.

Florian Eyben, Klaus R Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and Khiet P Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

Qi Fan, Haolin Zuo, Rui Liu, Zheng Lian, and Guanglai Gao. Learning noise-robust joint representation for multimodal emotion recognition under incomplete data scenarios. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 116–124, 2024.

Silin Gao, Jena D Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1656–1675, 2022.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pp. 3123–3128. Reykjavik, 2014.

Leslie S Greenberg. Emotion–focused therapy. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 11(1):3–16, 2004.

James J Gross. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26 (1):1–26, 2015.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 202–214, 2023.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2122–2132, 2018.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Karishma Hegde and Hemadri Jayalath. Emotions in the loop: A survey of affective computing for emotional support. *arXiv preprint arXiv:2505.01542*, 2025.

Clara E Hill. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*, 1999.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7837–7851, 2022.

Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, Ruichu Cai, Erik Cambria, and Hasti Seifi. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv e-prints*, pp. arXiv–2409, 2024.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021.

Juewen Hu, Yexin Li, Jiulin Li, Shuo Chen, and Pring Wong. Ecmf: Enhanced cross-modal fusion for multimodal emotion recognition in mer-semi challenge. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, pp. 8–12, 2025a.

Zhuozhao Hu, Kaishen Yuan, Xin Liu, Zitong Yu, Yuan Zong, Jingang Shi, Huanjing Yue, and Jingyu Yang. Feallm: Advancing facial emotion analysis in multimodal large language models with emotional synergy and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5677–5686, 2025b.

Jiahao Huang, Fengyan Lin, Xuechao Yang, Chen Feng, Kexin Zhu, Xu Yang, and Zhide Chen. Nano-emox: Unifying multimodal emotional intelligence from perception to empathy. *arXiv preprint arXiv:2603.02123*, 2026.

Zhaopei Huang, Jinming Zhao, and Qin Jin. Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains. *arXiv preprint arXiv:2405.10860*, 2024.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Takumi Karasawa, Satoki Hamanaka, Taiga Kume, Wataru Sasaki, Jin Nakazawa, and Tadashi Okoshi. Emogo: Emotion estimation on gazed object by using mobile eye-tracker. In *Proceedings of the 13th International Conference on the Internet of Things*, pp. 158–161, 2023.

Seulgi Kim and Harksoo Kim. A mental state extraction dataset for theory-of-mind-based reasoning in emotional support conversations. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pp. 2705–2723, 2026.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42 (11):2755–2766, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Victor AF Lamme. Towards a true neural stance on consciousness. *Trends in cognitive sciences*, 10 (11):494–501, 2006.

Yubeen Lee, Sangeun Lee, Chaewon Park, Junyeop Cha, and Eunil Park. Tagf: Time-aware gated fusion for multimodal valence-arousal estimation. *arXiv preprint arXiv:2507.02080*, 2025.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, 26:77–89, 2023.

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

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025a.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13837–13842, 2025b.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Hui Shen, Wendong Xu, Chaofan Tao, Min Yang, et al. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *arXiv preprint arXiv:2509.07403*, 2025a.

Yuanyuan Liu, Lin Wei, Kejun Liu, Zijing Chen, Zhe Chen, Chang Tang, Jingying Chen, and Shiguang Shan. Leveraging eye movement for instructing robust video-based facial expression recognition. *IEEE Transactions on Affective Computing*, 2025b.

Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao Fei, Mong-Li Lee, et al. Unveiling the cognitive compass: Theory-of-mind-guided multimodal emotion reasoning. *arXiv preprint arXiv:2602.00971*, 2026.

Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 53(4):1689–1696, 2021.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. *arXiv preprint arXiv:2602.19585*, 2026.

Rosalind W Picard. *Affective computing*. The MIT press, 1997.

Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pp. 3–33. Elsevier, 1980.

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 *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019a.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953, 2019b.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5370–5381, 2019.

Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3502–3506. IEEE, 2020.

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

David Sander, Didier Grandjean, and Klaus R Scherer. A systems approach to appraisal mechanisms in emotion. *Neural networks*, 18(4):317–352, 2005.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Erika H Siegel, Molly K Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, Jennifer Dy, Karen S Quigley, and Lisa Feldman Barrett. Emotion fingerprints or emotion populations? a meta-analytic investigation of autonomic features of emotion categories. *Psychological bulletin*, 144(4):343, 2018.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Changhao Song, Yazhou Zhang, Hui Gao, Kaiyun Huang, and Peng Zhang. Emotion-o1: Adaptive long reasoning for emotion understanding in llms. *arXiv preprint arXiv:2505.22548*, 2025.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, 2021.

Wenlong Tan and Bo Zhang. Dast-gan: an adversarial learning multimodal sentiment analysis model based on dynamic attention and spatio-temporal fusion. *Sensors*, 25(23):7109, 2025.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Ende Wang, Suyao Sun, Kejuan Xue, Yifan Ji, Taiyang Ma, Kuiming Jiang, and Jiageng Liu. A survey of representation learning for multi-modal emotion recognition. *Multimedia Tools and Applications*, 85(8):685, 2026.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844, 2022.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10): e2, 2016.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint arXiv:2505.20511*, 2025.

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1003–1012, 2019.

Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1642–1651, 2022.

Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*, 2025.

Gerard C Yeo and Desmond C Ong. Associations between cognitive appraisals and emotions: A meta-analytic review. *Psychological Bulletin*, 150(12):1440, 2024.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1103–1114, 2017.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2236–2246, 2018.

Raziyeh Zall, Alireza Kheyrkhah, Erik Cambria, Zahra Naseri, and M Reza Kangavari. Intelligent agents with emotional intelligence: Current trends, challenges, and future prospects. *arXiv preprint arXiv:2511.20657*, 2025.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13947–13966, 2024.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. Affective computing in the era of large language models: A survey from the nlp perspective. *Knowledge-Based Systems*, pp. 115411, 2026.

Yulin Zhang, Guozhen Zhao, Yezhi Shu, Yan Ge, Dan Zhang, Yong-Jin Liu, and Xianghong Sun. Cped: A chinese positive emotion database for emotion elicitation and analysis. *IEEE Transactions on Affective Computing*, 14(2):1417–1430, 2021.

Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. *Advances in Neural Information Processing Systems*, 38:142100–142129, 2026.

A Detailed Review Methodology

Literature search. We conducted systematic searches across IEEE Xplore, ScienceDirect, and Google Scholar, complemented by large-language-model-assisted supplementary retrieval. The primary search terms were "multimodal affective representation learning" and "multimodal emotion recognition." Additional queries used "feature disentanglement," "cross-modal alignment," and "missing modality" to broaden the coverage of studies on representation disentanglement, cross-modal semantic alignment, and robustness to missing modalities. Beyond database retrieval, ChatGPT, Claude, and Gemini were used to identify potentially relevant studies that might have been missed by the initial searches. Citation tracing of representative papers was further conducted to include studies on emotion theory and early affective representation methods. The initial search yielded 260 records. As shown in Fig. 3, removing 60 duplicates and excluding 50 records during title and abstract screening left 150 records for full-text assessment.

Full-text screening and eligibility assessment. The remaining records were examined with respect to their research objectives, modalities, methodological design, and evaluation procedures. Fifteen records were excluded because their methods were not relevant to the scope of this review, their evaluation or reporting was insufficient, or the full text was unavailable, leaving 135 articles. A further five articles were excluded because they did not address the core topics of affective representation, fusion, disentanglement, alignment, or robust learning. This screening process yielded 130 eligible papers.

Theoretical supplementation and classification. Citation tracing and snowballing were subsequently used to supplement the corpus with 16 references on emotion theory, early representation methods, and other relevant foundational work. The final review therefore included 146 references.

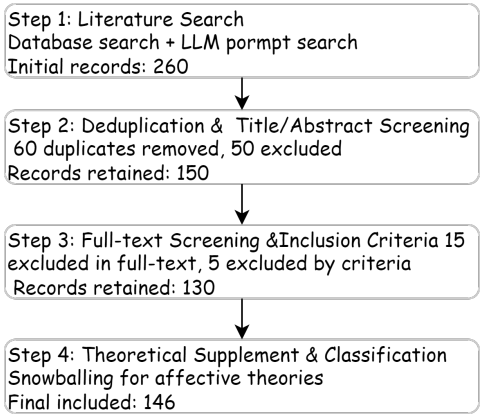

Figure 3: Flow diagram of literature search and selection.

B Detailed Foundations of Emotion Models

The design of an affective representation begins with how emotion itself is defined. Existing accounts can be organized into four complementary perspectives—categorical, dimensional, appraisal-based, and open-ended descriptions (Ekman, 1992; Russell, 1980; Sander et al., 2005; Barrett, 2017). These perspectives characterize emotion through basic categories, continuous valence–arousal spaces, event-appraisal processes, and context-dependent concepts expressed in natural language, respectively (Picard, 1997; Calvo & D'Mello, 2010). Their relationships are illustrated in Fig. 4. This section further compares their definitions, representative models, strengths, and limitations to clarify how different theories of emotion provide the semantic foundations for affective representation.

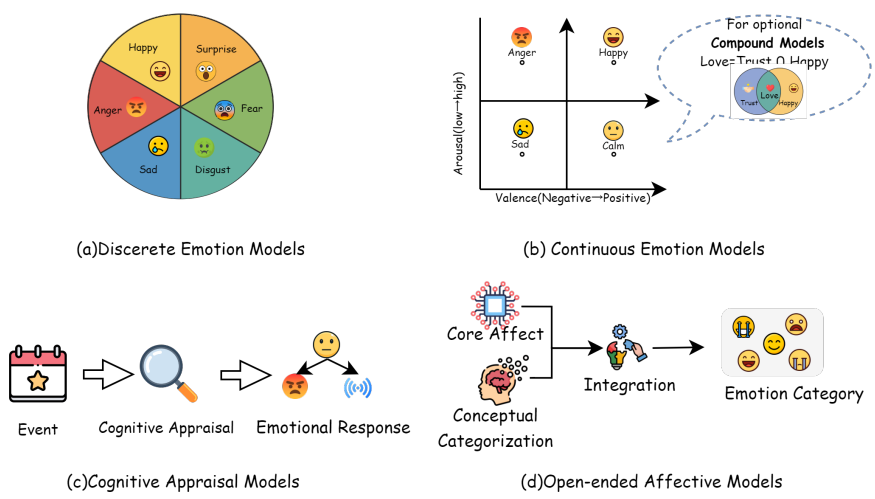

Figure 4: Models of emotion from categorical, dimensional, appraisal-based, and open-ended descriptive perspectives.

B.1 Categorical Models of Emotion

Categorical models represent emotion using a finite set of basic categories (Ekman, 1992; Ekman et al., 1999). Representative accounts include Ekman's six basic emotions—disgust, anger, fear, happiness, sadness, and surprise (Ekman, 1992; Ekman et al., 1999)—and Plutchik's eight emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation (Plutchik, 1980). These models provide an intuitive and explicit semantic space for emotion classification and support both single-label and multi-label descriptions (Ekman, 1992; Plutchik, 1980).

A major advantage of categorical models is that their label spaces and prediction targets are relatively well defined, making standardized annotation and evaluation comparatively straightforward (Ekman et al., 1999; Calvo & D'Mello, 2010). However, discretizing continuous emotional experience may obscure intensity, mixed states, and individual variation, while the granularity and coverage of predefined label sets constrain their ability to represent fine-grained, culturally specific, or open-set emotions (Russell, 1980; Barrett, 2017). Coarse categories tend to lose distinctions in intensity and composition, whereas increasingly fine-grained categories raise annotation costs and introduce greater ambiguity between class boundaries. Categorical models are therefore well suited to classification under predefined label spaces, but are insufficient on their own for comprehensively representing open-ended or fine-grained emotional experience (Barrett, 2017; Plutchik, 1980; Calvo & D'Mello, 2010).

B.2 Dimensional Models of Emotion

Dimensional models represent emotional states as positions in a continuous space rather than assigning them exclusively to discrete categories (Russell, 1980). Russell's two-dimensional circumplex model of valence and arousal is a representative example (Russell, 1980). Later formulations introduce dominance to form a valence–arousal–dominance (VAD) space, and these continuous dimensions are widely used for affect annotation and regression (Russell, 1980; Sander et al., 2005).

By preserving continuous variation and similarity among emotional states, dimensional models are well suited to intensity estimation, regression, mixed-emotion analysis, and modeling affective trajectories over time (Russell, 1980; Lee et al., 2025; Cowen & Keltner, 2017). Their limitations include cultural variation in the interpretation and annotation of affective dimensions, indistinct boundaries between regions of the continuous space, and differences in how annotators interpret numerical scores (Havaldar et al., 2023; Belay et al., 2025). More fundamentally, dimensional coordinates indicate the degree of an affective state but do not directly explain what caused it, who or what it concerns, or how the surrounding context shaped it (Sander et al., 2005; Barrett, 2017). Dimensional models are therefore particularly useful for representing affective intensity and change,

whereas categorical and open-ended descriptions can provide complementary information about causes, objects, and context (Russell, 1980; Sander et al., 2005; Barrett, 2017).

B.3 Appraisal-Based Models of Emotion

Appraisal-based models conceptualize emotion as arising from an individual's evaluation of events and provide a theoretical basis for representing emotional causes and contextual relations (Sander et al., 2005; Yeo & Ong, 2024). A representative account is Scherer's component process model, in which emotion emerges through a sequence of rapid appraisals concerning event relevance, implications for personal goals, and the individual's ability to cope with the situation (Sander et al., 2005). For affective computing, this perspective suggests that representations should capture not only an observable emotion label, but also a relational chain describing what happened, whose goals were affected, how the event was evaluated, and what response followed (Sander et al., 2005; Schölkopf et al., 2021). Research on emotion regulation further indicates that appraisal processes and regulation strategies jointly shape emotional experience and expression (Gross, 2015; Aldao et al., 2010).

Appraisal-based models can explain why an emotion arises, what it is directed toward, and how situational factors and individual responses jointly contribute to it (Sander et al., 2005; Poria et al., 2021). However, appraisal processes are often rapid, implicit, and difficult to observe directly, which makes them challenging to annotate and model; datasets that explicitly record the corresponding causal relations are also relatively limited (Sander et al., 2005; Schölkopf et al., 2021). This perspective therefore extends affective computing beyond category and intensity prediction toward modeling structured relationships among events, goals, appraisals, and responses (Sander et al., 2005; Schölkopf et al., 2021; Poria et al., 2021).

B.4 Open-Ended Descriptions of Emotion

Open-ended descriptions emphasize the roles of context and conceptualization in emotion, consistent with the theory of constructed emotion (Barrett, 2017). Rather than restricting annotations to a predefined label set, they allow emotional states, their causes, their objects, and their surrounding contexts to be expressed in natural language (Barrett, 2017; Schölkopf et al., 2021). In Barrett's account, emotion is not generated by a fixed set of hardwired modules; instead, the brain conceptualizes and categorizes internal and external sensory information, with core affect providing basic experiential information and concepts, language, social norms, and episodic memory jointly contributing to the construction of specific emotional experiences (Barrett, 2017). Cognitive behavioral therapy and emotion-focused therapy provide additional perspectives on the malleability and context dependence of emotion (Beck, 2020; Greenberg, 2004), while the exploration–insight–action framework of helping skills offers another reference for understanding stages of affective processing (Hill, 1999). Compared with fixed labels, open-ended descriptions can represent complex states such as "exhaustion tinged with relief," while also preserving mixed emotions, differences in intensity, causes, objects, contextual details, and fine-grained meanings that may otherwise remain ambiguous (Barrett, 2017; Poria et al., 2021).

Such descriptions provide richer supervision for affective understanding with foundation models, enabling models to learn not only emotion categories but also the linguistic relationships among emotion concepts, contextual semantics, and observable cues (Barrett, 2017; Huang et al., 2024; Song et al., 2025). At the same time, their flexibility introduces greater semantic ambiguity, annotator variation, and evaluation difficulty, making them more informative when used in conjunction with categorical and dimensional descriptions (Barrett, 2017; Havaldar et al., 2023; Belay et al., 2025). For affective computing, this perspective implies that representations should accommodate conceptual knowledge and context-dependent semantic variation, including underlying affective states, the organization of emotion concepts, and the influence of contextual expectations on emotion categorization (Barrett, 2017; Schölkopf et al., 2021; Lian et al., 2025a).

C Detailed Modality-Specific Affective Information

C.1 Textual Modality

Multimodal affective representations draw on heterogeneous information from text, speech, visual behavior, physiological signals, and eye movements (Koelstra et al., 2011; Baltrušaitis et al., 2018; Wang et al., 2026; Eyben et al., 2015; Baltrušaitis et al., 2016; Zhong et al., 2020; Karasawa et al., 2023), each of which captures emotion through a distinct observation and expression mechanism. Text primarily conveys semantic content and cognitive appraisal (Devlin et al., 2019; Socher et al., 2013); speech provides prosodic and intonational cues (Eyben et al., 2015; Baevski et al., 2020; Hsu et al., 2021); visual observations capture facial expressions and body posture (Baltrušaitis et al., 2016; He et al., 2016; Dosovitskiy, 2020; Kosti et al., 2019); physiological signals reflect internal bodily states (Koelstra et al., 2011; Zhong et al., 2020; Siegel et al., 2018); and eye movements provide complementary information about attention and cognitive load (Karasawa et al., 2023; Collins & Davies, 2023; Liu et al., 2025b). Because no single modality captures the full range of an affective experience, combining these heterogeneous sources can compensate for information that is unavailable from single-modality observations (Baltrušaitis et al., 2018; Wang et al., 2026).

C.2 Speech Modality

Speech conveys paralinguistic information through prosodic and vocal characteristics such as pitch, speaking rate, energy, rhythm, pauses, and voice quality (Eyben et al., 2015). These cues are associated with valence, arousal, and attitude, and can represent affective intensity and temporal variation that written language alone may not reveal (Russell, 1980; Eyben et al., 2015; Siegel et al., 2018). For example, sadness is often associated with a slower speaking rate, whereas anger may be accompanied by higher pitch; prosody also contributes to the perception of speaking style, emotional state, and social intent (Eyben et al., 2015; Siegel et al., 2018; Rizos et al., 2020). However, affective variation in speech is entangled with linguistic content, speaker identity, and habitual vocal characteristics. Similar prosodic patterns may arise from differences in speaker, accent, habitual speaking rate, or emotional state, making emotion-related intonation difficult to separate from speaker-specific vocal characteristics (Hazarika et al., 2020; Yang et al., 2022; Rizos et al., 2020).

Speech representation learning must therefore preserve affective cues while reducing interference from speaker identity and linguistic content, and must also establish correspondences with other modalities (Hazarika et al., 2020; Yang et al., 2022; Li et al., 2024). Speech is commonly paired with text to integrate prosodic and semantic information, or with visual observations to relate vocal delivery to facial expressions and body posture (Tsai et al., 2019; Baltrušaitis et al., 2018; Wang et al., 2026). Prosodic cues are widely used in speech emotion recognition, conversational emotion analysis, and depression detection (Eyben et al., 2015; Poria et al., 2019a; Gratch et al., 2014). Nevertheless, speech alone may provide insufficient semantic detail or visual context for interpreting an interaction (Tsai et al., 2019; Baltrušaitis et al., 2018), making multimodal integration particularly important when a task requires a more complete account of affect (Baltrušaitis et al., 2018; Wang et al., 2026).

C.3 Visual Modality

The visual modality captures embodied affective information through facial expressions, body posture, gestures, and movement, including expression-related valence and arousal, facial action units, and temporal changes in posture (Baltrušaitis et al., 2016; Kosti et al., 2019; Yan et al., 2014). These observations convey basic emotions, attitudes, and interactional intentions, while complementing linguistic and vocal cues (Ekman, 1992; Baltrušaitis et al., 2016; Kosti et al., 2019). Facial expressions and posture can signal states such as happiness, anger, and fear (Baltrušaitis et al., 2016; Kosti et al., 2019; Yan et al., 2014); however, their interpretation is highly context dependent, and the same expression may be perceived differently across situations and individuals (Kosti et al., 2019; Chen et al., 2022; Lamme, 2006). Cultural conventions further influence both emotional expression and recognition (Havaldar et al., 2023; Belay et al., 2025). In addition, microexpressions are brief and may be missed (Yan et al., 2014), while occlusion, lighting, and camera viewpoint can degrade the quality of visual observations (Baltrušaitis et al., 2016; Kosti et al., 2019). Facial and bodily signals

may also encode identity, movement, and interactional intent in addition to affective information (Hazarika et al., 2020; Yang et al., 2022; Meng et al., 2026).

Visual representation learning must therefore capture subtle and transient affective cues while accounting for contextual and identity-related variation (Hazarika et al., 2020; Yang et al., 2022; Meng et al., 2026). Visual observations are often combined with text or speech to provide semantic and prosodic information that images or videos cannot reveal on their own (Tsai et al., 2019; Baltrušaitis et al., 2018; Wang et al., 2026). Facial expression recognition and affective video analysis rely strongly on visual information (Baltrušaitis et al., 2016; Kosti et al., 2019; Yan et al., 2014), whereas conversational emotion analysis typically integrates text, speech, and vision (Poria et al., 2019a;b). Because each modality observes affect through a different mechanism, multimodal integration becomes particularly useful when affective evidence is distributed across multiple channels (Baltrušaitis et al., 2018; Wang et al., 2026).

C.4 Physiological Modality

Physiological signals characterize internal responses associated with the autonomic and central nervous systems. Common measurements include electrodermal activity, heart rate, heart rate variability, respiration, electromyography, and electroencephalography (EEG) (Koelstra et al., 2011; Zhong et al., 2020; Makowski et al., 2021). These signals are related to valence, arousal, and emotion regulation and, because they are less directly controlled through deliberate expression, can be useful for tracking changes in arousal and stress (Koelstra et al., 2011; Siegel et al., 2018). EEG may additionally provide information related to cognitive and affective appraisal (Zhong et al., 2020). However, physiological responses exhibit substantial individual variation, and measurements can be affected by baseline activity, age, health, medication, and task context (Koelstra et al., 2011; Siegel et al., 2018). Different emotions may also produce similar physiological patterns (Siegel et al., 2018), meaning that such signals do not directly specify emotional semantics or map uniquely onto emotion categories (Zhong et al., 2020; Siegel et al., 2018). Their collection also frequently requires contact-based equipment (Koelstra et al., 2011).

Physiological representation learning must therefore address individual variation, the gap between bodily responses and emotion semantics, and the practical constraints of naturalistic interaction. Personalized modeling, multimodal fusion, and interpretable representations provide relevant directions (Zhong et al., 2020; Siegel et al., 2018; Yang et al., 2022). Combining physiological signals with text, speech, or visual observations can further relate internal bodily responses to subjective meaning and observable behavior (Koelstra et al., 2011; Baltrušaitis et al., 2018; Wang et al., 2026). Physiological measurements are particularly relevant to stress recognition, emotion regulation, and clinical monitoring (Koelstra et al., 2011; Siegel et al., 2018; Gratch et al., 2014), whereas conversational emotion analysis in natural settings often relies more heavily on text, speech, and vision (Poria et al., 2019a;b). Their contribution to multimodal affective representation is therefore most informative when they provide complementary information that is not captured by the other available modalities (Baltrušaitis et al., 2018; Wang et al., 2026).

C.5 Eye-Movement Modality

Eye movements characterize visual attention and cognitive processing through measurements such as fixations, saccades, pupil diameter, blinks, and fixation duration (Karasawa et al., 2023; Collins & Davies, 2023). These measurements are associated with attention allocation, cognitive load, affective arousal, and attentional biases toward emotional stimuli (Collins & Davies, 2023; Liu et al., 2025b). Their high temporal resolution enables continuous observation of cognitive and affective processing, while pupil dilation and fixation patterns can serve as measurable indicators of cognitive and affective load (Collins & Davies, 2023). However, eye movements are influenced by task instructions, lighting, screen calibration, habitual viewing patterns, and fatigue. Their relationship with any particular emotional state is not one-to-one, and inter-participant differences further complicate generalization (Collins & Davies, 2023; Liu et al., 2025b).

Eye-movement representation learning must therefore account for individual variation, task dependence, and the interpretation of gaze behavior in relation to emotion. Context modeling, participant alignment, and multimodal fusion can help address these issues (Karasawa et al., 2023; Collins & Davies, 2023; Liu et al., 2025b). Eye tracking can also be combined with physiological signals,

visual observations, text, or speech to provide complementary information about attention and cognitive load (Karasawa et al., 2023; Collins & Davies, 2023). It is particularly relevant to affective attentional bias, advertising responses, and emotional experience in human–computer interaction (Karasawa et al., 2023; Collins & Davies, 2023; Liu et al., 2025b), but is less frequently used on its own in tasks such as speech emotion recognition (Eyben et al., 2015). Its value within a multimodal affective representation therefore depends on whether attention-related information adds evidence that cannot be captured by the other available observations (Baltrušaitis et al., 2018; Wang et al., 2026).

D Extended Method Taxonomy

This appendix provides an extended taxonomy of the five major families of multimodal affective representation learning methods discussed in Section 5. While the main text summarizes their primary objectives, representative mechanisms, and limitations, the extended taxonomy further describes their core characteristics and typical application scenarios. Table 3 presents the detailed comparison.

D.1 Representative Architectures

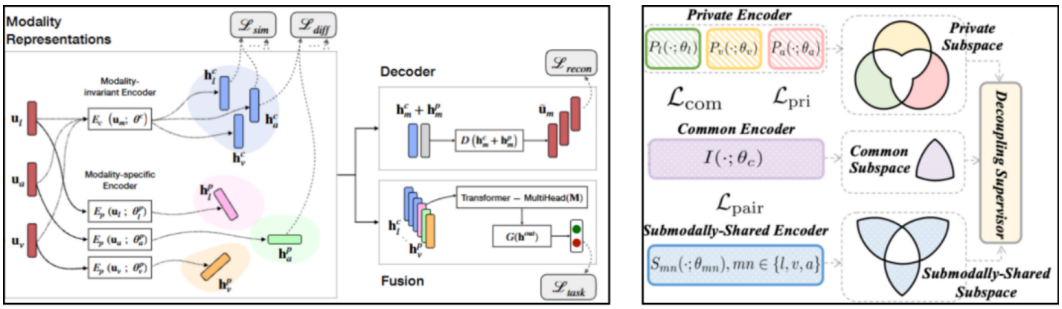

Figure 5: Representative architectures for multimodal representation disentanglement. The left panel illustrates MISA's modality-invariant and modality-specific representation learning (Hazarika et al., 2020), while the right panel illustrates tri-subspace disentanglement with private, common, and submodality-shared representations (Yang et al., 2022).

E Extended Application Analysis

This appendix provides an extended analysis of the four application areas introduced in Section 6, including task definitions, representative datasets, method families, functional mappings, and evaluation criteria for affect detection and recognition, affective reasoning, affective response and support, and cross-task generalization. Figure 7 summarizes their overall relationships, while Tables 4–7 provide detailed functional comparisons across method families. A complete comparison of the datasets is provided in Table 8.

E.1 Affect Detection and Recognition

Affect detection and recognition enable machines to identify emotion categories or affective polarity and estimate affective intensity from multimodal observations. These tasks constitute a central component of affective computing (Picard, 1997; Calvo & D'Mello, 2010). Depending on the annotation scheme, they can be formulated as either discrete classification or continuous regression. The former predicts predefined categories such as happiness, sadness, and anger, whereas the latter estimates continuous dimensions such as valence, arousal, and dominance. Related tasks further extend to conversational emotion recognition and emotion-cause identification (Russell, 1980; Poria et al., 2019a;b; 2021). Inputs may include text, speech, visual, physiological, and eye-tracking

Table 3: Extended taxonomy of multimodal affective representation learning methods.

Method Family	**Representation Function**	**Core Characteristics**	**Representative Applications**
Cross-modal Alignment and Disentangled Representation	Separate shared and modality-specific factors while establishing cross-modal semantic correspondence	Maps heterogeneous textual, acoustic, and visual features into comparable representation spaces; decomposes shared and modality-specific information; commonly employs contrastive learning, shared–private decomposition, and hierarchical alignment to reduce semantic gaps across modalities	Emotion recognition; fine-grained emotion analysis; cross-speaker and cross-cultural affect modeling
Cross-modal Interaction and Dynamic Fusion	Adaptively integrate complementary information across modalities	Models inter-modal interactions through cross-modal attention, gating, tensor fusion, and Transformer-based architectures; dynamically adjusts modality contributions according to samples, temporal states, or modality reliability rather than relying on fixed concatenation or static weighting	Multimodal emotion recognition; valence–arousal estimation; emotion intensity regression; dynamic affect monitoring
Context and Dialogue Structure Modeling	Encode temporal, speaker, and interaction dependencies into affective representations	Uses graphs, hypergraphs, memory mechanisms, or structured contextual modeling to capture speaker relations, inter-utterance dependencies, emotional evolution, and long-range dialogue context, extending utterance-level representation toward structured conversational modeling	Conversational emotion recognition; affective reasoning; empathetic dialogue; multi-party interaction analysis
Self-supervised and Robust Learning	Preserve representation stability under missing, noisy, or incomplete observations	Uses self-supervised pretraining, masking, reconstruction, consistency constraints, and contrastive objectives to reduce dependence on labeled data and maintain transferable affective information under non-ideal observation conditions	Missing-modality emotion recognition; cross-domain affect analysis; low-resource learning; wearable affect monitoring
Large Models and Prompt Learning	Unify heterogeneous affective cues within shared semantic and generative spaces	Exploits pretrained knowledge in large language and multimodal foundation models through prompting, instruction tuning, and multimodal interaction; uses natural language as a unified interface and extends affective modeling from discriminative prediction toward generative understanding and reasoning	Affective reasoning; empathetic response generation; emotional support; open-ended emotion description

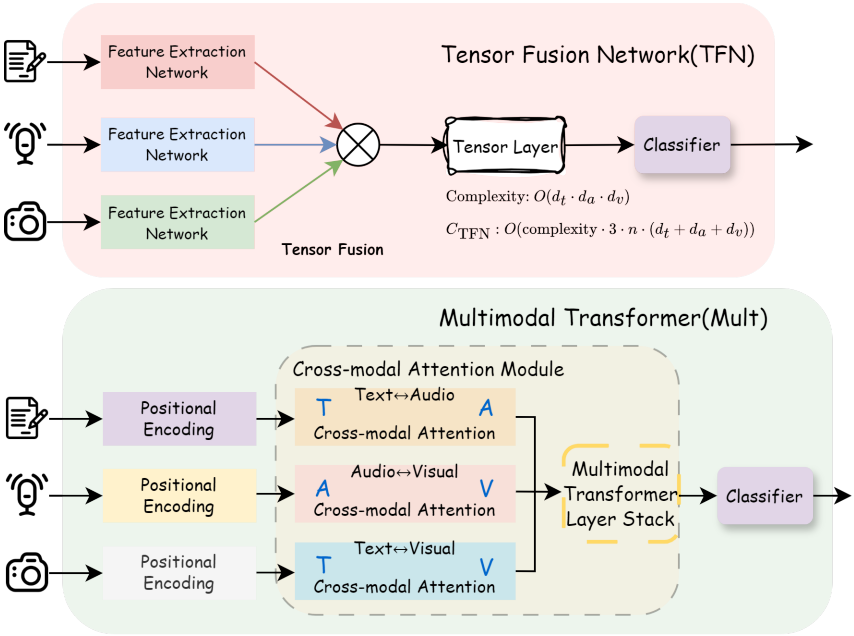

Figure 6: Representative architectures for multimodal interaction and dynamic fusion. TFN models explicit high-order cross-modal interactions through tensor fusion (Zadeh et al., 2017), whereas MulT captures cross-modal dependencies through directional cross-modal attention (Tsai et al., 2019).

data, while outputs can take the form of categorical labels, affective polarity, continuous scores, or open-vocabulary emotion labels (Baltrušaitis et al., 2018; Wang et al., 2026; Zhong et al., 2020).

Commonly used datasets include CMU-MOSI, CMU-MOSEI, CH-SIMS, IEMOCAP, MELD, DEAP, CPED, EMOTIC, CASME II, and MER-2024. CMU-MOSI (Zadeh et al., 2016) contains 2,199 video segments with textual, acoustic, and visual information and sentiment intensity scores ranging from -3 to $+3$. CMU-MOSEI (Zadeh et al., 2018) contains 22,856 segments and provides emotion annotations in addition to sentiment scores. CH-SIMS (Yu et al., 2020) contains 2,281 segments with both unimodal and multimodal labels. IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a) are widely used for conversational emotion recognition. DEAP (Koelstra et al., 2011) and CPED (Zhang et al., 2021) support the study of physiological affective responses, whereas EMOTIC (Kosti et al., 2019) and CASME II (Yan et al., 2014) provide visual affective data. MER-2024 (Lian et al., 2024) introduces open-vocabulary emotion recognition, and MER-2025 (Lian et al., 2025b) further includes tasks involving fine-grained emotions and affective descriptions.

These applications require models to extract affective cues from heterogeneous inputs while accounting for modality differences, temporal variation, and conversational context. For cross-modal alignment and representation disentanglement, MISA (Hazarika et al., 2020) and related methods (Yang et al., 2022) learn shared and modality-specific information, while UniMSE (Hu et al., 2022) uses contrastive learning to connect multimodal sentiment analysis with conversational emotion recognition. For cross-modal interaction and dynamic fusion, TFN (Zadeh et al., 2017) models modality interactions through tensor outer products, whereas MulT (Tsai et al., 2019) employs cross-modal attention to relate unaligned sequences. DialogueGCN (Ghosal et al., 2019), MMGCN (Hu et al., 2021), and GraphCFC (Li et al., 2023) use graph structures to model relationships among utterances and modalities; graph neural networks have also been used to learn relationships among EEG channels (Zhong et al., 2020). Finally, AffectGPT (Lian et al., 2025a), FEALLM (Hu et al., 2025b), Omni-Emotion (Yang et al., 2025), and Emotion-o1 (Song et al., 2025) extend large-model-based approaches to multimodal affect understanding, facial affect analysis, and affective reasoning. The functional contributions of these method families are summarized in Table 4.

For evaluation, discrete emotion classification commonly uses Accuracy, Precision, Recall, F1, Weighted-F1, and unweighted average recall (UAR) (Poria et al., 2019a; Busso et al., 2008).



Figure 7: Overview of applications of multimodal affective representations. Affect detection and recognition are illustrated with representative benchmarks and methods including CMU-MOSI, CMU-MOSEI, CH-SIMS, IEMOCAP, MELD, DEAP, EMOTIC, MER-2024, TFN, MulT, MISA, DialogueGCN, MMGCN, GraphCFC, UniMSE, AffectGPT, FEALLM, and Omni-Emotion (Zadeh et al., 2016; 2018; Yu et al., 2020; Busso et al., 2008; Poria et al., 2019a; Koelstra et al., 2011; Kosti et al., 2019; Lian et al., 2024; Zadeh et al., 2017; Tsai et al., 2019; Hazarika et al., 2020; Ghosal et al., 2019; Hu et al., 2021; Li et al., 2023; Hu et al., 2022; Lian et al., 2025a; Hu et al., 2025b; Yang et al., 2025). Affective reasoning is represented by RECCON, ECF, ECPE, ECR-Chain, HitEmotion, and Emotion-o1 (Poria et al., 2021; Wang et al., 2022; Xia & Ding, 2019; Huang et al., 2024; Luo et al., 2026; Song et al., 2025). Affective response and support include EmpatheticDialogues, ESConv, PsyQA, CPsyCoun, MENTOS, and MERCI (Rashkin et al., 2019; Liu et al., 2021; Sun et al., 2021; Zhang et al., 2024; Kim & Kim, 2026; Althubyani et al., 2025). Cross-task generalization is represented by UniMSE, MissModal, XLM-EMO, and AffectGPT (Hu et al., 2022; Lin & Hu, 2023; Bianchi et al., 2022; Lian et al., 2025a).

Table 4: Functional roles of multimodal affective representation methods in affect detection and recognition.

Method Family	Alignment	Complementarity	Disentanglement	Robustness
Cross-modal Alignment and Disentangled Representation	●	○	●	○
Cross-modal Interaction and Dynamic Fusion	○	●	○	●
Context and Dialogue Structure Modeling	○	●		○
Self-supervised and Robust Learning	○	○	○	●
Large Models and Prompt Learning	●	○	○	○

Note: ● denotes a dominant role; ○ denotes a secondary or auxiliary role; blank cells indicate a non-primary role.

Continuous-score and intensity prediction are commonly evaluated using mean absolute error (MAE), root mean squared error (RMSE), Pearson's correlation coefficient, and the concordance correlation coefficient (CCC) (Lee et al., 2025; Zadeh et al., 2018; 2016). Conversational emotion recognition typically reports utterance-level classification performance, with additional attention to the effects of class imbalance (Ghosal et al., 2019; Poria et al., 2019a).

E.2 Affective Reasoning

Affective reasoning extends emotion recognition toward inferring the causes, objects, context, and associated mental states underlying an emotion. It asks why an emotion arises, who or what it concerns, and how the surrounding situation shapes it, thereby moving affective understanding from classification toward explanation (Schölkopf et al., 2021; Poria et al., 2021; Luo et al., 2026). Inputs typically consist of multimodal dialogues or videos together with contextual information, while outputs may include identified causes, explanations, mental states, or reasoning chains. Whereas affect detection primarily predicts categories or intensity, affective reasoning emphasizes relationships among events, cognitive appraisals, and emotional responses (Sander et al., 2005; Poria et al., 2021; Luo et al., 2026).

Common datasets in this area include RECCON, ECF, and the ECPE corpus. RECCON (Poria et al., 2021) annotates emotion causes in conversations and supports both emotion-causing utterance identification and cause-span extraction. ECF (Wang et al., 2022) contains 1,374 multimodal dialogues and 13,619 utterances, annotated with emotion categories and emotion–cause utterance pairs. The ECPE corpus (Xia & Ding, 2019) contains 1,945 Chinese documents annotated with emotion-clause–cause-clause pairs.

Affective reasoning requires models to identify context relevant to an emotion and construct an interpretable inference process. DialogueGCN (Ghosal et al., 2019) and dialogue memory networks (Hazarika et al., 2018) model speaker relationships and prior emotional states. ComFact (Gao et al., 2022) provides links to relevant commonsense facts, while COMET (Bosselut et al., 2019) generates commonsense relations that can support emotion-cause inference. ECR-Chain (Huang et al., 2024) follows a stimulus–appraisal–emotion sequence, HitEmotion (Luo et al., 2026) introduces a mental-state-guided reasoning chain and the TMPO method, and Emotion-o1 (Song et al., 2025) employs adaptive reasoning for broader affective understanding. Generated reasoning chains nevertheless require careful evaluation because a verbal explanation may not faithfully reflect the basis of a model's actual judgment (Turpin et al., 2023). The functional emphasis of these method families in affective reasoning is summarized in Table 5.

Table 5: Functional roles of multimodal affective representation methods in affective reasoning.

Method Family	**Alignment**	**Complementarity**	**Disentanglement**	**Robustness**
Cross-modal Alignment and Disentangled Representation	●	○	●	○
Cross-modal Interaction and Dynamic Fusion	○	●	○	○
Context and Dialogue Structure Modeling	○	●		○
Self-supervised and Robust Learning	○	○	○	●
Large Models and Prompt Learning	●	○	○	○

Note: ● denotes a dominant role; ○ denotes a secondary or auxiliary role; blank cells indicate a non-primary role.

For evaluation, emotion-cause identification and cause-span extraction commonly use Precision, Recall, and F1 (Poria et al., 2021). ECR-Chain (Huang et al., 2024) reports Macro-F1 for emotion-causing utterance prediction and uses GPT-4 and Claude 3 scores to evaluate generated reasoning chains. HitEmotion (Luo et al., 2026) reports Accuracy, Weighted-F1, Micro-F1, and Exact Match F1 for its respective subtasks, together with semantic scoring for open-ended answers.

E.3 Affective Response and Support

Affective response and affective support are closely related interaction-oriented tasks. Affective response involves generating empathetic responses, emotional expressions, or multimodal affective descriptions appropriate to the user's state, with emphasis on emotional resonance and semantic coherence (Rashkin et al., 2019; Lian et al., 2025a; Blume et al., 2024). Affective support further extends to mental-health, counseling, and companionship settings, where a system may need to provide strategic reassurance, help users manage distress, or offer guidance while accounting for individual needs and safety (Zhang et al., 2024; Liu et al., 2021). Inputs may include dialogue history, user utterances, and multimodal signals, while outputs may take the form of empathetic responses, support strategies, or affective descriptions in text, speech, or visual form (Lian et al., 2025b;a).

Common datasets include EmpatheticDialogues (Rashkin et al., 2019), ESConv (Liu et al., 2021), MENTOS (Kim & Kim, 2026), PsyQA (Sun et al., 2021), CPsyCoun (Zhang et al., 2024), and MERCI (Althubyani et al., 2025). EmpatheticDialogues contains approximately 25,000 conversations, ESConv contains 1,053 multi-turn support conversations, MENTOS provides turn-level belief, emotion, and intent annotations, PsyQA provides approximately 22,000 Chinese mental-health support questions and 56,000 extended responses, CPsyCoun reconstructs Chinese multi-turn counseling conversations, and MERCI contains human–robot dialogues with audio, video, text, and user information.

These applications require models to transform affective understanding into appropriate responses and adapt support strategies across dialogue turns. Alignment and disentanglement separate affect-relevant signals from irrelevant factors, fusion integrates multimodal and conversational evidence, context modeling captures speaker- and turn-level dependencies, robust learning handles limited annotations and missing modalities, and large models support generation and strategy planning. EmpatheticDialogues (Rashkin et al., 2019) and ESConv (Liu et al., 2021) focus on empathetic responses and support strategies, PsyQA (Sun et al., 2021) and CPsyCoun (Zhang et al., 2024) support long-form and multi-turn counseling generation, and MERCI (Althubyani et al., 2025) supports personalized multimodal response modeling. The corresponding functional mapping is summarized in Table 6.

Table 6: Functional roles of multimodal affective representation methods in affective response and support.

Method Family	**Alignment**	**Complementarity**	**Disentanglement**	**Robustness**
Cross-modal Alignment and Disentangled Representation	●	○	●	○
Cross-modal Interaction and Dynamic Fusion	○	●	○	●
Context and Dialogue Structure Modeling		●		○
Self-supervised and Robust Learning	○	○	○	●
Large Models and Prompt Learning	●	○	○	○

Note: ● denotes a dominant role; ○ denotes a secondary or auxiliary role; blank cells indicate a non-primary role.

For evaluation, BLEU and ROUGE measure overlap between generated responses and reference texts, while Distinct metrics assess diversity of expression (Rashkin et al., 2019; Liu et al., 2021). Affective support additionally requires evaluation of empathy, contextual relevance, support-strategy appropriateness, and consistency across dialogue turns. Human evaluation can further assess whether a response adequately addresses the user's needs (Zhang et al., 2024; Liu et al., 2021).

E.4 Cross-task Generalization

Cross-task generalization examines whether affective representations and models can transfer across datasets, languages, cultures, and tasks. It asks whether affective knowledge acquired in a source

task or domain remains useful in an unseen target domain while preserving distinctions in emotion categories, intensity, and causes (Baltrušaitis et al., 2018; Wang et al., 2026; Belay et al., 2025). Inputs consist of multimodal samples from the source domain and observations from the target domain, while outputs may include target-domain emotion categories, intensity scores, explanations, or predictions for different tasks. Distribution shift, cultural variation, incompatible annotation schemes, and missing modalities constitute key challenges (Lin & Hu, 2023; Havaldar et al., 2023; Belay et al., 2025).

Common datasets include CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2019a). These resources support the evaluation of cross-corpus sentiment transfer and conversational emotion transfer.

UniMSE (Hu et al., 2022) uses contrastive learning to connect multimodal sentiment analysis with conversational emotion recognition. MissModal (Lin & Hu, 2023) constrains representations under complete and incomplete inputs. XLM-EMO (Bianchi et al., 2022) predicts emotions in multilingual social-media text, while AffectGPT (Lian et al., 2025a) addresses multimodal affect understanding. The corresponding functional profile is summarized in Table 7.

Table 7: Functional roles of multimodal affective representation methods in cross-task generalization.

Method Family	**Alignment**	**Complementarity**	**Disentanglement**	**Robustness**
Cross-modal Alignment and Disentangled Representation	●	○	●	○
Cross-modal Interaction and Dynamic Fusion	○	●	○	●
Context and Dialogue Structure Modeling		●		○
Self-supervised and Robust Learning	○	○	○	●
Large Models and Prompt Learning	●	○	○	○

Note: ● denotes a dominant role; ○ denotes a secondary or auxiliary role; blank cells indicate a non-primary role.

In cross-task generalization, alignment and robustness play particularly prominent roles. Alignment establishes transferable affective semantics, whereas robust learning addresses distribution shifts and missing modalities; fusion and disentanglement provide complementary support. Evaluation depends on the target task. Classification commonly uses Accuracy, F1, and Weighted-F1, whereas regression uses MAE, RMSE, and Pearson correlation. The performance gap between target-domain and in-domain testing can also be used to quantify generalization (Belay et al., 2025; Lian et al., 2026).

E.5 Datasets across Application Areas

The datasets used across the four application areas differ in scale, modality configuration, and annotation scheme. Detection and recognition datasets cover discrete categories, continuous affect dimensions, and open-vocabulary labels; reasoning datasets emphasize emotion causes and cause–emotion relations; response and support datasets focus on empathy, support strategies, and counseling interaction. The complete comparison is provided in Table 8.

Table 8: Datasets for applications of multimodal affective representations.

Applicable Task	Dataset	Scale	Annotation System	Modality
Emotion Detection and Recognition	DEAP (Koelstra et al., 2011)	32 subjects × 40 videos	Valence/Arousal/Dominance, etc.	P
	CPED (Zhang et al., 2021)	Database scale	Chinese positive emotion elicitation analysis	V/P
	EMOTIC (Kosti et al., 2019)	23,788 images	26 discrete emotion categories + VAD	V
	CASME II (Yan et al., 2014)	247 video clips	Micro-expression Action Units (AU)	V
	MER-2024 (Lian et al., 2024)	5,030 labeled + 115,595 unlabeled	6 categories + open vocabulary	T/A/V
	MER-2025 (Lian et al., 2025b)	369 labeled + 124,802 unlabeled	Fine-grained emotion + personality tags	T/A/V
Emotion Detection and Recognition; Cross-task Generalization	CH-SIMS (Yu et al., 2020)	2,281 clips	Continuous scores + single/multimodal tags	T/A/V
	CMU-MOSI (Zadeh et al., 2016)	2,199 clips	Continuous emotion scores [-3,+3]	T/A/V
	CMU-MOSEI (Zadeh et al., 2018)	22,856 clips	Continuous emotion scores [-3,+3] + 6 emotions	T/A/V
	IEMOCAP (Busso et al., 2008)	Approx. 12 hours	6 emotion categories	T/A/V
	MELD (Poria et al., 2019a)	13,708 utterances, 1,433 dialogues	7 emotion categories + 3 emotion polarities	T/A/V
Emotion Reasoning	RECCON (Poria et al., 2021)	1,122 dialogues	Emotion cause annotation	T
	ECF (Wang et al., 2022)	1,374 dialogues, 13,619 utterances	Emotion categories and 9,794 emotion-cause pairs	T/A/V
	ECPE corpus (Xia & Ding, 2019)	1,945 documents	Emotion clause–cause clause pairing	T
Emotional Response and Support	EmpatheticDialogues (Rashkin et al., 2019)	25k dialogues	32 emotion categories + situation description	T
	MENTOS (Kim & Kim, 2026)	1,300 dialogues, 12,917 turns	Supporter's Belief, Emotion, Intent; 22 categories	T
	MERCI (Althubyani et al., 2025)	Personalized dialogue data	Emotion perception + personalized dialogue	T/A/V
	CPsyCoun (Zhang et al., 2024)	Multi-turn dialogue reconstruction	Chinese psychological counseling report	T
	PsyQA (Sun et al., 2021)	22k Q&A pairs	Chinese psychological counseling long text	T
	ESConv (Liu et al., 2021)	1,053 dialogues	Emotion + 8 support strategies	T

Note: T=Text; A=Audio; V=Vision; P=Physiological signals; E=Eye tracking.