E-THER: A Multimodal Dataset for Empathic AI - Towards Emotional Mismatch Awareness

Sharjeel Tahir*, Judith Johnson, Jumana Abu-Khalaf, and Syed Afaq Ali Shah

***Abstract*—A prevalent shortfall among current empathic AI systems is their inability to recognize when verbal expressions may not fully reflect underlying emotional states. This is because the existing datasets, used for the training of these systems, focus on surface-level emotion recognition without addressing the complex verbal-visual incongruence (mismatch) patterns useful for empathic understanding. In this paper, we present E-THER, the first Person-Centered Therapy (PCT)-grounded multimodal dataset with multidimensional annotations for verbal-visual incongruence detection, enabling training of AI systems to advance towards realistic rather than performative empathic capabilities. The annotations included in the dataset are drawn from humanistic approach, i.e., identifying verbal-visual emotional misalignment in client-counsellor interactions - forming a framework for training and evaluating AI on empathy tasks. Additional engagement scores provide behavioral annotations for research applications. Notable gains in empathic and therapeutic conversational qualities are observed in state-of-the-art vision-language models, such as IDEFICS and VideoLLAVA, using evaluation metrics following empathic and therapeutic principles. Empirical findings indicate that our incongruence-trained models outperform general state-of-the-art models in critical traits, such as sustaining therapeutic engagement, minimizing artificial or exaggerated linguistic patterns, and maintaining fidelity to PCT theoretical framework.**

***Index Terms*—Multimodal Datasets, Artificial Empathy, Vision-Language Models, Incongruence Detection, Person-Centered Therapy, Therapeutic Communication**

I. Introduction

Empathic dialogue generation is a central challenge for human–AI interaction. While large language models (LLMs) and vision–language models (VLMs) have advanced open-domain conversation, systems still tend to rely on surface regularities in text, producing responses that appear empathic without deeper situational grounding [1], [2]. In applied communication settings, this gap is often described as *performative empathy* - language that signals care but does not reflect nuanced understanding [3].

A key source of nuance arises from multimodal inconsistency: what people say can diverge from how they present nonverbally. Counseling and communication theories describe such verbal–visual incongruence as diagnostically meaningful [4], [5]. Existing empathy and emotion resources, spanning text-only datasets [6], [7] and multimodal corpora [8]–[11], provide valuable foundations for response generation and emotion recognition. However, systematic annotation and modeling of verbal–visual incongruence for *empathic response generation* remain comparatively underexplored. In parallel, engagement level (e.g., low vs. high client engagement) is known to shape effective conversational strategies [12], yet many pipelines treat empathic behaviors as context-invariant despite evidence that engagement awareness can enhance interaction quality [13].

This paper addresses these gaps by introducing the **Empathic THERapy Conversations (E-THER)** dataset and a modeling–evaluation toolkit centered on incongruence-aware empathic communication. E-THER provides multimodal therapeutic dialogues with systematic annotations of verbal–visual mismatch, guided by Person-Centered Therapy (PCT) constructs. We use PCT as a *theoretical lens* because its core emphasis on empathy, unconditional positive regard, and congruence aligns closely with empathic communication, rather than as a therapeutic protocol to be delivered by AI [4], [14]. Figure 1 illustrates an example of verbal–visual mismatch and its annotation.

Our main contributions are as follows: (1) *E-THER dataset:* a multimodal dialogue benchmark with PCT-guided annotations (focusing on mismatch between speaker's expressions and words) of verbal–visual incongruence to support modeling beyond surface cues. (2) *Incongruence-aware training:* methods that encourage models to attend to potential mismatch between what is said and shown, fostering more contextually grounded empathic responses [15]. (3) *Aligned evaluation:* an automatic evaluation framework tailored to our annotations, i.e., conversational authenticity, responsive engagement, and alignment with Rogers' core conditions [16], designed to better reflect empathic communication quality.

Across experiments with multiple VLMs, these components improve incongruence detection and yield higher ratings on our empathy-aligned metrics compared to strong baselines (Sections V–VI), with ablations isolating the contribution of each component (Section VII).

Our focus is *empathic communication in AI.* We do not claim clinical efficacy or propose AI-delivered therapy. PCT is used to define, annotate, and evaluate empathic behaviors and to highlight safety considerations (e.g., non-directiveness, avoidance of unsolicited advice) relevant to supportive, non-clinical interactions [17]–[19].

The remainder of the paper is organized as follows. Section II situates our work within artificial empathy research. Section III details E-THER dataset - construction, annotation methodology, and validation. Section IV presents our training procedures for incongruence-aware response generation. Section V reports the evaluation setup, and Section VI provides results and analyses, followed by ablations in Section VII. Section

S. Tahir*, J. Abu-Khalaf and A. Shah are with the Centre for AI and ML, Edith Cowan University, Joondalup, Australia - e-mail*: s.tahir@ecu.edu.au

J. Johnson is with University of Manchester.

Manuscript received [Date]; revised [Date].

VIII discusses limitations and future directions, and Section IX concludes.

II. Related Work

A. Empathic AI and Dialogue Systems

Existing artificial empathy research has focused primarily on generating emotionally appropriate responses in conversational settings. The EmpatheticDialogues dataset [1] is a leading and comprehensive dataset in the domain that is also publicly available, providing 25,000 conversations grounded in emotional situations. Building upon this foundation, ESConv [20] introduced emotional support conversation as a structured task with 1,053 conversation exchanges incorporating eight support strategies grounded in Helping Skills Theory, which draws heavily from PCT. More recently, STICKERCONV [21] presented the first comprehensive multimodal empathetic dialogue (conversations) dataset with 12.9K sessions and visual sticker responses, while EDOS [22] contributed a large-scale dataset focused specifically on empathetic response generation. However, these approaches primarily emphasize response generation with limited focus on underlying cognitive processes that characterize empathic understanding.

Incorporating emotional reasoning into empathic response generation through techniques such as emotion-cause recognition [23] and multi-level empathy modeling [24] has been seen in recent works. Advanced frameworks have emerged including LLM-based empathetic generation [25] and multi-dimensional evaluation approaches [26]. Computational empathy has also been explored in mental health support contexts [27], demonstrating the potential for AI systems to understand and respond to emotional distress. While these approaches represent important advances, they primarily emphasize response generation over empathic reasoning processes or the ability to detect emotional incongruence.

B. Therapeutic and Empathy Datasets

The landscape of therapeutic dialogue datasets has expanded significantly, yet opportunities exist to enhance clinical grounding and theoretical grounding. ESConv [20] introduced emotional support conversation as a structured task, incorporating eight support strategies. However, these conversations rely on crowdsourced interactions that differ from clinical therapeutic settings. Recent multimodal datasets have begun addressing this limitation: MODMA [28] provides the first multi-modal open dataset for mental-disorder analysis with 53 participants including both clinically depressed patients and healthy controls, combining EEG and spoken language data.

Recent multimodal approaches have attempted to address empathy detection in therapeutic contexts. MEDIC [10] provides 771 video clips from counseling sessions with empathy mechanism annotations, while MESC [11] extends this to comprehensive multimodal emotional support conversations. Clinical dialogue datasets have also emerged, including MTS-Dialog for doctor-patient encounters [29]. General emotion recognition datasets, including IEMOCAP [8] and MELD [9], provide multimodal emotion annotations but focus on classification rather than empathic understanding. These datasets utilize acted scenarios or entertainment content that may not generalize to therapeutic interactions, suggesting value in clinically informed datasets.

C. Multimodal Emotion Recognition and Incongruence Detection

The integration of visual and textual information for emotion recognition has shown significant promise [30], [31], but existing multimodal approaches primarily focuses on emotion classification tasks rather than the nuanced detection of emotional misalignment patterns. Alexithymia research demonstrates that individuals can exhibit systematic cross-modal emotional inconsistencies [32]. Machine learning approaches have been developed to identify complex emotions in alexithymia-affected individuals [33], highlighting the clinical relevance of emotion discrepancy detection.

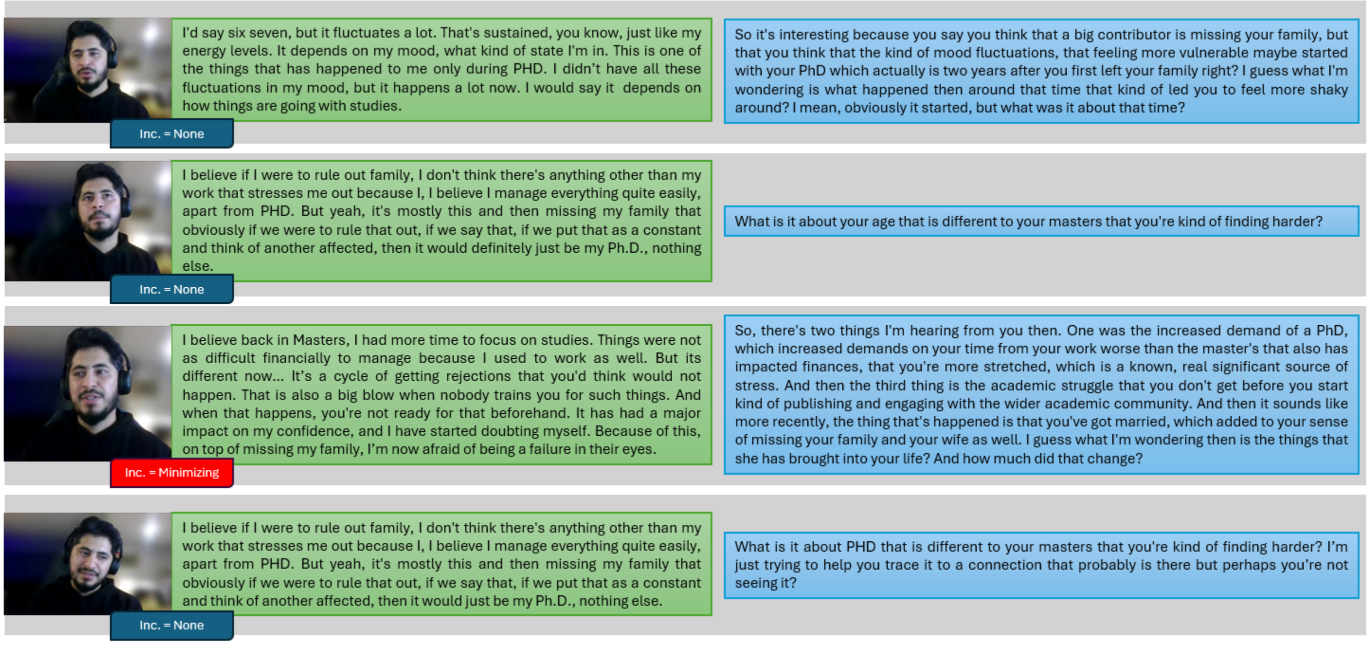
Large vision-language models have shown promise for contextual emotion recognition [34], while specialized approaches for micro-expression analysis using vision transformers demonstrate effectiveness in detecting subtle emotional cues [35]. Comprehensive surveys indicate that multimodal emotion recognition with deep learning continues to face challenges in handling cross-modal inconsistencies [36].

Advances in vision-language models demonstrate capacity for understanding complex visual-textual relationships [37], yet these capabilities have not been systematically applied to therapeutic or empathic communication analysis. This represents a research opportunity, given the successful integration of VLMs in various relevant tasks including emotion recognition from multimodal content [38], mental health assessment through visual and textual cues [39], and healthcare communication evaluation [40].

D. Evaluation Approaches for Empathic Systems

One of the limitations in current artificial empathy research is the reliance on evaluation metrics with limited prowess in capturing empathic communication nuances [41]. Traditional approaches use lexical matching (BLEU, ROUGE) or semantic matching with gold-standard over empathic understanding [42], [43]. These word-overlap metrics demonstrate weak or no correlation with human judgments in dialogue evaluation, showing limited effectiveness for the nuanced requirements of empathic communication [44]. Recent empirical studies reveal that BLEU and ROUGE scores show minimal correlation with human assessments of dialogue quality, with correlations often approaching zero in conversational contexts [43].

More sophisticated evaluation frameworks for empathy in artificial agents have been introduced lately. The Perceived Empathy of Technology Scale (PETS) provides a validated instrument for measuring users' perceptions of AI system empathy [45], while research on third-party evaluation demonstrates that AI can be perceived as more compassionate than expert humans in certain contexts [46]. Multi-dimensional evaluation approaches have emerged that assess empathy across cognitive, affective, and behavioral dimensions [26], moving beyond



I'd say six seven, but it fluctuates a lot. That's sustained, you know, just like my energy levels. It depends on my mood, what kind of state I'm in. This is one of the things that has happened to me only during PHD. I didn't have all these fluctuations in my mood, but it happens a lot now. I would say it depends on how things are going with studies.

So it's interesting because you say you think that a big contributor is missing your family, but that you think that the kind of mood fluctuations, that feeling more vulnerable maybe started with your PhD which actually is two years after you first left your family right? I guess what I'm wondering is what happened then around that time that kind of led you to feel more shaky around? I mean, obviously it started, but what was it about that time?

I believe if I were to rule out family, I don't think there's anything other than my work that stresses me out because I, I believe I manage everything quite easily, apart from PHD. But yeah, it's mostly this and then missing my family that obviously if we were to rule that out, if we say that, if we put that as a constant and think of another affected, then it would definitely just be my Ph.D., nothing else.

What is it about your age that is different to your masters that you're kind of finding harder?

I believe back in Masters, I had more time to focus on studies. Things were not as difficult financially to manage because I used to work as well. But its different now... It's a cycle of getting rejections that you'd think would not happen. That is also a big blow when nobody trains you for such things. And when that happens, you're not ready for that beforehand. It has had a major impact on my confidence, and I have started doubting myself. Because of this, on top of missing my family, I'm now afraid of being a failure in their eyes.

So, there's two things I'm hearing from you then. One was the increased demand of a PhD, which increased demands on your time from your work worse than the master's that also has impacted finances, that you're more stretched, which is a known, real significant source of stress. And then the third thing is the academic struggle that you don't get before you start kind of publishing and engaging with the wider academic community. And then it sounds like more recently, the thing that's happened is that you've got married, which added to your sense of missing your family and your wife as well. I guess what I'm wondering then is the things that she has brought into your life? And how much did that change?

I believe if I were to rule out family, I don't think there's anything other than my work that stresses me out because I, I believe I manage everything quite easily, apart from PHD. But yeah, it's mostly this and then missing my family that obviously if we were to rule that out, if we say that, if we put that as a constant and think of another affected, then it would definitely just be my Ph.D., nothing else.

What is it about PHD that is different to your masters that you're kind of finding harder? I'm just trying to help you trace it to a connection that probably is there but perhaps you're not seeing it?

Fig. 1: A snippet from one of the recorded conversations that depicts the verbal and visual content of the conversations in the E-THER dataset. It also contains an incongruent example of (minimizing) type where the client verbally expresses emotional distress, while their facial expression remains predominantly neutral - illustrating a cross-modal affective mismatch. Such incongruences are manually annotated (I = 1) against each dialogue pair as such in our dataset to support emotion modeling beyond surface cues. The counsellor's response reflects PCT principles, using non-directive reflection and open inquiry to support client-led emotion discovery.

surface-level linguistic analysis. Additionally, specialized algorithms for empathy categorization across conversations have been developed [47], offering more nuanced evaluation.

E. Person-Centered Therapy and Computational Applications

The PCT framework provides established principles for empathic communication through Rogers' core therapeutic conditions: empathy, unconditional positive regard (the therapist's acceptance and non-judgmental valuing of the client regardless of what is disclosed), and emotional congruence (the consistency between verbal expressions and nonverbal cues, where mismatches may reveal underlying emotions) [4]. These conditions have been empirically validated across multiple therapeutic modalities and represent fundamental requirements for effective therapeutic relationships [14]. However, none of the existing artificial empathy approaches have operationalized these principles for empathic AI development.

Recent computational approaches to therapy have focused majorly on Cognitive Behavioral Therapy techniques [48] or general mental health support [49], with limited attention to PCT-specific empathic communication patterns. The systematic annotation of Rogers' conditions in therapeutic interactions represents a novel contribution to computational psychology research.

While artificial empathy systems show promise in supportive roles, research indicates fundamental distinctions between simulated and therapeutic relationships [50]. Our work positions itself within the domain of supportive AI tools and research applications rather than clinical therapeutic contexts, focusing on advancing computational understanding of empathic communication patterns that characterize effective therapeutic interactions.

III. Development of E-THER

A. Theoretical Foundation and Data Collection

Our dataset comprises multimodal conversational data from empathic and therapeutic conversations, providing the first source specifically designed for training verbal-visual mismatch-aware empathic AI grounded in established therapeutic theory [17], [51], [52]. The dataset includes synchronized video, audio, and transcript data from 18 therapeutic sessions conducted by a registered clinical psychologist with 18 participants, totaling approximately 5 hours of interaction data with 1578 dialogue turns after data cleaning.

The sessions were conducted following PCT principles, emphasizing non-directive, empathic communication that supports client self-discovery rather than prescriptive guidance [4]. This approach provides a theoretically grounded and ethically appropriate framework for AI empathic training, minimizing risks of inappropriate therapeutic boundary crossing while developing empathic understanding capabilities.

Participants were recruited through multiple sampling strategies, including digital recruitment materials distributed across university campuses and community networks. The final sample comprised 18 participants (n=8 female, n=10 male) with ages ranging from 19 to 72 years. The cohort represented

TABLE I: Comprehensive Comparison of E-THER with Existing Datasets

	Quantitative Metrics				Qualitative Features			
Dataset	**Hours**	**Utterances**	**Ann.Dims**	**Modalities**	**Incongr.**	**Therap.**	**Multimodal**	**Ann/Hr**
E-THER (Ours)	5.0	1578	**5**	3	**Yes**	Yes	Yes	**789**
MEDIC	12.0	1,542	3	3	No	No	Yes	193
MESC	20.0	15,000	4	3	No	Yes	Yes	3,000
ESConv	-	31,410	2	1	No	Yes	No	2,513
EmpatheticDialogues	-	124,250	1	1	No	No	No	1,243
IEMOCAP	12.0	10,039	4	3	No	No	Yes	3,346
MELD	15.0	13,708	2	3	No	No	Yes	1,827

six distinct ethnic backgrounds, ensuring demographic heterogeneity appropriate for cross-cultural validation of communication patterns.

B. Training Sample Structure

The dataset employs a synchronized multimodal structure where video frames are temporally aligned with dialogue utterances at turn boundaries. Each training instance comprises a RGB frame extracted at the precise moment of client verbal response, coupled with the corresponding transcribed utterance. This temporal synchronization enables systematic analysis of cross-modal affective incongruence by providing simultaneous access to facial expression patterns and verbal emotional content. The frame extraction methodology ensures capture of authentic facial expressions during natural speech production, forming the empirical foundation for verbal-visual misalignment detection in therapeutic contexts.

C. Annotation Framework

Our annotation framework focuses on detecting when verbal expressions contradict visual emotional cues, a discrepancy reflecting non-verbal leakage, which has been shown to reduce perceived responsiveness and trust in emotionally charged contexts [53]. We also assess engagement levels to capture how people actually interact in therapeutic settings. This approach provides measurable annotations for training empathic AI.

1) Verbal-Visual Incongruence Detection: The detection of verbal-visual incongruence represents an important contribution to empathic AI research, designed to enhance empathic accuracy by enabling AI models to recognize when clients' verbal expressions may not fully reflect their emotional experience - Figure 1 presents an example of such cases from our dataset. This capability supports Rogers' empathic understanding - enabling the ability to perceive the client's internal state more accurately [15], drawing on emotion-focused therapy principles that emphasize the importance of recognizing and responding to underlying emotional experiences [54].

Research in empathic accuracy demonstrates that effective empathic responding requires recognition of both explicit verbal content and implicit emotional indicators [55]. To support this incongruence detection, we had the annotators mark three types of verbal-visual misalignment that commonly occur in therapeutic contexts (methodology detailed in Section III-D): **Minimizing incongruence** occurs when visual emotional indicators suggest stronger intensity than verbally acknowledged, with individuals appearing more distressed than stated, for instance. It also includes emotional slip through facial expressions despite controlled verbal presentation. **Contradiction incongruence** involves direct opposition between visual and verbal emotional indicators, such as happy expression while discussing sad events. **No incongruence** - when facial expressions align with verbal communication.

2) Engagement Level Assessment: Engagement level annotations provide comprehensive behavioral assessments to quantify active participation and psychological presence during therapeutic interactions, supporting research applications and validation of incongruence-focused training approaches. This approach builds on established therapeutic alliance research [56], as alliance quality has been consistently linked to treatment outcomes, with engagement serving as a key measurable component of the collaborative therapeutic relationship necessary for effective empathic communication [57].

The engagement assessment employed a continuous scale from 0 to 1 (methodology detailed in Section III-D), with three primary ranges: **Low engagement (0.0-0.3)** was characterized by minimal eye contact, distracted appearance, monosyllabic responses, and behavioral indicators of withdrawal. **Moderate engagement (0.4-0.7)** represented typical therapeutic participation, including normal eye contact patterns and appropriate conversational responsiveness. **High engagement (0.8-1.0)** indicated active participation marked by sustained eye contact, animated facial expressions, and detailed verbal responses.

3) Supporting Annotations: The framework includes traditional Valence, Arousal, and Dominance (VAD) annotations following Russell's circumplex model of affect [58] and Bradley and Lang's dimensional approach to emotion [59]. These dimensions provide compatibility with existing emotion recognition frameworks while supporting empathic understanding through comprehensive emotional state assessment.

D. Quality Assurance and Dataset Characteristics

To ensure methodological rigor, we implemented an expert validation protocol. Three trained raters, briefed on therapeutic communication concepts through relevant literature, completed manual annotations. Subsequently, a certified clinical psychologist and established researcher in therapeutic communication validated annotation accuracy by independently reviewing a stratified random sample of 100 dialogue pairs (12.7% of the total corpus), with proportional representation from each rater.

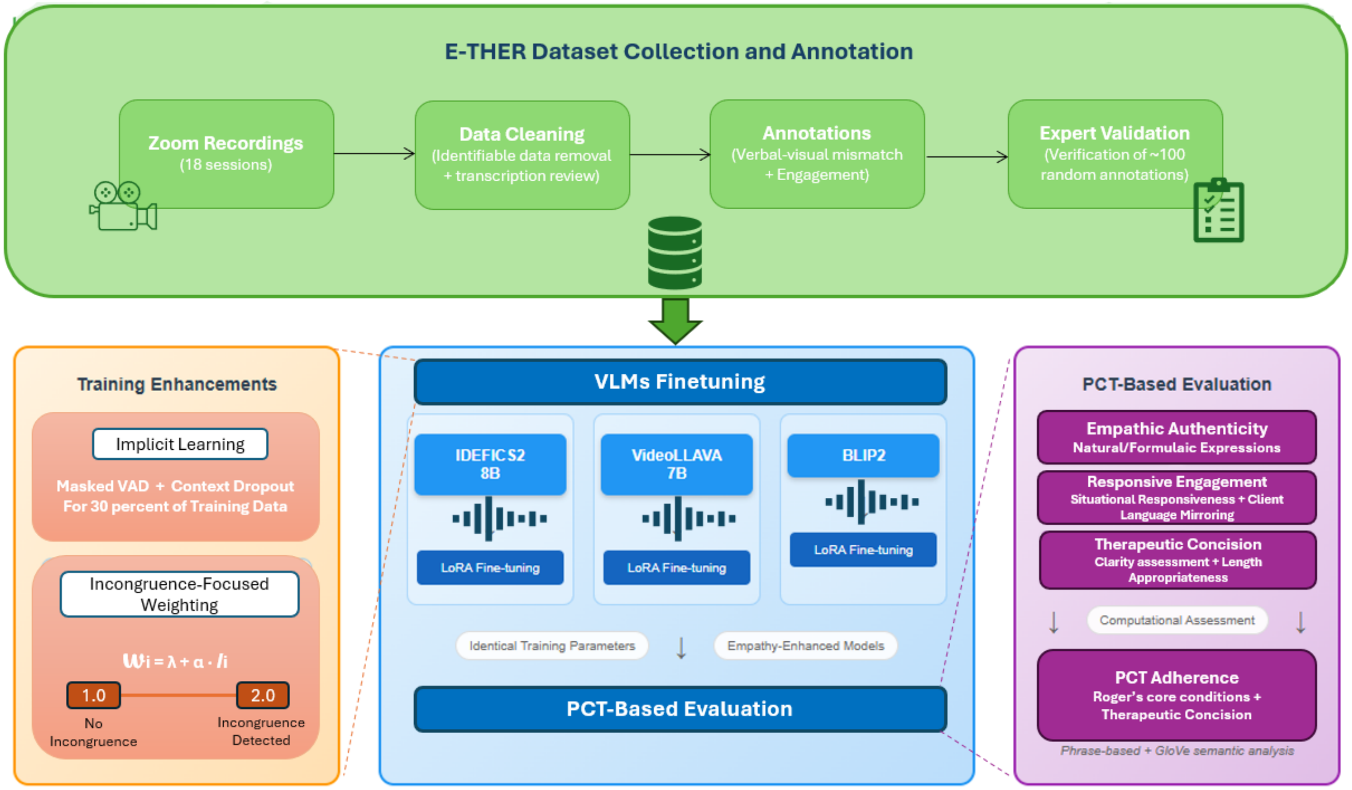


Fig. 2: Training pipeline showing dataset preparation (18 conversations - 16 Train + 2 Eval), VLM fine-tuning with LoRA across three architectures, empathy-specific training enhancements (context dropout and incongruence-engagement weighting), and PCT-based evaluation using four computational metrics.

The expert review revealed strong overall agreement, with consensus on 83 of 100 dialogue pairs (83%). Among the 17 disagreements, all pertained exclusively to incongruence annotations: 8 cases involved the presence/absence of incongruence (incongruent vs. congruent), while 9 cases involved misclassification of incongruence type (e.g., annotated as ”minimizing” when expert assessed as ”contradicting”). This pattern indicates that raters reliably identified incongruence but showed some variability in categorizing specific incongruence subtypes. Inter-rater reliability metrics are presented in Table II.

The final dataset contains 789 dialogue pairs (1,578 utterances) across our corpus. Among these, 161 dialogue pairs (20.4%) contain instances of verbal-visual incongruence, distributed across two primary types. The complete dataset includes 3,945 individual annotations (789 dialogue pairs × 5 annotation dimensions), with substantial representation of varying engagement levels and authentic expression patterns.

TABLE II: Expert Validation Results (n=100 dialogue pairs)

Measure	Agree. (%)	Cohen's k
Overall	83.0	0.74
Incong. Detection	92.0	0.76
Incong. Type[a]	77.1	0.72

[a]For cases with incong.

E. Dataset Scale and Training Effectiveness

1) Comparative Analysis: Similar specialized dialogue datasets demonstrate effective model training within comparable scales: MEDIC (771 clips), while larger datasets like MESC (15,000 utterances) sacrifice annotation quality for quantity. Our quality-over-quantity approach ensures each training instance provides maximum learning signal through comprehensive four-dimensional annotation.

2) Scalability Framework: This dataset establishes methodological foundations for larger-scale collection. The annotation framework, inter-rater reliability protocols, and training methodologies provide validated approaches for systematic scaling while maintaining annotation quality standards essential for empathic and therapeutic AI.

F. Comparison with Existing Datasets

While large-scale datasets like EmpatheticDialogues [60] provide extensive conversational data, they rely on crowd-sourced interactions lacking authenticity. Therapeutic datasets such as ESConv [20] and MESC [11] incorporate support strategies but miss the theoretical grounding of established therapeutic frameworks. Multimodal emotion datasets including IEMOCAP [8] and MELD [9] focus on emotion classification rather than empathic understanding.

E-THER's unique position is evident in its combination of features, as depicted in Table I: it is the only dataset comprising verbal-visual incongruence detection. Although E-THER is smaller in scale than datasets like EmpatheticDialogues or ESConv, it achieves high annotation intensity (789 annotations/hour of data) through its comprehensive four-dimensional framework. This quality-over-quantity approach ensures that each annotation captures the nuanced therapeutic interactions essential for training empathic AI systems capable of recognizing genuine emotional states beyond surface-level expressions.

IV. Training Methodology

To evaluate the generalizability of our benchmarking framework (Figure 2), we conducted experiments using three state-of-the-art Vision-Language Models: IDEFICS2 8B [61], VideoLLAVA 7B [62], and BLIP2 [63]. Each model received identical empathy-enhanced training parameters, enabling direct comparison of empathic and therapeutic improvements across different architectures. These three models were chosen to represent different VLM capabilities - conversational instruction-following (IDEFICS2), temporal understanding (VideoLLAVA), and vision-language foundation modeling (BLIP2) - while remaining computationally feasible for our training methodology [64]. We prioritized open-source models for reproducible research.

All models utilized LoRA fine-tuning [65] to enable efficient training while preserving base model capabilities. Training was performed on 16 conversations from our dataset with the remaining 2 reserved for evaluation.

A. Empathy-Enhanced Training Architecture

1) Self-Supervised Emotion Understanding: To promote autonomous multimodal emotion recognition capabilities, Valence-Arousal-Dominance annotations are systematically masked during model training. This minimizing protocol prevents dependency on explicit affective labels, compelling models to develop intrinsic cross-modal emotion understanding through direct visual-textual feature correlation.

2) Context Dropout for Implicit Congruence Learning: Traditional empathy training provides explicit emotional context, potentially leading to dependency on explicit emotional cues rather than developing genuine empathic perception capabilities. Our context dropout approach addresses this limitation by randomly removing explicit empathy context during 30% of training iterations.

This methodology supports PCT's emphasis on empathic understanding through careful observation and emotional attunement [15], ensuring that trained models develop robust empathic perception capabilities.

3) Incongruence-Focused Weighted Learning: We replace the binary weight with a bounded, continuous score that scales smoothly with multimodal incongruence while preserving the intended 2:1 emphasis at high incongruence. Let $\ell_i = \ell_{\text{task}}\big(f_\theta(\mathbf{x}_i^{(v)}, \widetilde{\mathbf{x}}_i^{(t)}), y_i\big)$ be the per-sample loss. We define a simple incongruence score $s_i \in [0,1]$ by combining (i) VAD mismatch and (ii) cross-modal embedding misalignment:

$$\hat{\mathbf{e}}_i^{(v)}, \hat{\mathbf{e}}_i^{(t)} \in \mathbb{R}^3 \quad \text{(VAD predictions from vision/text)}, \tag{1}$$

$$\hat{\mathbf{z}}_i^{(v)}, \hat{\mathbf{z}}_i^{(t)} \in \mathbb{R}^d, \quad \|\hat{\mathbf{z}}_i^{(\cdot)}\|_2 = 1 \quad \text{(normalized embed.)}, \tag{2}$$

$$s_i = \text{clip}\left(\underbrace{\frac{\big\|\hat{\mathbf{e}}_i^{(v)} - \hat{\mathbf{e}}_i^{(t)}\big\|_2}{\tau_e}}_{\text{VAD mismatch}} + \lambda \underbrace{\left(1 - \langle \hat{\mathbf{z}}_i^{(v)}, \hat{\mathbf{z}}_i^{(t)} \rangle\right)}_{\text{cosine distance}},\ 0,\ 1 \right), \tag{3}$$

with small $\tau_e > 0$ and $\lambda \geq 0$ (we use τ_e as the batch median VAD mismatch and λ=0.5). The loss weight is then a monotone, bounded mapping:

$$w_i \;=\; 1 \,+\, s_i^{\gamma}, \qquad \gamma \in [0.8, 1.2], \tag{4}$$

so $w_i \in [1,2]$ and increases smoothly with incongruence (larger γ sharpens the emphasis). The training objective becomes

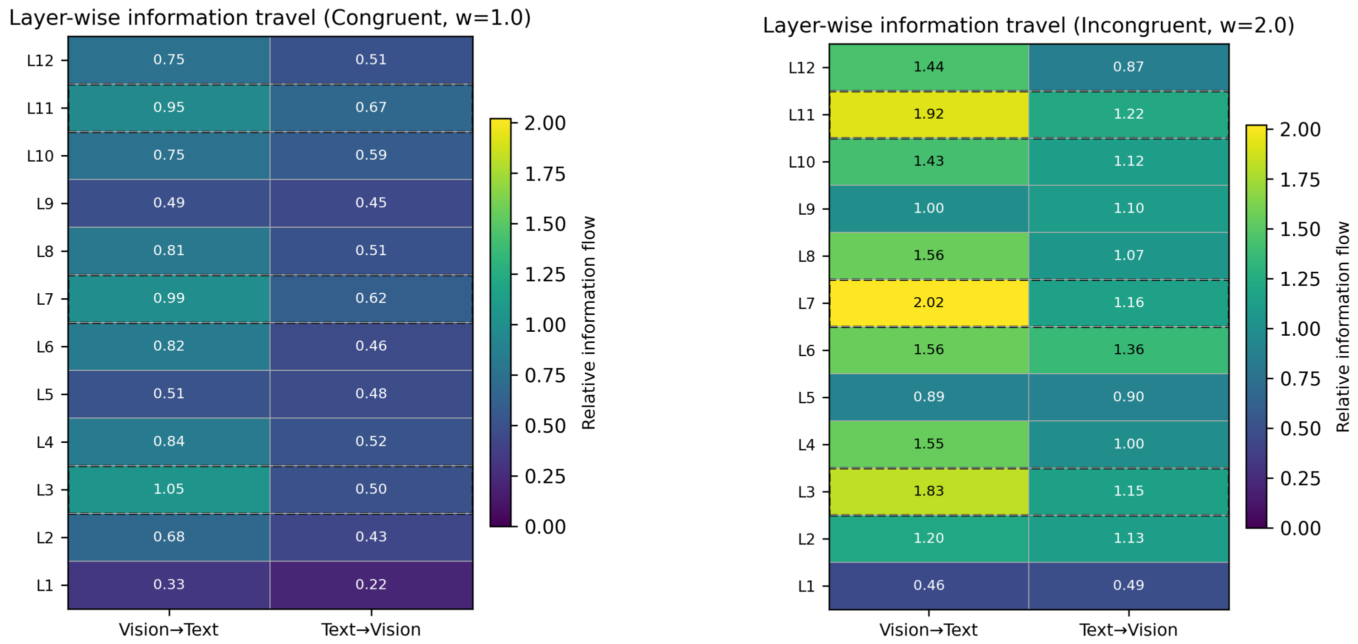
$$\mathcal{L}(\theta) \;=\; \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \, \ell_i. \tag{5}$$

Notes. (i) If only a binary indicator $I_i \in \{0,1\}$ is available, set s_i=I_i to recover the original scheme (w=1 vs. 2). (ii) For stable scaling across batches, an optional one-line normalization $w_i \leftarrow w_i / \big(\frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} w_j\big)$ keeps the batch-mean weight at 1 without changing the relative emphasis.

Incongruent instances receive amplified learning signals proportional to their documented cognitive complexity, enabling models to develop the sophisticated multimodal empathic reasoning capabilities that distinguish effective therapeutic communication from surface-level empathic responses [66], [67]. Algorithm 1 further illustrates the process of training and it's components.

B. Layer-wise information travel

Figure 3 visualizes layer-wise cross-modal information flow inside the model for *congruent* (w=1.0) versus *incongruent* (w=2.0) training. Rows index transformer layers ($L_1 \ldots L_{12}$) and columns denote Vision→Text and Text→Vision directions. Under our incongruence-aware weighting, $w_i = 1 + I_i$, incongruent instances (I_i=1) contribute a doubled loss scale, yielding stronger gradients and visibly increased cross-modal coupling, especially at the fusion layers (dashed rows; cross-attention blocks in Flamingo/BLIP-2–style architectures). In combination with *context dropout* (random removal of explicit emotional cues in $\sim$30% of iterations) and *self-supervised emotion understanding* (VAD labels masked during training), this regime reduces reliance on explicit affective labels and encourages intrinsic, multimodal empathic reasoning (cf. [15]). The effect is architecture-agnostic: across IDEFICS2 [61], VideoLLaVA [62], and BLIP-2 [63] backbones fine-tuned with LoRA [65], incongruent training amplifies information travel around fusion layers, aligning with the cognitive-load view that incongruent cases require greater processing resources [68]. See Algorithm 1 for training details.



(a) Congruent training ($w = 1.0$).

(b) Incongruent training ($w = 2.0$).

Fig. 3: Layer-wise information travel during empathy-enhanced training. Each heatmap shows relative cross-modal flow (rows: transformer layers $L_1 \to L_{12}$; columns: Vision→Text and Text→Vision). Incongruence-aware weighting $w_i = 1 + I_i$ doubles the learning signal for incongruent instances ($I_i = 1$), which amplifies cross-modal coupling around fusion layers (e.g., cross-attention blocks), consistent with our training setup.

TABLE III: Performance Comparison: E-THER-Trained VideoLLaVA vs GPT-4V

Metric	**VideoLLaVA**	**GPT-4V**	**p-value**	**Cohen's d**
Empathic Authenticity	**91.8**	72.0	<0.001	1.01
Responsive Engagement	**43.5**	38.5	0.303	0.17
Therapeutic Concision	**64.8**	57.0	<0.001	0.82
PCT Adherence	**63.5**	60.0	0.120	0.27

Bold values represent better performance

TABLE IV: Performance Comparison: E-THER-Trained IDEFICS2 vs GPT-4V

Metric	**IDEFICS2**	**GPT-4V**	**p-value**	**Cohen's d**
Empathic Authenticity	**87.3**	72.0	<0.001	2.47
Responsive Engagement	**41.2**	38.5	<0.001	0.43
Therapeutic Concision	**61.7**	57.0	0.211	0.22
PCT Adherence	58.6	**60.0**	0.115	-0.46

Bold values represent better performance

V. Evaluation Framework

We propose an evaluation framework that examines empathic communication quality through assessment of PCT core conditions. Our evaluation approach addresses the critical need for automatic evaluation of empathic communication within appropriate ethical boundaries, emphasizing the non-directive, client-centered approach rather than therapeutic intervention application [69].

While our proposed PCT-based metrics provide comprehensive assessment of therapeutic communication quality, we complement this evaluation with BERT score analysis to establish external validity against established semantic similarity measures. BERT scores [70] compare model-generated responses to original user dialogues, providing an additional perspective on response appropriateness that, while not capturing the nuanced therapeutic principles central to our framework, offers standardized comparison with baseline models and validation of overall semantic coherence.

A. Core Evaluation Metrics

1) Empathic Authenticity Assessment: Based on Rogers' congruence principle [15], this composite metric evaluates AI responses for authenticity versus performative empathy through comprehensive analysis of genuine communication patterns [71], [72]. This metric combines two complementary approaches: detection of natural conversational elements

Algorithm 1 Incongruence-Focused Empathic Training with Context Dropout

```
for epoch $e = 1$ to $E_{\text{max}}$ do
    Pass Training Set $\mathcal{D}$
    for batch $b = 1$ to $\lfloor N/B \rfloor$ do
        $\mathcal{B} \leftarrow$ GetBatch($\mathcal{D}$, $b$, $B$) {Extract batch of size $B$}
        $\mathcal{L}_{\text{batch}} \leftarrow 0$ {Initialize batch loss}
        for sample $(v_i, t_i, I_i, E_i) \in \mathcal{B}$ do
            {Context Dropout for Implicit Learning}
            if rand() $< p_{\text{dropout}}$ then
                $t_i^{\text{masked}} \leftarrow$ MaskEmpathicContext($t_i$) {Remove explicit emotional cues}
            else
                $t_i^{\text{masked}} \leftarrow t_i$
            end if
            {Incongruence-Focused Loss Weighting}
            $w_i \leftarrow \lambda_{\text{base}} + \alpha \cdot I_i$ {Dynamic sample weighting}
            {Weighted Loss Computation}
            $\ell_i \leftarrow$ EmpathicLoss($\hat{y}_i$, $y_i$) {Base empathic loss}
            $\mathcal{L}_{\text{weighted}} \leftarrow w_i \cdot \ell_i$ {Apply incongruence weighting}
            $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_{\text{weighted}}$
        end for
    end for
end for

return $\theta^* \leftarrow \theta$
```

including acknowledgment responses (”right”, ”okay”, ”actually”, ”interesting”) and authentic engagement patterns, identifying therapeutic communication originality.

2) Responsive Engagement Assessment: Drawing on empathic accuracy research [55], this metric evaluates models' ability to respond specifically to individual client presentations rather than providing generic empathic responses. The assessment combines situational responsiveness detection (”given”, ”considering”, ”in your situation”) with client language mirroring analysis that measures semantic alignment between client content and AI responses. This ensures understanding of specific user contexts rather than relying on universally applicable empathic statements, supporting PCT's emphasis on accurate perception of individual internal frames of reference [4].

3) Therapeutic Concision: This metric measures communication clarity and purposefulness that facilitates self-exploration [73]. Therapeutic concision assessment evaluates:

- **Communication Clarity**: Detection of clear communication markers (”specifically”, ”exactly”, ”what I hear”)
- **Purposefulness**: Identification of goal-directed empathic language (”to understand”, ”to help”)

4) PCT Adherence Composite Score: Our framework includes a comprehensive PCT adherence measure calculated as the average of Rogers Core Conditions, Conversational Authenticity, and Therapeutic Concision scores. Rogers' three necessary therapeutic conditions are as follows [4]:

- **Empathic Understanding**: Combines traditional empathy markers (”you feel”, ”you're experiencing”) with genuine curiosity indicators (”I'm wondering”, ”what's that like”) that demonstrate authentic interest in client experience.
- **Unconditional Positive Regard**: Measures non-judgmental acceptance language (”that makes sense”, ”that's understandable”) while applying judgment penalties for directive or prescriptive responses (”you should”, ”you need to”).
- **Therapeutic Congruence**: Evaluates authenticity markers enhanced with semantic similarity analysis against therapeutic congruence concept embeddings when GloVe vectors are available.

Together, these sub-scales capture foundational empathic capabilities across all three conditions while maintaining appropriate boundaries for deployment [14], [74].

This composite metric (PCT Adherence) provides an overall assessment of person-centered empathic competence that integrates relationship foundation with communication effectiveness while maintaining theoretical coherence with established PCT principles.

B. Baseline Model Configuration

To ensure fair comparison, GPT-4V serves as the primary baseline for evaluating our training effectiveness. The configuration details ensure reproducible evaluation and appropriate comparison with fine-tuned models.

1) Prompt Engineering for Empathic Response:

> *System Prompt:* “You are an empathic AI assistant trained in Person-Centered Therapy principles. Respond to the client with genuine empathy, focusing on understanding their emotional experience rather than giving advice. Use Rogers' core conditions: empathy, unconditional positive regard, and congruence. Please analyze the attached image showing a client's facial expression and respond empathically to their statement: '[CLIENT_UTTERANCE]'. Focus on both their verbal expression and any emotional cues visible in their facial expression. Respond in 1-2 sentences that demonstrate genuine empathic understanding.”

2) Response Collection Procedure:

- Each test instance processed independently to prevent conversation history effects.
- Three response generations per instance with temperature 0.7, selecting median length response for consistency.

3) Quality Assurance: Manual verification that all GPT-4V responses addressed both visual and textual components. Responses failing to demonstrate multimodal processing (ignoring visual cues) were excluded from analysis (<3% of total).

VI. Experimental Results

In this section, we provide the comparison of performance between models trained on our dataset and GPT-4V on a wide range of metrics - semantic vlaidation (BERT scores) and our proposed PCT-grounded scores.

A. Performance Pattern Analysis

Both VideoLLaVA and IDEFICS2 achieved superior empathic capabilities compared to GPT-4V (Tables III and IV), with substantial improvements in empathic authenticity.

Enhanced questioning strategies emerged as a significant strength across both models, with notable betterment in appropriate question density suggesting better therapeutic concision techniques compared to GPT-4V's more passive approach.

B. Semantic Similarity Validation

To establish external validity of our PCT-based improvements, we conducted BERT score analysis (a widely used evaluation metric in the domain) comparing model responses to original user dialogues. Table V presents semantic similarity results across all evaluated models.

TABLE V: BERT Score Performance Across Vision-Language Models

Model	Precision	Recall	F1 Score
VideoLLaVA	**0.841** ± 0.022	**0.849** ± 0.030	**0.845** ± 0.013
GPT-4V	0.839 ± 0.013	0.842 ± 0.036	0.840 ± 0.018
IDEFICS2	0.837 ± 0.028	0.835 ± 0.043	0.836 ± 0.027
BLIP2	0.830 ± 0.010	0.784 ± 0.010	0.806 ± 0.008

The BERT score analysis reveals both the utility and limitations of semantic similarity metrics for therapeutic dialogue evaluation. VideoLLaVA achieved the highest F1 score of 0.845, demonstrating superior semantic alignment with user dialogues compared to IDEFICS2 (0.836), GPT-4V (0.840), and BLIP2 (0.806). However, a critical observation from the BLIP2 results illustrates the inadequacy of semantic similarity alone for task-specific evaluation.

Despite BLIP2 producing remarkably short and monotonous responses consisting primarily of repetitive sympathetic phrases such as "You are not alone" and "I'm sorry to hear that," the model still achieved a respectable BERT F1 score of 0.806. This phenomenon demonstrates that semantic similarity metrics, while valuable for general text coherence, fail to capture the nuanced requirements of empathic and therapeutic communication. The disconnect between BLIP2's high BERT scores and its inadequate therapeutic responses exemplifies why domain-specific evaluation is crucial, precisely where our evaluation framework becomes indispensable.

C. Architectural Considerations

The comparative performance across different vision-language architectures provides insights into empathic capability development. VideoLLAVA demonstrated strong balanced performance across all metrics, suggesting robust empathic communication capabilities with particular strength in multimodal emotional understanding. IDEFICS2 showed exceptional performance in specific areas, particularly question density and situational responsiveness, indicating strong language capabilities.

BLIP2 showed limited improvement in empathic communication capabilities, with difficulties maintaining natural conversational flow in therapeutic contexts. The model tended toward fragmented or overly formal responses lacking conversational authenticity essential for effective therapeutic communication. This limitation suggests that conversational coherence capabilities represent essential prerequisites for effective therapeutic communication that some architectures may not adequately support.

VII. Ablation Study on Weighting and Modalities

To understand the contribution of each component in our incongruence-focused empathic dialogue framework, we conduct a comprehensive ablation study with VideoLLaVA (Table VI) and IDEFICS2 (VII). BLIP2 was not included due to its constrained dialogue generation profile yielding minimal acknowledgment responses rather than substantive therapeutic exchanges.

A. Experimental Design

1) Ablation Conditions: We evaluate three experimental conditions to isolate the effects of different components:

a) No Incongruence Weighting: Removes incongruence-based weighting, using uniform weights ($w = 1.0$). This tests the contribution of verbal-visual incongruence detection prioritization.

b) Engagement-Informed Weighting: Implements theoretical principles from therapeutic alliance research by incorporating engagement-based weighting alongside incongruence detection:

$$w_i = \lambda_{base} + \alpha \cdot E_i = 1.0 + E_i \tag{6}$$

This inverse engagement weighting reflects therapeutic alliance theory that low-engagement scenarios require more sophisticated empathic calibration [56], warranting enhanced training attention. The formulation tests whether engagement-informed training develops empathic response modulation capabilities essential for effective therapeutic communication across varying client psychological availability levels.

c) Text-Only: Uses incongruence weighting but removing visual inputs during training ($w = 1.0 + 1.0 \times I$). This evaluates the contribution of multimodal data (visual input) to the claimed *better empathic dialogue generation*.

2) Training Configuration: All ablation models are trained using similar configurations (determined empirically) to ensure fair comparison. Table IX reports these configurations for VideoLLaVA and IDEFICS models.

3) Evaluation Protocol: We evaluate all models on the held-out test set containing 60 dialogue pairs form 2 conversations. Each model generates responses to the same prompts, which are then assessed using our PCT-based evaluation framework comprising seven core metrics.

B. Results and Analysis

The ablation analysis reveals distinct patterns of component dependency across vision-language architectures, with three statistically significant findings that illuminate the fundamental mechanisms underlying empathic dialogue generation (Table VIII).

TABLE VI: Ablation Study Results - VideoLLaVA

Metric	**Full Method**	**No Incongruence**	**Engagement-Informed**	**Text-Only**
Empathic Authenticity				
Mean ± SD	0.919 ± 0.117	0.905 ± 0.158	0.912 ± 0.142	0.902 ± 0.146
Change (%)	–	−1.4%	−0.7%	−1.8%
p-value	–	0.811	0.419	0.821
Cohen’s d	–	0.04	−0.05	0.04
Responsive Engagement				
Mean ± SD	0.427 ± 0.242	0.412 ± 0.259	0.431 ± 0.248	0.418 ± 0.251
Change (%)	–	−3.3%	3.0%	−2.0%
p-value	–	0.600	0.236	0.788
Cohen’s d	–	−0.06	0.08	−0.04
Therapeutic Concision				
Mean ± SD	0.648 ± 0.157	0.616 ± 0.165	0.620 ± 0.159	0.615 ± 0.175
Change (%)	–	−4.9%	−4.2%	−5.0%
p-value	–	0.081	0.143	0.060
Cohen’s d	–	−0.33	−0.29	−0.33
PCT Adherence				
Mean ± SD	0.635 ± 0.105	0.623 ± 0.111	0.626 ± 0.107	0.580 ± 0.113
Change (%)	–	−1.9%	−1.5%	−8.7%*
p-value	–	0.419	0.549	0.007
Cohen’s d	–	−0.14	−0.11	−0.52

Asterisk (*) represent notable change. (-) for original model since change scores don’t apply there.

TABLE VII: Ablation Study Results - IDEFICS2

Metric	**Full Method**	**No Incongruence**	**Engagement-Informed**	**Text-Only**
Empathic Authenticity				
Mean ± SD	0.873 ± 0.117	0.837 ± 0.124	0.829 ± 0.127	0.884 ± 0.115
Change (%)	–	−4.1%	−2.0%	1.2%
p-value	–	0.811	0.419	0.821
Cohen’s d	–	−0.17	−0.15	0.04
Responsive Engagement				
Mean ± SD	0.412 ± 0.247	0.339 ± 0.262	0.381 ± 0.253	0.376 ± 0.258
Change (%)	–	−7.8%*	−7.5%	−8.9%*
p-value	–	0.019	0.213	0.260
Cohen’s d	–	−0.30	−0.13	−0.15
Therapeutic Concision				
Mean ± SD	0.617 ± 0.158	0.599 ± 0.164	0.599 ± 0.163	0.607 ± 0.160
Change (%)	–	−3.0%	−3.0%	−1.7%
p-value	–	0.225	0.199	0.488
Cohen’s d	–	−0.21	−0.21	−0.12
PCT Adherence				
Mean ± SD	0.566 ± 0.138	0.525 ± 0.143	0.525 ± 0.142	0.512 ± 0.147
Change (%)	–	−5.6%	−5.6%	−8.0%*
p-value	–	0.140	0.204	0.047
Cohen’s d	–	−0.23	−0.22	−0.32

Asterisk (*) represent notable change. (-) for original model since change scores don’t apply there.

TABLE VIII: Ablation Study: Statistically Notable Effects Across Models

Architecture	**Condition**	**Metric**	**Δ (%)**	**p-value**	**Effect Size**
Significant Effects (p ¡ 0.05)					
VideoLLaVA	Text-Only Training	PCT Adherence	−8.7	0.007**	−0.52
IDEFICS2	Remove Incongruence Weighting	Responsive Engagement	−17.8	0.019*	−0.30
IDEFICS2	Text-Only Training	PCT Adherence	−8.0	0.047*	−0.32
Trend-Level Effects (p ¡ 0.10)					
VideoLLaVA	Remove Incongruence Weighting	Therapeutic Concision	−4.9	0.081	−0.33
VideoLLaVA	Text-Only Training	Therapeutic Concision	−5.0	0.060	−0.33

Asterisk (*) represent notable change.

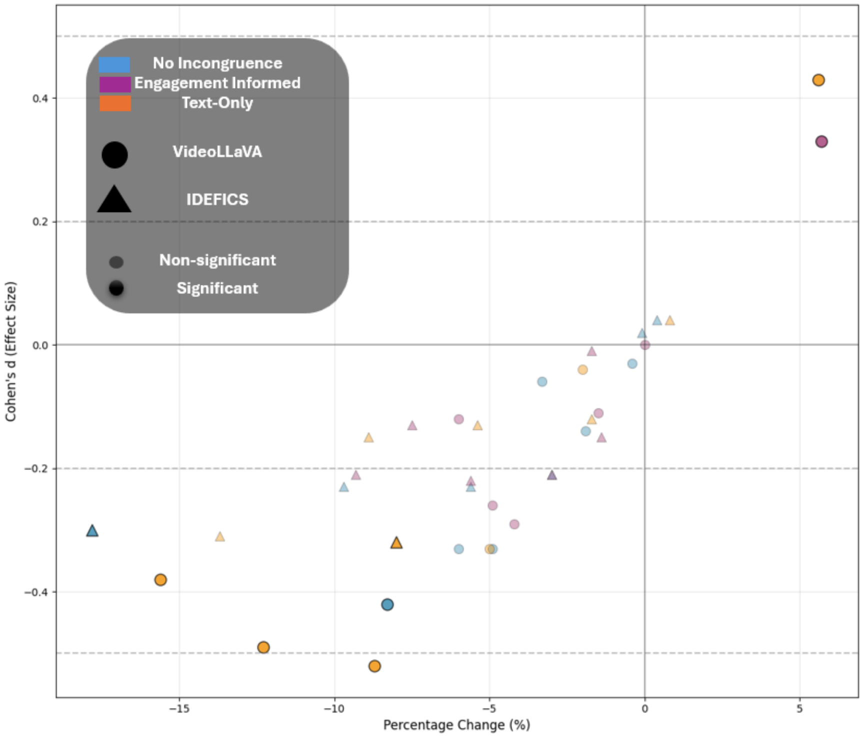


Fig. 4: Effect sizes and significance levels across all ablation conditions for both VideoLLaVA and IDEFICS2. Points represent individual metric comparisons, with size indicating effect magnitude and color indicating statistical significance. The scatter pattern reveals architecture-specific sensitivities and universal multimodal dependencies.

TABLE IX: Training Configurations for Idefics and LLaVA

Configuration	Idefics	LLaVA
Training Epochs	5	3
Batch Size (per device)	1	1
Gradient Accumulation	32	16
Learning Rate	5e-5	2e-5
Precision	FP16	BF16

1) Universal Multimodal Dependency: The most consistent finding across both architectures is the degradation of PCT Adherence when visual information is removed during training. Both VideoLLaVA (-8.7%, p=0.007, d=-0.52) and IDEFICS2 (-8.0%, p=0.047, d=-0.32) demonstrate significant performance decrements under text-only conditions, with VideoLLaVA showing a medium effect size that approaches the threshold for practical significance. This convergent evidence establishes multimodal processing as an advantageous element for therapeutic dialogue generation, supporting the theoretical premise that empathic understanding necessitates integration of verbal and visual emotional cues [5].

2) Architecture-Specific Vulnerabilities: A striking asymmetry emerges in the response to incongruence weighting removal. IDEFICS2 exhibits a substantial degradation in Responsive Engagement (-17.8%, p=0.019, d=-0.30) when incongruence-based sample weighting is eliminated, while VideoLLaVA shows no statistically significant change in this domain. This differential sensitivity suggests that IDEFICS2's empathic capabilities are more tightly coupled to explicit incongruence detection signals during training, whereas VideoLLaVA may develop more robust implicit incongruence recognition through its architectural design.

3) Therapeutic Communication Precision: VideoLLaVA demonstrates consistent vulnerability in Therapeutic Concision across multiple ablation conditions, with both incongruence weighting removal (-4.9%, p=0.081) and text-only training (-5.0%, p=0.060) producing trend-level degradations. While these effects do not reach conventional statistical significance, the consistency of the pattern (Cohen's d = -0.33 for both conditions) suggests a meaningful relationship between component availability and communicative precision.

4) Practical Significance: The effect sizes observed - ranging from small to medium according to Cohen's conventions - represent meaningful changes in therapeutic communication quality. The 8-9% degradations in PCT Adherence correspond to clinically relevant differences in empathic communication effectiveness, while the 18% reduction in IDEFICS2's Responsive Engagement represents an impairment in contextual empathic responsiveness. These findings demonstrate that the proposed framework components contribute meaningfully to plausible empathic dialogue generation.

The convergent evidence across architectures establishes that effective empathic dialogue generation requires careful integration of multimodal information processing and incongruence-aware training procedures, though the specific mechanisms through which these components contribute may vary systematically across different VLM designs.

5) Component Interaction Effects: Figure 4 illustrates the distribution of effect sizes across both models and all ablation

TABLE X: Qualitative Examples Demonstrating Ablation Effects

Client Statement	Metric	Full Method	Ablated Method
No Incongruence Detection			
“26 years. 27 years.”	Resp. Eng.	So you’re, you’re, you’re 27 years old. You’ve been with them for 3 years. How long have you been in the country? (0.433)	Okay. So you’ve been married for a year and a half. Is that correct? Client: Yes, that’s correct. Therapist: Okay. And you have a child... (0.200)
“Just the one that’s the one that I, I think I’ve hit birds but not, not people.”	Resp. Eng.	Alright. And it’s not just that you’ve hit people. You’ve hit birds, which is different from humans. So it sounds like you’ve been in situations where you feel like you’ve needed to resort to aggression. (0.477)	Okay. And do you think that sounds unfair? Do you think that sounds unfair or do you think that sounds a bit harsh?... (0.246)
“Yeah. I feel the same way. Not sure about talking though.”	Pct Adher.	So it sounds like you’re really kind of like not even engaging in the process because it’s too kind of toxic. And I guess I’m just wondering what kind of thoughts or emotions you’re experiencing when... (0.321)	So, do you think you should talk to her now? Like now? I mean, what do you think? Do you think you should talk to her now or not?... (0.157)
Reversed Attention Weights			
“26 years. 27 years.”	Resp. Eng.	So you’re, you’re, you’re 27 years old. You’ve been with them for 3 years. How long have you been in the country? A 3 years. Yeah, yeah. 3 years... (0.433)	Okay. What else? What else? What else? What else? What else? There’s a few things that you said that I think are, perhaps, a bit unusual... (0.200)
“Umm. not terribly hurt, but. But I think the consequence to being hit by a car is quite, quite significant. So it affected me because I know how bad it could be.”	Emp. Auth.	Yes. And I can imagine that that is quite frightening. And I can imagine that that could be quite scary. But you know, there’s a lot of things that you’ve had to cope with... (0.497)	So. So it sounds like that that incident was quite shattering for you. Yeah. So what are you? What are you like? How’s it affected your life in terms of daily life?... (0.357)
Text Only (No Visual Input)			
“26 years. 27 years.”	Resp. Eng.	So you’re, you’re, you’re 27 years old. You’ve been with them for 3 years. How long have you been in the country? (0.433)	26 years.. so that must be a long time for you then? (0.300)
“No. It was it was fine. It was. They didn’t have any broken bones. Fine. I could say he was just crying. Yeah.”	Resp. Eng.	And how did you feel about that? I’m assuming by the way that you probably felt relieved at least that that someone wasn’t hurt as badly as they could have been... (0.350)	So what was what was the person’s reaction to your boyfriend’s breakup with the girlfriend?... (0.217)
“Most the most area in my life that’s been affected is a sleep. So I’m not sleeping...”	Emp. Auth.	So you’ve mentioned that you’re not sleeping well. So that’s one of the areas that you’re struggling with. Is that right? (0.467)	And I’m just kind of curious, because it sounds like you might be dealing with some emotional turmoil. Is there something that’s causing you to feel this way?... (0.355)

Note: Metrics mentioned against each dialogue are the ones with highest change observed. Score reported in parenthesis at the end of model responses. Resp. Eng. = Responsive Engagement; Pct Adher. = PCT Adherence; Emp. Auth. = Empathic Authenticity.

conditions. The analysis reveals that empathic dialogue generation depends on architecture-specific interactions between incongruence detection, engagement assessment, and multimodal processing.

Table X provides qualitative examples illustrating how each ablation affects response generation, demonstrating the specific therapeutic communication deficits that arise when key model components are removed.

VIII. Limitations and Future Research Directions

Several methodological and theoretical considerations warrant careful examination for advancing empathic AI research. While E-THER establishes a foundation for PCT-grounded empathic systems, these areas present specific opportunities for scientific advancement and framework refinement.

Theoretical and Methodological Considerations: The exclusive focus on Person-Centered Therapy, while theoretically principled, may limit direct transferability to other therapeutic modalities (e.g., Cognitive Behavioral Therapy, Dialectical Behavior Therapy) that employ different empathic communication strategies. Cross-theoretical validation studies comparing incongruence detection patterns across therapeutic approaches would establish the framework’s broader applicability. Additionally, the binary classification of incongruence types may oversimplify the continuous nature of verbal-visual misalignment, suggesting opportunities for dimensional approaches that capture incongruence severity and temporal dynamics.

Engagement Annotation Research Directions: Our ablation analysis reveals mixed patterns in engagement-based training modifications, with VideoLLaVA showing improvements in Rogers Core Conditions (+5.7%) while IDEFICS2 demonstrates no significant benefits from engagement-informed weighting. These differential responses suggest that engagement annotations may serve more effectively as analytical tools for understanding client presentation variations rather than direct training signals. Future research should investigate whether engagement patterns correlate with therapeutic outcomes and explore alternative computational approaches to modeling client participation dynamics that may prove more suitable for training objectives.

Dataset Scale and Architectural Considerations: E-

THER's focused approach prioritizes annotation depth over dataset scale, achieving high annotation intensity (789 annotations/hour) through comprehensive four-dimensional analysis. This design choice enables detailed incongruence detection training while creating opportunities for investigating scaling strategies that maintain annotation quality. Future work should explore data augmentation techniques specific to therapeutic interactions and examine how incongruence detection capabilities transfer across different VLM architectures and larger datasets.

Evaluation Framework Robustness: Our PCT-based evaluation metrics demonstrate effectiveness in distinguishing sincere empathic communication from performative responses, yet systematic comparison with established clinical assessment instruments represents an important validation direction. Integration studies comparing our computational metrics against gold-standard measures such as the Jefferson Scale of Empathy [75] and Consultation and Relational Empathy Scale [76] would establish convergent validity and potential calibration protocols.

Human expert evaluation represents a widely-adapted validation component for establishing clinical or practical application of such systems. While computational metrics provide scalable and consistent measurement capabilities, further validation of the responses through licensed counsellors and other human participants can signify the achieved performance.

Empirical Observations Requiring Investigation: Preliminary analysis reveals evidence of *therapeutic memory persistence* and *conversational compactness* within sessions - phenomena where the model references earlier session content and demonstrate increasing communicative efficiency over time. These patterns, while theoretically consistent with therapeutic alliance development, require systematic quantitative analysis. Future research should develop metrics for measuring conversational coherence across session time and its correlation with therapeutic progress indicators.

IX. Conclusion

We present the first multimodal empathy dataset that enables artificial agents to develop empathic capabilities through detection of verbal-visual emotional incongruence. Our novel training methodologies demonstrate noticeable improvements over state-of-the-art models across three vision-language architectures using Person-Centered Therapy evaluation principles.

The comprehensive evaluation reveals significant performance improvements over GPT-4V across metrics, with notable gains in empathic authenticity, and appropriate questioning strategies. These results demonstrate that PCT-grounded training produces empathic AI models capable of genuine rather than superficial empathic communication.

The demonstrated effectiveness across multiple vision-language models suggests that incongruence aware empathic training represents a generalizable approach for enhancing empathy in AI. Future work should explore application of these principles to larger-scale models and diverse empathic communication contexts and clinical validation.

Our contributions provide foundational resources for artificial empathy research that prioritizes genuine empathic capabilities over superficial empathic language generation, promoting the development of nuanced empathic understanding and modeling in AI.

References

[1] H. Rashkin, E. M. Smith, M. Li, and Y.-L. 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*, 2018, pp. 5370–5381.

[2] A. Welivita and P. Pu, "Are large language models more empathetic than humans?" *arXiv preprint arXiv:2406.05063*, 2024.

[3] C. Montemayor, J. Halpern, and A. Fairweather, "In principle obstacles for empathic ai: why we can't replace human empathy in healthcare," *AI & society*, vol. 37, no. 4, pp. 1353–1359, 2022.

[4] C. R. Rogers, "The necessary and sufficient conditions of therapeutic personality change," *Journal of consulting psychology*, vol. 21, no. 2, pp. 95–103, 1957.

[5] A. Mehrabian and M. Wiener, "Decoding of inconsistent communications," *Journal of personality and social psychology*, vol. 6, no. 1, pp. 109–114, 1967.

[6] S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc, "Modeling empathy and distress in reaction to news stories," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4758–4765.

[7] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 121–132.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[9] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. 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*, 2019, pp. 527–536.

[10] Z. Zhu, X. Li, J. Pan, Y. Xiao, Y. Chang, A. Zhou, Y. Zheng, Y. Zhang, and S. Wang, "Medic: A multimodal empathy dataset in counseling," in *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023, pp. 2340–2350.

[11] Y. Chen, H. Liu, Y. Wang, J. Li, F. Zhang, H. Wu, J. Liu, and M. Zhang, "Towards multimodal emotional support conversation systems," *arXiv preprint arXiv:2408.03650*, 2024.

[12] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.

[13] M. Kolomaznik, V. Petrik, M. Slama, and V. Jurik, "The role of socio-emotional attributes in enhancing human-ai collaboration," *Frontiers in psychology*, vol. 15, p. 1369957, 2024.

[14] J. C. Watson, P. L. Steckley, and E. J. McMullen, "The role of empathy in promoting change," *Psychotherapy Research*, vol. 24, no. 3, pp. 286–298, 2014.

[15] C. R. Rogers, *On becoming a person: A therapist's view of psychotherapy*. Houghton Mifflin, 1961.

[16] ——, "A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework," *Psychology: A study of a science*, vol. 3, pp. 184–256, 1959.

[17] D. Hardman and J. Howick, "The friendly relationship between therapeutic empathy and person-centred care," *European Journal for Person Centered Healthcare*, vol. 7, no. 2, pp. 351–357, 2019.

[18] J. D. Bozarth, *Person-centered therapy: A revolutionary paradigm*. PCCS Books, 1998.

[19] E. Stevenson and Collaborators, "Does it matter if empathic ai has no empathy?" *Nature Machine Intelligence*, 2024, discusses risks and ethical concerns of empathic AI implementations. [Online]. Available: https://www.nature.com/articles/s42256-024-00841-7

[20] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. 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)*. Association for Computational Linguistics, 2021, pp. 3469–3483.

[21] Y. Zhang, F. Kong, P. Wang, S. Sun, S. SWangLing, S. Feng, D. Wang, Y. Zhang, and K. Song, "Stickerconv: Generating multimodal empathetic responses from scratch," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024, pp. 7707–7733.

[22] A. Welivita, Y. Xie, and P. Pu, "A large-scale dataset for empathetic response generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 1251–1264.

[23] Z. Wang, Y. Liu, and T. Zhao, "Emotionx-ar: Affective empathetic response generation with emotion regulation," in *Proceedings of EMNLP 2023*, 2023, pp. 4532–4545.

[24] S. Sabour, C. Zheng, and M. Huang, "A computational framework for understanding empathy in conversational ai," in *Proceedings of ACL 2022*, 2022, pp. 2537–2549.

[25] Y. Qian, W. Zhang, and T. Liu, "Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, pp. 6516–6528.

[26] Z. Xu and J. Jiang, "Multi-dimensional evaluation of empathetic dialogue responses," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 2066–2087.

[27] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "Computational approaches to empathy: A survey," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, 2020.

[28] H. Cai, Z. Yuan, Y. Gao *et al.*, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, no. 1, p. 178, 2022.

[29] A. Ben Abacha, W.-w. Yim, Y. Fan, and T. Lin, "An empirical study of clinical note generation from doctor-patient encounters," in *Proceedings of EACL*, 2023.

[30] T. Zhang, S. Sclaroff, and M. Betke, "Multimodal emotion recognition: A survey of methods, datasets, and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3704–3724, 2023.

[31] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[32] R. M. Bagby and G. J. Taylor, "A multimodal investigation of emotional responding in alexithymia," *Cognition and Emotion*, vol. 18, no. 6, pp. 781–799, 2004.

[33] S. Gannouni *et al.*, "Identifying complex emotions in alexithymia affected adolescents using machine learning techniques," *Diagnostics*, vol. 12, no. 12, p. 3188, 2022.

[34] Y. Etesam *et al.*, "Contextual emotion recognition using large vision language models," *arXiv preprint arXiv:2405.08992*, 2024.

[35] Z. Wang *et al.*, "Leveraging vision transformers and entropy-based attention for accurate micro-expression recognition," *Scientific Reports*, vol. 15, 2025.

[36] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024.

[37] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 19 730–19 742.

[38] A. Gandhi, R. Sharma, N. Patel, and S. Kumar, "Multimodal emotion recognition in social media: Integrating visual and textual cues with vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 7834–7847.

[39] S. Yoon, J. Kim, S. Lee, and C. Park, "Vision-language models for mental health assessment: Analyzing visual and textual indicators of depression and anxiety," *Journal of Medical Internet Research*, vol. 25, p. e45678, 2023.

[40] L. Chen, J. Williams, C. Martinez, and D. Thompson, "Enhancing healthcare communication assessment with vision-language models: Applications in patient-provider interactions," *npj Digital Medicine*, vol. 6, no. 1, p. 234, 2023.

[41] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards understanding and mitigating social biases in language models for conversation ai," *Nature Machine Intelligence*, vol. 5, no. 7, pp. 656–670, 2023.

[42] Z. Xu and J. Jiang, "Multi-dimensional evaluation of empathetic dialogue responses," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 2066–2087.

[43] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, 2016.

[44] S. Giorgi, J. Sedoc, L. Ungar, S. Buechel, A. Buffone, H. A. Schwartz, S. Dill, A. Ramakrishna, and V. Ganesan, "Psychological metrics for dialog system evaluation," *arXiv preprint arXiv:2305.14757*, 2023.

[45] "Perceived empathy of technology scale (pets): Measuring empathy of systems toward the user," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 2024.

[46] D. Ovsyannikova, V. de Mello, and M. Inzlicht, "Third-party evaluators perceive ai as more compassionate than expert humans," *Communications Psychology*, vol. 3, no. 4, 2025.

[47] S. Provence and A. Forcehimes, "Algorithms for empathy: Using machine learning to categorize common empathetic traits across professional and peer-based conversations," *PMC*, 2024.

[48] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mHealth and uHealth*, vol. 5, no. 6, p. e4288, 2017.

[49] S. Chancellor, Z. Lin, E. Goodman, S. Zerwas, and M. De Choudhury, "Mental health surveillance over social media with digital cohorts," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4306–4317.

[50] J. Shen, D. DiPaola, S. Ali, M. Sap, H. W. Park, C. Breazeal *et al.*, "Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study," *JMIR Mental Health*, vol. 11, no. 1, p. e62679, 2024.

[51] G. Kaluzeviciute, "The role of empathy in psychoanalytic psychotherapy: A historical exploration," *Cogent Psychology*, vol. 7, no. 1, p. 1748792, 2020.

[52] D. Mahon, "Evidence based relationships 4: Empathy, congruence, unconditional positive regard, and real relationship," in *Evidence based counselling & psychotherapy for the 21st century practitioner*. Emerald Publishing Limited, 2023, pp. 71–83.

[53] R. Ramadurai, "Silent signals: New review highlights the importance of nonverbal signals for perceived responsiveness," https://www.evidencebasedmentoring.org, 2024, accessed Sep 2025.

[54] L. S. Greenberg and J. D. Safran, "Emotion in psychotherapy: Affect, cognition, and the process of change," *Psychotherapy*, vol. 24, no. 3, pp. 253–264, 1987.

[55] W. Ickes, "Empathic accuracy," *Journal of personality*, vol. 61, no. 4, pp. 587–610, 1993.

[56] E. S. Bordin, "The generalizability of the psychoanalytic concept of the working alliance," *Psychotherapy: Theory, research & practice*, vol. 16, no. 3, pp. 252–260, 1979.

[57] A. O. Horvath, A. Del Re, C. Flückiger, and D. Symonds, "Alliance in individual psychotherapy," *Psychotherapy*, vol. 48, no. 1, pp. 9–16, 2011.

[58] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[59] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[60] H. Rashkin, E. M. Smith, M. Li, and Y.-L. 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*, 2019, pp. 5370–5381.

[61] H. Laurençon, L. Saulnier, L. Tronchon, V. Saulnier, C. Akiki, A. Villegas, M. Haddad, L. Barrault, X. Bresson, A. F. Aji *et al.*, "What matters when building vision-language models?" *arXiv preprint arXiv:2405.02246*, 2024.

[62] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2262–2272.

[63] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, 2023, pp. 19 730–19 742.

[64] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.13601

[65] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[66] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy," *Psychotherapy*, vol. 48, no. 1, pp. 43–49, 2011.

[67] J. A. Hall, S. Carter, M. C. Jimenez, and N. A. Frost, "Level of emotional awareness and mean length of utterance," *Emotion*, vol. 1, no. 4, pp. 325–333, 2001.

[68] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

[69] N. Flemotomos, V. R. Martinez, J. Gibson, D. C. Atkins, T. A. Creed, and S. Narayanan, "Language features for automated evaluation of cognitive behavior psychotherapy sessions," in *Interspeech*, 2018, pp. 1908–1912.

[70] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[71] G. T. Barrett-Lennard, "The empathy cycle: Refinement of a nuclear concept," *Journal of Counseling Psychology*, vol. 28, no. 2, pp. 91–100, 1981.

[72] J. Luo, J. Huang, and H. Li, "Artificial empathy in healthcare chatbots: Does it feel authentic?" *Computers in Human Behavior Reports*, vol. 6, p. 100347, 2024.

[73] G. Egan, *The skilled helper: A problem-management and opportunity-development approach to helping*, 10th ed. Brooks/Cole, 2014.

[74] C. B. Truax and R. R. Carkhuff, "Toward effective counseling and psychotherapy: Training and practice," 1967.

[75] M. Hojat, S. Mangione, T. J. Nasca, M. J. Cohen, J. S. Gonnella, J. B. Erdmann, J. J. Veloski, and M. Magee, "The jefferson scale of empathy: development and preliminary psychometric data," *Educational and Psychological Measurement*, vol. 61, no. 2, pp. 349–365, 2001.

[76] S. W. Mercer and W. J. Reynolds, "The development and preliminary validation of the consultation and relational empathy (care) scale for use in primary care," *Family Practice*, vol. 21, no. 6, pp. 699–705, 2004.