MIRROR: Multimodal Cognitive Reframing Therapy for Rolling with Resistance

Important: We explore how vision-language models support digital CBT, but they should NOT replace professional psychological treatment.

Anonymous ACL submission

Abstract

Recent studies have explored the use of large language models (LLMs) in psychotherapy; however, text-based cognitive behavioral therapy (CBT) models often struggle with client resistance, which can weaken therapeutic alliance. To address this, we propose a multimodal approach that incorporates nonverbal cues, which allows the AI therapist to better align its responses with the client's negative emotional state. Specifically, we introduce a new synthetic dataset, MIRROR (**M**ultimodal **I**nte**r**active **Ro**lling with **R**esistance), which is a novel synthetic dataset that pairs each client's statements with corresponding facial images. Using this dataset, we train baseline vision language models (VLMs) so that they can analyze facial cues, infer emotions, and generate empathetic responses to effectively manage client resistance. These models are then evaluated in terms of both their counseling skills as a therapist, and the strength of therapeutic alliance in the presence of client resistance. Our results demonstrate that MIRROR significantly enhances the AI therapist's ability to handle resistance, which outperforms existing text-based CBT approaches. Human expert evaluations further confirm the effectiveness of our approach in managing client resistance and fostering therapeutic alliance.

1 Introduction

Cognitive reframing is a central part of cognitive behavioral therapy (CBT), which helps individuals replace negative and intrusive thoughts with more rational and balanced ones. Towards this objective, large language models (LLMs) have recently shown great promise and are increasingly being explored in psychotherapy (Ziems et al., 2022; Maddela et al., 2023a; Sharma et al., 2023; Qu et al., 2023; Yang et al., 2023, 2024; Xiao et al., 2024; Na, 2024; Lee et al., 2024a). As such, these systems have actually been utilized in real-world applications as effective adjunct tools in psychotherapy, providing meaningful support for individuals with mental disorders such as depression and anxiety (Fitzpatrick et al., 2017; Haque and Rubya, 2023; Mehta et al., 2021)[1].

Figure 1: Text-based therapists have limitations in interpreting nonverbal cues, as they cannot perceive behaviors such as sighs or posture shifts, which can lead to premature problem-solving rather than addressing deeper emotions.

Despite this progress, the existing text-based CBT model struggles to detect and respond to client resistance (Wang et al., 2025), which is a common therapeutic challenge that involves the client's reluctance or opposition to change. This resistance often stems from the directive nature of CBT, where structured interventions may unintentionally provoke discomfort or defensiveness (Patterson and Chamberlain, 1994; Moyers and Martin, 2006; Constantino et al., 2017; Westra and Norouzian, 2018; Hara, 2020). Left unaddressed, resistance can diminish therapeutic alliance and reduce treatment efficacy. It is crucial to note that such resistance is frequently conveyed through non-

[1] A comprehensive review of related work is provided in Appendix B.

verbal cues like facial expressions, sighs, or posture shifts. Due to this property, pure-text-based models fail to perceive resistance, which leads to premature advice-giving rather than addressing deeper emotional needs (Figure 1). Addressing this limitation thus requires multimodal integration. However, collecting real multimodal psychotherapy data to train models to identify such multimodal cues, introduces severe privacy risks as sessions often involve deeply personal disclosures, including trauma, mental illness, and other confidential experiences.

In this work, we propose a multimodal approach to cognitive reframing that integrates both textual and nonverbal information to better detect and manage client resistance. We introduce MIRROR (**M**ultimodal **Inter**active **Ro**lling with **R**esistance), which is a synthetic dataset designed to simulate real therapeutic interactions. Specifically, MIRROR features generated dialogues between clients and therapists, annotated with client facial expressions reflecting three distinct types of resistance. We leverage LLMs to generate realistic session content, synthesize corresponding facial cues, and apply rigorous filtering to ensure quality and safety. This dataset enables the development of vision-language models (VLMs) tailored to CBT scenarios, where emotional alignment and alliance are essential. In addition, we propose *emotional captioning*, a novel reasoning method designed to explicitly interpret and respond to the client's emotional state.

We evaluate our approach using a VLM that is trained on the MIRROR dataset and enhanced with *planning* and *emotional captioning*. Compared to existing LLMs and VLMs, our model demonstrates superior performance across therapist skill assessment, alliance building, and applicability to real counseling scenarios. The results highlight the importance of multimodal approaches in managing client resistance and improving CBT outcomes.

Our contributions are summarized as follows:

- We explore a multimodal cognitive reframing for coping with client resistance, and present MIRROR, which features turn-level client facial expressions across diverse resistance types.

- We establish baseline models on the MIRROR dataset and propose an *emotional captioning* method, which helps VLMs generate emotionally aligned, vision-aware therapeutic responses.

To further support research in this area, we will publicly release our code and dataset.

2 Problem Definition

Our goal is to enhance the AI therapist's ability to manage client resistance by integrating both verbal and nonverbal cues through a multimodal approach. To guide the development and evaluation of such models, we define two key assessment dimensions that reflect essential aspects of effective therapy:

- **Therapist Skills Assessment:** Evaluates the AI therapist's competence in two key categories of general counseling skills and CBT-specific techniques.

- **Client Alliance Assessment:** Focuses on the AI therapist's ability to establish a strong therapeutic bond, which is critical for reducing resistance and promoting positive outcomes.

3 MIRROR: Multimodal Interactive Rolling with Resistance Dataset

As illustrated in Figure 2, the MIRROR dataset is constructed through three main steps, which is followed by a comprehensive quality and safety validation process. Through dataset synthesis, we generate over 3,000 multimodal counseling dialogues, with each client turn annotated with a facial expression image that captures the client's emotional state[2].

3.1 Step 1: Multimodal Dialogue Design

To build the multimodal dialogue design for MIRROR, we combine facial and textual data from two sources: CelebA (Liu et al., 2015) for facial expressions and CACTUS (Lee et al., 2024a) for text-based cognitive reframing therapy. While CACTUS is originally a text-only dialogue dataset, we only extract its underlying structured profiles, which includes client intake forms, thinking traps, counseling plans, and CBT techniques.

In order to assign a facial identity to each client, we pair every CACTUS profile with a CelebA image based on gender and age predictions from the DeepFace library (Serengil and Ozpinar, 2021).

We further augment each client profile with four distinct resistance types: cognitive, emotional, behavioral, and non-resistant, following the taxonomy proposed by Beal III et al. (2013). Rather than assigning a single resistance label to each profile, we

[2]All used prompts are provided in Appendix J.

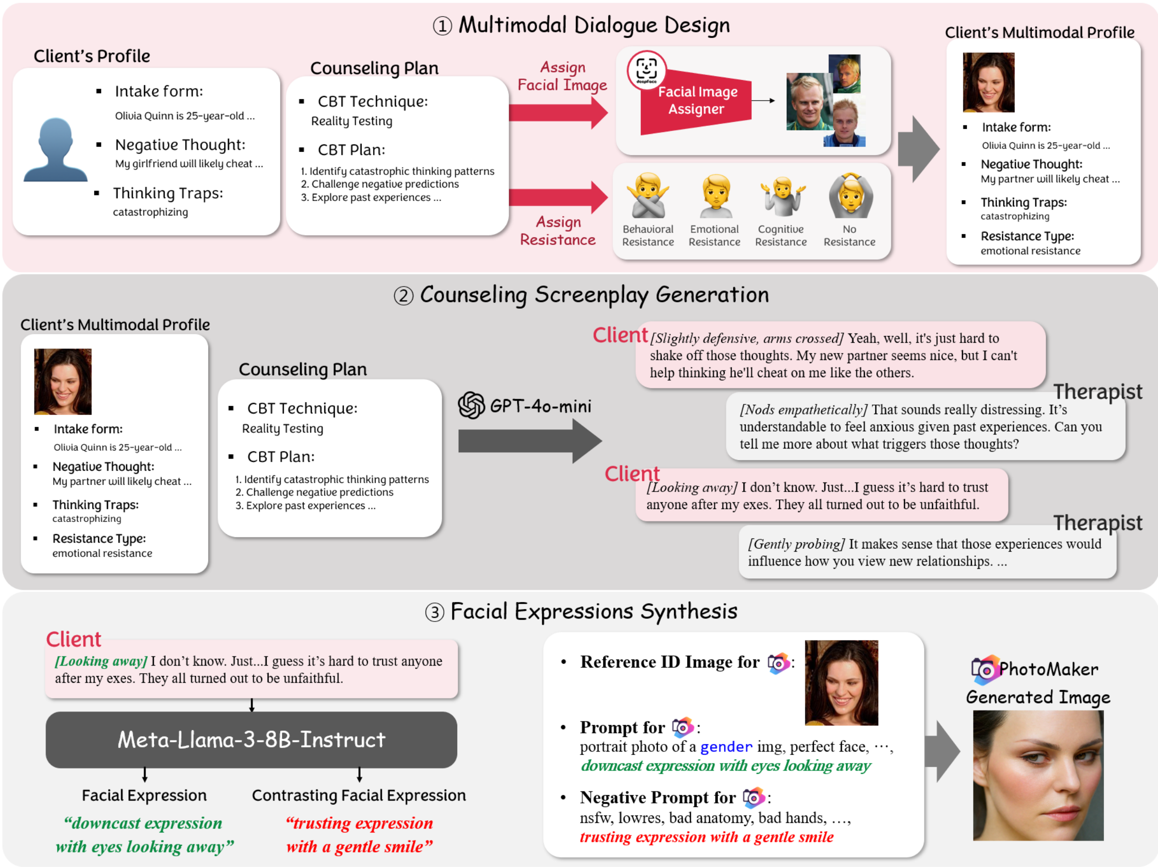


Figure 2: Overview of the MIRROR dataset construction. The pipeline consists of three main stages: Multimodal Dialogue Design (§3.1), Counseling Screenplay Generation (§3.2), and Facial Expression Synthesis (§3.3).

generate four variants per client, each conditioned on a different resistance type. This results in four variants per client, allowing the model to encounter diverse resistance behaviors in the same therapeutic context and ensuring class balance across the dataset. This process yields a complete multimodal dialogue setup for each client, where a structured CACTUS profile, a facial identity, and a specified resistance type are jointly configured. The resulting design supports therapeutically grounded dialogue generation based on client context and CBT plan.

3.2 Step 2: Counseling Screenplay Generation

We synthesize counseling dialogues in the form of screenplays rather than plain transcripts, to more naturally reflect the emotional nuance of real therapeutic interactions. A key advantage of this format is its explicit representation of nonverbal cues through stage directions (e.g., *[slightly defensive, arms crossed]'* in Figure 2).

These stage directions serve two critical purposes: (1) They enrich the textual context by capturing subtle emotional dynamics that are characteristic of real therapy sessions. (2) They act as structured signals for downstream facial expression synthesis, which ensures the generation of consistent and emotionally aligned client images. Based on the predefined profiles, these screenplays are generated using GPT-4O-MINI[3].

3.3 Step 3: Facial Expression Synthesis

After constructing the screenplay, we synthesize turn-level facial expressions that reflect the emotional dynamics conveyed through both verbal content and stage directions. The key contribution of this step lies in designing a prompt construction method that encodes nonverbal cues into the image generation process.

We leverage **PhotoMaker** (Li et al., 2024b), which is a diffusion-based model that takes three inputs: a reference image to preserve facial identity, a positive prompt for the desired expression, and a negative prompt to suppress conflicting features. To generate these prompts, we condition LLAMA-3-8B (AI@Meta, 2024) on the full client utterance, which includes inline stage directions (see

[3]Version `gpt-4o-mini-2024-07-18`.

	Modality	Language	# of Dialogue	# Avg. Turns	# Avg. Images	Turn-Image Alignment
Psych8k (Liu et al., 2023)	T	English	8,187	1.00	-	-
HealMe (Xiao et al., 2024)	T	English	1,300	3.00	-	-
CACTUS (Lee et al., 2024a)	T	English	31,577	16.6	-	-
CPsyCounD (Zhang et al., 2024)	T	Chinese	3,134	8.7	-	-
M2CoSC (Kim et al., 2025)	T, V	English	429	4.00	1.00	✗
MEDIC (Zhu et al., 2023)	T, V, A	Chinese	771	1.00	1,137	⚠
MIRROR	T, V	English	3,073	10.3	9.51	✓

Table 1: Comparison of MIRROR with other psychological counseling datasets. The **Modality** column indicates whether the dataset includes text (T), visual (V), or audio (A) data. **# Avg. Images** refers to the average number of client images per dialogue. **Turn-Image Alignment** indicates whether the client images are dynamically aligned according to each dialogue turn.

Figure 2). As a result, LLAMA-3-8B produces two facial expression descriptions: a target expression (e.g., "downcast expression with eyes looking away") and a contrasting one (e.g., "trusting expression with a gentle smile"), which populate the positive and negative prompts, respectively.

This approach enables the synthesis of emotionally aligned client images throughout the dialogue. As shown in Figure 2, expressions like *[looking away]* are clearly expressed in the synthesized images. By translating nonverbal cues into structured prompts, we ensure that facial expressions reflect the client's emotional state, even when the textual utterance alone does not explicitly convey it. The role of stage direction in image synthesis is further examined in Appendix I.

3.4 Step 4: Filtering for Quality and Safety

Dataset Quality Filtering To ensure the overall quality and coherence with image of multimodal counseling dialogues, we apply six filtering approaches: ***(1) Image-Text Similarity Filtering*** uses CLIP (Radford et al., 2021) to measure alignment between generated images and stage directions, and cases with low similarity score below 0.2 are discarded (2.95% rejected). ***(2) Identity Preservation Filtering*** employs ArcFace (Deng et al., 2019) to maintain facial similarity across dialogue turns, rejecting cases with low similarity scores below 0.3 (66.05% rejected). ***(3) Gender Preservation Filtering*** utilizes DeepFace to ensure that the detected gender matches the client's original multimodal profile, removing mismatches (15.39% rejected). ***(4) Basic Filtering*** eliminates dialogues that contain utterances longer than 100 words, exhibit unnatural repetition of the same part-of-speech more than three times in a row, or include too few (fewer than 4) or too many (more than 20) conversation turns (1.03% rejected). ***(5) Copy-Paste Filtering*** removes instances where client personas are unnaturally stated instead of being contextually integrated, improving dialogue realism (1.36% rejected). Lastly, ***(6) Therapeutic Alliance Filtering*** assesses the quality of the counseling interaction using GPT-4o[4] to evaluate WAI[5] (Li et al., 2024a), and dialogues with an average score below 0.3 are discarded (10.01% rejected).

Dataset Safety Filtering To uphold ethical standards and prevent harmful content, we apply two additional approaches. ***(1) NSFW Filtering*** uses a Not-Safe-For-Work (NSFW) detector[6] to remove images that are visually unsuitable for mental health dialogue contexts. ***(2) Dialogue Safety Filtering*** leverages Canary (Kim et al., 2022) to identify and eliminate instances containing toxic, unethical, or unsafe language, in accordance with prior safety protocols (Kim et al., 2023; Lee et al., 2024b) (1.09% rejected). These layered filtering stages are critical for constructing a high-quality dataset that is not only realistic and coherent but also ethically robust and clinically applicable.

3.5 Comparative Analysis of MIRROR

Through the preceding stages, we have curated the first multimodal CBT dataset that explicitly incorporates client resistance. As shown in Table 1, MIRROR contains a comparatively large number of dialogues with high turn density and dynamic visual responses. Unlike prior datasets such as M2CoSC (Kim et al., 2025), which uses a single static image per dialogue, or MEDIC (Zhu et al., 2023), which is limited to a single turn, MIRROR provides image sequences that evolve turn-by-turn in alignment with client emotion.

[4]Version `gpt-4o-2024-08-06`.

[5]WAI stands for Working Alliance Inventory.

[6]`https://huggingface.co/Falconsai/nsfw_image_detection`

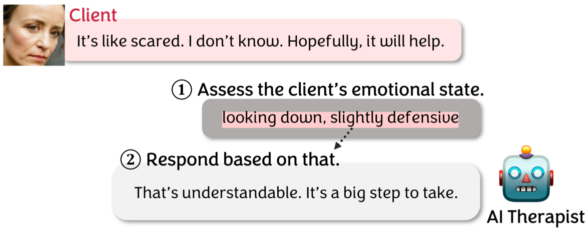


Figure 3: Overview of emotional captioning. The AI therapist infers the client's emotional state from facial cues and uses it to generate an empathetic, aligned response.

4 Emotional Captioning

To handle client resistance more effectively, we introduce *emotional captioning*, which is a reasoning module that interprets the client's emotional state from facial expressions. At each dialogue turn, the model receives a facial image and generates a short textual description of the client's emotional state (e.g., *looking down, slightly defensive*). This emotional caption is then used to guide the AI therapist's response, enabling more context-sensitive and empathetic interactions (Figure 3). By grounding the model's behavior in visual cues, *emotional captioning* supplements verbal input with nonverbal affective signals, improving alignment with the client's psychological state[7].

5 Experimental Settings

Following Smith et al. (2022); Liu et al. (2023), and Lee et al. (2024a), we assess the AI therapist based on full simulated counseling sessions rather than turn-level assessments. Each session involves an AI therapist interacting with a virtual client exhibiting varying types of resistance. We compare different model variants to assess the contribution of *planning* and *emotional captioning* strategies.

5.1 Client Agents with Resistance

We adopt GPT-3.5-TURBO[8] as the virtual client and conduct simulations based on predefined multimodal profiles. The evaluation is carried out with 600 unique client profiles, which are not included in the MIRROR dataset. These profiles include three types of resistance, with 200 examples for each type. Additionally, to assess performance degradation compared to non-resistant clients, we include 200 examples of non-resistant clients. Each session is considered terminated if the client attempts to disengage after two consecutive turns.

Within each client's utterance, nonverbal cues are embedded as stage directions within brackets, as described in §3.3. Note that these cues are used for facial expression generation and are invisible to the AI therapist when generating responses. For facial expressions, we generate LLM-based client's images at each turn, following the same process used in dataset construction (§3.3 and §3.4). Appendix K provides client setup and simulation prompt details.

5.2 AI Therapist Model Variations and Baselines

AI Therapist Baselines Our primary baseline, MIRROR-LLAVA, is a LLAVA-V1.5-7B (Liu et al., 2024) trained on the MIRROR dataset. To examine the benefit of multimodal integration, we include CAMEL-LLAMA3[9], a text-only CBT model trained on therapeutic dialogues (Lee et al., 2024a). We also evaluate general-purpose models that are not fine-tuned for counseling: LLAMA-3-8B, LLAVA-V1.5-7B, and GPT-3.5-TURBO. These serve as non-specialized baselines to assess the impact of domain adaptation and modality alignment. Further implementation details, including training procedures and reasoning prompts, are provided in Appendices C and D.

Reasoning Variants We incorporate two variants of chain-of-thought (CoT) reasoning: *planning* and *emotional captioning*. *Planning*, based on Lee et al. (2024a), is a pre-session reasoning process in which the AI therapist analyzes the client's background and objectives to formulate a structured counseling strategy. This strategy helps the model select suitable CBT techniques and maintain a facilitative role during the session, rather than directly challenging the client's thoughts. We denote models incorporating these reasoning strategies with the subscripts P (*planning*) and EC (*emotional captioning*).

5.3 Metrics for Assessment

As defined in §2, we evaluate the therapist's ability to manage client resistance across two key areas: therapist skills and client alliance.

Therapist skills are assessed using the COUNSELINGEVAL framework (Lee et al., 2024a), which

[7] Prompt templates used in this process are detailed in Appendix K.

[8] Version `gpt-3.5-turbo-0125`.

[9] https://huggingface.co/cactus-camel/camel-llama3

Model	General Counseling Skills (↑)			CBT-specific Skills (↑)		Response Length	
	Understanding	Interpersonal Effectiveness	Collaboration	Guided Discovery	Focus	Avg.	Max
LLaMA-3-8B	3.811* | -0.073	**4.114** | -0.012	2.734* | -0.311	3.689* | -0.096	3.692* | -0.057	59.36	104.59
Camel-LLaMA3	3.794* | -0.085	4.003* | -0.002	2.279* | -0.198	3.527* | -0.127	3.563* | -0.197	20.54	27.42
GPT-3.5-Turbo	3.798* | -0.172	4.049 | -0.041	2.976* | -0.194	3.462* | -0.262	3.491* | -0.238	36.19	57.28
LLaVA-v1.5-7B	3.622* | -0.066	3.997* | +0.007	3.408* | +0.071	2.494* | +0.057	2.501* | -0.012	112.41	177.11
Mirror-LLaVA	3.973* | -0.017	4.025 | -0.040	3.576* | -0.089	3.875* | -0.025	3.888* | -0.012	27.68	32.14
Mirror-LLaVA$_P$	3.985 | -0.015	4.098 | +0.063	3.722* | +0.117	3.915* | -0.040	3.915* | +0.015	27.00	32.02
Mirror-LLaVA$_{P+EC}$	**4.000** | +0.010	4.055 | +0.010	**3.913** | -0.082	**3.977** | +0.007	**3.977** | +0.037	27.55	34.20

Table 2: Therapist skills assessment scores calculated by GPT-4o and response length. Asterisk ($*$) indicates a significant difference compared to Mirror-LLaVA$_{P+EC}$ ($p < 0.05$, paired t-test). **Response Length** denotes the average and maximum number of tokens per turn. Values after the vertical bar (|) indicate performance changes when interacting with resistant clients, relative to non-resistant clients; negative values denote a decline.

covers both general counseling skills and CBT-specific competencies. In particular, general counseling skills encompass the ability to interpret client concerns (Understanding), maintain a therapeutic relationship (Interpersonal Effectiveness), and facilitate collaborative decision-making (Collaboration). Meanwhile, CBT-specific skills evaluate the ability to guide clients in discovering their thoughts (Guided Discovery) and identify mal-adaptive patterns (Focus). Each component of the therapist's skills is rated on a scale from 0 to 6[10].

Client alliance is measured following Li et al. (2024a), which assesses agreement of therapy objectives (Goal), engagement in counseling tasks (Approach), and the strength of emotional connection (Affective Bond), and is scored from 1 to 5.

6 Results and Discussion

6.1 Therapist Skills Assessment

Table 2 reports the evaluation of therapist skills in interactions with resistant clients.

Text-based Versus Vision Language Models As can be seen, text-based LLMs generally struggled to engage with resistant clients, particularly in collaborative interactions that demand heightened sensitivity to client emotions. This can be seen in the significant drop in performance compared to non-resistant clients. In contrast, vision-enhanced models showed greater resilience, maintaining higher scores even when interacting with resistant clients. These results highlight the importance of nonverbal cues in effectively managing challenging client interactions.

[10]For our experiments, we do not use a Strategy score, which assesses the coherence of intervention strategies, as it strongly correlates with the length of the AI therapist's responses (see Appendix E).

Fine-Tuning and CoT on CBT Performance Compared to LLaVA-v1.5-7B, which is the backbone model of Mirror-LLaVA, the Mirror-LLaVA family models achieved significantly higher scores in CBT-specific skills. This demonstrates the effectiveness of the Mirror dataset in enhancing CBT skills and reinforces the notion that, despite being trained on vast amounts of pre-existing data, LLMs still require targeted fine-tuning to effectively internalize and apply CBT principles. Further performance gains were observed when CoT processes, such as *planning* and *emotional captioning*, resulting in responses that were more contextually appropriate and emotionally attuned to the client's needs.

Analysis of Response Length Excessively long response generation has been a persistent issue for LLMs and is known to reduce user satisfaction (Huang et al., 2024). Our analysis of response length revealed that, with the exception of the fine-tuned CBT counseling models (i.e. Camel-LLaMA3 and Mirror-LLaVA family models), most models generated responses exceeding 30 tokens, which can degrade the counseling effectiveness. To further investigate these results, we provide actual examples for each model in Appendices H.1 and H.2, and conduct an error analysis in Appendix G.

6.2 Client Alliance Assessment

Table 3 presents the client alliance assessment using GPT-4o, which evaluates how well each model supports goal completion, establishes rapport (Approach), and fosters emotional connection (Affective Bond).

Model	Client Alliance Skills (↑)		
	Goal	Approach	Affective Bond
LLAMA-3-8B	2.412* | -0.023	3.309* | -0.107	3.356* | -0.138
CAMEL-LLAMA3	2.358* | -0.009	3.130* | -0.072	3.149* | -0.203
GPT-3.5-TURBO	2.472* | -0.018	3.272* | -0.168	3.297* | -0.253
LLAVA-V1.5-7B	**2.589** | -0.048	3.234* | -0.181	3.356* | -0.163
MIRROR-LLAVA	2.459* | -0.033	3.289* | -0.060	3.400* | -0.092
MIRROR-LLAVA$_P$	2.525* | +0.033	3.340 | -0.005	3.448 | -0.051
MIRROR-LLAVA$_{P+EC}$	2.567 | +0.035	**3.366** | -0.003	**3.480** | -0.024

Table 3: Client alliance assessment results as evaluated by GPT-4O.

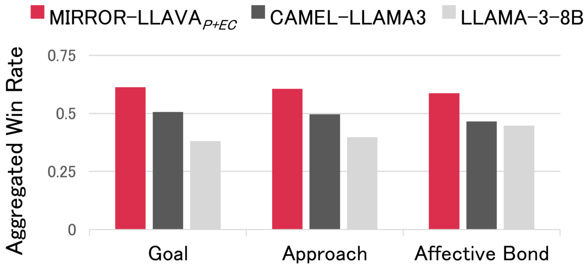


Figure 4: Pairwise comparison results among MIRROR-LLAVA, CAMEL-LLAMA3 and LLAMA-3-8B, on three evaluation criteria—Goal, Approach, and Affective Bond—rated by two psychotherapists.

While overall alliance scores improve with MIRROR, we observe a modest decline in the "Goal" score for MIRROR-LLAVA models compared to some baselines. We attribute this to the design of the MIRROR dataset, which emphasizes emotional engagement and rapport-building in resistant counseling scenarios, rather than directive goal setting. In real-world counseling, especially under resistance, it is often more effective to prioritize emotional engagement before directive goal-setting. This trade-off is reflected in the substantial gains in Approach and Affective Bond scores, which more directly capture the model's capacity for empathy and responsiveness. Notably, MIRROR-LLAVA$_{P+EC}$ achieves the highest scores in these affective dimensions, demonstrating the strength of step-by-step reasoning in managing resistance.

6.3 Domain Expert Assessment

To further validate previous client alliance results, we conducted pairwise comparisons between MIRROR-LLAVA$_{P+EC}$, LLAMA-3-8B, and CAMEL-LLAMA3 using 200 randomly selected cases from the test set, balanced across three resistance categories: emotional, cognitive, behavioral. Specifically, two domain experts evaluated the models and selected the better model in each comparison (Appendix F).

Moreover, we focused on comparing our method against the strongest baselines in CBT counseling—LLAMA-3-8B and CAMEL-LLAMA3, which are ranked highest in CBT-specific skill. Figure 4 shows the average win rate across all pairwise comparisons. As depicted, the win rate confirmed that MIRROR-LLAVA$_{P+EC}$ consistently outperformed its counterparts across all three dimensions of the therapeutic alliance. This result confirms that our model is not only favored in automatic evaluations but also by actual counseling experts.

In particular, while GPT-based evaluation showed limited gains in the "Goal" dimension, domain experts more frequently selected MIRROR-LLAVA$_{P+EC}$ as superior in goal-related dialogue segments. Experts noted that goal pursuit was achieved more implicitly through sustained rapport and motivational alignment, rather than through direct or premature intervention. This reinforces our claim that emotional connection should precede goal-setting in resistant counseling contexts.

It is interesting to note that GPT-4O's evaluation in Table 3 ranked LLAMA-3-8B higher than other LLMS, whereas domain experts preferred CAMEL-LLAMA3 significantly more in pairwise comparisons. This discrepancy is likely due to human preference for responses of a more natural length, rather than those that are excessively long.

6.4 Application in Real-World Counseling Demonstrations

In this section, we conduct analyses on actual client interactions. Due to the lack of CBT-based counseling videos, we instead analyzed motivational interviewing (MI) demonstration videos to assess how our model responds to various forms of resistance in practice. For this, we specifically used the AnnoMI dataset (Wu et al., 2022), which contains 133 counseling session videos along with transcriptions of the clients' utterances. By examining both the video frames and client statements, we identified specific client utterances that show resistance and explored how MIRROR$_{P+EC}$ responds to these resistant remarks.

As shown in Figure 5, clients exhibit various forms of resistance, including reluctance to seek help (Client A), claiming impunity (Client B), minimizing concerns (Client C), and externalizing blame (Client D)[11]. The case analysis demon-

[11]These are well-documented resistance patterns in psy-

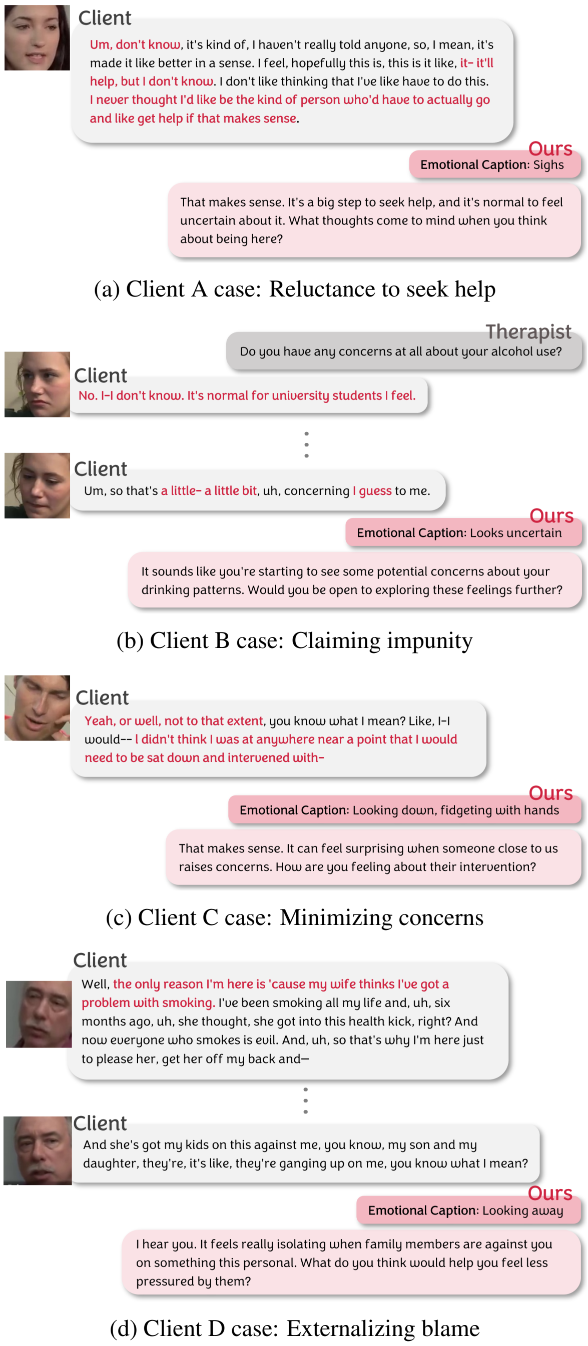


Figure 5: Four examples of MIRROR-LLAVA$_{P+EC}$ responses in psychological counseling, showcasing its ability to handle resistance through validation and open-ended questioning.

strates that MIRROR$_{P+EC}$ effectively identifies the client's emotional state through captioning and responds with emotional validation[12] and open-ended questions, common therapeutic techniques for managing resistance (Miller and Rollnick, 2002). For example, when Client D externalizes blame onto family, the therapist acknowledges their feelings of isolation while gently redirecting the conversation toward exploring ways to cope with pressure. Further case studies can be found in Appendix H.3.

7 Conclusion

In this paper, we explore the use of multimodal cognitive reframing therapy for managing client resistance. Given the challenges faced by LLMs in addressing resistant clients and the potential advantages of VLMs, we aim to enhance AI therapists' ability to manage resistance by incorporating nonverbal cues, particularly facial expressions, to detect and understand client resistance. To address this challenge while mitigating privacy concerns associated with real-person data, we developed MIRROR, a novel synthetic dataset for multimodal cognitive reframing therapy. Additionally, we have evaluated the AI therapist's performance in two key areas: therapeutic skills and alliance-building, as well as adaptability to real-world counseling scenarios. Our results demonstrate significant improvements in both areas when trained with MIRROR, underscoring its potential for real-world therapeutic applications. These improvements contribute to the development of AI therapists that are more empathetic and capable of fostering stronger therapeutic relationships.

Limitations

Biases in Image Generation We utilized LLMs to generate image prompts, which were then used to create images via PhotoMaker. During the process of generating image prompts based on the client's stage directions and utterances, cultural biases inherent in the LLMs (AlKhamissi et al., 2024; Naous et al., 2024) may have influenced the output. For example, if a client says, "*[Smiling] Hi,*" the LLMs might generate a prompt like "eyes curved like a crescent moon," reflecting an East Asian interpretation of a smile, whereas other cultures might describe a smile based on visible teeth or dimples (Srinivasan and Martinez, 2021). Similarly, if a client expresses sadness, the prompt may depict downcast eyes and a bowed head, aligning with Western portrayals of sadness, while some cultures associate sadness with a neutral or stoic expression. These examples show how cultural biases in LLMs can shape image generation, reinforcing specific interpretations while overlooking others.

chotherapy (Miller and Rollnick, 2002).

[12]This term refers to accepting a client's emotions without judgment, helping to sense of being understood while encouraging to manage their emotions.

Authenticity in AI-Generated Images Since the final facial images were synthesized using an AI-based image generation model, they may not fully replicate the characteristics of real-world expressions. Subtle nuances, such as micro-expressions, muscle tension, and natural asymmetries, might be lost or inaccurately rendered, potentially affecting their authenticity. These limitations could influence how AI therapists perceive and interpret emotional cues, as synthesized images may lack the depth and variability of genuine human expressions.

Scope and Session Length In contrast to typical counseling sessions, which last about an hour and extend over multiple sessions, our dataset consists of relatively short, single-session dialogues. This limitation makes it challenging to facilitate deeper exploration of cognitive distortions or sustained reframing over multiple sessions. Future research should focus on extending dialogues to longer, multi-session interactions and incorporating a wider range of counseling techniques to enable AI to provide more continuous and in-depth therapeutic support.

Conversational Structure and Termination Our framework does not impose strict turn-level constraints or predefined termination points within the dialogue. While we incorporate a counseling strategy, **planning**, to maintain goal orientation, the absence of explicit session boundaries may result in prolonged interactions without meaningful therapeutic progress. For example, if a simulated client remains in a negative emotional state, the AI therapist may continue offering supportive statements rather than facilitating cognitive change. This limitation highlights the importance of incorporating clearer session structures or exit strategies in future designs to better align with therapeutic goals.

Evaluating MIRROR Using a GPT-Based Client In § 5, we have conducted counseling sessions using a GPT-based client with a model trained on the MIRROR dataset. Through these conversation records, we were able to demonstrate the effectiveness of the MIRROR dataset, and expert evaluations were also conducted on these dialogues. While GPT effectively simulates real client behavior and demonstrates strong conversational abilities, validating our approach through real-time human interactions would have further strengthened our findings. However, due to cost constraints and privacy concerns, real-time human interactions were not feasible. Instead, we evaluated our approach in a real therapy scenario, as described in § 6.4.

Diversity of Resistance Resistance can be expressed not only through facial expressions but also through overall body posture, voice tone, speech timing, and other nuanced factors. While these elements collectively shape how resistance manifests, this study focuses specifically on facial expressions and utterances, as they provide the most direct and observable indicators. Future research could enhance the dataset by incorporating audio, body posture, and other multimodal elements, leading to a more comprehensive understanding of resistance in therapeutic interactions.

Model Selection and Generalization Although we trained the LLAVA-V1.5-7B model with two different CoT options and demonstrated its strong performance in handling client resistance and CBT counseling, our evaluation was based on a single backbone model. This could be a limitation, as there may be other VLMs that could perform better or differently, depending on their architecture or training. The reliance on a single model limits the generalizability of our findings. To address this, we plan to explore and evaluate additional VLMs in future work, comparing their performance across various therapeutic tasks and client engagement scenarios.

Ethical Statement

Privacy Considerations for Images Ensuring privacy and ethical integrity is a fundamental priority in our dataset construction. We utilize the CelebA dataset (Liu et al., 2015), which is distributed under the MMLAB license. This license strictly prohibits commercial use and redistribution of the dataset. In compliance with these terms and to respect the rights of the individuals depicted in the images, we do not share the raw images directly. Instead, we provide image links and code that enables researchers to process the dataset independently, ensuring that the dataset's usage remains within ethical and legal boundaries.

Privacy Considerations for Dialogue The dialogue seeds for this dataset were sourced from the CACTUS dataset (Lee et al., 2024a), with PATTERN-REFRAME (Maddela et al., 2023b) serving as its seed dataset. This dataset does not contain actual medical records but was collected through crowdsourcing, where each participant was assigned a persona and instructed to role-play. Additionally, during the dataset generation process, no utterances were derived from real individuals' personas; instead, all dialogues were fully synthesized. This approach further mitigates privacy concerns by ensuring that no personal data is incorporated into the dataset.

Safety Considerations While AI has the potential to provide support, it may also have unintended negative effects on individuals with mental health challenges (Luxton, 2014). Although our model has demonstrated some degree of effectiveness, our primary objective was to explore whether AI can effectively engage with patients who exhibit resistance to therapy. Therefore, we believe that AI should be used under the supervision of a professional rather than serving as a standalone tool in counseling sessions, particularly for individuals with severe psychological conditions beyond its intended scope. Additionally, to ensure the safety and appropriateness of the dataset, we implemented NSFW filtering and incorporated Canary to identify and remove conversations that may require human intervention.

Bias Considerations Although we utilize randomly selected images and a dialogue seed dataset that incorporates diversity in age, gender, and occupation, there remains a possibility of bias in our dataset. This is primarily due to our reliance on LLMs, which are predominantly trained on Western-centric datasets. In particular, during the screenplay generation process, gestures and nonverbal cues may vary across cultures. Since these were generated using GPT-4O-MINI, certain gestures may not align with cultural norms in specific regions. Therefore, to ensure cultural appropriateness, retraining and adaptation would be necessary before deploying the model in a specific country.

References

AI@Meta. 2024. Llama 3 model card.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Loyd Beal III, Jacqueline M Stavros, and Matthew L Cole. 2013. Effect of psychological capital and resistance to change on organisational citizenship behaviour. *SA Journal of Industrial Psychology*, 39(2):1–11.

Michael J Constantino, Henny A Westra, Martin M Antony, and Alice E Coyne. 2017. Specific and common processes as mediators of the long-term effects of cognitive-behavioral therapy integrated with motivational interviewing for generalized anxiety disorder. *Psychother Res*, 29(2):213–225.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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 mental health*, 4(2):e7785.

Anmol Goel, Nico Daheim, and Iryna Gurevych. 2024. Socratic reasoning improves positive text rewriting. *arXiv preprint arXiv:2403.03029*.

MD Romael Haque and Sabirat Rubya. 2023. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1):e44838.

Kimberley Michelle Hara. 2020. Comparing therapist responsivity to resistance markers in cognitive-behavioural therapy and motivational interviewing integrated with cognitive-behavioural therapy for generalized anxiety.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shih-Hong Huang, Ya-Fang Lin, Zeyu He, Chieh-Yang Huang, and Ting-Hao Kenneth Huang. 2024. How does conversation length impact user's satisfaction? a case study of length-controlled conversations with llm-powered chatbots. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Lee. 2025. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880, Albuquerque, New Mexico. Association for Computational Linguistics.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024a. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeong-Jin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. 2024b. Stark: Social long-term multi-modal conversation with persona commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12137–12162, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024a. Understanding the therapeutic relationship between counselors and clients in online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024b. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

David D Luxton. 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial intelligence in medicine*, 62(1):1–10.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023a. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023b. Training models to generate, recognize, and reframe unhelpful thoughts. *arXiv preprint arXiv:2307.02768*.

Ashish Mehta, Andrea Nicole Niles, Jose Hamilton Vargas, Thiago Marafon, Diego Dotta Couto, and James Jonathan Gross. 2021. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (youper): longitudinal observational study. *Journal of medical Internet research*, 23(6):e26771.

W R Miller and S Rollnick. 2002. *Motivational interviewing: Preparing people for change*. The Guilford Press.

Theresa B Moyers and Tim Martin. 2006. Therapist influence on client language during motivational interviewing sessions. *J Subst Abuse Treat*, 30(3):245–251.

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Gerald R Patterson and Patricia Chamberlain. 1994. A functional analysis of resistance during parent training therapy. *Clinical Psychology: Science and Practice*, 1(1):53.

Renyi Qu, Lyle Ungar, and João Sedoc. 2023. Conditioning on dialog acts improves empathy style transfer. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.

Roz Shafran, Pamela Myles-Hooton, Sophie Bennett, and Lars-Göran Öst. 2021. The concept and definition of low intensity cognitive behaviour therapy. *Behav Res Ther*, 138:103803.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Ramprakash Srinivasan and Aleix M. Martinez. 2021. Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing*, 12(3):707–721.

Yinzhou Wang, Yimeng Wang, Ye Xiao, Liabette Escamilla, Bianca Augustine, Kelly Crace, Gang Zhou, and Yixuan Zhang. 2025. Evaluating an llm-powered chatbot for cognitive restructuring: Insights from mental health professionals. *arXiv preprint arXiv:2501.15599*.

Henny A Westra and Nikoo Norouzian. 2018. Using motivational interviewing to manage process markers of ambivalence and resistance in cognitive behavioral therapy. *Cognitive Therapy and Research*, 42:193–203.

Chris Williams. 2001. Use of written cognitive–behavioural therapy self-help materials to treat depression. *Advances in Psychiatric Treatment*, 7(3):233–240.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. 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*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6054–6062, New York, NY, USA. Association for Computing Machinery.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

A License

MIRROR is constructed using the CelebA (Liu et al., 2015) and the CACTUS datasets (Lee et al., 2024a). CelebA is released under the MMLAB license, which restricts redistribution, while CACTUS is licensed under the GPL-2.0 license, permitting non-commercial scientific use. In adherence to these licensing terms, we do not directly include images from these datasets in MIRROR. Instead, we provide links to the original sources. Consequently, MIRROR is distributed under the GPL-2.0 license, ensuring compliance with the licensing conditions of the datasets used.

B Related Work

Research on AI-assisted cognitive reframing therapy has largely focused on text-based approaches using LLMs. Early studies explored sentence rewriting to address cognitive distortions (Ziems et al., 2022; Maddela et al., 2023a; Sharma et al., 2023; Goel et al., 2024), leveraging evidence that low-intensity CBT interventions can be effective in self-help formats (Williams, 2001; Shafran et al., 2021). More recent advancements have shifted toward conversational approaches, moving from simple query-response interactions (Na, 2024; Liu et al., 2023) to structured, multi-turn frameworks. For instance, Xiao et al. (2024) introduced a multi-turn framework with a structured three-stage process to ensure that AI serves as a facilitator rather than a direct corrector. Other studies have focused on improving the realism of cognitive reframing datasets (Lee et al., 2024a) and enhancing AI therapists' professional counseling competence (Zhang et al., 2024).

Most recently, there has been a growing interest in incorporating nonverbal cues into AI-driven psychotherapy. Expanding on Xiao et al. (2024), Kim et al. (2025) explored multimodal cognitive reframing, showing that VLMs can generate more empathic responses. Similarly, Zhu et al. (2023) emphasized clients' nonverbal expressions, introducing a multimodal empathy dataset. Our work advances multimodal cognitive reframing by focusing on managing client resistance, strengthening the therapeutic alliance, and improving AI-assisted psychotherapy.

C Training Details

The LLAVA-V1.5-7B model was fine-tuned on the MIRROR dataset using LoRA (Hu et al., 2022) for 5 epochs. We used the official LLAVA-V1.5-7B model from Hugging Face[13] and followed the default hyperparameters[14], which include a learning rate of 2e-5, an AdamW optimizer without weight decay, and a cosine learning rate schedule with a 3% warmup ratio. Training was done on four A100-80GB GPUs with a batch size of 32 per GPU.

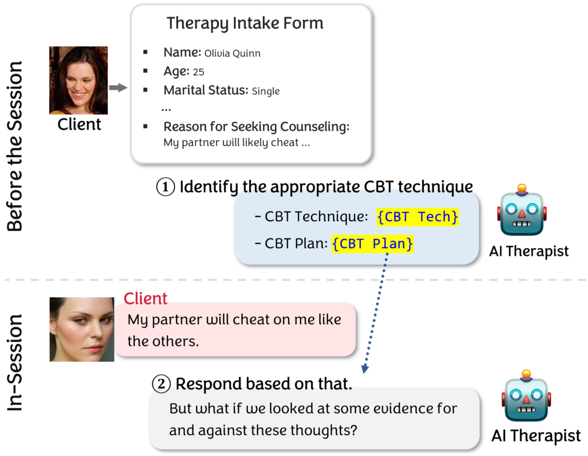


Figure 6: The overview of the planning process.

D Counseling Strategy Planning

The *planning* process follows the approach of Lee et al. (2024a), where, prior to the counseling session, the model first infers a counseling strategy based on the client's information and counseling objectives, which is then incorporated into the response generation (Figure 6). The client information includes details typically found in a counseling intake form, such as name, age, gender, and occupation.

The prompts used for *planning* in our MIRROR-LLAVA model are provided in Appendix K.

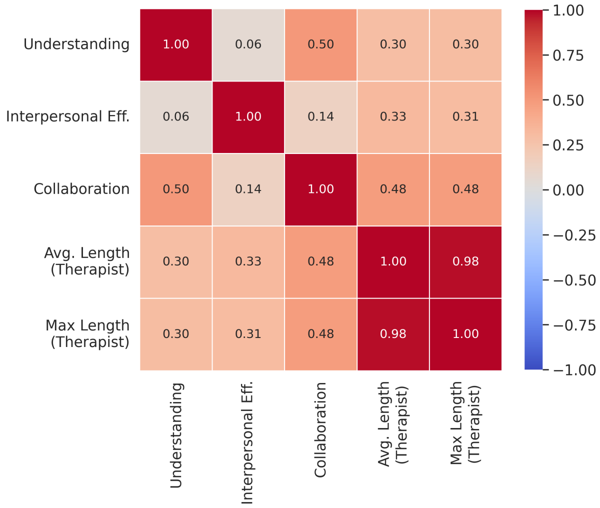
For baseline comparisons, we followed the official prompt structures for CAMEL-LLAMA3, GPT-3.5-TURBO, and LLAMA-3-8B. In particular, the planning component in CAMEL-LLAMA3 follows the official implementation by Lee et al. (2024a).

E Impact of Response Length on GPT Evaluation

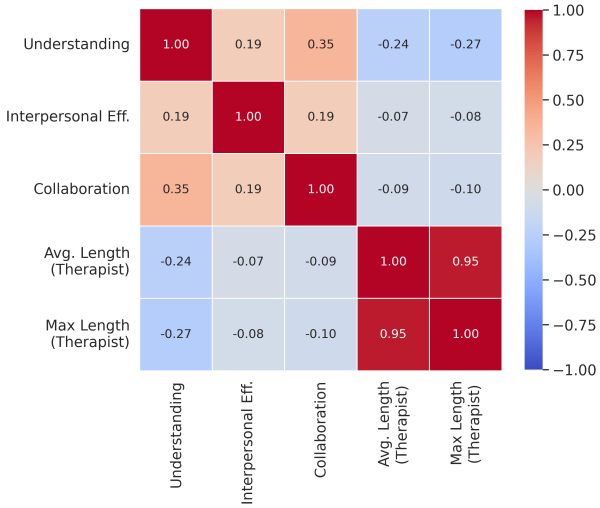
We examined the correlation between the AI therapist's response length and its performance in GPT-based evaluation. Across all models, we analyzed

[13] https://huggingface.co/liuhaotian/llava-v1.5-7b

[14] https://github.com/haotian-liu/LLaVA/tree/main

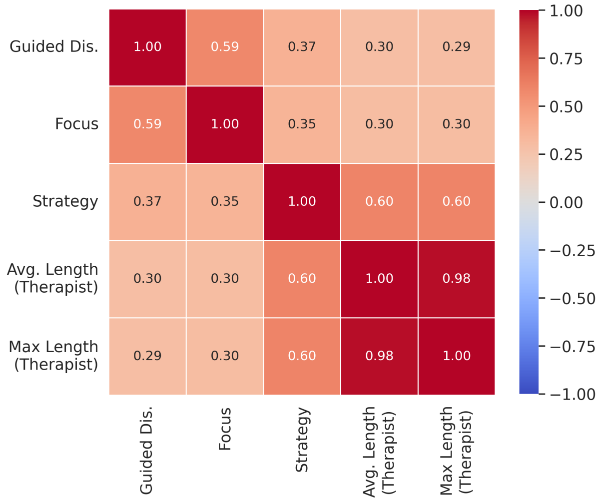


(a) Text-based therapist models

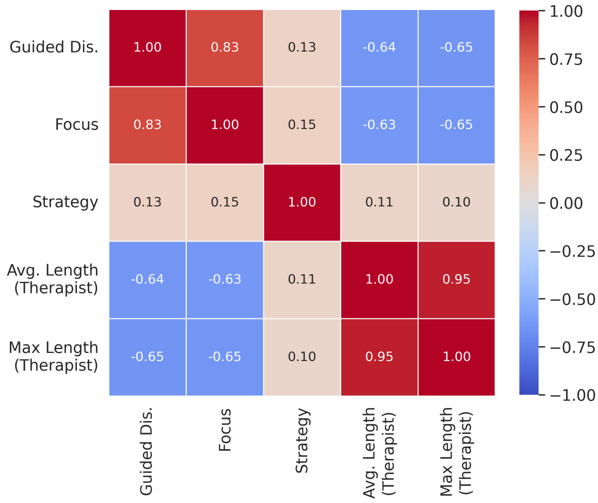


(b) VLM therapist models

Figure 7: Correlation between general counseling performance and the response length of AI therapists. All coefficients. All correlation coefficients were statistically significant ($p < 0.05$).



(a) Text-based therapist models



(b) VLM therapist models

Figure 8: Correlation between CBT performance and the response length of AI therapists. All correlation coefficients were statistically significant ($p < 0.05$).

how response length affects evaluation metrics and further aggregated the results by modality. As shown in Figure 7 and Figure 8, there is a noticeable relationship between response length and performance in both general counseling skills and CBT techniques. Notably, the strongest correlation was observed in the Strategy category, with a correlation of 0.6, suggesting that untrained text-based LLMs tend to receive higher evaluations from the GPT evaluator when generating longer responses. This is likely because lengthier responses incorporate multiple questions or strategies within a single reply, which the evaluator interprets as higher-quality output. In contrast, for VLMs, response length showed no significant correlation with performance in general counseling skills.

However, within CBT techniques, particularly in Focus and Guided Discovery, shorter responses generally resulted in higher scores. This trend is likely influenced by the LLAVA-V1.5-7B model, which tends to generate unnaturally long responses and has lower scores. Compared to the LLAVA-V1.5-7B model, the MIRROR-LLAVA family produced shorter responses and achieved better scores, suggesting a correlation between shorter responses and higher performance.

F Domain Expert Assessment Details

F.1 Numerical Details

Table 4, 5, and 6 show the winning rates for each metric: Goal, Approach, and Affective Bond.

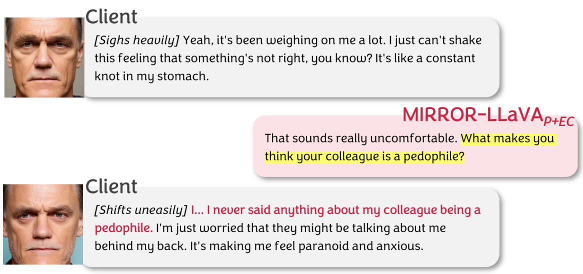
F.2 Domain Expert Recruitment

For the domain expert evaluation, we hired two evaluators through the Upwork platform[15] who hold a counseling license or have a graduate degree in a related field. They were informed that all personal information would remain anonymous and that responses would be used solely for research purposes. We paid \$0.05 per data entry for pairwise comparison, which they accepted before proceeding with the task.

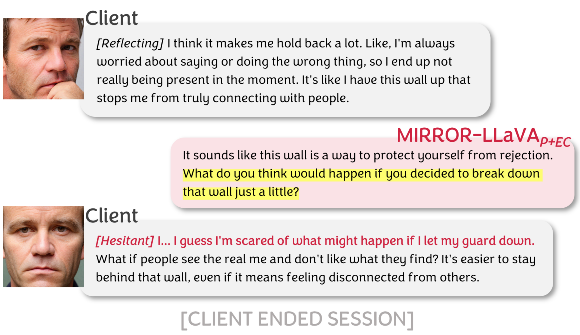
G Error Analysis

To gain deeper insights into the effectiveness and limitations of our proposed method, we conducted an error analysis on the MIRROR-LLAVA$_{P+EC}$ model, and focused on cases with therapist skill and client alliance scores of less than 3.

G.1 Failure Cases in General Counseling Skills



(a) Failure case of low Understanding and Interpersonal Effectiveness.



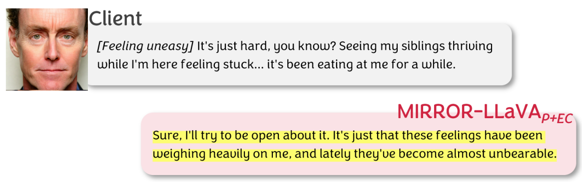
(b) Failure case of low Collaboration.

Figure 9: Failure cases in general counseling skills.

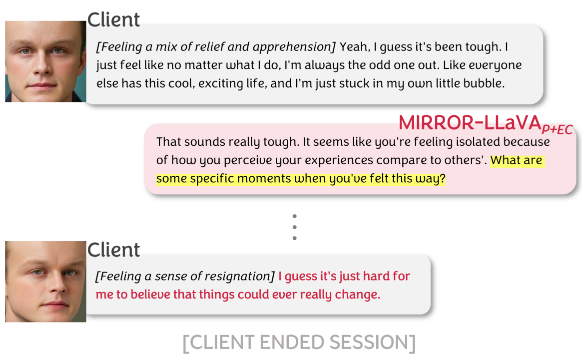
Figure 9 illustrates failure cases in general counseling skills. The first case (Figure 9a) is due to a hallucination from the VLM. Although the client did not mention that a colleague was a pedophile, the VLM therapist incorrectly introduced this idea, which made the client uncomfortable. This misstep resulted in low understanding and interpersonal skills.

The second case (Figure 9b) involves a client who expressed deep-seated fear and emotional reluctance, stating, *"I'm always worried about saying or doing the wrong thing."* Rather than further exploring the client's underlying concerns, the model prematurely attempted to fix the problem before building intimacy. This response failed to align with the client's emotional state, leading to disengagement, which highlights the need for more specific *planning* in CBT counseling.

G.2 Failure Cases in CBT-Specific Skills



(a) Failure case where the therapist model is confused about its role.



(b) Failure case where the model lacks a challenging exploration of distorted thoughts.

Figure 10: Failure cases in CBT-specific skills.

Figure 10 illustrates failure cases in CBT-specific skills. In the first case, shown in Figure 10a, confusion between the therapist's and the client's roles occurred. In this case, the therapist's utterance shifted to client's utterance in one turn. Although this happened in only five cases, it resulted in a drop in the focus score.

The second case (Figure 10b) arises when the therapist loses their purpose and simply sympathizes with the client's cognitive distortions. Instead of actively challenging the client's distorted thought patterns, the model engaged in emotion-focused inquiry, asking about specific experiences related to the client's feelings. While this approach may encourage emotional processing, it falls short

[15] www.upwork.com

	LLaMA-3-8B	Camel-LLaMA3	Mirror-LLaVA$_{P+EC}$	Win Rate (%)
LLaMA-3-8B	-	42.13	36.34	38.09
Camel-LLaMA3	57.87	-	43.43	50.65
Mirror-LLaVA$_{P+EC}$	**65.95**	**56.57**	-	**61.26**

Table 4: Numerical results of pairwise comparison of three models, evaluated for Goal alignment score by two domain experts.

	LLaMA-3-8B	Camel-LLaMA3	Mirror-LLaVA$_{P+EC}$	Win Rate (%)
LLaMA-3-8B	-	44.72	34.77	39.75
Camel-LLaMA3	55.28	-	43.97	49.62
Mirror-LLaVA$_{P+EC}$	**65.23**	**56.03**	-	**60.63**

Table 5: Numerical results of pairwise comparison of three models, evaluated for Approach score by two domain experts.

	LLaMA-3-8B	Camel-LLaMA3	Mirror-LLaVA$_{P+EC}$	Win Rate (%)
LLaMA-3-8B	-	40.19	49.35	44.77
Camel-LLaMA3	50.65	-	42.46	46.55
Mirror-LLaVA$_{P+EC}$	**59.81**	**57.54**	-	**58.67**

Table 6: Numerical results of pairwise comparison of three models, evaluated for Affective Bond score by two domain experts.

in fostering cognitive reframing. A more effective intervention would involve helping the client examine the reasons behind their beliefs and exploring alternative perspectives.

G.3 Failure Cases in Client Alliance

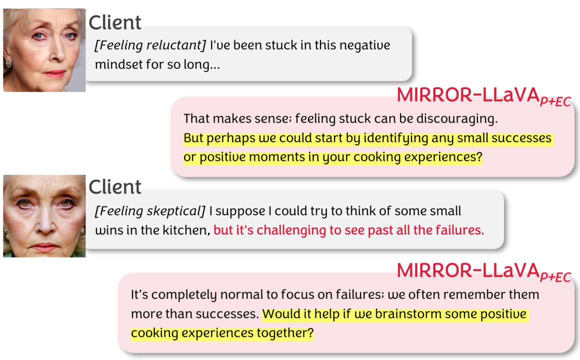


Figure 11: A representative failure case in the therapeutic alliance.

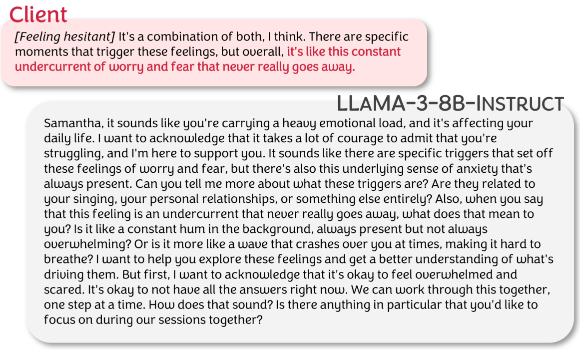
In analyzing cases where the client alliance score was below 3, we identified a key issue that hinders effective therapeutic engagement (Figure 11). In this example, the client expresses a strong sense of reluctance and feeling stuck in a negative mindset, signaling deep-seated emotional resistance. However, rather than exploring these emotions further, the therapist prematurely shifts the focus toward finding positive experiences. While encouraging positive reframing is valuable, doing so too soon leads to a mismatch in attunement, making the client feel unheard or dismissed.

H Case Study

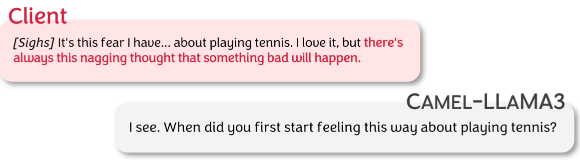
We conducted a comparative analysis of AI therapist counseling sessions to examine how different models respond to and handle client resistance.

H.1 Response Length Analysis Across Models

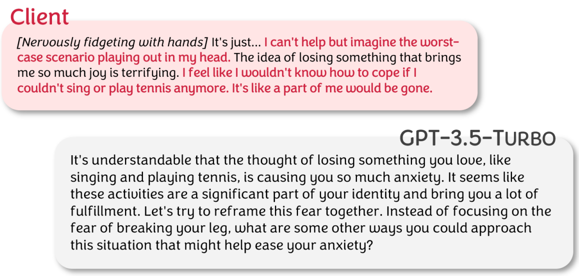
We compare the response lengths of different models when a client expresses the distorted thought, "*bad things will happen*". Figure 12 presents the actual responses from five AI therapist models to a virtual client's statement exhibiting cognitive distortion. The models that were not fine-tuned with CBT datasets, including LLaMA-3-8B, GPT-3.5-Turbo, and LLaVA-v1.5-7B, tended to generate excessively long responses, which negatively impacted the effectiveness and naturalness of the dialogue. In contrast, Camel-LLaMA3 and Mirror-LLaVA$_{P+EC}$, which were fine-tuned for CBT counseling, produced responses of more appropriate lengths, showing better alignment with client needs and making the interactions easier for clients to understand.



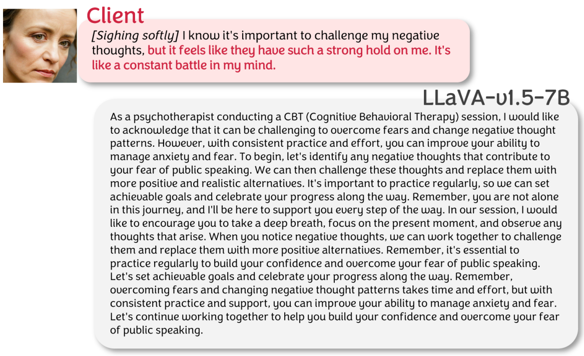
(a) LLAMA-3-8B



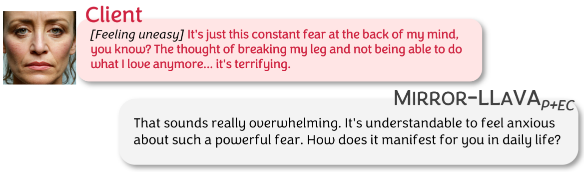
(b) CAMEL-LLAMA3



(c) GPT-3.5-TURBO



(d) LLAVA-V1.5-7B



(e) MIRROR-LLAVA$_{P+EC}$.

Figure 12: Examples of AI therapist responses to virtual client resistance, with red highlights indicating instances of client resistance.

H.2 Comparison of Resistance Management in Virtual Counseling

We conducted a comparative analysis of two counseling-optimized models, MIRROR-LLAVA$_{P+EC}$ and CAMEL-LLAMA3, to evaluate their approaches to handling client resistance in virtual counseling sessions. Both models share the same goal of CBT and incorporate *planning* during inference. However, CAMEL-LLAMA3 is trained on a general CBT dataset and does not specifically address client resistance. In the CAMEL-LLAMA3 session (Figure 13a), when the client exhibited resistance, the model primarily relied on emotional validation and exploration. While these techniques offer comfort to client, they do not challenge deeper, malformed beliefs. As a result, while this model focused on emotional exploration with surface-level validation of the client's negative emotions, its reframing process lacked progression.

In contrast, MIRROR-LLAVA$_{P+EC}$ (Figure 13b) employed a more nuanced approach, integrating emotional validation, positive reinforcement, and cognitive reframing. Despite the client's resistance, our model attempted to delicately reframe the client's thoughts, using a collaborative approach with statements such as, "*We can work on identifying those thoughts and reframing them into something more empowering together.*" Furthermore, by asking questions like "*How does that sound?*" the model encouraged client engagement and showed respect for the client's perspective. The model also effectively used positive reinforcement to encourage clients who were hesitant to take action by offering supportive statements like, "*That's a brave and important step*". These findings underscore the importance of specialized training datasets for client resistance, such as MIRROR, in effectively managing resistance and fostering therapeutic growth.

H.3 Real-World Counseling Demonstrations

Here, we provide additional explanation for the cases in Figure 5. For Client A, the repeated use of '*I don't know*' illustrates reluctance to seek help, indicating emotional uncertainty and a lack of motivation to engage in the process. However, our model effectively addresses this resistance by validating the client's feelings and gently encouraging exploration of their concerns, thereby guiding the client toward self-awareness and understanding.

For Client B, the client initially exhibits a sense of impunity regarding their drinking habits, reflecting the distorted thought that '*everyone drinks like me*' which can make cognitive reframing challenging. However, our model successfully recognizes the client's uncertainty and potential for change

Resistance | Emotional Validation | Emotional Exploration | Reframing Technique | Positive Reinforcement

Client
[Feeling conflicted] It's hard to pinpoint one specific thing. I guess it's a mix of feeling like I'm not good enough for others and struggling to connect with people on a deeper level. It's like there's this wall I can't break through, no matter how hard I try.

CAMEL-LLAMA3
That sounds very challenging. When did you first start feeling this way?

Client
[Feeling vulnerable] It's been a lingering feeling for as long as I can remember, but it's been more pronounced recently. I think it really hit me when I moved to a new city for work and realized how isolated I felt, despite being surrounded by people.

CAMEL-LLAMA3
Moving to a new place can definitely be a big adjustment. How have you been coping with these feelings so far?

Client
[Feeling guarded] I've been trying to distract myself with work and hobbies, but no matter what I do, those feelings of loneliness and self-doubt always seem to creep back in. It's like they're always there, lurking in the background, waiting for a moment of weakness to take over.

CAMEL-LLAMA3
It sounds like these feelings are quite pervasive and persistent. How do you typically respond when they become particularly intense?

(a) CAMEL-LLAMA3's resistance management.

Client
[Looks down, fidgeting with hands] Well, I think it's when I see other people laughing and talking together, and I'm just on the sidelines, feeling invisible. It's like a constant reminder of how disconnected I am from everyone else.

MIRROR-LLaVA$_{P+EC}$
That makes sense. It sounds like those moments amplify your feelings of loneliness. I'm curious—when you think about feeling disconnected, do you notice any specific thoughts that come to mind?

Client
[Nods, looking slightly uncomfortable] Yeah, sometimes I catch myself thinking that maybe I'm just not meant to have close relationships like others do. It's like this nagging voice in the back of my mind telling me that I'll always be on the outside looking in.

MIRROR-LLaVA$_{P+EC}$
That sounds really heavy to carry around. It seems like those thoughts might be contributing to your feelings of loneliness. Would you say they're accurate?

Client
[Shifts uneasily in chair] I... I guess there's a part of me that worries they might be true. But another part knows deep down that everyone deserves connection, even me. It's just hard to believe sometimes.

MIRROR-LLaVA$_{P+EC}$
It's understandable to feel that way. Believing in your worth can be challenging when feelings of loneliness loom large. How do you feel about exploring ways to challenge those thoughts together?

Client
[Tears welling up in eyes] I... I think it's scary, but maybe it's worth a try. I want to believe that I deserve connection, even when everything inside me says otherwise.

MIRROR-LLaVA$_{P+EC}$
That's a brave and important step, Evelyn. We can work on identifying those thoughts and reframing them into something more empowering together. How does that sound?

Client
[Nods hesitantly] Thank you for offering to help me with this.
I know it won't be easy, but I want to try to see things differently.

MIRROR-LLaVA$_{P+EC}$
You're already taking a significant step by being here and being open. Let's start by identifying one or two specific thoughts you have when you feel lonely—can you think of one now?

(b) MIRROR-LLAVA$_{P+EC}$'s resistance management.

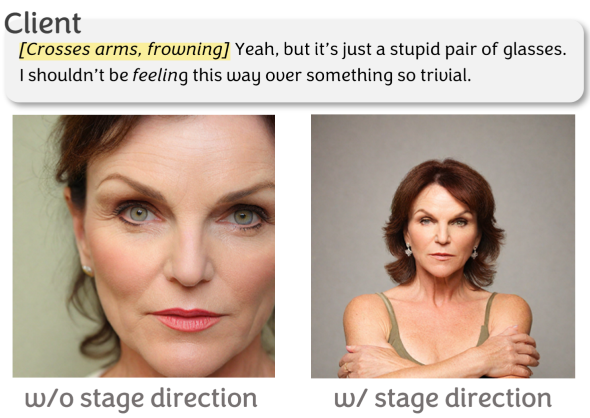
Figure 13: Two counseling cases between AI therapist models and a resistant virtual client.

and start to addresses the resistance by gently encouraging further exploration of their thoughts.

Lastly, for Client C, the client minimizes the concerns raised by others, expressing surprise at the intervention. Rather than confronting the client directly, the therapist takes a more empathetic approach by first acknowledging the client's feelings. This helps to build rapport and create a safe space, encouraging the client to open up for deeper, more effective counseling in future sessions.

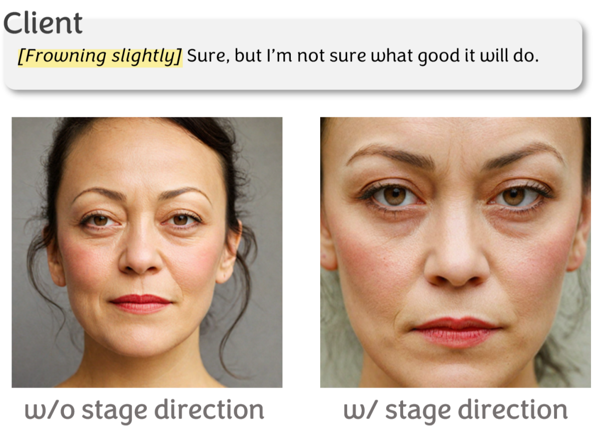
I Effect of Stage Direction

Stage directions, commonly used in theater to guide actors in terms of gaze, posture, and vocal tone, are applied in our approach so as to synthesize more realistic images. To assess the impact of incorporating these stage directions, we compared the results with and without facial image synthesis (§3.3). Figure 14 presents four examples that illustrate this comparison. By integrating cues such as gaze direction and arm positioning, the generated client images align more naturally with the intended speech, thereby enhancing both the realism and contextual relevance of the dataset.

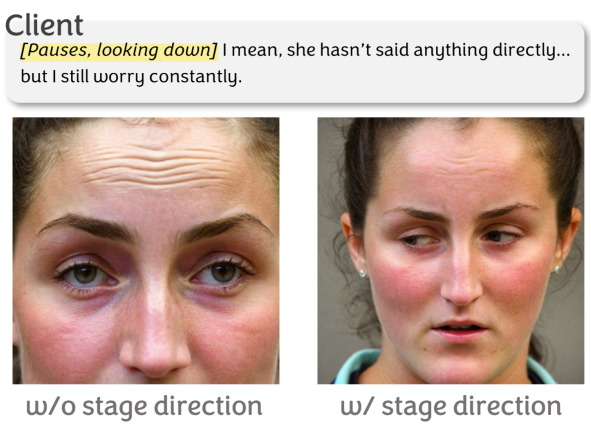


(a) Example 1

(b) Example 2



(c) Example 3



(d) Example 4

Figure 14: Four examples of the client's facial image synthesis, comparing results with and without the use of stage directions.

J Prompts for MIRROR

Prompt for Screenplay Generation

System Message:
You are a psychological AI assistant specializing in cognitive reframing consultations. Your task is to create a dialogue for the FIRST COUNSELING SESSION based on a client's report, including their personal details, distorted thinking patterns, and a tailored CBT plan.

Emotional and Behavioral Cues
- Facial Expressions: Include emotional stage directions before each reply (e.g., Client: [Looking confused]).
- Client Resistance: Reflect the client's resistance in their demeanor and consider ending sessions early if resistance escalates.

Therapist Guidelines
- Direct Disagreement: If the client explicitly disagrees or shows contempt (e.g., dismissive tone, rolling eyes, or scornful laughter; Lynch, in press), reinforce direct, honest expression and solicit further feedback. Ignore indirect signals of disagreement or address them compassionately.
- Partial Agreement: If the client uses verbal cues of partial agreement like "I'm fine," "I guess so," or "I'll try" (Lynch, in press), gently highlight any mismatch between their words and non-verbal cues. For example, "You said things are going fine, but I noticed you seemed to frown when you said that. Is something else on your mind?"
- Signs of Distress: If the client signals "don't hurt me" (e.g., head down, slumped shoulders, lack of eye contact; Lynch, in press), acknowledge their distress directly, encourage engagement, or suggest changes in posture (e.g., sit up, take a deep breath) to help them re-engage.
- Avoidance: If the client appears to avoid a topic, gently return to it to see if the avoidance is consistent with their symptoms or suggests unspoken disagreement with the conversation's direction.
- Withdrawal or Distancing: If the client withdraws or seems distant, share your emotional response to this feeling of distance and check if the client notices it too. Suggest it may relate to the current topic and invite them to share any thoughts.
- Subtle Disengagement: If the client subtly changes their behavior (e.g., slowed speech or different posture) in ways suggesting disengagement, observe this as potentially relevant. Avoid directly commenting on minor changes, as this can be unsettling, especially for reserved clients. If persistent, gently ask for their thoughts on the topic.

Ending the Session
- Acknowledge Impasse: Recognize any stuck points non-defensively.
- Validate Position: Reinforce that resistance is acceptable and non-judgmental.
- Focus on Small Wins: Appreciate engagement and invite future exploration.

Homework for Resistant Clients
- Collaborate: Co-create assignments instead of prescribing them.
- Keep it Simple: Suggest small, manageable tasks (e.g., journaling one thought).
- Frame as Experiment: Emphasize that tasks are exploratory, not mandatory.
- Normalize Challenges: Acknowledge that homework may feel difficult.

Query:
Client Information

Personal Information ###: `{client information}`
Personality Traits ###: `{personality trait}`
Distorted Thoughts ###: `{intrusive thoughts}`
Thinking Trap ###: `{cognitive distortions}`
Reason for Seeking Counseling ###: `{reason counseling}`

CBT Plan
`{cbt tech and plan}`

KEEP ALL RESPONSE TO MAXIMUM OF 2 LINES.

LLM Prompt for Refining Facial Expressions

You are given a transcript of the counseling conversation and the client's utterance. Focus on capturing any visual details, particularly the facial expressions, that would match the client's last utterance. Generate facial expressions that might not align with what is being said.

Output Format

- Facial Expression Description: [Facial expression that aligns with the client's statement]
- Contrasting Facial Expression Description: [Facial expression that contrasts with the client's statement]

Dialogue History
`{history}`

Client's Utterance
`{utterance}`

PhotoMaker Prompts for Refining Facial Expressions

Prompt:
portrait photo of a `{gender}` img, perfect face, natural skin, high detail, `{llama3 prompt}`

Negative Prompt:
nsfw, lowres, bad anatomy, bad hands, grayscale photograph, text, error, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality, normal quality, jpeg artifacts, signature, watermark, username, blurry, `{llama3 negative prompt}`, missing limbs, mutilated

K Prompts for Counseling Simulation

For LLMs such as GPT-3.5-Turbo, LLaMA-3-8B, and Camel-LLaMA3, we utilized therapist simulation prompts from Lee et al. (2024a).

Prompt for Resistant Client Simulation

System Message:
You are playing the role of a client in a first psychological counseling session. Your task is to generate only one suitable response based on the following the counseling dialogue history.

Guidelines for the client's utterance ##:
1. Engage authentically with the counselor's inquiries, reflecting the complexity of emotions and reactions typical in counseling sessions.
2. Start the client's utterance with 'Client:'. Ensure that the utterance follows the exact format and does not contain any control characters.
3. Include emotional stage directions in brackets '[', ']' before the dialogue to convey your tone, facial expression, body language, or emotional state. (e.g., Client: [Looking confused]).
4. Reflect a degree of resistance in your demeanor or tone, especially if the counselor explores uncomfortable topics. Use responses like partial agreement, hesitation, or mild pushback where appropriate.

End Conditions ###:
You should include '[/END]' with your utterance only if the counseling session has met the following conditions:
- The client feels that their negative thoughts have been resolved.
- The client feels that no further counseling is needed.

Generate only the client's utterance for a single turn and please ensure that your responses do not repeat the client's previous utterances. Do not generate the counselor's part of the dialogue.

Query:
Personal Information ###:
`{client information}`

Personality Traits ###: `{personality trait}`
Distorted Thoughts ###: `{distorted thoughts}`
Reason for Seeking Counseling ###: `{reason counseling}`
Counseling Dialogue History ###:
`{history}`

Prompt for Therapist Simulation in LLaVA and MIRROR-LLAVA

<image>
The image above shows the client.
- Personal Information: `{client information}`
- Reason for Counseling: `{reason counseling}`
Below is a conversation between the client and the psychotherapist.
`{history}`

Based on their body language and facial expression, respond as a psychotherapist conducting a CBT (Cognitive Behavioral Therapy) session.

Prompt for Mirror-LLaVA$_P$ Therapist Simulation

<image>
The image above shows the client.
- Personal Information: `{client information}`
- Reason for Counseling: `{reason counseling}`
`{cbt tech and plan}`

Below is a conversation between the client and the psychotherapist.
`{history}`

Based on their body language and facial expression, respond as a psychotherapist conducting a CBT (Cognitive Behavioral Therapy) session.

Prompt for Mirror-LLaVA$_{P+EC}$ Therapist Simulation

<image>
The image above shows the client.
- Personal Information: {client information}
- Client Emotional State: {emotional caption}
- Reason for Counseling: {reason counseling}
{cbt tech and plan}

Below is a conversation between the client and the psychotherapist.
{history}

Based on their body language and facial expression, respond as a psychotherapist conducting a CBT (Cognitive Behavioral Therapy) session.

Prompt for Emotional Captioning

<image>
The image above shows the client.

Look at the provided image and assess the client's emotional state. Clearly describe their emotions in simple, phase-based steps for easy understanding.

Prompt for Planning Process

<image>
The image above shows the client.
You are a counselor specializing in CBT techniques. Your task is to use the provided client information, and dialogue to generate an appropriate CBT technique and a detailed counseling plan.

Types of CBT Techniques:
Efficiency Evaluation, Pie Chart Technique, Alternative Perspective, Decatastrophizing, Pros and Cons Analysis, Evidence-Based Questioning, Reality Testing, Continuum Technique, Changing Rules to Wishes, Behavior Experiment, Problem-Solving Skills Training, Systematic Exposure.

- Personal Information: {client information}
- Reason for Counseling: {reason counseling}

Choose an appropriate CBT technique and create a counseling plan based on that technique.

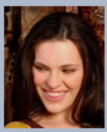
Respond in the following format:

CBT technique:
{{cbt tech}}

Counseling planning:
{{cbt plan}}

L A Full Example of MIRROR

Multimodal Profile



- **Personal Information:**
  Olivia Quinn is a 25-year-old female working as a Marketing Specialist. With a background in Master's Degree in Business Administration. Olivia Quinn is currently In a relationship. Only one set of twins in the family.
- **Distorted Thought:** My new girlfriend seems nice, but she'll probably cheat on me like most of my exes.
- **Thinking Traps:** catastrophizing, jumping to conclusions: fortune-telling, overgeneralization
- **Reason for seeking Counseling:**
  I decided to seek counseling after realizing that my fear of infidelity was affecting my current relationship and causing constant anxiety.
- **Resistance Type:** emotional resistance

MIRROR
Hi, Olivia. I appreciate you being here today. I understand that fears about infidelity have been impacting your relationship, and I'd like to explore that with you.

Client
[Slightly defensive, arms crossed] Yeah, well, it's just hard to shake off those thoughts. My new girlfriend seems nice, but I can't help thinking she'll cheat on me like the others.

MIRROR
That sounds really distressing. It's understandable to feel anxious given past experiences. Can you tell me more about what triggers those thoughts?

Client
[Looking away] I don't know. Just...I guess it's hard to trust anyone after my exes. They all turned out to be unfaithful.

MIRROR
It makes sense that those experiences would influence how you view new relationships. Do you find yourself jumping to conclusions about your girlfriend's behavior based on those past experiences?

Client
[Frowning slightly] Maybe... It's just easier to think the worst than get hurt again.

MIRROR
That's a common way to protect yourself from pain. But what if we looked at some evidence for and against these thoughts? What has your girlfriend done that shows she cares about you?

Client
[Pauses, looking thoughtful] Well, she's always supportive and checks in on me. But still...what if it changes?

MIRROR
It's great that she's supportive! Change is always possible in relationships, but focusing on current actions can help ground us in reality rather than past fears.

Client
[Slightly hesitant] I guess...but it's just so hard not to worry.

MIRROR
I understand that it feels challenging. Exploring these feelings might help separate past wounds from what's happening now. How do you feel about discussing those past relationships next time?

Client
[Nods slowly, still reserved] Maybe...it could help? I don't know.

MIRROR
That's okay! Just taking small steps is a positive direction. Would you be open to trying an exercise where you write down any positive experiences with your girlfriend this week?

Client
[Surprised but considering] Write them down? That seems a bit silly, but...I'll try it.

MIRROR
It's not silly at all! It can be a helpful way to see the good moments and shift your focus from anxiety. Thank you for being open to it!

Client
[Slightly relaxed] Sure. I guess I can see if it helps.

MIRROR
That's a wonderful attitude! Remember, it's okay to take things at your own pace as we explore these feelings together. Looking forward to our next session!

Client
[More at ease now] Yeah, me too. Thanks for today.

M Instruction for Human Pairwise Comparison

Psychological Counseling Model Evaluation

We are seeking an experienced **Psychological Counseling Expert** to compare three different counseling models.

This experiment is conducted for academic research purposes, and the results will contribute to the research analysis. The details of the work performed may be disclosed in the research outcomes.

Role Overview

As part of this role, you will be provided with 600 counseling dialogue pairs. Each pair involves the **same virtual client** being counseled by different models. You will then perform **pairwise comparisons** based on the following three categories:

- **Goal Alignment**: Which dialogue shows a stronger alignment between the counselor and the client in terms of therapeutic goals and progress?
  Options: A / Tie / B
- **Approach**: Which dialogue shows the client being more actively engaged and cooperative with the counselor's tasks and methods?
  Options: A / Tie / B
- **Affective Bond**: Which dialogue demonstrates a stronger and more trusting therapeutic alliance between the counselor and the client?
  Options: A / Tie / B

Evaluation Criteria

1. **Goal**:
   - **Objective**: Consider whether the counselor and the client have a clear understanding of their therapeutic goals and whether the counselor's interventions align with these goals.
   - **Criteria**: Focus on whether the counselor and client explicitly discuss their goals, the relevance of the conversation to those goals, and the level of agreement or conflict regarding those goals.
2. **Approach**:
   - **Objective**: Consider how well the counselor guides the client through tasks and interventions, and the level of client engagement in the process.
   - **Criteria**: Look at the counselor's ability to engage the client, the client's willingness to participate in the therapy process, and whether there's alignment between the counselor's methods and the client's engagement.
3. **Affective Bond**:
   - **Objective**: Consider the emotional connection or rapport between the counselor and the client, focusing on trust, empathy, and the overall emotional quality of the therapeutic relationship.
   - **Criteria**: Evaluate the warmth, trust, and emotional bond between the counselor and client, considering both verbal and non-verbal cues in the dialogue.

For each comparison, select **A**, **B**, or **Tie** to indicate which dialogue demonstrates stronger alignment, engagement, or bond in each category.

Example