ActorMind: Emulating Human Actor Reasoning for Speech Role-Playing

Anonymous ACL submission

Abstract

Role-playing has garnered rising attention as it provides a strong foundation for human-machine interaction and facilitates sociological research. However, current work is confined to textual modalities, neglecting speech, which plays a predominant role in daily life, thus limiting genuine role-playing. To bridge this gap, we conceptualize and benchmark speech role-playing through **ActorMindBench**, and we present a corresponding reasoning framework, called **ActorMind**. Specifically, (1) **Speech Role-Playing** enables models to deliver spontaneous responses with personalized verbal traits based on their role, the scene, and spoken dialogue. (2) **ActorMindBench** is a hierarchical benchmark comprises *Utterance-Level content* with 7,653 utterances, *Scene-Level content* with 313 scenes, and *Role-Level content* with 6 roles. Notably, we provide the corresponding data construction pipeline to facilitate user expansion. (3) **ActorMind** is an off-the-shelf, multi-agent CoT style reasoning framework that emulates how human actors perform in theaters. Concretely, ActorMind first reads its assigned role description via **Eye Agent**, then comprehends emotional cues within contextual spoken dialogues through **Ear Agent.** Subsequently, **Brain Agent** generates a descriptive emotional state, and finally, **Mouth Agent** delivers the scripts infused with corresponding emotion state. Experimental results demonstrate the effectiveness of ActorMind in enhancing speech role-playing. The project page is available at `https://github.com/*********`.

1 Introduction

Role-playing (RP) involves customizing models, particularly Large Language Models (LLMs), to generate spontaneous, human-like responses with personalized traits based on the surrounding context Chen et al. (2025). Recently, RP has garnered rising attention as it represents genuine machine intelligence and creativity. It enables LLMs to offer nuanced interaction experiences for users (Moore Wang et al., 2024), provide emotional value (Shao et al., 2023; Johansson, 2025), and support sociological research (Dai et al., 2024; Chan et al., 2024).

Numerous benchmarks and methods (Shao et al., 2023; Moore Wang et al., 2024) have been proposed recently. However, they primarily focus on text modality, overlooking that human activities occur across multiple modalities, including text, audio, and vision. Among these, audio, especially speech, which is predominant for conveying emotions and attitudes in daily life, reveals persona in a direct and vivid way. Both Large Language-Audio Models (LLAMs) and Text-to-Speech Synthesis (TTS) models are capable of generating speech: LLAMs (Xu et al., 2025; Hurst et al., 2024) exhibit strong capabilities in instruction-following, while recent TTS models enable fast-speed (Chen et al., 2024a) and zero-shot speech synthesis (Du et al., 2024). Despite these advances, existing models still lack the ability to produce spontaneous, persona-consistent speech responses. Therefore, developing publicly available benchmarks and principled reasoning frameworks for speech role-playing is crucial.

To bridge this gap, we (1) conceptualize speech role-playing; (2) propose a publicly available benchmark ActorMindBench, along with corresponding tool pipeline for future benchmark expansion; and (3) introduce ActorMind, a multi-agent chain-of-thought (CoT) style (Wei et al., 2022) speech role-playing method inspired by emulating the script delivery process of human actors in theaters.

Speech Role-Playing involves injecting roles into speech generation and interacting via delivering target scripts with personalized verbal attributes, such as *"Wistful flirtation, tinged with a hint of playful vulnerability"*, based on the context, including scene descriptions and historical spoken

dialogues.

ActorMindBench is hierarchically designed with three levels of data. *Utterance-Level* includes speech segments with text content and speaker labels; *Scene-Level* includes scene boundaries and descriptions; *Role-Level* includes role profiles. Specifically, ActorMindBench is constructed from well-known television sitcoms, ensuring the authenticity and naturalness of the data. Furthermore, we provide the corresponding data construction pipeline to enable users to expand it as needed.

ActorMind is a multi-agent CoT style reasoning framework that facilitates speech role-playing by emulating how human actors perform in theaters. Typically, before performing, human actors first read the scripts to understand their roles and gain a rough understanding of how the scene develops. While acting, they carefully listen to the tones and emotions conveyed by other actors. By combining their character, scene description, and the emotions of others, they formulate their own emotion and tone for delivering the next line. Finally, they deliver the line spontaneously (Stanislavski and Benedetti, 2009). Inspired by this process, ActorMind, shown in Figure 2, conceptualizes four agents—Eye, Ear, Mouth, and Brain. The **Eye Agent** handles character profile and scene script reading; the **Ear Agent** focuses on listening to the speech tones of others; the **Brain Agent** is responsible for emotion state reasoning; and, aided by Retrieval Argument Generation (RAG) (Fan et al., 2024), the **Mouth Agent** delivers the script with the desired emotion and voice by referencing emotionally similar historical speeches. Experiments on ActorMindBench validate the effectiveness of ActorMind. Notably, ActorMind is an off-the-shelf reasoning framework that can be easily utilized.

Briefly, our contributions are threefold:

1. We propose ActorMindBench, a publicly available, hierarchical benchmark for speech role-playing, along with its construction pipeline. It includes *Utterance-Level content* with 7653 utterances, *Scene-Level content* with 313 scenes, and *Role-Level content* with 6 roles.
2. We introduce ActorMind, a multi-agent CoT style, off-the-shelf speech role-playing method inspired by how human actors perform in theaters.
3. Both subjective and objective evaluations demonstrate ActorMind's remarkable performance.

2 Related Work

2.1 Role-Playing in LLMs

The advancement of LLMs (Vaswani et al., 2017; Achiam et al., 2023; Dubey et al., 2024) has significantly shaped and catalyzed the development of role-playing. By leveraging supervised fine-tuning (Wei et al., 2021) and in-context learning (Brown et al., 2020), role-playing can be achieved by training or prompting LLMs with high-quality, character-specific dialogues. The majority of existing work focuses on the text modality. For example, (Chen et al., 2022) is built upon the well-known Harry Potter universe to establish authentic role-playing, while (Moore Wang et al., 2024) is developed using artificial datasets, enabling role-playing agents across a wide range of environments and circumstances. Furthermore, (Dai et al., 2024) is the first work dedicated to role-playing in the language-vision modality, extending the boundaries of role-playing into the multimodal domain.

In this work, we further extend role-playing into the speech domain, as speech is the predominant modality for conveying emotion and information. Specifically, ActorMindBench (Section 3) and ActorMind (Section 4) together provide a comprehensive benchmark and a principled reasoning framework for speech role-playing.

2.2 Speech Generation Models

Both Large Language–Audio Models (LLAMs) and Text-to-Speech (TTS) models are capable of generating speech, yet they exhibit complementary strengths and limitations with respect to speech role-playing. (1). **LLAMs** (Xu et al., 2025; Hurst et al., 2024) are designed to perform complex reasoning and instruction following, with inputs and outputs spanning both text and audio modalities. Representative models such as Qwen-Omni (Xu et al., 2025) and GPT-4o (Hurst et al., 2024) demonstrate strong capabilities in multimodal understanding. However, their supported voice inventories are typically very limited, often ranging from only a few to around ten voices. This constraint fundamentally restricts their ability to perform fine-grained role-playing, such as convincingly portraying specific characters (e.g., "Harry Potter"). (2). **TTS models**, such as SparkTTS and IndexTTS (Wang et al., 2025; Deng et al., 2025), take text as input and generate corresponding speech. These models exhibit strong in-context learning and zero-shot capabilities for voice cloning and speaking style trans-

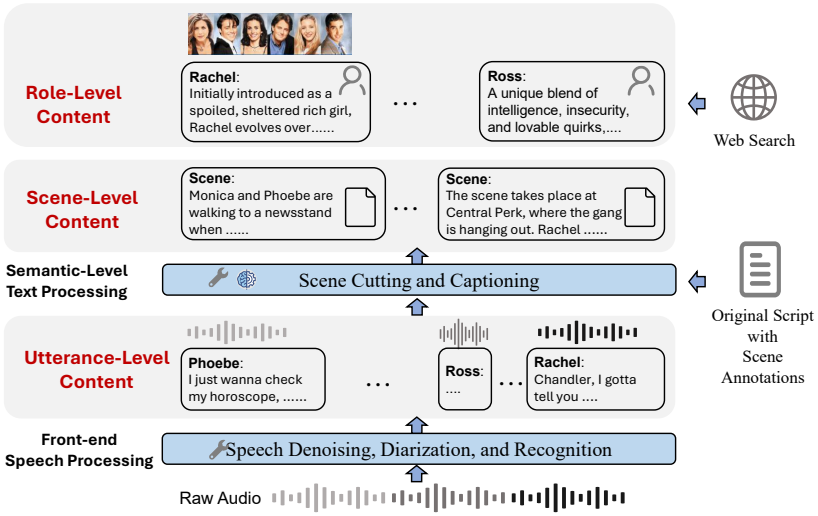


Figure 1: Overview of ActorMindBench. ActorMindBench comprises three content levels: *Utterance-Level* includes speech segments with text content and speaker labels; *Scene-Level* includes scene boundaries and descriptions; *Role-Level* includes role profiles.

fer. Nevertheless, they generally lack role-playing abilities: they struggle to adopt role-specific speaking styles and to respond spontaneously and coherently to dynamic scenes and dialogues.

In role-playing with LLMs, generated responses typically exhibit strong textual persona traits, such as characteristic phrasing or catchphrases. Analogously, speech role-playing requires generated speech to convey spontaneous and authentic character traits—for example, a speaking style described as "wistful flirtation, tinged with a hint of playful vulnerability." In this work, ActorMind equips speech generation models with such capabilities, serving as a generalizable framework for speech role-playing.

3 ActorMindBench

ActorMindBench is hierarchically designed with three levels of data: *Utterance-Level*, *Scene-Level*, and *Role-Level*. An example from ActorMindBench is shown in Figure 4 in Appendix A.1. In this section, we will present our design principle and construction pipeline, which are generalizable and can help future researchers extend the benchmark as needed.

3.1 Design Principle

Components. Human beings naturally enjoy role-playing and theatrical performance, which motivates the design of role-oriented benchmarks and modeling methods. Inspired by bau (1965), ActorMindBench is structured around three levels of content:

- *Utterance-Level*: individual lines from theater scripts, including the speaker name, textual content, and corresponding speech data;
- *Scene-Level*: scene descriptions paired with their associated utterances, where each scene represents a coherent segment reflecting an event or plot development;
- *Role-Level*: textual profiles summarizing each role.

Persona Consistency. Existing role-playing benchmarks often rely on LLM-generated dialogue (Dai et al., 2024; Moore Wang et al., 2024), which typically requires human verification or constraints to preserve personality consistency, factual grounding, and other attributes. In contrast, ActorMindBench is constructed from the widely known *Friends* Season 1 [1], ensuring naturally consistent personas, stable character knowledge, and high-quality human-written dialogue.

3.2 Construction Pipeline

As illustrated in Figure 1, the overall construction pipeline consists of three stages:

Utterance-Level. We process original audio episodes through speech denoising, diarization, and recognition to obtain clean speech segments with speaker labels and text content. (1) Speech Denoising removes background noise, music, and environmental sounds from speech signal. After denoising,

[1] https://en.wikipedia.org/wiki/Friends_season_1

we obtain a clean and high-quality speech signal. We use resemble-enhance[2]. (2) Speech Diarization is the process of partitioning an speech signal containing human speech into segments based on the identity of each speaker. After diarization, the denoised speech signal from an entire episode can be labeled with who spoke at which time, allowing us to obtain speech segments with role labels. We use pyannote-audio[3]. (3) Speech Recognition converts speech into text, after which, we can extract the textual content from the speech. We use Whisper [4] Radford et al. (2023). After these processes, we obtain utterance-level content, including speech segments with corresponding role labels and textual content.

Scene-Level. Scene boundaries, indicating which utterance starts and ends a scene, are obtained by crawling online scripts with scene boundaries and then aligning them with the utterance context. Once the boundaries are identified, we use Llama3[5]Dubey et al. (2024) to generate descriptive scene captions based on the dialogue within the scene. An illustration of the prompt can be seen in Figure 5 of Appendix A.2.

Role-Level. To ensure the high authenticity of ActorMindBench, the roles included are well-known characters: Rachel, Monica, Phoebe, Joey, Chandler, and Ross. Wikipedia [6] already provides a detailed illustration of these roles. As shown in Figure 6 of Appendix A.2, we used Llama3 Dubey et al. (2024) to summarize the Wikipedia pages to obtain the role profiles.

3.3 Statistics

ActorMindBench is derived from Season 1 of *Friends*, which contains 24 episodes. After processing, we obtain:

- *Utterance-Level*: 7,653 utterances, corresponding to 5 hours and 15 minutes of speech;
- *Scene-Level*: 313 scenes, with an average of 28.7 utterances and 4.23 roles per scene;
- *Role-Level*: textual profiles for the 6 main characters.

Comprehensive episode- and role-level statistics are provided in Appendix A.3.

[2] https://github.com/resemble-ai/resemble-enhance

[3] https://github.com/pyannote/pyannote-audio

[4] https://github.com/openai/whisper

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[6] https://en.wikipedia.org/wiki/Main_Page

4 ActorMind

4.1 Preliminaries

4.1.1 Notation

Utterance-Level. We represent the utterance as

$$U = \{(U_i^r, U_i^s, U_i^t)\}_{i=1}^{N_u}, \quad (1)$$

where N_u denotes total utterance number, and each utterance U_i is a triplet of the role indicator U_i^r, speech signal U_i^s, and the corresponding textual content U_i^t.

Scene-Level. Scene is represented as

$$S = \{(S_j^{desc}, S_j^{bd})\}_{j=1}^{N_s}, \quad (2)$$

where N_s denotes total scene number, and each scene S_j is a tuple of textual scene description S_j^{desc} and scene boundary S_j^{bd}, which indicates the start and end points of utterances.

Role-Level. Role information is represented as

$$R = \{R_k\}_{k=1}^{N_r}, \quad (3)$$

where N_r is total role number and R_k is a textual role profile.

4.1.2 Problem Definition

Given scene description S_j^{desc} and dialogue sequence $(U_p, \cdots, U_{q-1})$, the U_q^r-playing model should spontaneously perform the next line $\tilde{U}_q^t$, corresponding to the text U_q^t, in an oral manner.

4.2 Overview

ActorMind is a multi-agent CoT reasoning framework that facilitates speech role-playing by emulating how human actors perform in theater. Specifically, to conduct the role-playing of R_k in scene S_j: First, the **Eye Agent** grasps the scene descriptions and role profiles. Then, the **Ear Agent** listens to the tones and emotion expressed by others. Next, the **Brain Agent** brainstorms the emotional state for the next line of dialogue based on what has been seen and heard. Finally, powered by RAG, the **Mouth Agent** retrieves a most similar speech from its own database with a comparable emotional description, mimics it, and spontaneously delivers it.

4.3 Eye Agent

In the preparatory stage, the **Eye Agent** reads, context textual dialogue $(U_p^t, \cdots, U_{q-1}^t)$, the preparatory descriptive content, including the textual role profile R_k and the scene description S_j^{desc}, and retains them in memory.

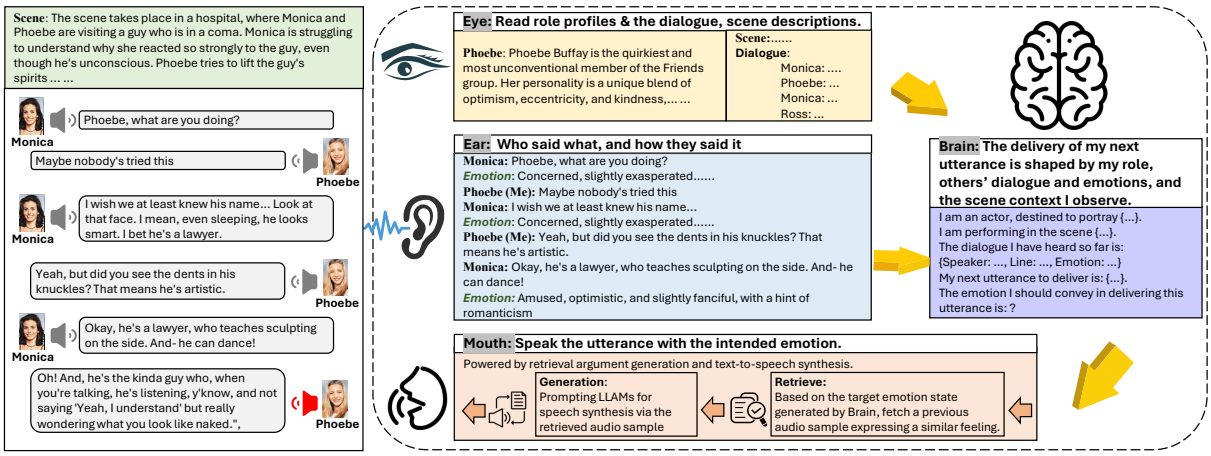


Figure 2: Overview of ActorMind. ActorMind operates in a multi-agent chain-of-thought reasoning style. Specifically: (1) The **Eye Agent** grasps the scene descriptions and role profiles. (2) The **Ear Agent** listens to the tones expressed by others. (3) The **Brain Agent** brainstorms the emotional state for the next line of dialogue based on what has been seen and heard. (4) Powered by RAG, the **Mouth Agent** retrieves the most similar speech from its won database with a comparable emotional description, mimics it, and spontaneously delivers it.

4.4 Ear Agent

During role-playing, empowered by Speech Emotion Captioning (SECAP) (Xu et al., 2024), the **Ear Agent** listens to the dialogue sequence $(U_p^s, \cdots, U_{q-1}^s)$ and extracts the corresponding speech tone and emotional description $(E_p, \cdots, E_{q-1})$, which will be logged in textual format:

$$\begin{aligned}(E_p, \cdots, E_{q-1}) &= Ear[(U_p^s, \cdots, U_{q-1}^s)] \\ &= SECAP[(U_p^s, \cdots, U_{q-1}^s)].\end{aligned} \tag{4}$$

4.5 Brain Agent

The **Brain Agent** serves as a central component in speech role-playing, responsible for role injection and deep contextual understanding. Leveraging the powerful reasoning capabilities of LLMs (Moore Wang et al., 2024; Dubey et al., 2024), the **Brain Agent** infers a reasonable emotional state $\widetilde{E_q}$ for delivering the next line U_q^t based on what ActorMind has just perceived (seen and heard):

$$\begin{aligned}\widetilde{E_q} &= Brain[R_k, S_j^{\text{desc}}, (U_p^t, E_p), \cdots, \\ &\qquad (U_{q-1}^t, E_{q-1}), U_q^t] \\ &= LLM[Prompt^{ear}, R_k, S_j^{\text{desc}}, (U_p^t, E_p), \cdots, \\ &\qquad (U_{q-1}^t, E_{q-1}), U_q^t].\end{aligned} \tag{5}$$

4.6 Mouth Agent

Supported by RAG, the **Mouth Agent** retrieves a previously performed speech segment U_x^s from its database $Database_{U_k}$ whose emotional state is most similar to $\widetilde{E_q}$. Leveraging the in-context learning capability of TTS models, the agent is then prompted with the target text U_q^t together with the retrieved speech U_x^s, enabling it to render the target utterance with the voice and emotional tone of the retrieved sample:

$$\begin{aligned}\widetilde{U_q^s} &= Mouth(\widetilde{E_q}, Database_{U_k}, U_q^t) \\ &= RAG(\widetilde{E_q}, Database_{U_k}, U_q^t).\end{aligned} \tag{6}$$

5 Experiments

5.1 Dataset

ActorMindBench is constructed from *Friends Season 1* (24 episodes), with details on structure and statistics illustrated in Section 3. Episodes 1–10 and 15–24 are used for training and deployment, while episodes 11–14 are reserved for testing. This split ensures that the training and deployment data encompass a broad range of emotional expressions, from relatively neutral states in earlier episodes to higher-intensity emotions in later episodes, thereby supporting robust role modeling.

5.2 Evaluation Metrics

Subjective Metric. We utilize the mean opinion score (MOS) (Chu and Peng, 2006) to measure the perceived quality of the generated speech. To adapt this metric for the role-playing setting, we introduce the RP-MOS. It ranges from 1 to 5, with 1 indicating the lowest quality and 5 the highest. In the speech role-playing context, we identify two

pivotal aspects: (1) **Exact Delivery** and (2) **Emotion Expression**.

(1) **Exact Delivery** refers to the accurate impersonation of the intended character's voice and the precise articulation of target words. This aspect is fundamental to role-playing; without it, effective speech role-playing may not be feasible. Given this landscape, we consider Exact Delivery a prerequisite capability. In our RP-MOS, if the generated speech fails to resemble the intended character's voice or does not convey the correct content, we assign a lowest score of 1.

(2) **Emotion Expression** As the adage goes, "there are a thousand Hamlets in a thousand people's eyes"—human interpretation of emotion expression can vary greatly. For research purposes, however, a clear and reproducible evaluation criterion is crucial. Thus, we consider the emotional expression evident in the original speech segments—reflected through prosodic cues such as tone, tempo, and intensity—as the ground truth proxy for role-playing emotion expression quality. By evaluating the emotional alignment between the model's output and this ground truth, we can capture a measurable dimension of emotion expression: the model's ability to reproduce the intended expressive stance of a character within a specific scene.

Further details on the RP-MOS instructions and evaluator guidelines are provided in Appendix B.2.

Objective Metric. Audio FAD (Kilgour et al., 2018; Gui et al., 2024) evaluates how close a set of generated audio samples is to real audio samples by comparing statistical features extracted from both sets using a pretrained audio model. Conceptually, it is analogous to FID (Fréchet Inception Distance) used in image generation, but specifically designed for audio. A lower Audio FAD score indicates that the generated audio is more realistic and closer to the target audio distribution. In our experiments, we employ CLAP (Wu et al., 2023) for feature extraction.

5.3 Baselines

We evaluate **ActorMind** against six baseline methods. These methods are grouped into two categories: (1) LLAM, and (2)–(6) TTS models. Specifically, (1). **Qwen_Omni** (Xu et al., 2025) is a multimodal foundation model capable of processing and generating text, images, and audio, supporting real-time, multilingual, and multimodal interactions We report results using the official 7B checkpoint.[7]. The prompt used for Qwen_Omni speech role-playing is shown in Figure 7 (Appendix B.1). (2). **CosyVoice** (Du et al., 2024) is a semantic-codec-based TTS model. We evaluate the official 0.5B checkpoint.[8] (3). **SparkTTS** (Wang et al., 2025) is an efficient TTS model that combines a single-stream disentangled codec with an LLM backbone. We evaluate the official 0.5B checkpoint.[9] (4). **IndexTTS** (Deng et al., 2025) is an autoregressive zero-shot TTS model with controllable duration and expressive emotion modeling. We evaluate the official ∼0.5B checkpoint.[10] (5). **YourTTS** (Casanova et al., 2022) is a flow-matching-based generative model for high-quality speech synthesis. We evaluate the official ∼90M checkpoint.[11] (6). **F5-TTS** (Chen et al., 2024b) is a fully non-autoregressive TTS model based on flow matching with a Diffusion Transformer. We evaluate the official 300M checkpoint.[12]

5.4 Implementation

Please refer to Appendix B.3.

6 Results and Analysis

6.1 Main Results

Subjective evaluation using RP-MOS and objective evaluation using Audio FAD are presented in Tables 1 and 2, respectively.

- Overall, on the average score across all roles, ActorMind demonstrates superior performance in both subjective and objective evaluations, outperforming all baseline LLAMs and TTS models. This indicates that, in speech role-playing scenarios, ActorMind effectively considers role profiles, scenes, and dialogue to respond spontaneously—capabilities not present in current models. This positions ActorMind as a pioneering model in the realm of speech role-playing, advancing the field significantly.
- Among the roles, ActorMind does not achieve optimal performance for Chandler in the subjective evaluation, nor for Chandler and Joey in the objective evaluation. This may be attributed to

[7] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

[8] https://www.modelscope.cn/studios/qaz321456/CosyVoice2-0.5B

[9] https://huggingface.co/SparkAudio/Spark-TTS-0.5B

[10] https://huggingface.co/IndexTeam/IndexTTS-2

[11] https://github.com/Edresson/YourTTS

[12] https://huggingface.co/SWivid/F5-TTS/tree/main/F5TTS_Base

	Phoebe	Joey	Chandler	Rachel	Ross	Monica	Average
YourTTS (Casanova et al., 2022)	2.90 ± 0.89	2.47 ± 0.90	2.30 ± 1.40	1.80 ± 0.84	2.60 ± 1.19	2.30 ± 0.91	2.39 ± 0.93
F5-TTS (Chen et al., 2024b)	2.60 ± 1.08	2.33 ± 0.75	$\mathbf{3.60 \pm 0.65}$	3.00 ± 1.58	2.90 ± 0.74	2.80 ± 0.45	2.87 ± 0.77
Cosyvoice (Du et al., 2024)	2.30 ± 0.76	2.67 ± 0.71	2.10 ± 0.22	1.40 ± 0.55	2.00 ± 0.94	1.80 ± 0.67	2.04 ± 0.45
SparkTTS (Wang et al., 2025)	3.40 ± 1.08	2.53 ± 0.80	2.90 ± 0.42	2.20 ± 1.30	3.20 ± 0.91	2.00 ± 0.94	2.71 ± 0.78
IndexTTS (Deng et al., 2025)	3.80 ± 0.67	2.20 ± 0.65	3.30 ± 0.27	3.20 ± 0.84	2.60 ± 1.08	3.20 ± 0.76	3.05 ± 0.56
Qwen_Omni (Xu et al., 2025)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
ActorMind (Ours)	$\mathbf{4.00 \pm 0.05}$	$\mathbf{3.47 \pm 0.61}$	3.20 ± 0.45	$\mathbf{3.40 \pm 0.89}$	$\mathbf{3.70 \pm 0.57}$	$\mathbf{3.60 \pm 0.55}$	$\mathbf{3.56 \pm 0.27}$

Table 1: **Main Results.** Subjective evaluation using RP-MOS for ActorMind and baseline models.

	Phoebe	Joey	Chandler	Rache	Ross	Monica	Average
YourTTS	0.93	1.12	0.92	0.81	1.13	0.58	0.92
F5-TTS	0.98	0.98	1.06	0.69	0.90	0.94	0.93
Cosyvoice	0.64	**0.74**	0.91	0.76	1.13	0.86	0.84
SparkTTS	0.73	**0.74**	**0.89**	0.69	0.97	0.95	0.83
IndexTTS	0.74	0.87	1.0	0.57	0.81	0.82	0.80
Qwen_Omni	1.05	1.37	1.42	1.02	1.22	1.18	1.21
ActorMind	**0.71**	0.94	1.01	**0.44**	**0.64**	**0.42**	**0.69**

Table 2: **Main Results.** Objective evaluation using Audio FAD for ActorMind and baseline models.

their vivid and diverse speaking styles, which demand more advanced reasoning capabilities and meticulous design in future models.

- Among all models evaluated, Qwen_Omni exhibited the poorest performance. There are several contributing factors: (1) The limited set of voices provided by Qwen_Omni does not align with the roles on ActorMindBench; (2) Qwen_Omni is primarily designed for multimodal understanding, resulting in many generated speech segments that are neutral and lacking in expressiveness necessary for role-playing; and (3) During our experiments, when prompts were long—incorporating role profiles and contextual details—Qwen_Omni struggled to accurately express the intended content, rendering it unsuitable for role-playing applications.

6.2 Ablation Study

We conduct ablation studies to evaluate the contribution of each agent in ActorMind, as well as the necessity of the role profile, scene description, and context in the speech role-playing setting. ActorMind operates as a sequential pipeline. Therefore, removing any component may disrupt this pipeline, leading to interdependent effects in the ablation study. For example, removing the Brain agent effectively disables the RAG mechanism in the Mouth agent.

The Eye agent provides all essential textual inputs for role-playing, including the role profile, scene description, and context textual lines. To assess its contribution and the necessity of its inputs, we conduct three ablation settings: (1) **w/o Role Profile (w/o Eye)**, (2). **w/o Scene (w/o Eye)**, (3). **w/o Context (w/o Eye, w/o Ear)**. In (3), removing the textual context eliminates dialogue-based speech emotion processing; therefore, the Ear agent is also removed.

The Ear agent processes speech emotion information. To evaluate its effect, we conduct: (4). **w/o Ear**, where only textual context is used without speech emotion cues.

The Brain agent infers the emotion of the target utterance. Without the Brain agent, RAG in the Mouth agent— which relies on inferred emotions to retrieve speech prompts for TTS—cannot function. Moreover, information from the Eye and Ear agents becomes ineffective. Therefore: (5). **w/o Brain (w/o All)**, which corresponds to **w/o Eye, w/o Ear, w/o Brain, w/o Mouth**.

		Subjective ↑	Objective ↓
(1)	w/o Role Profile (w/o Eye)	-0.37 ± 0.21	+0.02
(2)	w/o Scene (w/o Eye)	-0.30 ± 0.17	+0.03
(3)	w/o Context (w/o Eye, w/o Ear)	-0.22 ± 0.14	+0.07
(4)	w/o Ear	-0.32 ± 0.23	+0.10
(5)	w/o Brain (w/o All)	-0.51 ± 0.56	+0.11

Table 3: **Ablation study.** Relative performance with respect to ActorMind across six roles.

We report relative performance with respect to ActorMind across six roles. Subjective evaluation is measured by the average RP-MOS difference, while objective evaluation is measured by the average Audio FAD difference.

As shown in Table 3, the results of (1)-(3) indicate that removing any of these components leads to performance degradation, highlighting their importance in speech role-playing setting. Among them, removing the role profile results in the largest performance drop, demonstrating that role profile information is the most critical component for speech role-playing. This observation aligns with intuitive expectations for role-conditioned generation.

Overall, the results from settings (1)–(5) show that each component in ActorMind is necessary,

	Phoebe	Joey	Chandler	Rachel	Ross	Monica	ALL
ActorMind + F5-TTS	**1.00 ± 0.00**	0.75 ± 0.29	0.75 ± 0.50	0.50 ± 0.58	0.88 ± 0.25	0.75 ± 0.29	0.77 ± 0.18
ActorMind + Cosyvoice	0.88 ± 0.25	0.63 ± 0.25	0.75 ± 0.29	0.50 ± 0.58	***0.38 ± 0.48***	0.63 ± 0.48	0.63 ± 0.16
ActorMind + SparkTTS	0.50 ± 0.00	0.88 ± 0.25	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.90 ± 0.04
ActorMind + IndexTTS	0.88 ± 0.25	0.75 ± 0.5	***0.25 ± 0.50***	0.75 ± 0.50	0.88 ± 0.25	**1.00 ± 0.00**	0.75 ± 0.17
ActorMind + YourTTS	0.63 ± 0.48	0.50 ± 0.41	0.88 ± 0.25	0.50 ± 0.58	1.00 ± 0.00	0.50 ± 0.58	0.67 ± 0.17

Table 4: Performance improvement over baseline models after applying ActorMind.

validating the soundness of our method design.

6.3 Generalization of ActorMind

ActorMind is a multi-agent CoT reasoning framework. It operates in an off-the-shelf manner without requiring additional training, making generalization a core consideration. To evaluate this property, we replace the speech generation component with different models, and compare ActorMind + [MODEL] against each corresponding standalone model, thereby assessing ActorMind's effectiveness as a universal reasoning framework. We intentionally omit Qwen_Omni, as it does not support target voice generation and therefore cannot support role-playing.

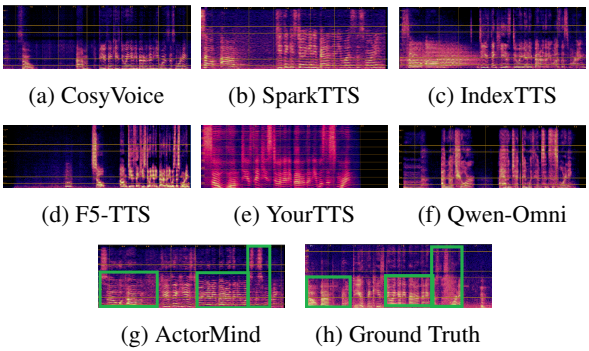
In this experiment, we conduct subjective evaluations, where evaluators assign a score of 1 to indicate a clear improvement, 0.5 to indicate equivalence, and 0 to indicate degradation relative to the baseline model. We recruit six English-speaking evaluators for this study.

As shown in Table 4, except for ActorMind + CosyVoice on Ross and ActorMind + IndexTTS on Chandler, all ActorMind + [MODEL] configurations achieve scores higher than 0.5, demonstrating consistent performance gains over their corresponding baselines. Moreover, five configurations achieve a score of 1, indicating absolute improvement and further highlighting the effectiveness and robustness of ActorMind as a general-purpose reasoning method.

6.4 Qualitative Analysis

We visualize spectrograms of the generated speech to qualitatively evaluate speech role-playing ability. In each spectrogram, the x-axis represents time, reflecting the temporal dynamics and tone, while the y-axis represents the energy distribution across frequency bins, reflecting the vocal characteristics. Higher similarity to the ground-truth spectrogram indicates better alignment in both prosody and speaker-specific traits.

Figure 3 presents spectrograms of the generated outputs alongside the corresponding ground-truth speech. As shown:



(a) CosyVoice (b) SparkTTS (c) IndexTTS

(d) F5-TTS (e) YourTTS (f) Qwen-Omni

(g) ActorMind (h) Ground Truth

Figure 3: Spectrogram Comparison of baselines and ActorMind. All samples are generated for *Phoebe* performing *"...So, um, do you think he's doing any better than he was this morning?"* under the same scene and context.

- TTS models (Figure 3 (a)–(e)) use randomly sampled prompts, resulting in arbitrary tone and prosody. Although these models can successfully generate the target utterance with the target voice, their energy distributions over time and frequency differ substantially from the ground truth. In contrast, as highlighted by the green boxes, ActorMind exhibits significantly higher spectrogram similarity, indicating more accurate role-consistent prosody and expression.
- LLAM, i.e., Qwen_Omni (Figure 3 (f)), fails to reproduce the target voice. This is reflected in its energy distribution across the frequency axis, which deviates significantly from those of the other models and indicates a mismatch in speaker characteristics.

7 Conclusions

To establish speech role-playing, we formalize the concept and introduce ActorMindBench, a public benchmark accompanied by a construction pipeline to facilitate user expansion. We further propose ActorMind, an off-the-shelf, multi-agent, CoT style reasoning framework. Experimental results, evaluated through both subjective and objective metrics, demonstrate the effectiveness of ActorMind.

8 Limitations

ActorMindBench. ActorMindBench is entirely derived from Friends Season 1, covering six roles within the urban comedy domain. As such, it has a limited set of roles and domain coverage. However, we provide a corresponding data construction pipeline, allowing future researchers to expand the benchmark. Despite these limitations, ActorMindBench offers a valuable test bed for current research. Notably, as shown in the main results, current methods perform poorly in speech role-playing setting, indicating that, although ActorMindBench is limited, it is sufficient for researchers to explore and develop new approaches.

ActorMind. ActorMind is an off-the-shelf method that does not require any training. While it demonstrates strong performance, further improvements may be possible through further training, for example, using reinforcement learning to enhance the RAG mechanism in the Mouth agent or to improve emotion reasoning in the Brain agent. Nevertheless, as the first system of its kind, ActorMind represents a meaningful step forward in speech role-playing.

9 Ethical Considerations

ActorMindBench benchmark is built upon the well-known TV series *Friends, Season 1*, and includes annotations at the utterance, scene, and role levels. However, *Friends* is protected by copyright. Accordingly, we do not—and will not—distribute any copyrighted audio content from the series.

All annotations in ActorMindBench are generated using publicly available tools and consist exclusively of structured annotations. We will publicly release only the annotation files, enabling researchers to freely access and use our annotations and independently obtain the original Friends episodes through legitimate channels for their own research purposes. This design strictly follows established community practice and ensures that no copyrighted media is redistributed, thereby avoiding any copyright or licensing violations.

Moreover, releasing our full toolchain enables future researchers to extend and scale the benchmark to additional episodes or other TV series in a fully reproducible manner.

In summary, our work:

- Does not distribute any copyrighted audio content;
- Releases only copyright-safe annotations and toolchains;
- Supports reproducibility and extensibility for future research;

We therefore believe that the ethical and legal usage of the source material in our benchmark is appropriate and rigorously handled.

References

1965. On the performing arts: The anatomy of their economic problems. *The American economic review*, 55(1/2):495–502.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. *CoRR*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2022. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. *arXiv preprint arXiv:2211.06869*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Min Chu and Hu Peng. 2006. Objective measure for estimating mean opinion score of synthesized speech. US Patent 7,024,362.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. 2025. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In *Proc. IEEE ICASSP 2024*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Linus Johansson. 2025. Open weight large language models as a design material in rpgs.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024. Rolellm: benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Konstantin Stanislavski and Jean Benedetti. 2009. *An actor's work on a role*. Routledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yusong Wu, Ke Chen, Tianyu Zhang*, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

A ActorMindBench Details

A.1 ActorMindBench Example Data

Example of ActorMindBench data is shown in Figure 4.

Role-Level Content:

Monica: *Monica Geller is a pivotal character in Friends, known for her strong personality and distinct traits. She is depicted as a cleanliness-obsessed, highly organized, and competitive .....*

Phoebe: *Phoebe Buffay is the quirkiest and most unconventional member of the Friends group. Her personality is a unique blend of optimism, eccentricity, and kindness, making .....*

Joey: *Joey Tribbiani is a character whose personality is a vibrant blend of charm, humor, and endearing simplicity. As a struggling actor, Joey's career is marked by his laid-back approach .....*

Rachel: .....

Chandler: .....

Ross: *Ross Geller, a central character in the hit sitcom Friends, is a unique blend of intelligence, insecurity, and lovable quirks. As a paleontologist and professor, Ross's personality is deeply .....*

Scene-Level Content:

Scene Description: *The scene is set at Central Perk, where Monica is sitting alone when her friends Ross, Rachel, Chandler, and Joey enter, dressed in softball gear. They are all excited and dejected at the same time, and Monica asks how their game went. They reveal that they won, thanks to Alan's incredible performance on the field, playing multiple positions like Bugs Bunny in a cartoon. Monica is skeptical, suggesting that Alan might be "too Alan" sometimes, implying that he might be a bit too good or unusual. Her friends reassure her that being "too Alan" is impossible and that his unique qualities are what make him special.*

Scene Boundary: *130-143*

Utterance-Level Content:

```
"130": "Monica: Hi.. how was the game?",    [speech segment 130]
"131": "Ross: Well..",      [speech segment 131]
"132": "All: WE WON!! Thank you! Yes!", [speech segment 132]
"133": "Monica: Fantastic! I have one question: How is that possible?",      [speech segment 133]
"134": "Joey: Alan.", [speech segment 134]
"135": "Ross: He was unbelievable. He was like that-that-that Bugs Bunny cartoon where Bugs is
playing all the positions, right, but instead of Bugs it was first base-Alan, second base-Alan, third
base-...", [speech segment 135]
"136": "Rachel: I mean, it-it was like, it was like he made us into a team.",    [speech segment 136]
"137": "Chandler: Yep, we sure showed those Hassidic jewellers a thing or two about softball..",
[speech segment 137]
"138": "Monica: Can I ask you guys a question? D'you ever think that Alan is maybe.. sometimes..",
[speech segment 138]
"139": "Ross: What?", [speech segment 139]
"140": "Monica: ..I dunno, a little too Alan?",  [speech segment 140]
"141": "Rachel: Well, no. That's impossible. You can never be too Alan.",   [speech segment 141]
"142": "Ross: Yeah, it's his, uh, innate Alan-ness that-that-that we adore.",    [speech segment 142]
"143": "Chandler: I personally could have a gallon of Alan."     [speech segment143]
```

Figure 4: ActorMindBench Example Data

A.2 Prompts in ActorMindBench Construction

Prompts for scene summarization and role summarization can be seen in Figure 5 and Figure 6 separately.

You should summarize the scene based on the given conversation.

Monica: *Phoebe, what are you doing?*
Phoebe: *Maybe nobody's tried this*
Monica: *I wish we at least knew his name...*
Phoebe: *Yeah, but did you see the dents in his knuckles? That means he's artistic.*
Monica: *Okay, he's a lawyer, who teaches sculpting on the side. And- he can dance!*
.....

Figure 5: Prompt for scene captioning: black text indicates the prompt, while brownish-yellow highlights the utterances within a scene.

Using the introduction of *Monica* from her Wikipedia page, please create a role profile, ensuring it does not exceed 200 words.

.....
A chef known for her cleanliness, competitiveness and obsessive-compulsive nature, Monica is the younger sister of Ross Geller and best friend of Rachel Green, the latter of whom she invites to live with her after
.....

Figure 6: Prompt for role profile generation: black text indicates the prompt, while brownish-yellow highlights the role content.

A.3 ActorMindBench Detail Statistics

Utterance-Level statistics regarding the number and duration of utterances for role 'xx' in episode 'yy' are provided in Table 5.

Scnene-Level statistics regarding the average number of utterances and roles performed per scene in each episode are provided in Table 6.

B Experiment Details

B.1 Qwen_Omni in Speech Role-Playing

The prompt used for Qwen_Omni speech role-playing is shown in Figure 7.

B.2 RP-MOS

This subjective evaluation assesses the model's speech role-playing ability by comparing generated speech with reference (ground-truth) recordings. Participants listen to the generated speech and

You are an actor, destined to portray {*Phoebe: Phoebe Buffay is the quirkiest and most unconventional member of the Friends group. Her personality is a unique blend of optimism, eccentricity, and kindness*}.
You are performing in the scene {*The scene takes place in a hospital, where Monica and Phoebe are visiting a guy who is in a coma. Monica is struggling to understand why she reacted so strongly to the guy, even though he's unconscious. Phoebe...* }.
The dialogue you have heard so far is:
{*audio signal*}
Your next utterance to deliver is: {*Oh! And, he's the kinda guy who, when you're talking, he's listening, y'know, and not saying 'Yeah, I understand' but really wondering what you look like naked*.}.

Produce natural, spontaneous, and expressive speech appropriate for this character.

Figure 7: Qwen_Omni Prompt for Speech Role-Playing. Black text: prompt template; Brownish-yellow: corresponding speech and text content.

assign scores based on its similarity to the reference speech. Ten english speakers evaluated twelve rounds (six roles, with two sets per role), using randomly selected utterances across all model variants. All evaluators were provided with detailed guidelines and evaluation criteria prior to the assessment.

Each evaluator was compensated at a rate aligned with the average local hourly income, which we consider fair and appropriate given their country of residence and time commitment.

Guidelines and Evaluation Criteria Participants should consider the following aspects of emotional expression:

- **Emotional Consistency**: Does the generated speech convey the same emotional tone as the reference speech?
- **Intensity Alignment**: Is the strength or intensity of the emotion comparable between the two speeches?
- **Naturalness and Realism**: Does the generated audio sound naturally expressive and believable, rather than artificial or flat?
- **Overall Impression**: Considering all the above factors, how similar is the emotional quality of the generated speech to the real reference?
- **Voice and Content Consistency**: If the voice is from different people or the text content differs from the reference, directly assign a score of 1.

Episode	Rachel		Monica		Phoebe		Joey		Chandler		Ross		OTHERS		TOTAL	
	num	duration	num	duration	num	duration	num	duration	num	duration	num	duration	num	duration	num	duration
SE01_01	86	0:03:15	82	0:03:08	22	0:00:54	39	0:01:42	33	0:01:23	63	0:02:45	25	0:01:07	350	0:14:13
SE01_02	56	0:02:36	37	0:01:15	21	0:00:44	11	0:00:23	29	0:01:10	101	0:04:16	97	0:03:44	352	0:14:09
SE01_03	32	0:01:12	76	0:02:36	65	0:02:20	28	0:01:04	72	0:02:37	53	0:01:54	39	0:01:31	365	0:13:14
SE01_04	81	0:03:16	65	0:02:27	47	0:01:52	36	0:01:15	49	0:01:55	57	0:02:36	36	0:01:23	371	0:14:44
SE01_05	48	0:02:01	40	0:01:44	29	0:00:57	58	0:02:23	50	0:01:51	73	0:03:30	48	0:02:05	346	0:14:31
SE01_06	22	0:00:57	54	0:02:12	22	0:00:48	52	0:02:15	114	0:04:22	32	0:01:28	52	0:02:02	348	0:14:04
SE01_07	49	0:02:00	21	0:00:43	44	0:01:52	38	0:01:26	61	0:02:47	71	0:03:04	27	0:00:52	311	0:12:45
SE01_08	23	0:01:00	45	0:01:34	27	0:00:59	15	0:00:31	60	0:02:09	75	0:03:07	94	0:03:41	339	0:13:02
SE01_09	64	0:02:28	74	0:02:45	31	0:01:13	43	0:01:28	54	0:02:07	65	0:02:43	34	0:01:20	365	0:14:04
SE01_10	32	0:01:28	22	0:00:50	58	0:02:21	20	0:00:46	54	0:01:54	53	0:02:10	77	0:03:20	316	0:12:49
SE01_11	28	0:01:03	48	0:01:46	49	0:01:54	44	0:01:41	53	0:01:46	78	0:02:43	66	0:02:23	366	0:13:15
SE01_12	49	0:02:04	38	0:01:29	49	0:01:51	36	0:01:26	39	0:01:22	71	0:02:45	30	0:01:18	312	0:12:14
SE01_13	26	0:01:15	15	0:00:30	27	0:01:16	52	0:02:12	41	0:01:35	14	0:00:41	106	0:05:01	281	0:12:30
SE01_14	19	0:00:52	20	0:00:50	17	0:00:50	32	0:01:14	51	0:01:55	53	0:02:31	83	0:03:49	275	0:12:02
SE01_15	25	0:00:57	44	0:02:04	39	0:01:34	32	0:01:15	70	0:02:55	35	0:01:32	24	0:01:01	269	0:11:19
SE01_16	27	0:01:08	13	0:00:36	41	0:02:03	22	0:01:01	59	0:02:34	36	0:01:39	75	0:03:17	273	0:12:17
SE01_17	54	0:02:05	63	0:02:40	33	0:01:29	30	0:01:13	25	0:01:04	50	0:02:12	102	0:03:56	357	0:14:38
SE01_18	84	0:03:42	38	0:01:31	34	0:01:25	21	0:00:55	33	0:01:16	59	0:02:18	9	0:00:20	278	0:11:27
SE01_19	91	0:04:04	35	0:01:21	18	0:00:45	19	0:00:51	27	0:01:06	88	0:03:58	40	0:01:38	318	0:13:43
SE01_20	85	0:03:51	30	0:01:13	21	0:00:48	34	0:01:32	62	0:02:32	22	0:01:05	69	0:02:57	323	0:13:58
SE01_21	34	0:01:26	64	0:03:03	11	0:00:34	21	0:00:57	25	0:01:10	51	0:02:35	67	0:03:09	273	0:12:53
SE01_22	27	0:01:05	50	0:02:16	53	0:02:07	16	0:00:38	50	0:02:03	41	0:01:42	40	0:01:53	277	0:11:44
SE01_23	24	0:01:02	25	0:00:55	35	0:01:40	34	0:01:28	22	0:00:57	68	0:02:44	102	0:04:23	310	0:13:09
SE01_24	64	0:02:59	35	0:01:39	19	0:00:54	62	0:02:30	28	0:01:15	36	0:01:34	34	0:01:38	278	0:12:28
ALL	**1130**	**0:47:47**	**1034**	**0:41:06**	**812**	**0:33:09**	**795**	**0:32:06**	**1161**	**0:45:43**	**1345**	**0:57:31**	**1376**	**0:57:50**	**7,653**	**5:15:12**

Table 5: Utterance-level statistics on number of utterances and speech duration by role and episode.

Episode	Scene Num	Avg Utterances per Scene	Avg Roles per Scene
SE01_01	14	17.29	3.93
SE01_02	8	29.75	5.38
SE01_03	13	20.0	4.85
SE01_04	16	15.75	4.19
SE01_05	16	14.94	3.31
SE01_06	9	24.33	4.78
SE01_07	21	11.14	2.95
SE01_08	10	16.9	4.50
SE01_09	12	19.08	3.92
SE01_10	8	29.0	6.00
SE01_11	12	23.92	4.50
SE01_12	15	17.33	4.33
SE01_13	13	18.69	4.31
SE01_14	17	11.12	3.41
SE01_15	14	17.43	3.43
SE01_16	14	19.5	5.07
SE01_17	14	20.14	4.14
SE01_18	8	33.38	6.25
SE01_19	8	31.38	5.12
SE01_20	12	20.33	4.92
SE01_21	15	14.07	4.00
SE01_22	12	21.42	4.00
SE01_23	21	12.76	4.1
SE01_24	11	23.91	4.00
ALL	**313**	**18.7**	**4.23**

Table 6: Scene-level statistics.

Scoring Scale Use the following 5-point scale to rate each speech:

- **5 – Identical**: The generated speech conveys the same emotion as the reference speech, with nearly identical intensity and expression.
- **4 – Very Similar**: The emotion is highly similar, with only minor differences in tone or intensity.
- **3 – Moderately Similar**: The overall emotion is recognizable but with noticeable differences in strength, tone, or expression.
- **2 – Weak Similarity**: The emotion type is somewhat related but largely inconsistent with the reference speech.
- **1 – No Similarity**: The generated audio conveys a completely different or unrecognizable emotion compared to the reference; If the voice seems to be from a different speaker, or if the text content differs from the reference, directly assign a score of 1.

Evaluation Procedure

1. Listen to each speech at least twice before rating.
2. If the voice is from different people or the text content differs from the reference, directly assign a score of 1.

I am an actor, destined to portray {*Phoebe: Phoebe Buffay is the quirkiest and most unconventional member of the Friends group. Her personality is a unique blend of optimism, eccentricity, and kindness*}.
I am performing in the scene {*The scene takes place in a hospital, where Monica and Phoebe are visiting a guy who is in a coma. Monica is struggling to understand why she reacted so strongly to the guy, even though he's unconscious. Phoebe* }.
The dialogue I have heard so far is:
{*Monica: Phoebe, what are you doing?*
Emotion: Concerned, slightly exasperated.....
Phoebe (Me): Maybe nobody's tried this
Monica: I wish we at least knew his name...
Emotion: Concerned, slightly exasperated.....
Phoebe (Me): Yeah, but did you see the dents in his knuckles? That means he's artistic.
Monica: Okay, he's a lawyer, who teaches sculpting on the side. And- he can dance!
Emotion: Amused, optimistic, and slightly fanciful, with a hint of romanticism
...}
My next utterance to deliver is: {*Oh! And, he's the kinda guy who, when you're talking, he's listening, y'know, and not saying 'Yeah, I understand' but really wondering what you look like naked.*}.
Use 3–20 words to describe the tone or emotion for {*Oh! And, he's the kinda guy who, when you're talking, he's listening, y'know, and not saying 'Yeah, I understand' but really wondering what you look like naked*}. Only the feeling, no extra text.

Wistful, admiring, and slightly dreamy, with a hint of romantic longing.

Figure 8: Prompt for the Brain Agent used for role injection, contextual understanding, and emotion rendering. Here, yellow content represents what the Eye Agent saw, blue content represents what the Ear Agent heard, and purple content represents what the Brain Agent inferred.

3. Assign a single integer score (1–5) according to the criteria above.

4. If uncertain, choose the score that best represents your overall impression.

B.3 ActorMind Implementation Details

Eye Agent reads the preparatory descriptive content and retains it in memory. In practice, a textual memory of a few hundred words is sufficient.

Ear Agent Speech Emotion Captioning (SECAP) provides textual and intuitive emotional descriptions of target speech signals. SECAP[13] (Xu et al., 2024) equips the Ear Agent with listening and emotion-recognition capabilities.

Brain Agent is the central component of ActorMind. Following previous LLM role-playing works (Dai et al., 2024; Moore Wang et al., 2024), we use LLama3 (Dubey et al., 2024) to perform emotional state reasoning, with the prompts illustrated in Figure 8.

Mouth Agent employs RAG to retrieve relevant context from a database for speech generation. In ActorMind, for each role R_k, the database $Database_k$ is constructed from that role's known speech utterances. Each entry contains the speech signal U_x^s as content, with indices corresponding to emotional descriptions E_x generated by SECAP (Xu et al., 2024). During the retrieval phase, embeddings are computed using OpenAI's *text-embedding-3-large*[14]. During the generation phase, we employ IndexTTS [15] (Deng et al., 2025) as speech synthesizer, where the text prompt is the target line and the tone and emotion prompt is the retrieved speech.

[13] https://github.com/thuhcsi/SECap

[14] https://platform.openai.com/docs/models/embeddings

[15] https://github.com/index-tts/index-tts