PACHAT: Persona-Aware Speech Assistant for Multi-party Dialogue

Dongjie Fu[1,2*] **Xize Cheng**[1*] **Linjun Li**[2*]
Xiaoda Yang[1] **Lujia Yang**[1] **Tao Jin**[1†]
[1] Zhejiang University [2] Meituan
{fudongjie,chengxize,xiaodayang,yanglujia,jint_zju}@zju.edu.cn
lilinjun05@meituan.com

Abstract

Extensive research on LLM-based spoken dialogue systems has significantly advanced the development of intelligent voice assistants. However, the integration of role information within speech remains an underexplored area, limiting its application in real-world scenarios, particularly in multi-party dialogue settings. With the growing demand for personalization, voice assistants that can recognize and remember users establish a deeper connection with them. We focus on enabling LLMs with speaker-awareness capabilities and enhancing their understanding of character settings through synthetic data to generate contextually appropriate responses. We introduce Persona-Dialogue, the first large-scale multi-party spoken dialogue dataset that incorporates speaker profiles. Based on this dataset, we propose PAChat, an architecture that simultaneously models both linguistic content and speaker features, allowing LLMs to map character settings to speaker identities in speech. Through extensive experiments, we demonstrate that PAChat successfully achieves speaker-specific responses, character understanding, and the generation of targeted replies in multi-party dialogue scenarios, surpassing existing spoken dialogue systems. For more details, please visit our demo page at https://persona-dialogue.github.io/.

1 Introduction

The impressive natural language understanding and dialogue generation capabilities of large models have quickly been applied to the field of chat assistants, significantly improving the quality of human-computer interaction. With the introduction of multimodal technology (Cheng et al., 2023; Huang et al., 2023; Li et al., 2023; Fu et al., 2024; Yan et al., 2025; Yang et al., 2024), researchers have built a bridge between the text input of large language model and human speech, enabling more direct communication between large models and users (OpenAI, 2024a). Compared to cascade models that used ASR to transcribe speech and then generated responses from text input models (Hoy, 2018; Hachman, 2019), spoken dialogue models capable of accepting speech can extract more information beyond the word to support more intelligent interaction.(Ji et al., 2024; Cheng et al., 2025c,b,a)

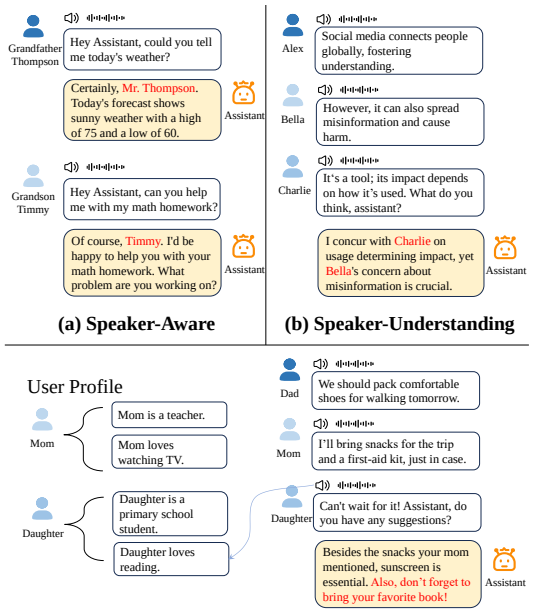

Figure 1: **Examples of identity information used in spoken conversations. (a)** Speaker-Aware: The model can accurately recognize the user and address them correctly. **(b)** Speaker-Understanding: The model is capable of identifying the identities of different speakers in a conversation. **(c)** Persona-Aware: The model can generate personalized responses by linking the identity information in the speech to the user's profile.

Initial speech-related MLLM research focused on content-centered spoken language modeling (Zhang et al., 2023a), and later some researchers explored speech audio processing (Kong et al., 2024;

*Equal contribution.
†Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29314–29331
November 4-9, 2025

Chu et al., 2023; Tang et al., 2023), while others focused on mining style and emotional factors in speech (Lin et al., 2024). Furthermore, considering the complexity of real-world dialogue scenarios where there may be more than two participants, the ability to identify voices is an important measure of the intelligence of a spoken dialogue model. Some research tested their models' capabilities on speaker recognition Q&A tasks, achieving impressive results (Chu et al., 2023; Tang et al., 2023). However, these models do not have the ability to handle voice and speaker information in dialogue, missing out on two types of real-world applications: understanding multi-role chat scenarios, and generating personalized responses for users.

Considering that voice assistants are gradually expanding from personal assistants to more complex scenarios, we believe that the model's role understanding ability should be reflected in three aspects, as shown in Figure 1. (a) Speaker-Aware, can be exemplified in the scenario where the model, upon receiving the user's voice signal, can return the correct address, thereby ensuring the fundamental etiquette of an assistant. This function can be completed under the training of most models' speaker verification tasks. (b) Speaker-Understanding, which means in a multi-party conversation, the model can establish a mapping with the role according to different timbres to understand the content of the conversation. For example, in a meeting scenario, the voice assistant can organize meeting minutes based on different people's speeches, completing tasks like meeting records and summaries. (c) Persona-aware ability, which means the model needs to understand the personal information behind the speaker's identity, linking voice and personal background, and maintaining the intelligence of personal voice assistants in multi-party scenarios to meet users' personalized needs. For instance, in a family scenario, when faced with dinner suggestions, the voice assistant needs to generate personalized responses based on the background information of different family members.

Current voice models have not paid attention to the latter two types of user needs, resulting in a lack of related ability. At the same time, due to the high cost of voice data collection, the privacy of spoken language, and copyright issues, there are very limited dialogue corpus datasets related to personalization, limiting the progress of dialogue models on personalization issues.

Table 1: Comparison of Spoken Dialogue Datasets. In the 'Source' column: **Env** means controlled environments, **Wild** means in-the-wild collection, and AI generation for (**AI-Gen**). **#Dialogues** represents the number of dialogues.

Dataset	Multi-Party	Persona	Source	Dialogues
Text-Based Dialogue Dataset				
Ubuntu IRC Logs	✔	✗	Wild	665k
Reddit Dataset	✔	✗	Wild	120k
Synthetic-Persona-Chat	✗	✔	AI-Gen	10,906
Live-Chat	✔	✔	Wild	1.33m
Spoken Dialogue Dataset				
MELD	✔	✗	Wild	1,433
DailyTalk	✗	✗	Env	2,514
SpokenWOZ	✗	✗	Env	5,700
StyleTalk	✗	✗	AI-Gen	2,967
Persona-Dialogue	✔	✔	AI-Gen	21,760

In view of the above problems, we choose to use large-scale synthetic data to simulate personalized needs scenarios to enhance the oral dialogue model's understanding of the identity information implied behind the voice. With the powerful data generation capabilities of large models (OpenAI et al., 2024), we have established a comprehensive text data generation process, including character information generation, background dialogue generation, and chat generation. Through high-fidelity, controllable TTS models (Du et al., 2024b), we generate oral dialogues in a zero-shot manner using open-source timbres and verify the speaker's identity to ensure the validity of the dataset. As shown in Table 1, we propose Persona-Dialogue, the first large-scale synthetic oral dialogue dataset that includes multi-party dialogue and role information, generating personalized responses for users from the perspective of voice assistants, covering the understanding of both dialogue content and role information. For the task of personalized voice assistant, we introduce PAChat, the first voice dialogue system designed to handle multi-party conversations and role understanding. PAChat explicitly models voice information and user representation and aligns with LLM to achieve natural personalized responses. We establish a benchmark for personalized responses in multi-party conversations, and comprehensively measure the performance of the model from both objective and subjective perspectives. Extensive experiments show that our model achieves state-of-the-art performance. In addition, we conduct more experiments on available real datasets, and the excellent performance of PAChat proves the effectiveness of synthetic data in dialogue system training. Our contributions are as follows:

• We propose Persona-Dialogue, the first large-

scale multi-role spoken dialogue dataset that includes detailed user profiles, with the aim of advancing the development of personalized voice assistants.

- We introduce PAChat, the first spoken dialogue model focused on personalized responses in multiparty conversations. The model simultaneously models semantic and speaker information, aiding the model in understanding the role identity behind the speech.
- We establish a benchmark for assessing whether the model understands the user's identity behind the speech, comprehensively measuring the model's persona understanding ability in various task forms.
- PAChat achieve the state-of-the-art performance on both our established benchmark and real-world data, demonstrating the effectiveness of Persona-Dialogue.

2 Related Work

The powerful semantic understanding and dialogue generation capabilities of large models have made them efficient chat assistants. Meanwhile, the development of multimodal technology (Cheng et al.; Yang et al., 2025; Xu et al., 2024; Shi et al., 2024; Xie et al., 2025; Yan et al., 2024; Yang et al., 2025) has aligned the sound modality to the input of large models, enabling large models to achieve extensive audio understanding. SpeechGPT (Zhang et al., 2023a) is the first to incorporate discrete speech units into the LLM framework, with subsequent work such as AudioPaLM (Rubenstein et al., 2023) continuing research on speech content. Some researchers are dedicated to using a unified model to handle speech and audio issues. SALMONN (Tang et al., 2023) enables LLMs to directly process and understand general audio inputs, which can be seen as a step towards AI with generic hearing abilities. Qwen-Audio1 (Chu et al., 2023) and Qwen-Audio2 (Chu et al., 2024) established the first comprehensive large-scale audio model capable of handling over 30 audio-related tasks, achieving excellent results in tasks such as speech recognition, speech translation, and audio event detection. As the audio understanding ability of the model improves, more intelligent spoken dialogue models emerge. StyleTalk (Lin et al., 2024) focuses on mining the style factors of spoken dialogue to generate responses with different emotional tones. While recent end-to-end speech dialogue models have developed rapidly (Fang et al., 2024; Zeng et al., 2024), they are more focused on the generation end compared to generating text through TTS to generate speech responses, only focusing on the semantic content of speech at the input end, ignoring the need for emotion information or identity information mining in spoken dialogue.

Some spoken dialogue models already have identity recognition capabilities, but are limited to simple Q&A tasks and have not achieved personalized dialogue based on identity information, overlooking the huge application value in this field. One obstacle to achieving personalized dialogue is the lack of datasets. The high cost of speech data collection means that some models can only experiment on small-scale data (Lin et al., 2024), and personal information is difficult to record. The text dialogue field has in-depth research on personalized dialogue (Jandaghi et al., 2023; Gao et al., 2023), but there is still a certain difference between text dialogue data such as online chat data and spoken dialogue. Some researchers use publicly available resources such as TV shows to compile spoken dialogue datasets (Poria et al., 2018; Chen et al., 2020), providing valuable resources. However, these resources, despite their value, are inherently constrained by their scale and the format of dialogues, which potentially limits their applicability in the context of voice assistant training. With the advancement of LLM technology, the synthesis of data leveraging its robust generative capabilities has emerged as a significant source for dialogue data training materials. StyleTalk (Lin et al., 2024) used a large language model (OpenAI, 2024b) combined with controllable text-to-speech (TTS) technology (Du et al., 2024a) to enhance the model's ability to capture varied speaking styles and respond properly in spoken conversations.

To endow the model with the ability to understand identity information and respond accordingly, we introduce Persona-Dialogue, the first large-scale synthetic spoken dialogue dataset with persona information annotations. By leveraging synthetic data, Persona-Dialogue significantly enhances the conversational capabilities of spoken dialogue systems, pushing their application boundaries in diverse and challenging environments.

3 Dataset: Persona-Dialogue

Unlike personal voice assistants, persona-dialogue focuses on assistant scenarios serving multiple

users, such as a home steward or shopping mall assistant, fulfilling roles such as chatting, answering, and guiding. Users register their voice identities in the corresponding scenarios and engage in spoken dialogue with other users or the assistant. The assistant then uses the users' voice information to understand the dialogue and generate personalized responses for answers or chat. We have established a comprehensive data construction process to ensure the naturalness and realism of the dataset, as shown in Figure 2. More details can be found in Section A.

3.1 User Profile Construction.

Taking into account the possibility of multiple user registrations within the scenario, we systematically collected and designed 21 representative multi-party conversational situations that necessitate the involvement of an intelligent assistant. The detailed scenario information was subsequently input into a large-scale language model, which, by virtue of its advanced inference capabilities, is employed to generate a diverse set of user profiles corresponding to each scenario. Each user may be present in the given scenario, and is characterized by a five-sentence background description encompassing identity, personal interests, and other scenario-relevant information. All details collectively depict an independent user profile, with each piece of information potentially serving as a contextual clue that the model must consider when generating responses.

3.2 Textual Dialogue Scripts Generation.

Each user may participate in the corresponding scenario and engage in relevant dialogues. Accordingly, for each scenario, users are randomly paired, and the selected user profiles are utilized as pre-set backgrounds to facilitate the generation of potential dialogues between users with the assistance of a large language model. Each set of dialogues encompasses both user-to-user communication and interactions between users and the voice assistant.

To ensure data diversity, we employ multiple data generation strategies. From the perspective of user-assistant interaction, we employ two primary methodologies. In the first approach, a large language model is utilized to generate complete dialogues encompassing both user and assistant utterances in a single process. While this method tends to produce highly coherent conversations, it may blur the distinction between user profile information and dialogue history. In the second approach, dialogues are initially generated exclusively between users, after which a selected user engages in subsequent interaction with the assistant. This sequential generation process more closely mirrors conversational patterns commonly observed in real-world scenarios. Furthermore, with respect to the inclusion or masking of user information, we generated dialogues under two settings: one that incorporates user profile information and another that omits it. Omitting user information helps prevent potential interference of pre-defined profiles with the assistant's understanding during dialogue, while incorporating user information enables the generation of more diverse and persona-consistent conversations.

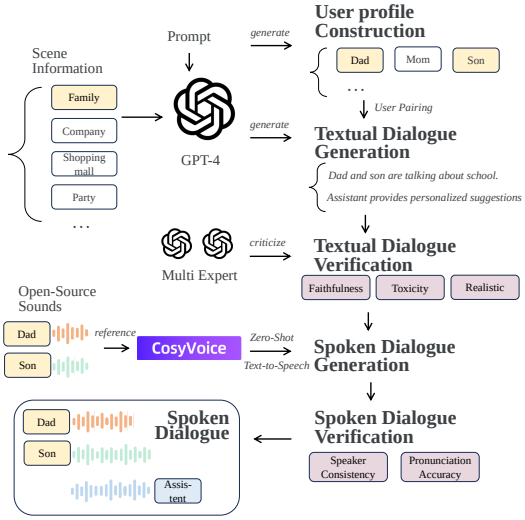

Figure 2: Spoken dialogue generation pipeline of Persona-Dialogue. Before constructing the dialogue, the user profile is first designed to ensure a close connection between the two parties. Additionally, multiple verification processes are employed to ensure the dataset's realism and naturalness.

3.3 Textual Dialogue Verification.

We utilize large language model as automated referee to ensure the quality and appropriateness of generated dialogues. Given the close relationship between user profiles and dialogue content in our scenario, we implement a multi-faceted evaluation framework comprising three distinct criteria. First, the faithfulness evaluation examines whether user utterances are consistent with their assigned profile attributes and flags any unreasonable or inconsistent statements, thereby reducing the risk of hallucinations and contradictions. Second, the relevance evaluation assesses, given a set of user profiles

and two candidate dialogues, which dialogue more appropriately corresponds to the provided user information. Third, the toxicity evaluation screens for the presence of prejudiced or hateful content, ensuring that harmful dialogues are excluded. This comprehensive approach ensures that the generated dialogues are not only coherent and contextually appropriate, but also safe and aligned with the intended user characteristics.

3.4 Spoken Dialogue Generation.

Since voice serves as the sole means for the model to identify users, we employ an open-source timbre corpus (Chung et al., 2018) as reference and leverage the powerful zero-shot generation capabilities of the TTS model CosyVoice2 (Du et al., 2024b) to synthesize voice data for each user. This approach ensures that each user's timbre is both unique and consistent across multiple dialogue turns, thereby minimizing the risk of voice confusion that could compromise the model's ability to accurately distinguish and understand individual speakers.

3.5 Speaker Verification

Due to the stringent requirements of the TTS model's zero-shot generation capability for reference audio, short or ambiguous reference samples may result in synthesized speech that is completely unintelligible. Therefore, we conduct rigorous validation of the generated results to ensure the quality and reliability of the spoken dialogue data. A speaker verification model (Plaquet and Bredin, 2023) is employed to confirm that the synthesized speech for each individual user maintains a consistent timbre across all utterances. In addition, we apply an automatic speech recognition (ASR) model (Radford et al., 2023) to control the word error rate of the generated samples within a low range, thereby ensuring the semantic accuracy of the synthesized speech. For each dialogue, multiple rounds of generation were performed until the synthesized data met the required standards.

4 Spoken Dialogue System: PAChat

The framework of PAChat is shown in Figure 3, including Llama 3.1 (Dubey et al., 2024) as large language model, Whisper model (Radford et al., 2023) as semantic encoder, identity encoder, the corresponding alignment layer Q-Former (Zhang et al., 2023b), and the TTS system CosyVoice2 (Du et al., 2024b). Considering the complexity of multi-role dialogue, we limit the model's response

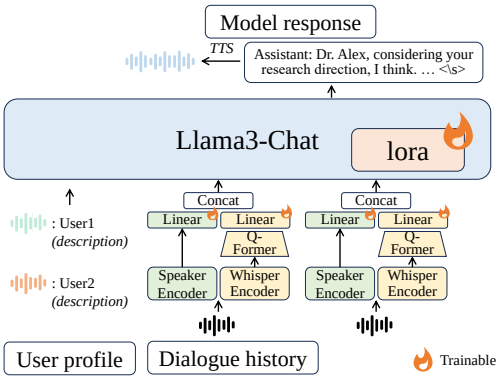

Figure 3: Overview of PAChat. PAChat explicitly models both semantic features and speaker representations. The semantic features employ Whisper and Q-former for more delicate feature extraction, aligned to the input of LLM. The model comprehends both the user profile and dialogue history information, generating personalized responses for specific users.

to interaction with a single user. We define the problem as: given a group of user information $U = \{U_1, U_2, \ldots, U_M\}$ and dialogue history $S = \{\mathbf{S}_{speaker_x,1}, \mathbf{S}_{speaker_y,2}, \ldots, \mathbf{S}_{speaker_z,T}\}$, where T represents the total number of dialogue turns and $speaker_x$ could be any speaker, potentially including the assistant. The current round is initiated by a user to the assistant, and the model generates a corresponding personalized text response $R_{assistant}$, and the efficient TTS system generates a real-time voice response $S_{assistant}$.

4.1 Base Large Language Model

We adopt Llama 3.1-Instruct 8B as our base large model for two considerations. From the model perspective, the Llama series, as the leading open-source model, possesses strong dialogue generation capabilities. Secondly, the 8B size serves as the golden dimension for dialogue assistants, exhibiting intelligent performance without high training costs. During the training process, the Llama 3.1-Instruct model remains frozen. We use the trainable LoRA adapter (Hu et al., 2021) for efficient parameter fine-tuning.

4.2 Speech Encoder

Compared to cascade models that use ASR to directly obtain text format for large model input, our model adopts feature extraction and alignment approach to understand the hidden information behind the acoustic features of speech. We simultaneously model semantic features and speaker features in speech, enabling the model to understand what the speaker is expressing while recognizing

the speaker's identity, thus acquiring the ability to understand dialogue and user identity information. For semantic features, we employ the Whisper encoder, which is trained with weak supervision on large-scale speech corpora, for feature extraction. The process of obtaining semantic features $\mathbf{F}_i^w \in \mathbb{R}^{N \times D_w}$, where $\mathbf{N}$ is the number of frames in each audio feature, can be represented as follows:

$$\mathbf{F}_i^w = \text{Whisper-Encoder}(\mathbf{S}_{user,i}) \quad (1)$$

For the whisper-encoded feather, we employ a window-level Q-Former to synchronize audio and language between frozen audio encoders and a frozen large language model, which can be formulated as:

$$\mathbf{F}_i^{w'} = \text{Q-Former}(\mathbf{Q^w}, \mathbf{F}_i^w) \quad (2)$$

where $\mathbf{Q^w}$ represents trainable queries and $\mathbf{F}_i^{w'} \in \mathbb{R}^{[N \times K/L] \times D_w}$ represent window-level attribute features.

For the extraction of speaker identity features, we employ the efficient PyAnnote.audio library (Plaquet and Bredin, 2023). The process can be represented as follows:

$$\mathbf{F}_i^s = \text{Speaker-Encoder}(\mathbf{S}_{user,i}) \quad (3)$$

The obtained speaker representation is a sequence-level feature, which no longer requires further encoding. We use a simple linear layer structure to align the extracted features with the input of the large model. The speaker identity information is concatenated with the frame-level semantic representation, and combined with the user profile and dialogue history as the input sequence to the large model.

4.3 Training

During the model training process, we froze the large model and encoder, and trained the LoRA adapters and alignment module separately. The overall training objective of the model can be denoted as:

$$L = -\sum_{t=1}^{N}\sum_{j=1}^{m} \log p(\mathbf{R}_t^j | \mathbf{Z}_{1:t}, \mathbf{R}_{1:t-1}, \mathbf{R}_t^{1:j-1}), \quad (4)$$

where N is the total number of dialogue turns, m is the number of tokens in the t-th turn's response, $\mathbf{R}_t^j$ is the j-th token in the response for the t-th turn, $\mathbf{Z}_{1:t}$ represents the audio features up to the t-th turn, and $\mathbf{R}_{1:t-1}$ refers to the tokens from all previous turns, while $\mathbf{R}_t^{1:j-1}$ denotes the preceding tokens within the same turn. This training objective guarantees that the model acquires the ability to produce contextually suitable responses across several dialogue exchanges, utilizing both the dialogue history and the audio features.

5 Experiment

5.1 Evaluation

For the metric, we draw on previous work (Lin et al., 2024) and adopt objective and subjective evaluations to evaluate the text output of models, which is aligned with the output format of other audio large language models. For objective evaluation, we utilized widely recognized text generation metrics, including vocabulary level scores such as BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), as well as semantic-level metrics like BERTScore (Zhang et al., 2019). For subjective evaluation, we used GPT-eval and Human-eval. We established detailed scoring criteria for the quality of questions as the standard for referee judgment, and each sample was scored by multiple referees and averaged to ensure the fairness of the results. More details can be found in Section C.1

For the evaluation dataset, we partitioned 1000 dialogues in Persona-Dialogue as the test set, including scenarios and tasks not found in the training set to verify the model's robustness to test the model's ability to understand dialogue and identity information. Furthermore, we employed zero-shot evaluation of Persona-Chat (Zhang et al., 2018), a classic persona dialogue dataset, to test the model's generalization performance on out-of-domain datasets. Since it is a text dialogue and consists of dialogues between two parties, which does not align with our assistant scenario, we constructed a spoken dialogue format using a TTS approach similar to Persona-Dialogue and added a round of dialogue between the speaker and the assistant for evaluation with the same way mentioned in Section 3.

For comparison, we selected FunAudioLLM (SpeechTeam, 2024), Salmonn (Tang et al., 2023), Qwen-audio (Chu et al., 2023), and Qwen-audio2(Chu et al., 2024), some of the most powerful dialogue models, as baselines. As they were not designed to handle multi-party conversations, we modified their data input format: before the dia-

Table 2: Performance comparison of various methods for spoken dialogue systems on the Persona-Dialogue and Persona-Chat datasets. The content metrics include **@B** (BLEU), **@R** (ROUGE-L), **@M** (METEOR), **@BS** (BERTScore), **@GPT** (GPT_{eval}) and **@H** (Human_{eval}). FunAudioLLM* represents the result of providing the speaker identity as a text prompt to FunAudioLLM.

Methods	*Persona-Dialogue*						*Persona-Chat*					
	@B	**@R**	**@M**	**@BS**	**@GPT**	**@H**	**@B**	**@R**	**@M**	**@BS**	**@GPT**	**@H**
FunAudioLLM	2.1	7.2	9.2	82.1	1.7	1.9	1.48	7.89	10.02	82.91	2.12	2.22
FunAudioLLM*	4.3	15.2	15.2	85.1	3.1	3.2	2.44	11.29	12.11	84.33	3.99	3.44
Qwen-Audio	3.1	9.2	11.2	85.1	2.8	2.8	2.11	9.23	11.23	84.12	3.02	2.98
Salmonn	2.9	12.4	14.2	86.3	2.9	2.8	2.34	10.19	12.09	85.32	3.12	3.21
Qwen2-Audio	3.4	14.1	15.4	86.5	3.1	3.0	2.41	11.23	13.02	84.91	3.15	3.18
PAChat (ours)	**5.2**	**19.2**	**17.4**	**88.9**	**3.8**	**3.6**	**2.51**	**12.66**	**13.49**	**85.33**	**4.04**	**3.45**

logue, the character's information and a speech segment were input as prompts into the model. They perform zero-shot generation on both evaluation datasets.

5.2 Implementation Details

All audio data are resampled to 16 kHz for consistency. In the windowed Q-Former, we set K = 1 , resulting in a single trainable query, and use L = 17 , which corresponds to approximately 0.33 seconds per window. The models are trained for 30,000 steps with a batch size of 48 on eight A800 GPUs. CosyVoice2 (Du et al., 2024b) is used for speech generation during data construction, as well as for generating spoken responses after the text replies produced by PAChat, with all parameters kept at their default settings.

5.3 Main Results

5.3.1 Persona-Dialogue Results

As shown in Table 2, we evaluated the performance of spoken dialogue models on the Persona-Dialogue dataset. PAChat achieves the best results across all metrics, demonstrating significant improvements in personalized response generation and particularly winning favor with both GPT and human judges. Among the models, FunAudioLLM, which receives input through ASR transcription, performs poorly due to its inability to access speaker identity information from the audio. When provide with speaker identity as a text prompt, its performance improved considerably, however, its ability to understand user information and dialogue history still lagged behind PAChat. PAChat outperformed the best audio-input model by 0.7 points in the GPT evaluation and by 0.6 points in the human evaluation, indicating that the model not only understands the dialogue but is also able to generate responses conditioned on the persona. Other models also achieved relatively high scores due to their identity recognition capabilities, but based solely on evaluation metrics, their understanding remains at a basic level and does not meet the requirements for practical application.

5.3.2 Persona-Chat Results

PAChat, trained on the Persona-Dialogue dataset, naturally excels on the test set generated in the same manner. It is expected to have higher BLEU scores under similar training corpus conditions. For a fairer comparison, we conducted more experiments on the out-of-domain dataset: Persona-Chat, which focuses more on personalized chat results. As shown in the Table 2, PAChat still achieved the best performance. Comparing results on two distinct datasets, PAChat and other models have closer BLEU scores on Persona-Chat, indicating that PAChat does not have a significant advantage in text comparison. However, PAChat achieves excellent results in subjective evaluations, scoring 4.02 and 3.45, respectively, widening the gap with other models. This suggests that PAChat generates more reasonable responses that align with the persona of the dialogue user.

Training on large-scale synthetic data has endowed PAChat with a stronger understanding of multi-party conversations and a greater ability to capture user personas. Its robust performance across different tasks indicates that Persona-Dialogue promotes personalized responses in multi-party conversations.

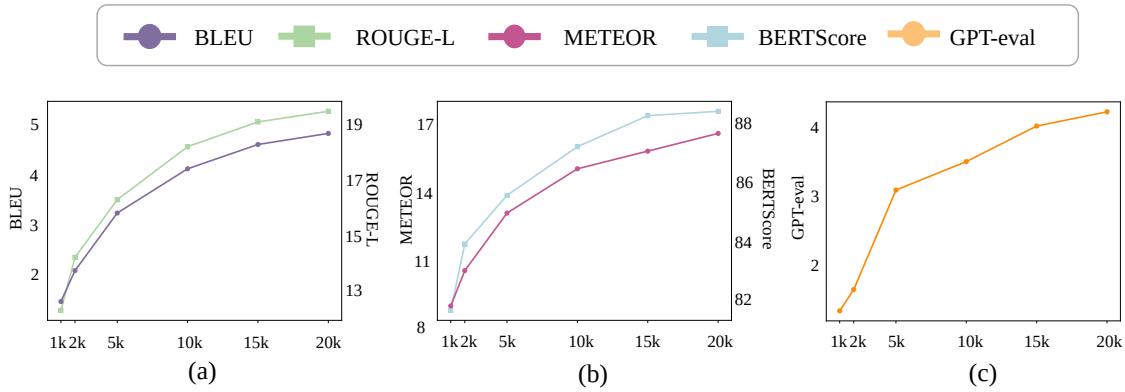

Figure 4: Performance comparison of different data scale on chat task.

5.4 Impact of data scale

Training data is essential for large models to learn and generalize in dialogue tasks, with data scale directly impacting performance. While large-scale audio data benefits model training as we previously discussed, the optimal amount needed remains unclear. To explore this, we conducted ablation experiments on training data size, as shown in Figure 4. Results indicate that with small datasets (1k, 2k), model performance is poor and it can hardly complete the task. As the data volume increases, the model's performance exhibits a steep-then-flat effect. With 20k data for training, the model's performance still improves, but gradually less noticeably. From a gradient perspective, continuing to increase the data scale can still enhance the model's performance, but excessive data, with a lot of text repetition, may increase the risk of overfitting and also raise costs. Therefore, 20k data appears to be a reasonable scale for personalized assistant tasks.

Table 3: Performance comparison of different settings on Persona-Dialogue. U-P stands for user profile and H-D stands history dialogue

setting		@B	@R	@M	@BS	@GPT
$U-P$	$H-D$					
✔		2.7	17.5	11.3	86.4	2.3
	✔	3.2	18.2	14.2	86.5	2.8
✔	✔	**5.2**	**19.2**	**17.4**	**88.9**	**3.8**

5.5 More dicussion about the task setting.

As the first work in the spoken dialogue domain aimed at generating personalized responses in multi-party conversations, we first need to ascertain the rationale of this approach. In real life, multi-party conversations are widespread, and a voice assistant participating in such conversations must have the ability to perceive information about the dialogue participants. The assistant needs to identify the speaker to understand the entire dialogue and generate targeted, more natural responses based on the speaker's information.

Thus, one might question whether our current modeling approach, which simultaneously considers user information and dialogue history, is reasonable if we aim to enhance the assistant's understanding of user information. It is worth noting that scenarios requiring responses based on both these types of information are not common. However, they complement each other in a dialogue. The dialogue topic and user information are like a query and key; we need the dialogue to know what to say next, and through the topic, we connect to the user's information to generate more targeted and reasonable responses, bringing satisfaction to the user. We tested scenarios where the generation does not include dialogue history or user information. From the results shown in Table 3, some model responses deviated significantly from the user's expected outcome, reflecting a decline in evaluation metrics. This corroborates from another perspective that this modeling approach helps enhance the model's information capture capabilities.

6 Conclusion

The exploration of speaker identity in multi-party dialogues has significant application potential but remains under-researched due to a lack of high-quality data. In an attempt to address this issue, we propose the utilization of large-scale synthetic data to augment models' ability to capture and understand multi-party dialogues and user-specific information. I n this paper, we introduce Persona-Dialogue, the first spoken dialogue dataset fea-

turing multi-party conversations and user profiles. Leveraging this dataset, we develop PAChat, a spoken dialogue framework that jointly models speech identity and semantic information to deliver personalized responses by interpreting user profiles and dialogue history. Extensive experiments show that our system can recognize speaker identity from speech and generate high-quality responses, providing valuable insights for advancing personalized voice assistants.

Limitations

From the experimental data on data scale, the performance gain brought by data scale has not yet reached its peak. The extent to which the scale of data no longer improves the performance of the model is still unknown, and the boundary of data volume needs to be further explored in the future.

Furthermore, the focus of this paper is on whether the text of the response is consistent with the speaker's identity information. Similarly, for different speakers, the model should also have the ability to reply in different tones, such as a gentle tone for children and a caring tone for the elderly, which we have not yet explored. We plan to introduce an end-to-end speech model, unify the speech generation module into the model, and achieve a more natural spoken dialogue system.

In addition, as AI assistants become increasingly human-like, it is crucial for them to acquire the ability to perceive speaking turns, similar to human participants in a conversation. This includes knowing when to speak and when to interrupt others. This paper focuses primarily on the textual quality of model responses and does not investigate these turn-taking abilities. If PAChat is to be applied in real-world dialogue scenarios in the future, equipping the model with turn-taking perception capabilities remains an important issue for further research.

Acknowledgements

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2025C02110), the Public Welfare Research Program of Ningbo (Grant No. 2024S062), and the Yongjiang Talent Project of Ningbo (Grant No. 2024A-161-G).

This research was also supported by Meituan.

Ethical Discussion

In this paper, strict control over the generation of dialogue content is implemented to avoid ethical risks. Moreover, this paper is solely for academic research and has not been used for commercial purposes. In the future, we will further explore how to reduce potential ethical risks in voice dialogue systems.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025a. AHabench: Benchmarking audio hallucinations in large audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, et al. 2025b. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, et al. 2025c. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*.

Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *Preprint*, arXiv:2303.05309.

Xize Cheng, Xiaoda Yang, Zehan Wang, Dongjie Fu, Rongjie Huang, Huadai Liu, Tao Jin, and Zhou Zhao. Noise-robust audio-visual speech-driven body language synthesis.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng

He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*.

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

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

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 preprint arXiv:2407.21783*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Dongjie Fu, Xize Cheng, Xiaoda Yang, Wang Hanting, Zhou Zhao, and Tao Jin. 2024. Boosting speech recognition robustness to modality-distortion with contrast-augmented prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3838–3847.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *Preprint*, arXiv:2306.08401.

Mark Hachman. 2019. The microsoft-amazon deal leaves cortana speakers with one advantage: Skype. *PCWorld Retrieved October*, 23.

Matthew B Hoy. 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Av-transpeech: Audio-visual robust speech-to-speech translation. *Preprint*, arXiv:2305.15403.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *Preprint*, arXiv:2312.10007.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*.

Linjun Li, Tao Jin, Xize Cheng, Ye Wang, Wang Lin, Rongjie Huang, and Zhou Zhao. 2023. Contrastive token-wise meta-learning for unseen performer visual temporal-aligned translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10993–11007.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. 2024. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. *arXiv preprint arXiv:2402.12786*.

OpenAI. 2024a. Chatgpt can now see, hear, and speak. *https://openai.com/index/chatgpt-can-now-see-hear-and-speak/*.

OpenAI. 2024b. Gpt-4o system card. *https://cdn.openai.com/gpt-4o-system-card.pdf*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve

Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Chaoyu Shi, Pengjie Ren, Dongjie Fu, Xin Xin, Shansong Yang, Fei Cai, Zhaochun Ren, and Zhumin Chen. 2024. Diversifying sequential recommendation with retrospective and prospective transformers. *ACM Transactions on Information Systems*, 42(5):1–37.

Tongyi SpeechTeam. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Zequn Xie, Haoming Ji, and Lingwei Meng. 2025. Dynamic uncertainty learning with noisy correspondence for text-based person search. *Preprint*, arXiv:2505.06566.

Jimin Xu, Tianbao Wang, Tao Jin, Shengyu Zhang, Dongjie Fu, Zhe Wang, Jiangjing Lyu, Chengfei Lv, Chaoyue Niu, Zhou Yu, et al. 2024. Mpod123: One image to 3d content generation using mask-enhanced progressive outline-to-detail optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10682–10692.

Weicai Yan, Wang Lin, Zirun Guo, Ye Wang, Fangming Feng, Xiaoda Yang, Zehan Wang, and Tao Jin. 2025. Diff-prompt: Diffusion-driven prompt generator with mask supervision. In *The Thirteenth International Conference on Learning Representations*.

Weicai Yan, Ye Wang, Wang Lin, Zirun Guo, Zhou Zhao, and Tao Jin. 2024. Low-rank prompt interaction for continual vision-language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8257–8266.

Xiaoda Yang, Xize Cheng, Minghui Fang, Hongshun Qiu, Yuhang Ma, JunYu Lu, Jiaqi Duan, Sihang Cai, Zehan Wang, Ruofan Hu, et al. 2025. Multimodal conditional retrieval with high controllability. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3577–3585.

Xiaoda Yang, Xize Cheng, Dongjie Fu, Minghui Fang, Jialung Zuo, Shengpeng Ji, Zhou Zhao, and Jin Tao. 2024. Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8149–8158.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2023b. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Preprint*, arXiv:1801.07243.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A More Details about Persona-Dialogue

A.1 Scenarios

In constructing the Persona-Dialogue dataset, we drew upon both everyday knowledge and insights from large language models to identify 21 potential application scenarios for voice assistants, as detailed in Table 5. We used these scenarios as categories to generate user profiles and dialogues. The proportion of data for each scenario, measured by the number of dialogues, is shown in Figure 5. Among these, scenarios such as Family Life, Company Meeting, and School Classroom are more common and exhibit a greater demand for voice assistants, therefore, we generated more dialogues for these scenarios. [1].

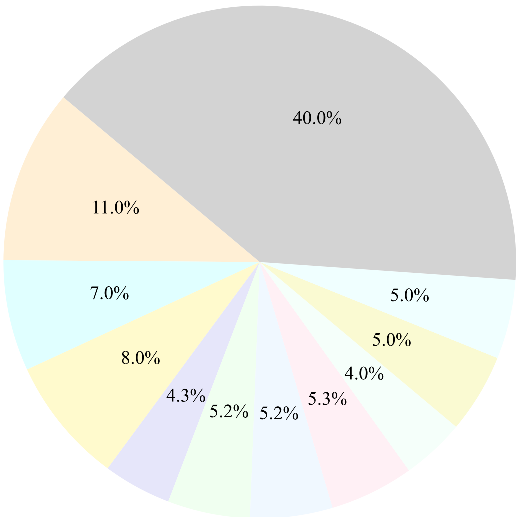

Figure 5: Proportion of Dialogue Numbers for Each Scenario.

A.2 Dialogue Topics

To avoid similarity among dialogues generated by the large language model within the same scenario, we not only randomly selected interlocutors but also created multiple dialogue topics for each scenario. This approach ensures diversity in user interactions and comprehensive coverage of various aspects within each scenario. Examples of scenarios and their corresponding dialogue topics are presented in Table 6.

[1]All data will be made publicly available after the paper is accepted.

A.3 User Profiles

We input scenario information into a large language model to generate possible dialogue users, along with their identities and descriptive information about their hobbies. For each scenario, we generate 3-4 groups of users according to different categories. Dialogue participants are selected from users within these groups. The distribution of the generated user information can be visualized using a word cloud, as shown in Figure 6.

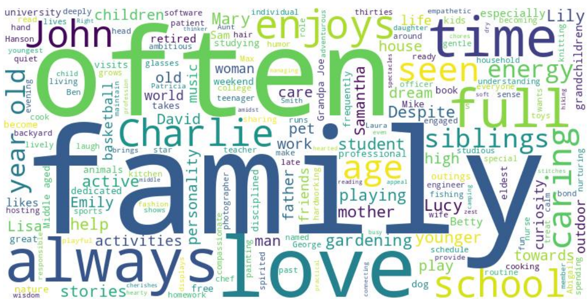

Figure 6: Word Cloud of User Profile Descriptions

A.4 Dataset Statistics

We have collected some statistics of Persona-Dialogue and presented them in the Table 7.

B More Details about Data Construction

All text generation tasks and evaluation tasks were completed by GPT-4-0613 (OpenAI et al., 2024)

B.1 Prompt Template for User Generation

The prompt template for user generation can be found in Table 8.

B.2 Prompt Template for Textual Dialogue Scripts Generation

In order to produce a wider variety of textual dialogue forms, we adopt two distinct approaches. In the first approach, direct dialogue generation, the LLM directly reads user information and, with reference to the scenario and dialogue topic, generates a complete dialogue segment. In the second approach, contextual interaction generation, the LLM first generates historical dialogue information between users, and then selects one user to generate an interaction turn with the voice assistant. The corresponding prompts for each approach can be found in Table 9. Note that the term "housekeeper" is used to create an assistant that expresses itself in a human-like manner, and this designation

may be replaced with different forms depending on the scenario.

B.3 Prompt Template for Textual Dialogue Verification

As a judge, the LLM is categorized into three types: faithfulness evaluation, relevance evaluation, and toxicity evaluation. The corresponding prompt templates can be found in Table 10.

C More Details about Evaluation

C.1 Evaluation Setting and Criteria

Here, we provide a detailed description of the evaluation procedures discussed in Section 5. For both GPT-based and human evaluations, the assessment criteria are the same, as shown in Table 11. For GPT-based evaluation, we used GPT-4-0613(OpenAI et al., 2024). For human evaluation, we invited five volunteers to score the responses, and the final score is the average of their ratings.

C.2 More Comprehensive Evaluation

As a dialogue model, PAChat ultimately outputs speech. However, since we employ an open-source TTS model and the current mainstream audio large language models primarily generate text outputs, the main experiments evaluate results from the perspective of text generation only. In this section, we further evaluate the model's end-to-end speech output performance. We use the MOS metric to assess the quality of the generated speech. Additionally, volunteers are asked to perform role-playing based on a given user profile or a custom profile. After freely interacting with each other, they communicate with the model to test its real-world ability to understand dialogue and user information. Each group consists of 3 to 5 participants, who jointly evaluate the model's responses and their scores are averaged. This process is repeated for 10 trials, and the overall average score is calculated. The evaluation criteria are consistent with those described in Section C.1. A comparison with the current open-source SOTA models, Qwen2-Audio + CosyVoice2, is shown in Table 4. The results indicate that PAChat achieves stronger real-world dialogue understanding capabilities, while the quality of generated speech is similar due to the use of the same TTS module.

Table 4: The Evaluation for Generated Speech.

Model	Mos	Human Score
Qwen2-Audio	3.88	2.81
PAChat	3.87	3.03

Table 5: Scenarios list

Family life	Company meeting	School classroom
Friends gathering	Shopping center	Travel group
Restaurant	Library/Bookstore	Stadium/Sports game
Concert/Music festival, Concert/Music festival, Art gallery, Gym, Park, Technology exhibition Sports club, Amusement park, Public transportation, Hospital, Pet shop, Cafe		

Table 6: Dialogue Topics in Main Scenarios

Family life	Upcoming Family Vacation Plans, Strategies for Challenging Homework Assignments, Discussion on Surprising News Headlines, Surprise Birthday Party Organization, Weekly Grocery Shopping List, Book Club Discussion, Debate on Household Chore Allocation, Living Room Redecoration Ideas, Adoption of a Healthy Lifestyle Plan, Family Movie Night Selection, Financial Budgeting Discussion, Recent Local Events Impact, Daily Work and School Event Sharing, Weekly Meal Planning and Preparation, Pet Ownership Challenges and Rewards, Bullying Incident Advice and Concerns, Electronics Usage Rules during School Week, Updates on After-School Sports Activities, Home Renovation Project Plans, Family Reunion Planning and Preparation
Company meeting	Annual Company Performance, Upcoming Marketing Strategies, New Product Launch, Future Investment Ideas, Quarterly Financial Report, Improving Customer Service, Expansion Plans, Employee Performance Reviews, Introduction of New Technologies, Charity/Community Involvement Opportunities, Remote Work Policies, Employee Benefits and Compensation, Workplace Safety Measures, Company Culture and Values, Company's Sustainability Initiatives, Fostering Innovation and Creativity, Team Building Activities, Upcoming Training Opportunities, Regulatory Compliance Issues, Competitive Analysis
School classroom	Mathematics Class Review, Cultural Event Planning, Political Subjects Debate, Literature Book Discussion, Music Album Release Discussion, Science Research Developments Conversation, Sports Game Strategies, Art Project Brainstorming, Computer Class Coding Strategies, Economics Class Financial Strategies, Science Experiment Discussion, Exam Preparation Tips, School Trip Recap, School Policy Changes Debate, TV Shows and Movies Discussion, Climate Change and Environmental Issues Talk, Homework Discussion, Fashion Trends Chat, School Charity Event Planning, Health and Wellness Tips Discussion
Friends gathering	Upcoming holiday plans, Recent movies or TV shows, Workplace stories and experiences, Personal fitness and health, Current political events, The latest technology trends, Food and restaurant recommendations, Home improvement and decoration ideas, Traveling experiences and dream destinations, Sports and recent games, Books and literature, New music or concerts, Mutual friends' life updates, Family and children, Pets and their funny stories, Childhood memories, Latest fashion trends, Cooking and recipes, Personal hobbies and interests, Wedding and special events preparations
Shopping center	Comparing prices between different stores, New fashion trends, Where to get the best deals, Recommendation for a good restaurant in the shopping center, Upcoming sales or events, The best time to shop to avoid crowds, Quality of customer service at a particular store, Recent purchases and their reviews, Newly opened stores in the shopping center, Current promotional offers or discounts, Comparing online shopping vs in-store shopping, Issues with parking space in the shopping center, Kids-friendly stores and amenities in the shopping center, Navigating through the shopping center, Health and safety measures in the shopping center due to COVID-19, Topics related to latest gadgets or electronics, Sharing experiences about a movie recently watched in the shopping center's cinema, Different payment methods accepted by stores, The effectiveness of a product they recently purchased, Whether or not a certain store accepts returns or exchanges
Travel group	Trip Itinerary Discussion, Personal Travel Experiences Sharing, Local Food and Cuisines Conversation, Best Time to Visit Debate, Travel Insurance and Safety Discussion, Cultural Heritage Thoughts Sharing, Weather-based Clothing Packing Discussion, Pick Up and Drop Locations and Timings Planning, Trip Expenses Division Planning, Local Attractions Decision, Destination City Transportation Options Discussion, Favorite Global Travel Destinations Conversation, Local Language Key Phrases Learning, Destination Country Travel Regulations and Requirements Discussion, Personal Preferences and Special Needs Discussion, Local Customs and Traditions Talk, Group Travel versus Solo Travel Pros and Cons Debate, Physical Demands of the Trip Chat, Photography and Site Specifications Conversation, Trip Extension or Next Trip Planning Discussion

Table 7: Statistics of Persona-Dialogue

Turns	Dialogues	Average Number of Turns per Dialogue	Duration(h)	Average Number of User Profiles per Scenarios
159933	21760	7.34	217	27

Table 8: Prompt Template for User Generation

Imagine a {DIALOGUE SCENARIO} scene with many individuals, and there may be a dialogue between two or more people, please construct 8 possible characters, and describe each character in 5 sentences, in the following format:

```
[{
  "name": ,
  "identity": ,
  "description": ["", "", "", "", ""]
}]
```

Table 9: Prompt Template for Dialogue Generation

Direct Dialogue Generation	Consider that in a {DIALOGUE SCENARIO} scenario, the following speaker and a human housekeeper are having a multi-person multi-turn conversation, and the interlocutor information is: {USER PROFILE} The topic of the conversation is {DIALOGUE TOPIC} please organize a conversation, the length of the dialogue is 4-10 rounds, each speaker's speech needs to match his or her identity, towards the end of the conversation, there needs to be a speaker to communicate with the housekeeper, it may be to chat or ask for advice, the housekeeper needs to reply to each person's identity characteristics and historical conversation data. Make sure that the whole conversation is natural and reasonable, and each speech should not be too long. Your response should be guaranteed in the following format: [{"speaker":<name>, "utterance": }, {"speaker":<name>, "utterance": }]
Contextual Interaction Generation	Consider that in a {DIALOGUE SCENARIO} scenario, the following speakers and a human housekeeper have already had a multi-person, multi-turn conversation. The interlocutor information is: {USER PROFILE}. The topic of the conversation is {DIALOGUE TOPIC}. The dialogue history is as follows: {DIALOGUE HISTORY} Now, please select one speaker from the participants to initiate a new conversational turn with the housekeeper. The speaker can either chat with the housekeeper or ask for advice. The housekeeper needs to reply according to the speaker's identity characteristics and the historical conversation data. Make sure the interaction is natural and reasonable, and each utterance should not be too long. Your response should be guaranteed in the following format: [{"speaker": <name>, "utterance": }, {"speaker": "assistant", "utterance": }]

Table 10: Prompt Template for Textual Dialogue Verification

Faithfulness Evaluation	Given users profiles respectively, does the following conversation between the users contradict either of their profiles? {USER PROFILE} {DIALOGUE}
Relevance Evaluation	Which one of Conversation 1 and Conversation 2 between users {USER PROFILE}? Conversation 1: {DIALOGUE 1} Conversation 2:{DIALOGUE 2}
Toxicity Evaluation	Is this conversation toxic? {DIALOGUE}

Table 11: The Evaluation Criteria for Human and GPT-4.

1	The response is completely irrelevant to the context, showing no understanding of user information or historical dialogue.
2	The response is partially relevant to the context, but unnatural and awkward. There is limited use of user information or historical dialogue.
3	The response is partially relevant to the context and demonstrates some understanding of user information or historical dialogue, but overall lacks fluency and naturalness.
4	The response is highly relevant to the context, expressed in a natural and fluent manner, and accurately references and utilizes user attributes or historical information.
5	The response is fully relevant to the context, expressed very fluently and naturally, makes full and skillful use of user attributes and historical dialogue knowledge, and demonstrates a strong understanding of the conversation.