# MultiActor-Audiobook: Zero-Shot Audiobook Generation with Faces and Voices of Multiple Speakers

*Kyeongman Park*[1], *Seongho Joo*[1], *Kyomin Jung*[1]

[1]Seoul National University, South Korea,

zzangmane@snu.ac.kr, seonghojoo@snu.ac.kr, kjung@snu.ac.kr

## Abstract

We introduce MultiActor-Audiobook, a zero-shot approach for generating audiobooks that automatically produces consistent, expressive, and speaker-appropriate prosody, including intonation and emotion. Previous audiobook systems have several limitations: they require users to manually configure the speaker's prosody, read each sentence with a monotonic tone compared to voice actors, or rely on costly training. However, our MultiActor-Audiobook addresses these issues by introducing two novel processes: (1) MSP (**Multimodal Speaker Persona Generation**) and (2) LSI (**LLM-based Script Instruction Generation**). With these two processes, MultiActor-Audiobook can generate more emotionally expressive audiobooks with a consistent speaker prosody without additional training. We compare our system with commercial products, through human and MLLM evaluations, achieving competitive results. Furthermore, we demonstrate the effectiveness of MSP and LSI through ablation studies.

**Index Terms**: human-computer interaction, audiobook generation, speech synthesis, face-to-voice

## 1. Introduction

Stories are complex literary works characterized by a wide range of emotions and multiple speakers. Transforming such texts into speech, as in the case of audiobooks, resembles human actors performing a script, where they convey the most fitting prosody and emotions for each scene [1, 2, 3]. For instance, reading a novel in a monotone voice would be unnatural; instead, the emotion, tone, and pitch must dynamically adapt to the context [4, 5, 6, 7, 8]. Furthermore, stories feature not only narrators but also diverse characters, each requiring a voice that reflects their distinct characteristics such as personality and physical traits like face [9, 10, 11, 12, 13, 14, 15, 16] to achieve natural-sounding dialogue.

Traditional audiobook generation systems have generally relied on two approaches. The first approach involves collecting extensive, high-cost datasets, such as over 60 hours of professional voice actor recordings [1], to predict dynamic prosody for each word embedding[1, 17, 18]. However, this method is limited to the specific domain in which the data was collected and incurs significant costs for data acquisition and model training. For example, [1] can only produce natural speech within the scope of traditional Chinese oral art forms.

The second approach is manual annotating, where humans meticulously select cadence, emotions, pitches, and tones for each speaker and each sentences [3, 19, 20]. This annotated data is then used with pre-trained prompt-TTS systems [2, 5, 6, 7, 12, 21]. While this method reduces costs for data collection and model training, it still requires substantial human labor and time to produce a single audiobook. For instance, NarrativePlay [19] involved manually selecting from over 3,000 predefined prosody patterns for each character.

To sum up, previous works often need costly data collection, model training, or manual annotations. To address these challenges, we propose **MultiActor-Audiobook** : *Zero-Shot Audiobook Generation with Faces and Voices of Multiple Speakers*. MultiActor-Audiobook introduces two innovative processes: (1) **Multimodal Speaker Persona Generation** and (2) **LLM-Based Script Instruction Generation**. Although both processes operate in a fully automated and zero-shot manner, MultiActor-Audiobook produces more emotionally expressive and speaker-appropriate audiobooks.

During Multimodal Speaker Persona Generation, we create a multimodal persona for each speaker. Specifically, LLM first identifies all speaker entities within the novel, including both dialogue and narration. It then extracts descriptive features from the text related to each character. Based on this textual persona information, we utilize a text2image model to generate an AI-generated face image that visually represents the character. Using this face image and its corresponding caption, we employ a pretrained Face-to-Voice system [14] to produce a unique voice sample that reflects each character's distinctive prosody. By leveraging visual information for each character, we can generate a more characteristic and fitting voice, and we can maintain speaker consistency throughout the story by anchoring each speaker to their unique voice samples.

Then the LLM-based Script Instruction Generation process employs GPT-4o [22] to create dynamic instructions for each sentence in the novel. These instructions mainly include emotional cues, tone, and pitch. During generation, we not only provide the target sentence but also give the surrounding context and each character's persona information to LLM. By providing this additional information, the LLM can predict more appropriate and continual emotional expression for each sentence. With these detailed emotional and contextual script instructions, the TTS model can read each sentence more naturally and expressively, leading to a better listening experience.

To evaluate our system's performance, human annotators and the MLLM (Multi-modal Large Language Model) compare it with baselines, including powerful commercial products. As a result, our system achieves comparable MOS scores to the baselines in human evaluation and demonstrates an average 0.225-point improvement in MLLM evaluation. Additionally, our ablation studies validate the effectiveness of each process by showing consistent improvements across all metrics, highlighting its respective contribution to overall performance.

## 2. MultiActor-Audiobook

MultiActor-Audiobook aims to create audiobooks that deliver speaker-aligned voice and natural emotional expressions. To generate each speaker-algined voices in the story, it performs **Multimodal Speaker Persona Generation**, and to ensure natural emotional expression, it utilizes **LLM-Based Script Instruction Generation**.

### 2.1. Multimodal Speaker Persona Generation

In this process, the LLM identifies each character in the story and generates audio samples and facial images that match each character. The process includes three steps: (1) Extract all speakers from the story and create captions, (2) Create Facial personas for each character, and (3) Create audio personas for each character.

#### *2.1.1. Extract all speakers and their characteristics from story*

We first input the entire story into the LLM and extract all distinct characters with speaking lines, including the narrator. The LLM also makes captions about the physical appearance and personality traits of each character. Even if the story does not explicitly describe their appearance or intonation, we guide the LLM to infer these details through reasonable imagination based on indirect descriptions in the narrative. An example prompt is as follows :

```
Analyze the following story and
extract:
1. The narrator: provide a
detailed description of the
narrator's external features that
could be used to create a portrait.
If the narrator is a third-person
narrator, analyze the main character
in the story and treat them as the
narrator.
2. A list of characters with
speaking lines (dialogue) and
for each character, provide their
name or role in the story and a
description of their external
features that could be used to
create a portrait in one sentence.
If sufficient information is not
provided, use reasonable imagination
to infer their features based on
their role and context.)
```

#### *2.1.2. Create facial personas*

Using the captions generated in the previous step, we create AI-generated face images with a State-of-the-Art text-to-image model, the Stable Diffusion Model. Specifically, considering that the primary dataset used to train the backbone TTS model is sourced from real human speech scenes, we employ a model specialized in generating photorealistic human images[1]. Additionally, we filter out any samples that do not depict a human face.

[1]The model is able to freely download from `https://huggingface.co/SG161222/Realistic_Vision_V2.0`

#### *2.1.3. Create audio personas*

Using the face images and captions generated in the previous step, we create audio samples that match each character. Specifically, we use the Masked Generation feature of FleSpeech to synthesize short voice samples for each character using only face images and captions.

### 2.2. LLM-Based Script Instruction Generation

In this process, we generate individual instructions for every sentence in the story. This process consists of two steps: (1) Identifying the speaker of each sentence, (2) Generating appropriate emotional instructions for each sentence.

#### *2.2.1. Identify the speaker of each sentence*

In this process, we input the entire story into the LLM to identify which character speaks each sentence. Since many dialogue lines are spoken by characters other than the narrator, accurately distinguishing the speaker of each line is crucial for the reader's experience. Once this process is complete, we match each sentence with the correct speaker's ID and persona information. An example prompt is as follows :

```
You are analyzing a story. Based
on the full story and the current
sentence, determine who is speaking
the current sentence. Use the
following rules: 1. Refer to
the full context of the story to
identify the speaker. 2. Only
assign the speaker from the provided
list of characters below. 3.
Consider dialogue attribution
markers (e.g., 'said Alice') and
indirect clues in the story. 4. If
the sentence seems to be narration,
return the main character's name.
```

#### *2.2.2. Create Text Description of Each Sentence*

In this process, we input the entire story, the target sentence, and the speaker information of the target sentence into the LLM to generate appropriate narration descriptions (e.g., *Use a calm and reassuring tone.*). We instruct the model to avoid descriptions unrelated to speaking and to ensure that the emotional flow transitions naturally from the previous sentence to prevent abrupt emotional shifts. An example prompt is as follows:

```
[Full Story] [Target Sentence] Based
on the full context of the story
and the current sentence, provide
a single, concise instruction
on how the sentence should be
read aloud emotionally. Make the
tone, pitch, pacing, and emotional
delivery needed for a professional
narration. If applicable, make a
smooth transition from the previous
emotion, and consider the emotions
between the characters.
```

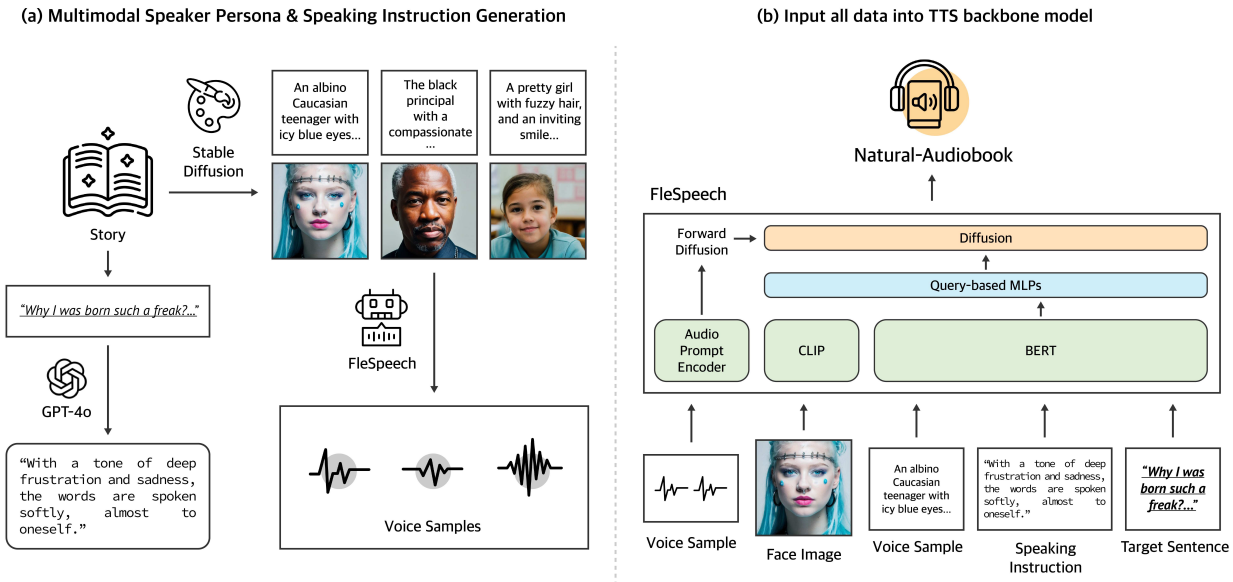


Figure 1: ***The MultiActor-Audiobook***. *At the left side of figure (a), we perform Multimodal Speaker Persona Generation to create each speaker's AI-generated face images and voice samples, and LLM-based Script Instruction Generation to annotate every sentence's speaking instructions. At the right side of figure (b), we input all the multimodal input to our backbone TTS model, the FleSpeech, to generate speaker-aligned emotional audiobook.*

### 2.3. Integration All Input Data

Finally, we input all the data generated from the previous two processes—face images, face captions, audio samples, text descriptions, and target texts—into FleSpeech [14], the backbone multimodal TTS system, to complete the audiobook on a sentence-by-sentence basis. FleSpeech includes a unified multimodal prompt encoder, which is trained to map representations from various modalities using query-based MLPs and a diffusion process, allowing us to input text, audio, and visual data all at once. The face images, face captions, and audio samples remain fixed for each character within a single story sample generation, while only the target text and text description are updated for each sentence.

## 3. Experiments

### 3.1. Experimental Setup

#### *3.1.1. Story Dataset*

We use the story dataset ReedsyPrompts[23] in this paper. The ReedsyPrompts contains appropriate length of stories, allowing for diverse and distinctive speakers to appear within a single story. We use only partial samples of the total training samples since the heavy generation costs. In 12 samples, each story has an average of 4.3 speakers and consists of an average of 175 sentences. The average audio length is 749.56 seconds.

#### *3.1.2. MultiActor-Audiobook Implementation Details*

For generating face images, we use the State-of-the-Art photorealistic Stable Diffusion model, the *SG161222/Realistic_Vision_V2.0*. We also utilize gpt-4o-2024-08-06 for the LLM capabilities, and FleSpeech served as the backbone TTS model, with the same parameter values as reported in the original paper. None of the models perform any additional training, and for the inference we use NVIDIA RTX A5000 GPU settings.

#### *3.1.3. Baselines*

The baselines are ElevenLabs, FakeyouTTS, F5-TTS, w/o Persona, and w/o Instruction. ElevenLabs and FakeyouTTS are well-known commercial TTS systems that do not support the input of speaking descriptions or multimodal persona information for speakers. Instead, they offer predefined character intonations that users can manually select. Therefore, we choose the most suitable character intonation for each story. We generate each audiobook using the same stories as our system. Note that ElevenLabs supports audiobook generation in a life-like and emotionally rich mode.

F5-TTS allows specifying the speaking voice style of each character but does not utilize visual persona information or speaking descriptions. We use each character's audio persona to determine the voice style for each sentence.

w/o `MSP` is a version that uses the same backbone TTS model of ours but masks the Multimodal Speaker Personas which include face images, face captions, and audio samples, relying solely on the target text and text descriptions.

w/o `LSI` refers to a version that utilizes all input data except for the LLM-based Script Instruction.

### 3.2. Main Experiments Results

We conduct a Mean Opinion Score (MOS) listening test with five human evaluators and an MLLM evaluator. Character-Voice Consistency measures how well the audiobook's voices align with the characters' personality in the story, including factors such as gender, pitch, speed, volume, and intonation. MOS-Q evaluates the overall audio quality, including aspects like

| | **Ours** | w/o MSP | w/o LSI | ElevenLabs | FakeYou | F5-TTS |
|---|---|---|---|---|---|---|
| Char-Con | $2.9_{\pm 0.12}$ / **3.9** | $2.6_{\pm 0.17}$ / 3.4 | $2.9_{\pm 0.10}$ / 3.8 | **4.2**$_{\pm 0.60}$ / 3.7 | $3.2_{\pm 0.31}$ / 3.2 | $2.6_{\pm 0.21}$ / 3.7 |
| MOS-Q | $2.4_{\pm 0.29}$ / 3.3 | $2.2_{\pm 0.21}$ / 3.1 | $2.8_{\pm 0.20}$ / 3.4 | **4.6**$_{\pm 0.27}$ / **4.0** | $3.6_{\pm 0.25}$ / 3.7 | $3.4_{\pm 0.25}$ / 3.7 |
| MOS-E | $2.9_{\pm 0.21}$ / **4.6** | $2.5_{\pm 0.19}$ / 4.0 | $2.6_{\pm 0.12}$ / 3.7 | **4.2**$_{\pm 0.52}$ / 4.2 | $2.5_{\pm 0.27}$ / 4.4 | $1.8_{\pm 0.31}$ / 3.9 |
| MOS-S | $2.6_{\pm 0.24}$ / **3.4** | $2.4_{\pm 0.15}$ / 3.3 | $2.5_{\pm 0.12}$ / 3.1 | **4.3**$_{\pm 0.44}$ / 2.6 | $2.4_{\pm 0.23}$ / 3.2 | $2.0_{\pm 0.30}$ / 2.6 |

Table 1: ***Human / MLLM average scores of*** *Char-Consistency(Char-Con), MOS-Quality(MOS-Q), MOS-Emotion(MOS-E), and MOS-Speaker(MOS-S) across our system and baselines. For each score, the value to the left of the slash (/) represents the average MOS scores of human annotators, while the value to the right represents the MOS scores of the gpt-4o-audio-preview. We bold the best values and underline the second best values.*

| | **Speaker Similarity** | **Turning Points** |
|---|---|---|
| **Ours** | **51.334** | **146885.1** |
| ElevenLabs | 40.473 | 125309.6 |
| FakeYou | 48.640 | 108087.5 |
| F5-TTS | 51.332 | 57737.7 |

Table 2: *The speaker embedding similarities and number of turning points of our system and baselines. We bold the best values and underline the second best values.*

clarity, high-frequency, and naturalness. MOS-E assesses how appropriately emotions are conveyed in each sentence, while MOS-S measures how accurately the speaker is identified for each sentence. For MLLM evaluations, we used the same question provided to human evaluators and employed the audio-to-text QA feature of gpt-4o-audio-preview.

The quantitative analysis involves the analysis of speaker embedding similarity and the number of pitch turning points. To evaluate the voice consistency of the audiobooks, we measure the speaker embedding every 10 seconds in a sample using an open-source speech embedding model[2], and calculate the average similarity. To assess emotional expressiveness, we count the number of pitch turning points. A higher speaker embedding similarity indicates better maintenance of voice consistency within the sample, while a greater number of pitch turning points suggests stronger emotional fluency.

#### *3.2.1. Character-Voice Consistency*

As shown in Table 1, ElevenLabs achieves the highest Character-Consistency score in human evaluation, followed by FakeYou. In the MLLM evaluation, our system scores the highest, while w/o LSI achieves the second-highest score. Although ElevenLabs' audiobook generation does not support automatic selection of an optimistic narrator voice, its substantial additional training cost may contribute to its superior performance compared to other baselines.

For FakeYou, we manually selected the best-matched famous actors' voices, such as Morgan Freeman, for each story, which likely led to strong human evaluator preferences in the Character-Consistency metric. However, the MLLM scores indicate that our system also performs well compared to other baselines, thanks to Multimodal Speaker Persona Generation.

[2] https://github.com/douglas125/SpeechIdentity?tab=readme-ov-file

#### *3.2.2. MOS-Q, MOS-E, MOS-S*

As shown in Table 1, ElevenLabs achieves the highest MOS-Q scores in both human and MLLM evaluations, followed by FakeYou. Compared to the zero-shot nature of our system, there commercial products have a significant advantage in this metric due to their substantial additional training costs.

For MOS-E and MOS-S scores, our system ranks second-highest in human evaluation and achieves the highest scores in MLLM evaluation. Its superior performance compared to w/o MSP and w/o LSI demonstrates that both of our novel processes—the Multimodal Speaker Persona Generation and LLM-based Script Instruction Generation—are crucial for conveying appropriate emotional expressions.

Since w/o LSI outperforms w/o MSP across all metrics, particularly in Char-Con and MOS-Q scores, we conclude that multimodal inputs play a central role in audiobook quality, primarily by enhancing the consistency of speaker identities.

#### *3.2.3. Quantitative Analysis*

As shown in Table 2, our system achieves the highest Speaker Similarity and number of pitch turning points. From this, we can conclude that our system maintains voice consistency across each sample very well while also producing various levels of emotional expressions, such as joy, love, and laughter, which result in frequent pitch variations.

## 4. Conclusion

We introduce MultiActor-Audiobook, a zero-shot system for generating emotionally expressive and speaker-appropriate audiobook without extra training or manual annotations. Our approach leverages two key processes: **Multimodal Speaker Persona Generation**, and **LLM-Based Script Instruction Generation**. Experimental results show that our system achieves mostly best or second highest scores in human and MLLM evaluations, and the best results in quantitative analysis, while eliminating the need for costly data collection and manual labeling.

## 5. Limitation

MultiActor-Audiobook exhibits lower quality than some commercial systems, mainly due to the backbone TTS model, FleSpeech, which is trained on a smaller, less specialized dataset. Large-scale audiobook datasets by professional voice actors could improve performance. Additionally, since FleSpeech was trained on specific face samples (e.g., TED lecturers), it struggles with unseen AI-generated faces. Future multimodal TTS systems with more diverse face samples could enhance MultiActor-Audiobook's performance.

# 6. References


[1] S. Ge, C. Xuan, R. Song, C. Zou, W. Liu, and J. Zhou, "From text to sound: A preliminary study on retrieving sound effects to radio stories," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 865–868.

[2] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.

[3] I. Ramli, N. Seman, N. Ardi, and N. Jamil, "Rule-based storytelling text-to-speech (tts) synthesis," in *MATEC Web of Conferences*, vol. 77. EDP Sciences, 2016, p. 04003.

[4] X. Wei, J. Jia, X. Li, Z. Wu, and Z. Wang, "A discourse-level multi-scale prosodic model for fine-grained emotion analysis," *arXiv preprint arXiv:2309.11849*, 2023.

[5] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song *et al.*, "Prompttts 2: Describing and generating voices with text prompt," *arXiv preprint arXiv:2309.02285*, 2023.

[7] R. Shimizu, R. Yamamoto, M. Kawamura, Y. Shirahata, H. Doi, T. Komatsu, and K. Tachibana, "Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 672–12 676.

[8] D. Diatlova and V. Shutov, "Emospeech: Guiding fastspeech2 towards emotional text to speech," *arXiv preprint arXiv:2307.00024*, 2023.

[9] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," *arXiv preprint arXiv:2403.16973*, 2024.

[10] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[11] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 682–689.

[12] W. Guan, Y. Li, T. Li, H. Huang, F. Wang, J. Lin, L. Huang, L. Li, and Q. Hong, "Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 117–18 125.

[13] J. Lee, Y. Oh, I. Hwang, and K. Lee, "Hear your face: Face-based voice conversion with f0 estimation," *arXiv preprint arXiv:2408.09802*, 2024.

[14] H. Li, Y. Li, X. Wang, J. Hu, Q. Xie, S. Yang, and L. Xie, "Flespeech: Flexibly controllable speech generation with various prompts," *arXiv preprint arXiv:2501.04644*, 2025.

[15] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, and Y. Ding, "Transformer-based multimodal information fusion for facial expression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2428–2437.

[16] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, Y. Li, Z. Long, H. Gao, K. Li *et al.*, "Vita-1.5: Towards gpt-4o level real-time vision and speech interaction," *arXiv preprint arXiv:2501.01957*, 2025.

[17] J. Bae, S. Jeong, S. Kang, N. Han, J.-Y. Lee, H. Kim, and T. Kim, "Sound of story: Multi-modal storytelling with audio," *arXiv preprint arXiv:2310.19264*, 2023.

[18] C. Pethe, B. Pham, F. D. Childress, Y. Yin, and S. Skiena, "Prosody analysis of audiobooks," *arXiv preprint arXiv:2310.06930*, 2023.

[19] R. Zhao, W. Zhang, J. Li, L. Zhu, Y. Li, Y. He, and L. Gui, "Narrativeplay: Interactive narrative understanding," *arXiv preprint arXiv:2310.01459*, 2023.

[20] H. Zhang, Y. Guo, S. Liu, X. Chen, and K. Yu, "Expressive tts driven by natural language prompts using few human annotations," *arXiv preprint arXiv:2311.01260*, 2023.

[21] M. Lee, E. Park, and S. Hong, "Fvtts: Face based voice synthesis for text-to-speech," *Proc. Interspeech 2024*, pp. 4953–4957, 2024.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[23] K. Park, N. Yang, and K. Jung, "Longstory: Coherent, complete and length controlled long story generation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2024, pp. 184–196.