Dean&Francis
ISSN 2959-6157

Research and Prospects of Multimodal Technology of AIGC in NPC Dialogue Generation

Yuanman Li[1, *]

[1]Facility of Applied Science, Macao Polytechnic University, Macao, China
*Corresponding author: P2215808@mpu.edu.mo

Abstract:

In today's technologically advanced world, in-game NPC dialogue with players is a crucial component of modern game production, directly impacting player immersion and engagement. The creation of large language models such as GPT-4V and VIMA, along with other advances in multimodal learning methods, has significantly enhanced the ability of game NPC systems to recognize and respond to diverse signals and complex scenarios. This paper examines the current state of the art in multimodal dialogue production in great detail, covering common methods, widely used datasets, and evaluation criteria. This work proposes solutions to these issues, including leveraging lightweight model structures, developing effective methods for aligning data from various modalities, and creating improved multimodal datasets specifically designed for gaming environments. This paper aims to inform the next generation of NPC dialogue systems using multimodal AIGC by examining the strengths and weaknesses of current approaches. This will help in-game dialogue become more meaningful, characters more aware of their surroundings, and interactions feel more realistic.

Keywords: Multimodal Learning; NPC Dialogue Generation; AI-Generated Content.

1. Introduction

Generating dialogue for non-player characters (NPCs) is crucial for virtual worlds and role-playing games. Rule-based dialogue has shown promising results in the past, but has shown shortcomings over time, such as not adapting well, not remembering context, and not allowing for personalization [1]. The emergence of artificial intelligence generated content (AIGC), particularly multimodal approaches, has filled these gaps. These approaches have made dialogue systems more flexible and dynamic, resulting in a better player experience [2, 3].

NPC dialogue is no longer a static design. Instead, interactive approaches are becoming the next design trend. Moreover, it grows in a continuous „perception-understanding-generation" cycle that is influenced by player behavior, clues in the environment, and the story [4]. Multimodal AIGC allows NPCs

ISSN 2959-6157

to understand emotions and context and respond appropriately by placing dialogue, visual, and textual inputs into a semantic space [5, 6]. This makes it easier to apply content such as contextualized training and adaptive story guidance, and makes the experience more immersive and engaging.

There has been a recent shift towards this design. Flamingo achieves effective few-shot multimodal learning [3], and the CLIP model shows strong vision-language alignment [2]. PaLM-E is all about embodied reasoning in diverse sorts of media [6]. OpenAI's GPT-4V, on the other hand, combines text and visuals into one framework, which makes NPC communication even better [5]. Even with these advancements, there are still challenges, like not having enough domain-specific datasets, not being able to undertake real-time reasoning, and fairness issues [7]. This paper builds on previous progress and describes popular methods, datasets, and evaluation criteria for creating NPC dialogues. It also discusses possible ways to improve them in the future.

2. Overview of Mainstream Methods

2.1 Classification and Comparison of Different Types of Methods

According to the development of NPC dialogue generation technology, mainstream methods can be divided into the following three categories (table 1).

Table 1. Types of mainstream methods

Category	Features	Representative Methods	Advantages	Limitations	Applicable Scenarios
Rule-based Methods	Rely on manually designed rules or dialogue trees, unable to handle complex scenarios.	Finite State Machines (FSM), AIML (Artificial Intelligence Markup Language).	Strong controllability, suitable for fixed scenarios.	Limited and repetitive output, poor scalability.	NPC dialogues in early linear plot games.
Statistical Methods	Generate dialogue using statistical learning techniques, emphasizing context modeling.	Hidden Markov Models (HMM), N-gram based methods.	More flexible in content generation.	Limited ability to understand context, cannot handle long-term dependencies.	Small- to medium-scale dialogue systems.
Deep Learning Methods	Use neural networks (especially deep learning models) to generate diverse dialogues.	Seq2Seq, Transformer, large language models (e.g., GPT, BERT).	High generation quality, strong contextual understanding.	Require large-scale data, may produce bias.	Dialogue generation in open-world and highly interactive games.

2.2 The Progressive and Evolutionary Relationship between Methods of the Same Type

Demonstrate the model's evolution from infrastructure to modern multimodal technology.

2.2.1 Seq2Seq Model (2014, NIPS)

In principle, the encoder compresses input semantics, and the decoder generates responses through autoregression, often in conjunction with attention and teacher-forced training. Its advantages lie in its end-to-end implementation simplicity, but representation bottlenecks limit its ability to model long-term context, world knowledge, and personality consistency. In NPC scenarios, it exhibits a „short-term" focus and short-term memory, making it difficult to adapt to task chains and environmental changes, and prone to templated responses, laying the groundwork for the subsequent introduction of stronger contextual mechanisms [8].

2.2.2 The Transformer model (2017, NIPS)

Replaces recurrent structures with multi-head self-attention, significantly improving parallelization capabilities and long-range dependency modeling, enabling more stable context tracking and coreference resolution [9]. When used with NPCs, it can better integrate plot context and dialogue history, but without explicit constraints, it can still lead to repetition and hallucinations. Therefore, it needs to be combined with retrieval, constraint decoding, or a rule layer to strike a balance between open-domain fluency and controllable game logic.

2.2.3 GPT Series Models (2018-2024, NeurIPS/ICLR)

GPT2: This model acquires general language priors

through self-supervised learning from massive amounts of text, enabling the generation of more natural continuations and responses with few samples [10]. Its contribution to NPCs lies in significantly improving language fluency and stylistic diversity. However, it lacks explicit alignment between task objectives and world settings, making it prone to „sounding good but not doing good." It requires retrieval augmentation (RAG) or a plot state machine to inject correct information and establish boundaries.

GPT-3: The leap in parameter and data scale enables powerful few-shot generalization, enabling the simulation of character tone and situational pragmatics with minimal prompts [11]. It can quickly adapt to different character cards and task styles in NPC scenarios, but inference costs are higher and the risk of hallucinations remains. Engineering efforts require a combination of memory (long-term character profiles/plot blackboards), tool calls, and cached inference to ensure real-time and consistency.

InstructGPT/RLHF: Using human-feedback reinforcement learning, model behavior is aligned to instructions and value constraints, significantly improving usability and security [12]. For NPCs, this means more robust character boundaries and content filtering (such as the suppression of violent/discriminatory language), and facilitates the implementation of a „plot safety strategy." The downside is that the migration of alignment data beyond the domain is still limited, and alignment samples and review criteria that fit the game's worldview need to be established.

2.3 Current Developments in Multimodal Input

Current developments combine and process information from multiple modalities, such as a combination of text, visual, and voice, to produce more realistic and engaging NPC dialogues. The rapid development of multimodal input technology has also expanded the application fields of NPC dialogue generation, enabling more effective player-NPC interaction in virtual worlds and highly immersive game environments, significantly improving the player's experience.

2.3.1 Text and Visual Integration

In the examination of text-visual fusion, conventional methodologies often employ a two-stage architecture: visual encoder-alignment adapter-language decoder. To start, a visual encoder (like ViT or CLIP [2]) changes the picture into a set of visual features. An alignment module, like linear mapping, MLP adapter, or Q-Former/Resampler, then turns these features into "pseudo-words." After that, these features are mixed with text words and sent to a big language model that uses a single attention mechanism to do modeling and autoregressive generation. Different research uses this concept in different ways.

People don't know all the nuances behind GPT-4V, but it's generally known that the image is broken up into visual words and put directly into the context. This lets the decoder model both text and image in the same attention space [5]. Flamingo [3], on the other hand, freezes the visual encoder and language model and puts a Perceiver Resampler and a gated cross-attention module between them. This makes it easy to add a limited number of condensed visual elements to the language stream, which shows that the model can generalize well even with few examples. The visual extension version of Llama 2, like LLaVA or BLIP-2 [13, 14], is another example. A projection adapter is frequently used with this kind of approach. LoRA technology can be used to quickly align many modes by freezing or partially freezing the Llama 2 settings and just fine-tuning the mapping layer or adapter.

2.3.2 Text and Speech Fusion

In terms of the integration of text and speech, the mainstream method usually adopts a "three-stage" process, namely the speech front-end, language model, and speech synthesis module. At first, the original audio is converted into a log-Merkel spectrum and fed into the speech encoder to generate the corresponding acoustic markers. Subsequently, the language model analyzes and understands these markers, completing text understanding and dialogue generation, and finally converting the output into natural speech through text-to-speech (TTS) technology. In order to make the interaction closer to human communication, emotion recognition modules are often introduced in research, which can use acoustic characteristics such as pitch, energy, and duration, and can also use special emotion encoders to provide prompts or control signals for language models, so as to make the generated dialogue more vivid and expressive. In terms of implementation, Whisper is one of the most representative solutions. It is based on an end-to-end Transformer encoder-decoder architecture that maps the Mercer spectrum directly into a sequence of subwords, with strong performance, and can handle long speech and multilingual inputs, so it is often used as a front-end module for speech-to-text systems. Another typical route is to integrate voice capabilities into Llama 2 [13]. There are two main implementation methods: one is a pipeline solution, which is first completed by Whisper and then handed over to Llama 2 for dialogue modeling; The second is a multi-modal alignment scheme, which directly generates speech markers with the help of audio encoders and jointly inputs them with text tags into Llama 2 for joint modeling. At the output side, these methods are usually combined with vocoders such as

ISSN 2959-6157

VITS or HiFi-GAN to achieve low-latency speech synthesis, thus building a complete interactive closed loop of "listening-understanding-speaking".

2.3.3 Full-modal Fusion of Text, Video and Voice

In the study of full-modal fusion, common methods often adopt the overall architecture of „multiple encoders in parallel - unified word space - single decoder". Specifically, image or video data is usually processed by a visual encoder, and speech is extracted through an audio encoder. The two generate modal words respectively; these modal information are then mapped to a shared embedding space through a projection layer or adapter and combined with text words to input the same Transformer decoder. In this unified framework, the model can jointly model multimodal information and output text (such as conversation content), speech (with the help of TTS headers), or action instructions according to demand during the decoding stage, thereby achieving highly integrated generation.

In practical applications, the extended form of GPT-4V can be considered a typical solution [5]. It can process visual and language information simultaneously in the same context, while the audio part is usually connected through a front-end encoder or an external ASR/TTS module to support cross-modal dialogue and understanding. Another representative exploration is PaLM-E [6]. This model is aimed at „embodied intelligence" scenarios. It maps images, text, and sensor or state information into a unified word sequence, and a single large Transformer performs conditional generation and reasoning. This design enables the model to integrate „what it sees, hears, and knows" into executable dialogues and action plans in complex environments, demonstrating stronger situational understanding and task adaptation capabilities.

2.4 Challenges and Prospects of Multimodal Input

Although multimodal technologies have made significant progress in NPC dialogue generation, some challenges remain. The following are the main issues and corresponding solutions:

2.4.1 Efficiency Issues in Modal Fusion

In NPC dialogue generation, multimodal models must simultaneously process multiple input sources, including text, images, and speech. This results in significant computational overhead and inference speeds that struggle to meet the real-time demands of highly interactive scenarios. To address this issue, researchers have proposed various optimization solutions. Lightweight methods (such as LoRA and Low-Rank Adaptation) significantly reduce computational costs by reducing trainable parameters. Efficient cross-modal attention mechanisms help models quickly focus on relevant information between modalities, avoiding inefficient computations and, to a certain extent, alleviating latency issues.

2.4.2 Insufficient Information Consistency Between Modalities

When conflicts arise between voice, visual, and text modal inputs, models often struggle to generate coherent and believable dialogue. For example, a player's voice and visual cues may give different instructions, leading to logical errors in NPC responses. To improve consistency, researchers have introduced cross-modal alignment algorithms, such as CLIP (Contrastive Language-Image Pre-training), which strengthen the semantic connections between different modalities through contrastive learning. Furthermore, introducing large-scale, diverse multimodal data during training can effectively enhance the model's robustness to conflicting information, thereby improving the stability of dialogue generation.

2.4.3 Insufficient Diversity and Scale of Datasets

At the moment, most multimodal datasets that are routinely utilized (such VQA and COCO) focus on general vision-language tasks and don't have many high-quality resources that are specific to gaming. This constraint makes it harder to use models in real-life situations. To solve this problem, academics and industry are working together to make multimodal datasets for gaming, like ones that combine screenshots of games, character dialogue, and player speech. Research is also employing generative adversarial networks (GANs) to add to training data, which makes the samples more diverse and covers a wider range of topics. This makes NPC dialogue generation perform better.

2.4.4 Bias and fairness issues

When people process visual and audio information, multimodal models may inherit or amplify social biases. When dealing with tasks related to gender or racial tone recognition, the model may exhibit discriminatory biases. This problem is more prominent in NPC game scenarios because it will reduce the player's interactive experience and directly affect the fairness and credibility of the entire system. To address this risk, researchers have proposed several improvement strategies to improve this risk, such as adding bias test datasets such as BanStereoSet in training and evaluation, and using „thinking chain" prompts to more effectively limit potential biases in the generation process. This enhances the fairness and reliability of NPC dialogue output, reduces bias, and helps to bring a positive gaming experience to players.

3. Technological Advantages, Bottlenecks, and Future Directions

At present, generative technology has been able to significantly improve the dynamics and interactivity of non-player character (NPC) dialogues, such as with the help of advanced multimodal models such as GPT-4V and PaLM-E, NPCs can achieve a higher degree of freedom in virtual worlds and immersive game environments. However, this technology still faces key bottlenecks such as low modal fusion efficiency and limited coverage of high-quality datasets. Looking ahead, research will focus on developing lightweight multimodal frameworks to optimize real-time inference performance while expanding high-quality, diverse multimodal datasets for game scenarios. In addition, the contextual memory mechanism needs to be strengthened to support more natural and coherent long-distance dialogue generation capabilities, so that players have a better experience.

4. Conclusion

This article takes a closer look at the effects of multimodal technology on NPC dialogue. Natural interaction, contextual understanding, and dynamic generation have made significant strides in this field, evolving from initial rule-based and statistical methods to contemporary innovations employing deep learning and multimodal frameworks. This article showcases the unique advantages of multimodal dialogue systems in making NPCs more immersive and adaptable. The survey also reveals some of the issues and challenges that need to be addressed. Limited real-time inference efficiency, challenges in maintaining consistency across modalities, lack of sufficient data resources in game scenarios, and potential bias and fairness issues are major obstacles to the continued development of multimodal NPC technology. This article provides several possible solutions that are expected to help the technology overcome the problems it currently faces. In summary, the creation of multimodal NPC dialogues is advancing rapidly, showing great application potential, but also bringing new challenges to data resources and algorithm design.

References

[1] Ammanabrolu P, et al. Learning to speak and act in a fantasy text adventure game. ACL. 2021.

[2] Radford A, et al. Learning transferable visual models from natural language supervision. ICML. 2021.

[3] Alayrac J B, et al. Flamingo: a visual language model for few-shot learning. NeurIPS. 2022.

[4] Bubeck S, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv: 2303.12712. 2023.

[5] OpenAI. GPT-4 technical report. arXiv.2303.08774. 2023.

[6] Driess D, et al. PaLM-E: an embodied multimodal language model. ICML. 2023.

[7] Hendricks L A, et al. Measuring bias in multimodal models. CVPR. 2021.

[8] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. NeurIPS. 2014.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. NeurIPS. 2017.

[10] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Technical Report. 2019.

[11] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. NeurIPS. 2020.

[12] Ouyang L, et al. Training language models to follow instructions with human feedback. NeurIPS. 2022.

[13] Touvron H, et al. LLaMA: open and efficient foundation language models. arXiv.2302.13971. 2023.

[14] Li Y, et al. LLaVA: large language and vision assistant. arXiv.2304.08485. 2023.