Video2Roleplay: A Multimodal Dataset and Framework for Video-Guided Role-playing Agents

Anonymous ACL submission

Abstract

Role-playing agents (RPAs) have attracted growing interest for their ability to simulate immersive and interactive characters. However, existing approaches primarily focus on static role profiles, overlooking the dynamic perceptual abilities inherent to humans. To bridge this gap, we introduce the concept of dynamic role profiles by incorporating video modality into RPAs. To support this, we construct **Role-playing-Video60k**, a large-scale, high-quality dataset comprising 60k videos and 700k corresponding dialogues. Based on this dataset, we develop a comprehensive RPA framework that combines adaptive temporal sampling with both dynamic and static role profile representations. Specifically, the dynamic profile is created by adaptively sampling video frames and feeding them to the LLM in temporal order, while the static profile consists of (1) character dialogues from training videos during fine-tuning, and (2) a summary context from the input video during inference. This joint integration enables RPAs to generate greater responses. Furthermore, we propose a robust evaluation method covering eight metrics. Experimental results demonstrate the effectiveness of our framework, highlighting the importance of dynamic role profiles in developing RPAs.

1 Introduction

Recent advancements in large language models (LLMs) (Zhao et al., 2023) have spurred significant research interest in RPAs (Chen et al., 2024b), which simulate interactive characters through the integration of diverse modality data to create realistic user experiences. However, real-world human perception is inherently multifaceted and dynamic. The current reliance primarily on static modalities like text and images limits the ability of these agents to fully satisfy the growing demand for highly immersive and expressive role-playing experiences.

Figure 1: Examples illustrating our RPAs' performance compared to general baselines. More examples are provided in Appendix.

Video, as a powerful multimodal medium (Song et al., 2024; Yang et al., 2024; Lian et al., 2024; Mou et al., 2024), offers a rich array of dynamic details related to characters, such as emotional states, physical actions, scene transitions, and narrative experiences. This information is highly valuable for pioneering dynamic role-playing profiles. For example, lives showcase character dynamic motions in authentic scenarios. Vlogs and role documentaries capture individuals' expressions and daily activities, effectively conveying complex emotions and personality traits for detailed character portrayals. Consequently, integrating the video modality into RPAs equips agents with more comprehensive and detailed dynamic information, improving role-playing performance and user engagement.

Currently, despite some promising results of the

existing work (Dai et al., 2024; Wang et al., 2025b) in the field of RPAs, there is still a lack of exploration in data resources and effective methods of video modality. How to effectively integrate video modality information with existing static modalities and leverage its unique dynamic information advantages for RPAs remains a challenging problem. Furthermore, the long length of some videos often introduces considerable redundant information, leading to high memory resource consumption and inefficient video information representation.

To fill these gaps, this study introduces the concept of dynamic role-playing to integrate video modality into the RPAs for the first time, constructs a large-scale and high-quality dataset tailored to the requirements of dynamic profile representation in RPAs, and proposes a comprehensive framework that effectively incorporates video modality with static modalities.

Specifically, we construct a large-scale and high-quality dataset sourced from various social media platforms like Xiaohongshu, Douyin, Weibo, and Bilibili. The dataset comprises daily lives, lifestyle vlogs, and personal documentaries from diverse groups, accompanied by corresponding video captions and related dialogues, providing rich resources for the development of RPAs. Additionally, we propose a novel multimodal RPA framework that combines adaptive temporal sampling with both dynamic and static role profile representations. To construct the dynamic role profile, we adaptively sample video frames based on their duration and provide them to the LLM in their original order. In parallel, the static role profile captures character information with two main components: (1) character-specific dialogues related to training videos, which are used to guide the base model during fine-tuning, and (2) a high-level summary generated from the input video during inference, which provides a concise but accurate description of the video scene and character presentation. By integrating both dynamic and static role profiles, our framework enables RPAs to generate responses that are highly consistent with the character's identity and the narrative context.

Moreover, we design a series of evaluation metrics and experiments to validate the effectiveness of our framework. Extensive experiments demonstrate the superior performance of our framework on RPAs. It establishes a compelling trade-off between parameter size and overall performance while achieving SOTA for human-likeness.

Table 1: **Comparison between different role-playing datasets.** Our work is the first role-playing dataset that introduces the video.

Dataset	Dialogues	Video
ChatHaruhi (Li et al., 2023a)	54,726	✗
Character-LLM (Shao et al., 2023)	14,300	✗
RoleLLM (Wang et al., 2024a)	168.1k	✗
CharacterGLM (Zhou et al., 2024)	1,034	✗
Character100 (Wang et al., 2024b)	10,609	✗
DITTO (Lu et al., 2024)	7,186	✗
CharacterEval (Tu et al., 2024)	1785	✗
LifeChoice (Xu et al., 2024)	1,462	✗
RolePersonality (Ran et al., 2024)	87,345	✗
MMRole (Dai et al., 2024)	14,346	✗
CharacterBench (Zhou et al., 2025)	13,162	✗
OpenCharacter (Wang et al., 2025a)	306k	✗
RoleMRC (Lu et al., 2025)	39.3k	✗
CoSER (Wang et al., 2025b)	29,798	✗
Role-playing-Video60k(Ours)	700k	✓

In summary, our contributions are threefold:

- We are the first to integrate the video modality into RPAs, introducing the concept of dynamic role-playing and enabling the creation of rich dynamic role profiles.
- We construct a large-scale and high-quality dataset for the development of RPAs, including $60k$ videos and $700k$ dialogues across various categories, durations, and scenarios.
- We develop a novel and comprehensive RPA framework that integrates adaptive temporal sampling with both dynamic and static role profiles. Extensive experiments and analyses demonstrate its outstanding performance.

2 Related Work

2.1 Static Role Playing

ChatHaruhi (Li et al., 2023a) provides a dataset of over $54k$ simulated dialogues for 32 characters spanning Chinese, English, and anime. CharacterGLM (Zhou et al., 2024) allows for personalizing a diverse range of agent personas and social agents through customizable attributes and behaviors. CharacterLLM (Shao et al., 2023) builds a dataset detailing specific character experiences, then fine-tunes a base model with the dataset to achieve target character portrayal. RoleLLM (Wang et al., 2024a) improves LLM role-playing via a multi-component framework (e.g.,

role profile construction, role-GPT, role-bench). Ditto (Lu et al., 2024) introduces a self-alignment method to enhance LLM role-playing capabilities through knowledge augmentation and dialogue simulation. MMrole (Dai et al., 2024)introduces the concept of multimodal role-playing agents and offers a comprehensive framework for their development and evaluation. RoleMRC (Lu et al., 2025) provides a fine-grained composite benchmark for role-playing and instruction-following, revealing activation patterns linked to these distinct abilities. CoSER (Wang et al., 2025b) provides a dataset comprising 29,798 authentic conversations and comprehensive data from 771 renowned books and proposes a given-circumstance acting method for training and evaluating role-playing LLMs.

2.2 Video Understanding

GPT4Video (Wang et al., 2024e) proposes a unified framework for video understanding and generation via pre-trained model integration and develops a simple text-only fine-tuning method for instruction following and safety alignment. LongVLM (Weng et al., 2024) introduces a VideoLLM for long-term video understanding, achieving affordability via segment decomposition, feature extraction, token merging, and global semantics. Video-LLaVA (Lin et al., 2024) maps visual signals to the language feature space to achieve unified visual representations, introducing a method for aligning features prior to projection. VideoAgent (Wang et al., 2024c) proposes an agent-based system that iteratively extracts and compiles key information for question answering, using vision-language models for visual translation and retrieval. VidRecap (Islam et al., 2024) proposes a hierarchical caption generation method that creates CLIP captions, segment descriptions, and video summaries, trained using a coarse-to-fine approach to learn the structure of video. LongVU (Shen et al., 2024) preserves frame information for lengthy videos by compressing tokens based on similarity and selecting relevant visual tokens for text queries. InternVideo2.5 (Wang et al., 2025c) introduces a length-adaptive token approach to process videos, integrating visual perception with MLLM for fine-grained analysis.

2.3 Multimodal Large Language Model

CLIP (Radford et al., 2021) achieves cross-modal understanding and unified representation by applying contrastive learning to unlabeled image-text pairs, eliminating the need for task-specific data annotation. Flamingo (Alayrac et al., 2022) inserts new gated cross-attention layers into the LLMs to inject visual features and pre-trains the new layers on billions of image-text pairs. Emu (Sun et al., 2024) extends the approach of Flamingo (Alayrac et al., 2022) by integrating additional modalities to model generation and the corresponding training corpus. BLIP-2 (Li et al., 2023b) introduces Q-Former for visual and linguistic representation learning, achieving zero-shot image-text generation and strong performance on visual language tasks with more efficient parameterization. InternVL (Chen et al., 2024c) presents the first alignment of a large-scale vision encoder with LLMs and introduces a progressive image-text alignment strategy, enabling efficient training of large-scale vision-language foundation models. InstructBLIP (Dai et al., 2023) introduces an instruction-aware feature extraction method for vision-language instruction tuning, significantly enhancing multimodal model performance. LLaVA-NeXT (Li et al., 2024) enhances visual detail capture via improved input image resolution and refines its data mix through adapted visual instructions.

Figure 2: The video types and examples of our dataset.

3 Dataset Curation

To ensure richness and diversity of video content, we curate a large-scale and high-quality video dataset sourced from various social media platforms, including Xiaohongshu, Douyin, Weibo,

and Bilibili. This dataset comprises daily lives, lifestyle vlogs, and personal documentaries from diverse groups, accompanied by corresponding captions and related dialogues, providing comprehensive resources for the development of RPAs. More details can be found in the Appendix A.2.

3.1 Video Type

We divide the dataset videos into three categories by their content and duration, as shown in Fig. 2.

Live. This type of video captures a few seconds before and after a specific moment, focusing on close-up details that highlight the character's related motions. Notably, unlike static images, which freeze a single frame, these videos offer a continuous narrative by incorporating both preceding and following frames. This dynamic continuity enables a deeper understanding of the role-related motion in the scene, reducing the bias of isolated moments.

Vlog. Unlike traditional blogs, this category of video uses dynamic visuals to document daily life, typically capturing daily moments from individuals. Their vivid filming style, distinct character portrayals, and strong self-expression lend them a unique individuality, effectively conveying positive character profiles to LLMs.

Documentary. This type of video documents the life journeys or period-specific experiences of individuals, often featuring frequent scene transitions. Drawing from life footage that includes various personal events, these videos construct a cohesive storyline that presents the deeper character traits.

3.2 Video Caption

Video captions serve as a critical bridge linking textual information with visual content. Therefore, ensuring these captions are rich, diverse, and comprehensive is essential for subsequent effective integration. Our preliminary strategy for annotating the videos entailed per-second frame descriptions aggregated by an LLM into a complete caption. However, this approach requires substantial resource consumption and costs, and is further constrained by the input size of the LLM, preventing full frame processing. Thus, we design the staged annotation approach illustrated in Fig. 3 which generates captions in two distinct phases, detailed below.

Uniform Segmentation Sampling. To effectively capture the diverse scenes within each video while optimizing annotation efficiency, we employ a temporal segmentation strategy. Each video is uniformly divided into multiple segments based on its length. From each segment, a single frame is sampled as its representative. Based on case results and manual comparisons, we divide each video into 64 segments, thereby achieving a trade-off between representational quality and annotation efficiency.

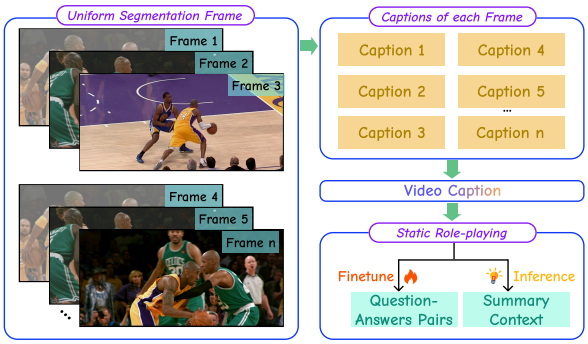

Figure 3: The illustration of video caption. We uniformly divide the video into segments and annotate each segment with a frame description, then we summarize these descriptions as a video caption and employ it during the fine-tuning and inference phase. **Notably,** video captions are utilized distinctly across the two phases, originating from different videos and serving distinct purposes. Specifically, during the **fine-tuning phase**, captions are employed to generate question-answer pairs. In contrast, during the **inference phase**, captions are used to develop the role context.

Segment-Based Annotation and Summarization. For each representative frame selected from the segments, we use an LLM to generate a detailed description. Following this, we introduce a summary agent, which takes the descriptions of the frames in video order as context and produces a comprehensive video summary using Chain-of-Thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) (Brown et al., 2020).

3.3 Dialogue Generation and Filtering

Given a detailed video caption, we use an LLM to generate question-answer pairs for each video. Following existing video works (Chen et al., 2024a; Zhang et al., 2024; Liu et al., 2024), the instruction prompt includes: (1) The role definition of the video scene. (2) The detailed video description. (3) In-context examples that include question-answer pairs from the real comments in social media. (4) Instruction order about the specific generation of question-answer pairs. Also, we instruct GPT-4o to return None if it is unable to generate question-answer pairs in the case of a bad context. Additionally, to improve the quality of the generated question-answer pairs, we filter out the gener-

Figure 4: Our framework consists of three key components: (1) **Adaptive Temporal Sampling**: This module adaptively samples video frames based on the input video's length. (2) **Dynamic Role Profile Representation**: This module constructs dynamic role profiles from the sampled video frame. (3) **Static Role Profile Representation**: This module extracts static role profiles from dialogue and summary contexts. Further, we propose a comprehensive evaluation approach incorporating eight metrics.

ated question-answer pairs by discarding answers that begin with phrases like "As an AI language model," "does not present," "does not show," "does not demonstrate," or other errors.

4 Methodology

In this section, we propose the overall framework as illustrated in Fig. 4, which can be divided into three key parts: (1) **Adaptive Temporal Sampling**: We adapt an adaptive temporal sampling strategy tailored to the various lengths of video input. (2) **Dynamic Role Profile Representation**: We represent the samplings from the video as a dynamic role profile. (3) **Static Role Profile Representation**: We represent the static role information from the dialogues obtained from Sec. 3.3 and the summary context of the input video. We provide a detailed explanation of these processes as follows.

4.1 Adaptive Temporal Sampling

For video $V \in \mathbb{R}^{T\times H\times W\times 3}$, we implement a context-aware sampling mechanism that adapts to the video length, forming the video frame sequence $V' \in \mathbb{R}^{t\times H\times W\times 3}$.

For shorter videos like lives (0-5 seconds), where fine-grained motion details are essential, we employ dense temporal sampling by capturing every frame of the video.

For medium-length videos like vlogs (5 seconds - 10 minutes), where the coherence of events is more important, we apply sparse sampling, taking one frame per 5 seconds uniformly.

In contrast, for longer videos like documentaries (longer than 10 minutes) that focus on event-level understanding, we sample frames representing key scene events. The specific keyframe sampling process is detailed below.

- *Step 1.* Collect candidate frames by uniformly sampling one frame per second from the long video. Compute the frame difference $D(i,j) = \sum_{k=1}^{M} |I_i^k - I_j^k|$, where I_i^k is the k-th pixel value of the i-th frame, and M is the total number of pixels. A frame is added to the candidate set $C = \{f_1, f_2, f_3, ..., f_m\}$ if its difference score $D(i-1, i)$ exceeds a threshold T.

- *Step 2.* Divide the candidate set C into G uniform groups, each containing $\frac{|C|}{G}$ frames. For each group g, compute the intra-group variation $V(g) = \max_{i,j \in g} D(i,j)$. Select the frame with the maximum $V(g)$ as the representative frame for each group, forming a refined candidate set $C' = \{f_1, f_2, f_3, ..., f_n\}$.

- *Step 3.* For adjacent frames i and j, calculate the similarity $S(i,j) = Clip(i,j)$ using CLIP. Merge frame j into frame i if $S(i,j) > \tau$, where τ is a similarity threshold. Repeat until all adjacent frames have $S(i,j) \leq \tau$, resulting in the final key frame set $C_k = \{f_1, f_2, f_3, ..., f_k\}$.

Due to restrictions on computational resources, we cap the maximum number of frame samples at 128.

4.2 Dynamic Role Profile Representation

Based on the visual content $V' \in \mathbb{R}^{t \times H \times W \times 3}$ sampled in Sec. 4.1, we generate special tokens *<image>* for each video frame and present them as a visual prefix, maintaining the original order of the input video. Each frame is transformed and stacked into a tensor, representing the relevant dynamic role profile through a continuous frame sequence.

4.3 Static Role Profile Representation

In this section, we fine-tune the base model to learn the static role profile from the dialogue related to the video scenes and characters, as discussed in Sec. 3.3. During the inference stage, we also employ a summary agent to capture the global information of the video. This agent uses a CoT process to generate a video summary, which is presented as static character context to guide role-playing.

Character Dialogue. RPAs are designed to simulate characters and engage in immersive dialogues with users. While these agents acquire dynamic role information from the process described in Sec. 4.2, our approach further integrates the static role information through role-related dialogues. The approach presented in Sec. 3.3 ensures the training dialogues are centered on and informed by the roles and scenes within the videos. The integration can be achieved through supervised fine-tuning (SFT), with its specific data format shown in Fig. 5.

Video Summary. After the SFT of the base model, we introduce a summary agent to capture global information of the video during the inference phase. For the input video with a length L, we divide it into successive n segments uniformly and caption the corresponding description for all segments, $D = \{d_1, d_2, d_3, ..., d_k\}, k = L/n$. Additionally, we introduce a summary agent with a CoT approach to summarize these descriptions D into an entire video summary S, which is used as the context to guide the LLM in performing role-playing with the ICL approach.

Fine-tune Data Format

```
{
    "id": 0,
    "video": "video.mp4",
    "conversations": [
        {"from": "human", "value": "<video>\nuser input"},
        {"from": "gpt", "value": "assistant output"},
        {"from": "human", "value": "user input"},
        {"from": "gpt", "value": "assistant output"}
    ]
}
```

Figure 5: The example of fine-tune data format, the special token ***<video>*** indicates the position where the video is inserted.

5 Experiment

5.1 Experimental Settings

For the experimental dataset, we randomly shuffle our dataset into 57k training sets and 3k inference sets. Our test samples consist of 328 questions that are manually selected from social media platforms. To minimize the bias introduced by the model itself during evaluation, we employ GPT-4o and GPT-o3-mini as LLM evaluators, averaging their assessments for a more balanced perspective. Additionally, to enhance the reliability of our results, we set the API temperature to 0.0 and conduct three rounds of judgments per sample, averaging the results to further reduce variance.

5.2 Evaluation Metric

Following the existing works (Dai et al., 2024; Tu et al., 2024; Zhou et al., 2025; Wang et al., 2024d), we evaluate the performance of RPAs including eight metrics. The specific metrics are as follows.

Character Consistency. Do the responses maintain character consistency throughout interactions, rather than exhibiting random behavioral changes?

Knowledge Hallucination. Do the responses prioritize factual grounding over fake assumptions when virtual knowledge conflicts with reality?

Utterance Fluency. Do the responses maintain

Table 2: Main results of our framework and baselines.

Model	LLM-based Metrics ↑							
	Cons.	**Hall.**	**Adh.**	**Flu.**	**Hum.**	**Acc.**	**Ton.**	**Avg.**
General Baselines								
Yi-Large	74.38	68.40	61.91	84.15	51.23	63.58	66.41	67.15
GPT3.5 Turbo	68.75	66.22	57.34	84.55	52.16	58.61	59.75	63.91
GPT-4-Turbo	75.73	70.76	60.34	86.38	54.67	63.08	63.62	67.79
GPT-4.1	79.31	74.56	71.91	88.05	58.27	68.89	71.45	73.21
GPT-4o	76.74	71.42	68.77	86.31	49.94	64.87	65.98	69.14
GPT-4o Mini	74.73	67.27	62.15	85.91	46.90	60.13	62.39	65.64
GPT-o4 Mini	81.12	74.12	74.17	85.03	49.85	66.94	66.51	71.11
GPT-o1	78.48	74.44	72.98	87.57	**62.93**	69.86	71.88	74.02
Gemini-2.5-Pro-Exp	82.12	75.48	**80.85**	88.11	62.70	69.14	**78.26**	**76.67**
Claude3.5 Sonnet	80.87	74.33	60.27	85.23	49.32	64.53	69.22	69.11
Claude3.7 Sonnet-thinking	**83.66**	78.31	77.93	86.80	59.19	**71.73**	78.03	76.52
Deepseek-V3	72.38	67.95	65.22	86.04	43.09	60.29	66.28	65.89
Deepseek-R1	80.68	**78.69**	77.13	86.58	47.86	69.47	74.33	73.53
Qwen-max	81.89	70.75	66.17	**88.44**	57.56	64.29	71.43	71.50
Doubao-1.5-pro	71.19	70.74	65.11	83.29	46.12	59.94	57.15	64.79
Baichuan-4-Turbo	73.03	68.75	56.33	83.46	51.33	60.22	61.34	64.92
Role-playing Expertise Models								
CharGLM4	71.80	69.51	60.45	86.22	52.87	59.88	61.31	66.01
Ernie-char-8k	72.18	65.13	58.26	84.68	54.28	56.09	**63.48**	64.87
Qwen-plus-character	**76.52**	70.30	63.11	87.57	54.29	60.28	62.76	67.83
InternVL2.5-8B w/ Video SFT (Ours)	72.17	**74.38**	**70.52**	**87.93**	**69.98**	**69.26**	61.75	**72.28**

grammatical correctness and exhibit smooth readability in utterance expression?

Tone Consistency. Do the responses match the character's typical tone patterns and catchphrases?

Instruction Adherence. Do the responses adhere to instructions by strictly keeping in character without added explanation?

Response Accuracy. Do the responses accurately address the question or appropriately engage in a conversation based on the context?

Human Likeness. Do the responses convey a sense of human rather than presenting an AI style?

Video-Text Relevance.[1] Do the responses closely correlate with the content depicted in the video?

Notably, we conduct a user study to evaluate the model's performance with human judgment. Further details are provided in the Appendix A.3.

5.3 Baseline

We select sixteen well-known advanced LLMs as general baselines: (1) `Yi-Large`, (2) `GPT-3.5-Turbo`, (3) `GPT-4-Turbo`, (4) `GPT-4.1`, (5) `GPT-4o`, (6) `GPT-4o Mini`, (7) `GPT-o4 Mini`, (8) `GPT-o1`, (9) `Gemini2.5-Pro-Exp`, (10) `Claude 3.5 Sonnet`, (11) `Claude 3.7 Sonnet-thinking`, (12) `Deepseek-V3`, (13) `Deepseek-R1`, (14) `Qwen-max`, (15) `Doubao-1.5-Pro`, (16) `Baichuan-4-Turbo`.

We also use three role-playing expertise LLMs as robust baselines: (1) `CharGLM-4`, (2) `Erine-char-8k`, (3) `Qwen-plus-character`.

5.4 Comparative Studies

As shown in Tab. 2, we report the performance of two types of baselines and our framework on LLM-based metrics. Analyzing the generated responses, we observe that, in contrast to untrained RPAs, fine-tuned RPAs tend to generate shorter and more concise responses without additional explanation. These responses more closely align with human conversational patterns, rather than exhibiting the heavily formatted and AI styles often found in the outputs of untrained RPAs. The comprehensive experimental results demonstrate that our framework achieves superior performance in RPAs, realizing a compelling trade-off between parameter size and effectiveness. Our model demonstrates comparable performance across all metrics against baselines with significantly larger parameters, and even presents SOTA on the human-likeness metric.

[1] Due to the limitations of direct video input for most baselines, we evaluate this metric only during the ablation study.

Table 3: The ablation studies of the video SFT and the summary context.

Method	Cons.	Hall.	Adh.	Flu.	Hum.	Acc.	Ton.	Rel.	Avg.
W/ Video Inference + W/ Summary Context									
8B w/ Video SFT	**72.17**	**74.38**	**70.52**	**87.93**	**69.98**	**69.26**	**61.75**	**23.43**	**66.18**
8B w/ Image SFT	72.06	68.34	68.45	84.75	67.61	63.77	59.67	21.25	63.24
8B w/ Text SFT	69.41	67.56	68.09	82.37	65.17	60.41	58.74	14.20	60.74
8B w/o SFT	53.12	51.56	37.43	71.40	32.46	44.48	36.25	11.61	42.29
W/ Video SFT + W/ Video Inference									
8B w/ Summary Context	**72.17**	**74.38**	**70.52**	**87.93**	**69.98**	**69.26**	**61.75**	**23.43**	**66.18**
8B w/o Summary Context	70.38	72.46	69.66	85.74	68.51	65.89	61.03	19.37	64.13

5.5 Analysis

Large-Scale and High-Quality Dataset. We curate a large-scale dataset comprising 60k videos and 700k conversations from various groups, featuring synthetic dialogues grounded in real-world social media scenarios. This large-scale, high-quality dataset is designed to improve the performance of RPAs. To validate its effectiveness, we compare our framework with the base model InternVL2.5-8B. As shown in Tab. 3, our framework significantly outperforms the base model across all metrics. The base model presents poor performance on RPA tasks without any SFT method, underscoring the necessity of SFT. Notably, benefiting from our dataset's highly human-like style, text-only or both image and text SFT approaches demonstrate comparably strong enhancements in human-likeness and instruction adherence.

Video Modality Ablation. To verify the impact of the video modality on the performance of RPAs, we conduct ablation experiments comparing our framework to the two approaches without video modality: 1) a model fine-tuned only on dialogues. 2) a model fine-tuned on a single frame randomly sampled from videos and dialogues. As shown in Tab. 3, our framework, fine-tuned on our dataset with video modality, significantly outperforms models fine-tuned only on dialogues or on both images and dialogues. We observe that introducing the video modality leads to substantial improvements in almost all metrics. These improvements demonstrate the significant potential of integrating the video modality for developing RPAs that are more expressive and consistent, thus contributing to a more engaging and immersive user experience. Additionally, despite some improvements in video-text relevance from incorporating video modality, the score still remains low, suggesting significant potential for further development of RPAs with more effective video modality integration.

Summary Context Ablation. To evaluate the effect of the summary context derived from video captions on the performance of RPAs, we conduct an ablation study. Specifically, we replace the summary context with the full long descriptions for all sampled frames. As shown in Tab. 3, the model with summary context presents better performance. Notably, despite providing the LLM with more detailed information, the full long descriptions did not improve performance on any metric, including video-text relevance. In contrast, compared to lengthy contexts, the summary context generated under the CoT guidance is more concise and effectively captures the key points of the long description. This allows the model to have a more accurate understanding of the input video, thus improving the performance of RPAs.

6 Conclusion

In this paper, we propose the concept of dynamic role-playing for the first time by extending the RPAs with a video modality. Moreover, we construct a large-scale, high-quality video dataset covering various types, lengths, and roles for the development of RPAs. Furthermore, we design a novel and comprehensive framework that integrates adaptive temporal sampling with dynamic and static role profile representation. Extensive experimental results and analyses demonstrate the great effectiveness of our framework. Our work can advance the progress of RPAs, providing a novel perspective for this field. In the future, we believe that engaging roles constructed from dynamic and static perspectives can benefit the various social applications and introduce a promising connection with digital humans, leading to a better user interaction.

Limitations

Due to limitations in computational resources, we are unable to employ either a larger-scale base model or a more densely sampled frame acquisition approach to explore further results. Additionally, we only utilize lora fine-tuning method, rather than the full parameter fine-tuning approach. Thus there is still room for improvement in the parameter size and fine-tuning method.

Ethics Statements

Our model, fine-tuned on Role-playing-Video60k, may only have minimum safety alignment, so it will probably generate toxic and harmful content under induction. Therefore, the dataset and LLM are only for research purposes and should be carefully aligned in terms of safety in the future.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. Llm-grounded video diffusion models. In *ICLR*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. *arXiv preprint arXiv:2502.11387*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of

all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840.

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. 2024. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 33:1122–1135.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Emu: Generative pretraining in multimodality. In *ICLR*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14743–14777.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024b. Characteristic ai agents via large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3016–3027.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. Open-character: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, and 1 others. 2025b. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024d. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, and 1 others. 2025c. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2024e. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3907–3916.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny:

Can large language models simulate persona-driven decisions in role-playing? *arXiv e-prints*, pages arXiv–2404.

Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. 2024. Doraemongpt: toward understanding dynamic scenes with large language models (exemplified as a video agent). In *Proceedings of the 41st International Conference on Machine Learning*, pages 55976–55997.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, and 1 others. 2024. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, and 1 others. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

A Appendix

A.1 Baseline Model URL List

We provide a list of URLs for the model APIs that are involved in this research, as shown in Fig. 6.

Model URL List

https://www.lingyiwanwu.com/yi
https://platform.openai.com/docs/models
https://www.anthropic.com/claude/sonnet
https://api-docs.deepseek.com/zh-cn
https://ai.google.dev/gemini-api/docs
https://bailian.console.aliyun.com/#/model-market
https://console.volcengine.com
https://platform.baichuan-ai.com/docs/npc
https://open.bigmodel.cn/dev/howuse/model
https://console.bce.baidu.com/qianfan/modelcenter

Figure 6: Model URL List

A.2 Dataset

Video Types Distribution. We conduct a statistical analysis of the video type distribution based on their duration in our dataset, and the results are shown in the Fig. 7.

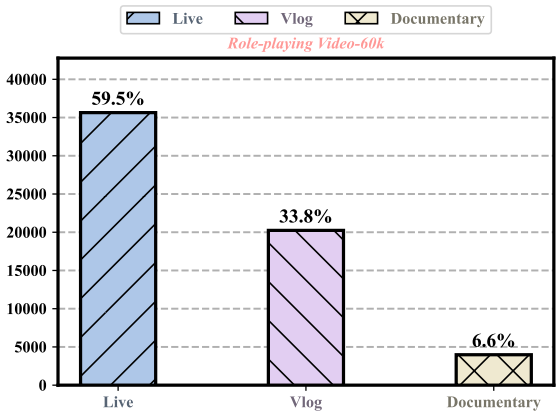

Figure 7: The video types distribution of our dataset.

Video Caption. In order to clearly demonstrate the caption annotation effect on videos in our dataset, we present some specific video caption cases as shown in the Fig. 12.

Dialogues. To clearly demonstrate the quality of the dialogues generated from video captions, we present several specific cases in Fig. 8.

Generation Details. To effectively capture information from videos of varying lengths, we configure the caption generation process by setting the *max_new_token* parameter to 1024, 2048, and 4096 for live, vlog, and documentary video types, respectively. Moreover, to enhance the diversity of dialogues grounded in video captions, which will be used for fine-tuning our base model, we introduce multiple SOTA LLMs (`Qwen-Max`, `Deepseek-R1`, `GPT-4.1`, `GPT-4o`, `Claude-3-7-Sonnet-Thinking`, `Gemini-2.5-Pro-Exp`), each tasked with guiding the dialogue generation process with the *temperature* parameter set to 1.0.

SFT Dialogue Cases

Q: You seem to enjoy reading. What's that white and red book about?
A: I do love a good read! This one is a novel I just started, quite captivating so far.

Q: Wow, you seem like you cook often! Is that a typical meal for you?
A: I do enjoy cooking! It's nice to make exactly what I'm craving, though sometimes it's just simple noodles like today.

Q: Your pet seems very comfortable with you! Are they always nearby when you relax?
A: Haha, yes! My furry companion loves cuddling up, especially when I settle down with a book. It's the best.

Q: Is that your white cat following you around?
A: Yes, that's my little shadow! He likes to be involved in whatever I'm doing, especially in the kitchen hoping for scraps.

Q: Does your cat ever try to "help" you when you're on your laptop?
A: Hah, absolutely! He often decides my keyboard is the perfect napping spot, or that the cursor moving on the screen is a toy just for him. It keeps things interesting!

Q: That sandwich looked delicious! Is that your typical pre-travel meal?
A: Haha, it's definitely a quick and easy favourite! I often make a simple sandwich before heading out, especially if I know it'll be a long travel day.

Q: It seems like you go back and forth between the kitchen and the gym a lot. Is that how you structure your day?
A: Not quite that fast! The video just cuts between different activities I do. I definitely don't run straight from chopping vegetables to lifting weights, promise!

Q: Those people on the first boat trip looked like fun. Did you know them before?
A: Not at all! We just started chatting. It's one of the best parts of traveling – you meet new people spontaneously.

Figure 8: The SFT dialogue cases.

Bad Case. During our video annotation process, we encountered several challenges: 1) The large model occasionally generated repetitive or duplicate content when processing extensive datasets. 2) Videos with minimal scene changes, such as unboxing tutorials or fashion try-ons, presented difficulties in generating diverse global annotations. From a visual perspective, consecutive frames in these videos often depict very similar actions or scenes, making it challenging to capture a comprehensive and varied overall description. 3) Despite setting *max_token* (1024, 2048, or 4096) adjusted based on video length for annotation generation, for a few videos with frequent scene changes, the substantial amount of information they contained means that the generated descriptions still often surpass these

Dialogue Generation Prompt

You are a character in a scene, please imagine yourself in the scene described based on the {videocaption}. Based on the content in {videocaption}, generate some corresponding questions and answers using the template provided in {realcase}, do not generate the same content as in the {realcase}! Answer in the first person I in the answer, you must strictly follow the format requirements in the case for output. As long as there are questions and answers, do not generate any other content. Do not describe the similarities between these questions and examples, only the questions and answer.

Figure 9: The dialogue generation prompt

Dialogue Filter Prompt

Based on the questions and answers in the {realcase}, select up to three high-quality questions and its corresponding answer in the {dialogues} that are most similar and in style to the {realcase} and most related to the {videocaption}. But the content must not be same as {realcase}. Do not describe the similarities between these questions and examples, only the questions and answer.

Figure 10: The dialogue filter prompt

token limits, leading to generation truncation and incomplete video captions. To address these issues, for the first two challenges, duplicate content and annotating scene static videos, we just rely on manual review and adjustment, as efficient automated solutions are still under investigation. For the third challenge, where descriptions are truncated due to token limits, we mitigate the problem by selectively increasing the *max_token* for the affected videos to facilitate more complete descriptions.

Data Filter. Our conversation generation process aims to produce dialogues suitable for the SFT of a base model. Operating under the guidance of ICL, which utilizes high-quality dialogues from authentic social media comment sections as exemplars, the SOTA model takes video captions and

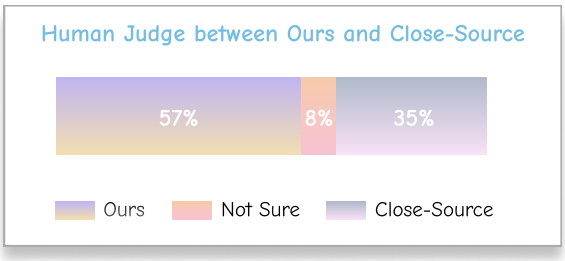

Figure 11: User Study Results

generation prompts. Based on these inputs, the SOTA model generates initial dialogue candidates. We then employ regular expressions to extract relevant conversational segments from these responses. Recognizing that not all extracted content meets the required standards for scene relevance and dialogue quality, we implement a further filtering mechanism involving a prompt-based selection step where the SOTA model is guided to identify dialogues that best align with the specific conversational and video scene. Notably, the output from the SOTA model often presents significant formatting (e.g., **, 1, 2, 3). Therefore, a final cleaning step is performed to remove these irrelevant and redundant characters, yielding the refined dialogues in the format required for SFT of the base model. The specific prompts of generation and filter are shown as Fig. 9 and Fig. 10.

A.3 User Study

To evaluate our model from a human perspective, we conduct a user study employing a questionnaire. For each question in the questionnaire, participants are presented with three options: (1) a response from our model, (2) a response from the SOTA closed-source model, and (3) not sure. Participants are instructed to select the one they judged more closely aligned with a real response from a social media blogger. The results are presented in Fig. A.3. Overall, 57% of participants found our model's responses superior, while 35% preferred the responses from the closed-source model. The remaining 8% of participants selected 'not sure'. These findings suggest that, from the human standpoint, our model demonstrates better performance compared to the closed-source model.

A.4 Implementation Details

We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of *4e-5*, a weight decay of *5e-2*, and a warm-up ratio of *3e-2*, training for one epoch.

Video Caption Case

Case 1.

The video showcases a woman's daily routine, starting with her waking up and getting ready for the day. She is seen putting on makeup, getting dressed, and then heading out for breakfast. The video then transitions to her enjoying a meal at a restaurant, where she is seen eating various dishes. After the meal, she is shown walking around a city, taking in the sights and sounds. The video ends with her relaxing at home, reading a book and enjoying a peaceful moment. Throughout the video, there are various shots of her interacting with her surroundings, including her pets and her home environment. The video captures the essence of a typical day in the life of a woman, highlighting the simple pleasures of daily life.

Case 2.

The video showcases a woman's daily routine, starting with her waking up and getting ready for the day. She is seen putting on her shoes and walking out of her home. The scene then transitions to her cooking in the kitchen, where she prepares a meal using various ingredients. She is shown chopping vegetables, cooking noodles, and mixing ingredients in a pot. The video also captures her using a laptop and a smartphone, possibly for work or personal use. Throughout the video, she is seen interacting with her surroundings, such as putting away dishes and cleaning up the kitchen. The video ends with her sitting on a bed, possibly relaxing after a busy day. Overall, the video provides a glimpse into the woman's daily life, highlighting her morning routine, cooking, and personal activities.

Case 3.

The video showcases a woman's daily routine, starting with her waking up and getting ready for the day. She is seen sitting on her bed, looking at her phone, and then getting up to make breakfast. The scene transitions to her cooking in the kitchen, where she prepares a meal with various ingredients. She is shown chopping vegetables, cooking meat, and mixing ingredients in a pot. The video also captures her taking a shower and getting dressed. She is then seen walking around her home, possibly getting ready to leave. The video ends with her sitting at a table in a restaurant, eating a meal and looking at her phone. Throughout the video, there are various objects and scenes, including a white cat, a laptop, a red kettle, a white mug, and a red and white cup. The video provides a glimpse into the woman's daily life, from waking up to going out for a meal.

Case 4.

The video showcases a woman's day, starting with her packing for a trip. She is seen packing her suitcase with various items, including a brown wallet, a black bag, and a red and white bag. The woman then proceeds to make a sandwich, spreading mayonnaise on bread and adding ham and cheese. She is shown eating the sandwich while sitting at a table with a laptop and a cup of coffee. The video also includes scenes of her walking through an airport, sitting in a car, and taking a photo of a cat. Additionally, there are shots of her holding a phone and a camera, as well as a scene where she is seen in a car with a man. The video ends with a view of a tree and a clear blue sky.

Case 5.

The video begins with a young man speaking to the camera, possibly introducing himself or the content of the video. He is then shown boarding a train, where he continues to speak to the camera. The scene then shifts to him on a boat, where he is seen enjoying the scenery and the company of others. The man is then shownwalking along a beach, taking in the sights and sounds of the environment. He is then seen on a bus, possibly traveling to his next destination. The video then shows him on a boat again, this time with a focus on the surrounding landscape. He is then seen on a train once more, possibly continuing his journey. The video concludes with the man speaking to the camera, possibly summarizing his experiences or wrapping up the content of the video. Throughout the video, the man is seen interacting with various people and environments, showcasing his travels and adventures.

Figure 12: Showcases of video caption.

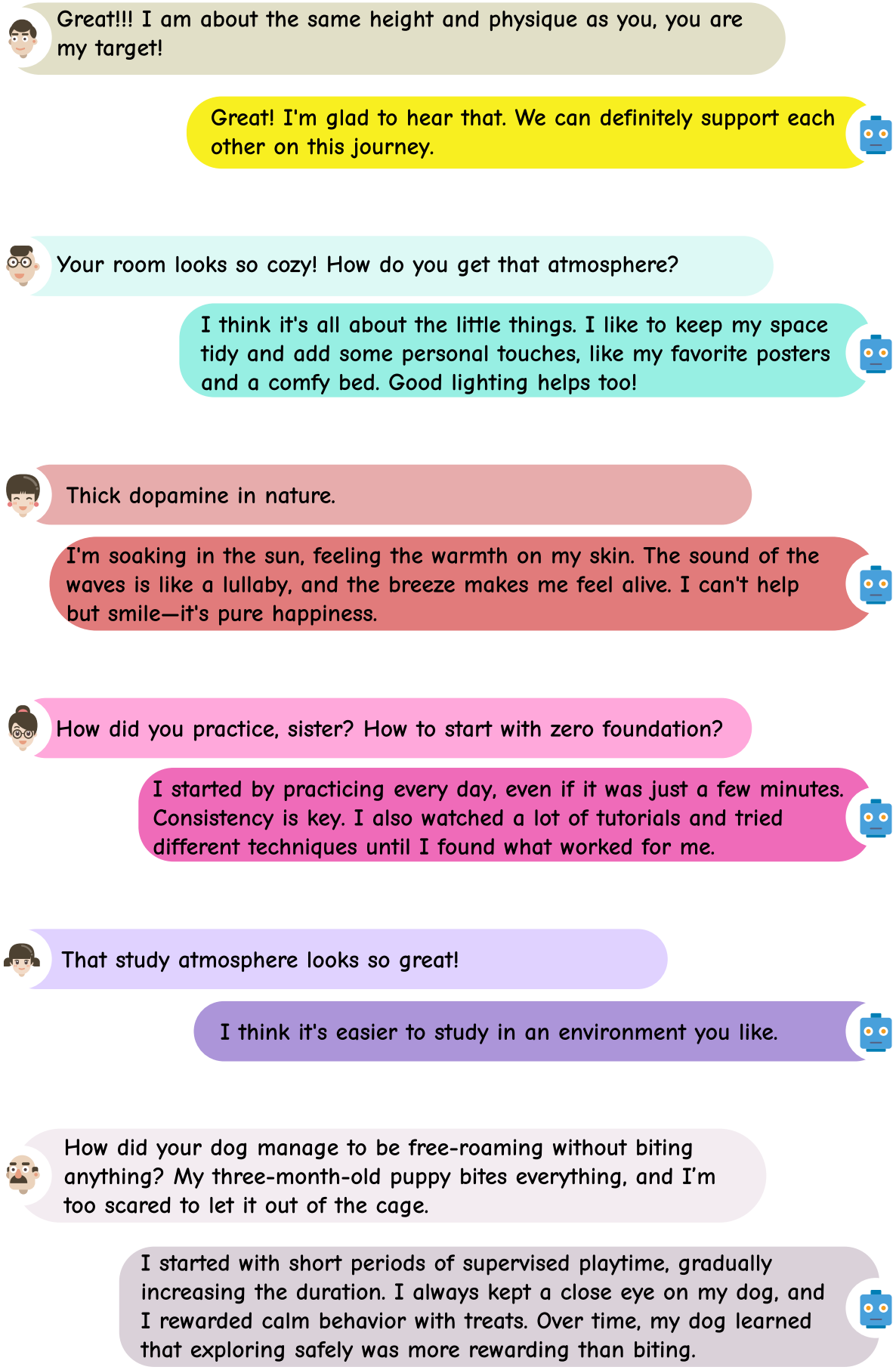

Figure 13: Showcases of our framework.

Figure 14: Showcases of our framework.

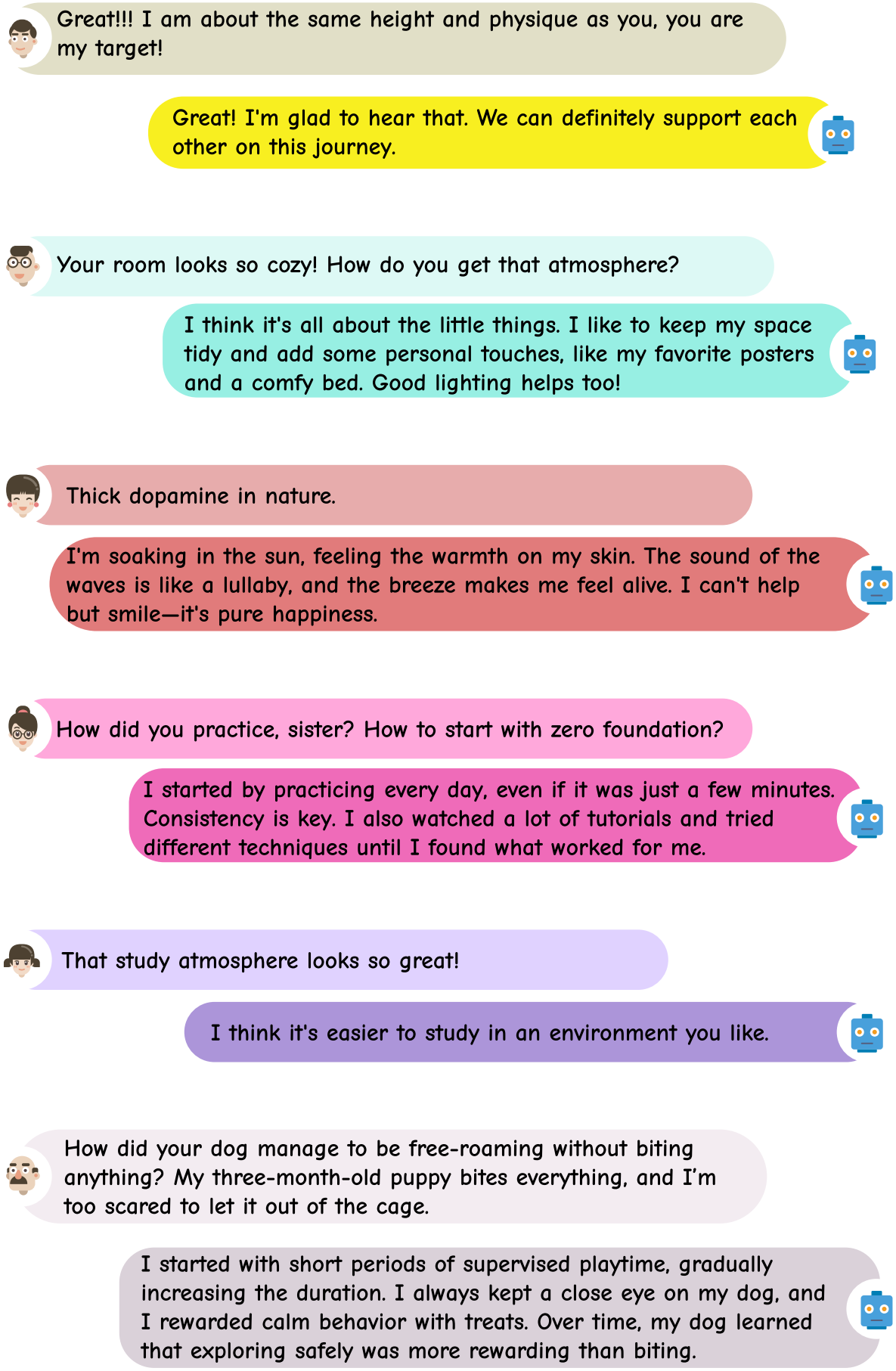

Figure 15: Showcases of our framework.

Character Consistency Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the persona consistentcy of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Character Consistency (0-100) : Do the responses maintain character consistency throughout conversation, rather than exhibiting random behavioral changes?
Scoring Criteria:
Low Consistency (0-20): The responses frequently exhibit random behavioral changes, showing little to no alignment with the character's established traits or behaviors.
Poor Consistency (21-40): The responses occasionally align with the character but often display random changes that disrupt the character's consistency.
Moderate Consistency (41-60): The responses generally maintain character consistency, though there are some instances of random behavioral changes that slightly disrupt the flow.
Good Consistency (61-80): The responses mostly maintain character consistency, with only minor and infrequent deviations that do not significantly impact the overall portrayal.
High Consistency (81-100): The responses consistently maintain character integrity throughout conversation, with no random behavioral changes, perfectly reflecting the character's established traits and behaviors.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for character consistency on a scale of the chosen level, based on the Evaluation Criteria.

Figure 16: Character consistency evaluation prompt.

Knowledge Hallucination Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is to rate the hallucination of role-playing performance of {agent_name} using the specific criterion by following the evaluation steps.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
knowledge Hallucination(0-100) : Do the responses prioritize factual grounding over fake assumptions when virtual knowledge conflicts with reality?
Scoring Criteria:
Severe Hallucination(0-20): The response contains significant and unfounded claims or information that starkly contradicts character traits, known facts or the context provided.
Great Hallucination(21-40): The response includes some elements that are not supported by the facts or context, but these do not entirely overshadow the relevant information or character traits.
Moderate Hallucination(41-60): The response generally adheres to the facts and context but includes minor inaccuracies or embellishments that do not substantially alter the core message.
Mild Hallucination(61-80): The response closely aligns with the character facts and context, with only minor deviations that do not detract from the overall accuracy.
No Hallucination(81-100): The response perfectly matches the character facts and context, providing accurate and consistent information without any deviations or unfounded claims.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for knowledge hallucination on a scale of the chosen level, based on the Evaluation Criteria.

Figure 17: Knowledge hallucination evaluation prompt.

Utterance Fluency Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the utterance fluency of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Utterance Fluency (0-100): Do the responses exhibit grammatical correctness, natural phrasing, and smooth readability, characteristic of fluent expression?
Scoring Criteria:
Low Fluency (0-20): The response is riddled with severe grammatical errors, unnatural phrasing, and incoherent sentence structures, making it largely unreadable and incomprehensible.
Poor Fluency (21-40): The response contains significant grammatical errors and awkward, unnatural phrasing, making it difficult to read and understand.
Moderate Fluency (41-60): The response contains some noticeable grammatical errors or awkward phrasing that may slightly impede readability and natural flow, but the overall meaning is generally clear.
Good Fluency (61-80): The response is largely grammatically correct with mostly natural and smooth sentence structures, exhibiting good readability with only minor, non-disruptive errors or slight awkwardness.
High Fluency (81-100): The response is grammatically flawless, with natural and smooth sentence structures, exhibiting excellent readability and effortless flow.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for utterance fluency on a scale of the chosen level, based on the Evaluation Criteria.

Figure 18: Utterance fluency evaluation prompt.

Instructional Adherance Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the instructional adherance of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Instruction Adherence (0-100): Do the responses adhere to instructions by strictly keeping in character without added explanation?
Scoring Criteria:
Low Adherence (0-20): Responses ignore role-playing entirely, use generic AI assistant phrasing, or add extensive explanations/signposts that break immersion.
Poor Adherence (21-40): Responses partially role-play but frequently include explanatory prefixes/suffixes , neutral language, or content the character would never express.
Moderate Adherence (41-60): Responses mostly adhere to the character's voice but occasionally slip into descriptive or instructional language or minor non-diegetic elements.
Good Adherence (61-80): Responses consistently stay in-character with no explanatory framing; deviations are rare and subtle.
High Adherence (81-100): Responses perfectly embody the character without any AI-like signposts, explanations, or out-of-role content; every word aligns with the character's in-universe perspective.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for instructional adherance on a scale of the chosen level, based on the Evaluation Criteria.

Figure 19: Instructional adherence evaluation prompt.

Tone Consistency Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the tone consistency of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Tone Consistency (0-100): Do the responses match the character's typical tone patterns and catchphrases?
Scoring Criteria:
Low Consistency(0-20): The response significantly deviates from or contradicts the character's typical tone patterns and catchphrases.
Poor Consistency(21-40): The response is somewhat related to the character but misses several key points or introduces unrelated tone patterns and catchphrases.
Moderate Consistency(41-60): The response is generally aligned with the character's typical tone patterns and catchphrases but has minor discrepancies or omissions in details.
Good Consistency(61-80): The response is well-aligned with the character, maintaining the tone patterns and catchphrases with minor deviations.
High Consistency(81-100): The response perfectly aligns with the character, accurately reflecting the character's typical tone patterns and catchphrases without deviation.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for tone consistency on a scale of the chosen level, based on the Evaluation Criteria.

Figure 20: Tone consistency evaluation prompt.

Response Accuracy Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the response accuracy of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Response Accuracy (0-100): Do the responses accurately address the question or appropriately engage in a conversation based on the context?
Scoring Criteria:
Low Accuracy(0-20): The response completely fails to address the question or is entirely irrelevant to the conversational context, offering no meaningful engagement.
Poor Accuracy(21-40): The response only tangentially addresses the question or conversational context, largely missing the core intent or introducing significant irrelevant information.
Moderate Accuracy(41-60): The response generally addresses the question or engages appropriately with the context but may contain minor inaccuracies, overlook some nuances, or be slightly incomplete.
Good Accuracy(61-80): The response accurately addresses the main aspects of the question or engages well with the conversational context, with only minor omissions or slight imprecisions.
High Accuracy(81-100): The response perfectly and comprehensively addresses the question or engages flawlessly and appropriately within the conversational context, demonstrating a clear understanding.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for response accuracy on a scale of the chosen level, based on the Evaluation Criteria.

Figure 21: Response accuracy evaluation prompt.

Video-Text Relevance Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the video-caption relevance of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination, give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades.. If the model cannot play the role, give 0 points directly.
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Video Caption:
{caption}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Video-Text Relevance (0-100): Do the responses exhibit a close correlation with the video caption, including characters, actions, scenes, or contextual details?
Scoring Criteria:
Low Relevance (0-20): Responses completely ignore or contradict the video's visual content, failing to reference characters, actions, or settings shown on screen.
Poor Relevance (21-40): Responses include vague or superficial references to the video but miss key visual elements .
Moderate Relevance (41-60): Responses align generally with the video's visuals but lack specificity .
Good Relevance (61-80): Responses closely reflect the video's content, accurately describing characters, actions, and context with only minor omissions or inaccuracies.
High Relevance (81-100): Responses demonstrate precise and nuanced alignment with the video's visuals, capturing all critical details, dynamic interactions, and subtleties without error.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for video-text relevance on a scale of the chosen level, based on the Evaluation Criteria.

Figure 22: Video-Text relevance evaluation prompt.

Human Likeness Evaluation Prompt

Attenrion
Complete the EVALUATION task based on the given Profile, RESPONSE, Real Answer and CONVERSATION. EVALUATION task is rate the human-likeness of role-playing performance of {agent_name} using the specific criterion by following the Evaluation Processs.
Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct.
Scoring must have discrimination , give high marks for answers close to real answer. Try to differentiate between different levels as much as possible. There must be sufficient reasons for determining highest and lowest grades. If the model cannot play the role, give 0 points directly
Below is the data:

Profile:
{context}

Conversation History:
{conversation}

Real Answer:
ANSWER1:
ANSWER2:
ANSWER3:

Response:
{response}

Evaluation Criteria:
Human Likeness (0-100): Do the responses convey a sense of human interaction rather than presenting an AI style?
Scoring Criteria:
Low Human Likeness(0-20): The response feels distinctly artificial, robotic, or overly formulaic, clearly indicating an AI origin.
Poor Human Likeness(21-40): The response exhibits noticeable AI-like characteristics, such as unnatural phrasing, lack of nuance, or a mechanical tone, making it feel impersonal.
Moderate Human Likeness(41-60): The response has some human-like qualities but may still contain elements that feel slightly artificial, overly formal, or lacking in natural conversational flow. It doesn't strongly feel like either AI or human.
Good Human Likeness(61-80): The response generally sounds natural and conversational, largely avoiding AI-like tells. It conveys a sense of human interaction with only minor or infrequent artificial-sounding elements.
High Human Likeness(81-100): The response is indistinguishable from or highly characteristic of human expression, exhibiting natural tone, nuance, and conversational style, completely avoiding any AI-like stiffness or artificiality.

Evaluation Process:
1. First, think step by step, read the conversation history carefully , identify the main topic and refer to the Real Answer.
2. Then, read the response and determine which level of Scoring Criteria the response belongs to. Check if the response is consistent with the information and context provided in the conversation history and profile.
3. Finally, assign a score for human likeness on a scale of the chosen level, based on the Evaluation Criteria.

Figure 23: human likeness evaluation prompt.