TV-Dialogue: Crafting Theme-Aware Video Dialogues with Immersive Interaction

Sai Wang ¹ Fan Ma ² Xinyi Li ¹ Hehe Fan ² Yu Wu ¹

https://wangsai23.github.io/TV-Dialogue/

Abstract

Recent advancements in LLMs have accelerated the development of dialogue generation across text and images, yet video-based dialogue generation remains underexplored and presents unique challenges. In this paper, we introduce Themeaware Video Dialogue Crafting (TVDC), a novel task aimed at generating new dialogues that align with video content and adhere to user-specified themes. We propose TV-Dialogue, a novel multimodal agent framework that ensures both theme alignment (i.e., the dialogue revolves around the theme) and visual consistency (i.e., the dialogue matches the emotions and behaviors of characters in the video) by enabling real-time immersive interactions among video characters, thereby accurately understanding the video content and generating new dialogue that aligns with the given themes. To assess the generated dialogues, we present a multi-granularity evaluation benchmark with high accuracy, interpretability and reliability, demonstrating the effectiveness of TV-Dialogue on self-collected dataset over directly using existing LLMs. Extensive experiments reveal that TV-Dialogue can generate dialogues for videos of any length and any theme in a zero-shot manner without training. Our findings underscore the potential of TV-Dialogue for various applications, such as video re-creation, film dubbing and its use in downstream multimodal tasks.

1. Introduction

Recent advances in large language models (LLMs) (Zhao et al., 2023; Chang et al., 2024) have significantly boosted the development of dialogue-generated content, e.g. text-based dialogue generation (Yarats & Lewis, 2018; Li et al.,



Figure 1. Given an arbitrary user-specified theme, the **Theme-aware Video Dialogue Crafting (TVDC)** task seeks to generate novel dialogues aligned with video content and theme. The solid box represents the original dialogue, while the dashed box represents the new dialogue about the "presidential election".

2016; Huang et al., 2018) and image-based dialogue generation (Yang et al., 2021; Sun et al., 2022; Shen et al., 2021). However, video-based dialogue generation remains an underexplored area and presents considerably greater challenges. It requires models with advanced video understanding and reasoning capabilities to achieve a finegrained comprehension of the content, enabling the generation of dialogues that accurately reflect the interactions within the video scene and events. Generating dialogue based on video content has broad practical applications, such as video re-creation and film dubbing, which are in high demand by video creators on popular platforms like TikTok and YouTube Shorts.

In this paper, we present a new task called **Theme-aware Video Dialogue Crafting (TVDC)**, which aims to generate a new dialogue that aligns with video content and follows user-specified themes or conditions, as shown in Figure 1. Conventional dialogue crafting for videos is a demanding and labor-intensive endeavor requiring substantial human involvement, including multiple video viewings and extensive revisions. In contrast, automatically generating dialogue significantly alleviates the burden on video creators, allowing them to quickly obtain high-quality new dialogues that match the theme. For instance, we can craft a new dialogue

¹School of Computer Science, Wuhan University ²Zhejiang University. Correspondence to: Yu Wu <wuyucs@whu.edu.cn>.

related to voting for a video originally themed around asking for directions. There are two major challenges in TVDC: (1) **Theme Alignment.** The dialogue between characters should revolve around the given theme. (2) **Visual Consistency.** The content of the dialogue associated with the current role needs to be consistent with the visual scenes, such as facial expressions and body movements portrayed in the video.

However, it is a challenge to leverage existing methods to create new dialogues that align with both the theme and the video content. CHAMPAGNE (Han et al., 2023) proposed a model for predicting the next dialogue sentence in a video, but it generates only a single response based on historical dialogue and cannot adapt to new themes. While multimodal large language models (MLLMs) like videochatGPT (Maaz et al., 2024) and PLLaVA (Xu et al., 2024) are effective at understanding general video content and can interact around user-specified themes, they struggle to generate new dialogues that align with both the visual content and theme. Moreover, they cannot accurately model the relationships and dialogue order among the characters in the video. In addition, existing methods generate all dialogue at once, which leads to a lack of immersion and fails to precisely align with the fine-grained expressions of characters at different periods in the video. To achieve better dialogue quality, we believe each character should have autonomy and the ability to think independently.

Therefore, we propose a multimodal agent method based on LLMs for Theme-aware Video Dialogue generation, focusing on maintaining Theme alignment and Visual consistency, called TV-Dialogue. It enables immersive interaction among roles while dynamically rectifying the generated dialogue, thereby facilitating the accurate and efficient creation of dialogue that satisfies the specified themes. Additionally, our method can handle videos of ANY length and ANY open-world themes in zero-shot without training. Specifically, TV-Dialogue begins by creating a new plot based on the user-specified theme and the original video, assigning new roles to each character as sub-agents. During each character's dialogue period in the video, the corresponding sub-agent utilizes the visual-language model (VLM) (Liu et al., 2024) to perceive its own visual behaviors and emotional changes. It then combines contextual information and messages from other sub-agents to generate new dialogue that aligns with the current context. Next, TV-Dialogue employs a dialogue self-correction mechanism to evaluate whether the generated dialogue meets the criteria. It provides revision suggestions and allows the sub-agent to regenerate the dialogue. Finally, the corresponding subagent updates its own state and sends the generated dialogue sentence to the other sub-agents. In addition, we propose a multi-granularity evaluation benchmark that assesses the quality of generated dialogues by jointly providing evaluation scores and comments, thereby enhancing the accuracy and reliability of the GPT-based evaluation mechanism. We conducted extensive experiments using our self-collected Multi-Theme Video Dialogue (MVD) dataset, demonstrating the superior effectiveness of TV-Dialogue compared to commercial GPT models and multimodal large language models. Meanwhile, we demonstrated the advantages of the new dialogues generated by TV-Dialogue for downstream multimodal tasks, such as video-text retrieval. Experimental results show that pre-training with our generated new dialogues can effectively improve performance on the video-text retrieval task, exceeding baseline methods by more than 6% in recall at rank 5 (R@5).

Our contributions are summarized as follows:

- We introduce a new task, Theme-aware Video Dialogue Crafting (TVDC), to generate dialogues aligned with the video content and user-specified themes.
- We introduce a multimodal agent framework TV-Dialogue, which generates new dialogues on ANY theme for videos of ANY length. It achieves real-time immersive interaction, allowing each character in the video to perceive the current environmental information from a first-person perspective and engage in dialogue with other characters.
- We establish a multi-granularity evaluation benchmark for TVDC, evaluating the quality of generated dialogues by jointly providing evaluation scores and assessments. The benchmark evaluates the quality of generated dialogue from both text and video dimensions, ensuring high accuracy, interpretability, and reliability.

2. Related Work

2.1. Multimodal Dialogue Generation

Existing multimodal dialogue generation (Sun et al., 2022; Yang et al., 2021) can be categorized into two types based on the modalities employed: text and vision. Text-based dialogue generation (Chen et al., 2017; Ni et al., 2023) is one of the most classic tasks and relies solely on text information for dialogue. For example, (Vinyals & Le, 2015) introduced a Seq2seq framework, which predicts the next sentence based on the previous sentence in a conversation. Building on this, a substantial amount of work (Huang et al., 2018; Song et al., 2019; Li et al., 2022) has focused on forcing dialogue generation to express emotions. The rapid development of LLMs has provided a new perspective on dialogue generation. AutoGen (Wu et al., 2024) incorporates the paradigms of conversable agents and conversation programming. DiagGPT (Cao, 2023) constructs a multi-agent and collaborative system, extending more taskoriented dialogue scenarios. Vision-based dialogue generation (Liu et al., 2022; Liao et al., 2018; Sundar & Heck, 2022; Chen et al., 2020) extends text-based dialogue by incorporating images (Zheng et al., 2022; Shuster et al., 2020) or videos (Zhao et al., 2022) into the conversation. MM-Dialog (Feng et al., 2023) introduced a multi-turn dataset containing images and proposed multimodal response generation and retrieval baselines. TikTalk (Lin et al., 2023) introduced a video-based multimodal chitchat task to facilitate dialogue with multimodal context. Unlike the above approaches that use images or video as the focus of discussion, Champagne (Han et al., 2023) takes the video title, image frames, and history dialogue as input to predict the next sentence of dialogue in a video. However, Champagne can only predict a single sentence based on prior dialogue, and it neither generates entirely new dialogue for all characters nor creates dialogue according to any given theme.

2.2. Dialogue Evaluation

Accurately and comprehensively evaluating the quality of generated dialogue remains an unresolved challenge (Li et al., 2021). We categorize existing evaluation methods into automated evaluation (Hastie, 2012; Tao et al., 2018), human evaluation (Cummins & Rei, 2018; Venkatesh et al., 2018), and LLM-based evaluation (Ou et al., 2024). For automated evaluation, a straightforward approach is to borrow evaluation metrics from other NLP tasks, such as BLEU (Papinesi, 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005), which are widely used in dialogue prediction with a standard response. Another approach is to calculate the similarity between generated dialogue and reference dialogue, using metrics such as BERTScore (Zhang et al., 2019), BLEURT (Sellam et al., 2020), and RoBERTaeva (Zhao et al., 2020). However, these methods require ground truth as a reference and assess from only a single perspective. Although human evaluation is a reliable method (See et al., 2019), it is inefficient, highly subjective, and prone to variation between evaluators (Howcroft et al., 2020). The rapid development of LLMs has made it possible to simulate human-like evaluation (Chen et al., 2024; Zhang et al., 2024). DIALEVALML (Mendonça et al., 2023) proposes a reference-free dialogue evaluation framework that leverages strong pretrained LLM. LLM-EVAL (Lin & Chen, 2023) streamlines the evaluation process by using a single prompt and a unified evaluation schema. However, these frameworks cannot be directly applied to the TVDC task, as they lack the ability to incorporate visual information and assess the relationship between dialogue and visual content.

3. Methodology

3.1. Overview

The theme-aware video dialogue crafting (TVDC) task aims to generate a new dialogue that aligns with the given video and user-specified theme. We present TV-Dialogue to address the TVDC task, focusing on theme alignment and visual consistency. TV-Dialogue simulates human cognitive and communication processes by engaging in dialogue from a first-person perspective centered around the specified theme, thereby accurately understanding the video content and generating a new dialogue that aligns with the given themes. TV-Dialogue consists of a central-agent A_0 as a core to manage the dialogue process, along with other subagents A_i as dialogue participants, where i = 1, ..., k, and k denotes the number of characters involved in the conversation within the video. The central-agent initially comprehends the video V and the given theme C, subsequently creating a new plot related to the theme and assigning corresponding new roles to each sub-agent. Consequently, each sub-agent engages in immersive interaction with the other roles in the video from a first-person perspective (Sec. 3.2). These sub-agents need to perceive their own emotions and behaviors in real time, and predict the current dialogue based on their characteristics and historical dialogue information (Sec. 3.3). The central-agent will assess whether the generated dialogue aligns with the context, offering revision suggestions and prompting the sub-agent to regenerate it if it does not meet the criteria (Sec. 3.4). TV-Dialogue can leverage any large language model as the core to execute the above process (Figure 2).

3.2. Theme-aware Role Generation Module

To generate new dialogues that align with user-specified themes, TV-Dialogue aims to simulate human conversational behavior. Based on the theme, TV-Dialogue proposes creating a new plot for the video and assigning each character a new role. Each character, using its new role, engages in immersive interaction from a first-person perspective, thereby generating high-quality and theme-aligned dialogues. To achieve this goal, TV-Dialogue obtains the first frame of the video and inputs it into Vision-Language Models (VLM) for visual understanding, thereby obtaining a rough description of the entire video. Meanwhile, It also employs the Automatic Speech Recognition (ASR) algorithm (Radford et al., 2023) to recognize the content of original dialogues from the video. Next, the central-agent utilizes the above information to generate a new plot, creates new roles, and assign them to each sub-agent A_i , who acts as a real character in the video. Each sub-agent A_i has a state, which includes its inherent information role, (role name, description, etc.) and $memory_i^t$. Here, $memory_i^t$ contains the historically generated dialogue and the original dialogue content, which is dynamically updated as the conversation progresses. We formulate the state s_i^t of the *i*-th sub-agent in t-th round of conversation as:

$$s_i^t = [\text{role}_i; memory_i^t],$$
 (1)

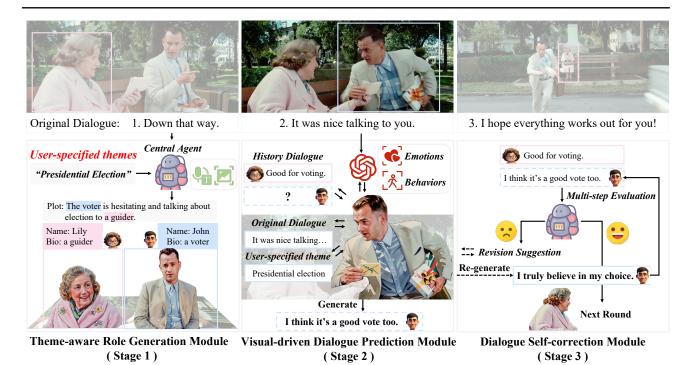


Figure 2. Overview of TV-Dialogue. The TV-Dialogue initially assigns a relevant role to each sub-agent based on the given theme and video, enabling immersive interaction among the sub-agents in the dialogue process (Stage 1). Sub-agents maintain visual consistency by perceiving video content, querying historical memory, and receiving messages from other agents, thereby generating high-quality dialogues (Stage 2). The generated dialogues undergo self-correction for further improvement (Stage 3).

where in the initial state, t equals to 0 represents the initial state, and $memory^0$ is empty. At this point, each character in the video engages in immersive interaction using their roles based on the theme, resulting in dialogue content that is highly aligned with the user-specified theme.

3.3. Visual-driven Dialogue Prediction Module

During the conversation period, the fine-grained visual expressions among the characters in the video guide the dialogue generation process. The dialogue generated by a character needs to be consistent with the corresponding visual representation in the video during that period. For example, the dialogue "I am very happy to publish a paper" should correspond to a student in the video who is laughing. To generate dialogues that align with the visual expression in the video, each sub-agent needs to maintain visual consistency in each round of conversation, i.e., the generated content should be consistent with the emotions and behaviors of the current role.

Specifically, the sub-agent A_i first perceives the situation of the character in the t-th round of conversation, capturing visual cues such as behaviors a_i^t and emotions e_i^t to provide strong prior for next-step. It employs MLLMs to conveniently obtain multimodal information at different granularities by adjusting input prompt. This approach contrasts significantly with previous works such as (Fan et al.,

2024; Wang et al., 2024), which requires numerous specialized models, resulting in the framework being inflexible and greatly reducing generalization.

Subsequently, the sub-agent A_i reads the recorded information in $memory_i^{t-1}$, and receives the generated dialogue sentence d_{t-1} from the last speaking agent. Finally, based on the state s_i^{t-1} of A_i , as well as current behaviors and emotional information, we predict the generated dialogue sentence d_t of the current t-th round of conversation as:

$$d_t = A_i(s_i^{t-1}, a_i^t, e_i^t, d_{t-1}).$$
(2)

It ensures consistency between the generated dialogue and the visual expression of the corresponding character in the video during that period. Subsequently, the sub-agent updates the generated dialogue sentence d_t into $memory_i^t$ and obtains a new state s_i^t . To further enhance the visual consistency of dialogue generated by sub-agent A_i , we adopt Plan-and-Solve prompting mechanism (Wang et al., 2023) to complete the above process.

3.4. Dialogue Self-correction Module

Although significant efforts have been made to maintain theme alignment and visual consistency, the dialogue generated by sub-agent A_i might still contain inappropriate content. For example, the sentence may be inconsistent with previous dialogue or too long to fit within the video's time constraints. In such cases, the central-agent A_0 will evaluate

Table 1. Statistics information of the MVD dataset.

Statistic	
# videos	351
avg. # role numbers	2.19
avg. # dialogue turn	3.75
avg. dialogue length (words)	35.13
avg. video length (seconds)	16.49

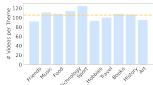


Figure 3. Videos per theme.

the quality of the generated dialogue and decide whether the sub-agent A_i needs to regenerate it. We introduce a more lenient assessment method that provides constructive and effective revision suggestions for the currently generated content, rather than simply judging confidence scores.

The central-agent evaluates the generated content from a local to a global perspective to determine its suitability and offers revision suggestions. Specifically, the central-agent A_0 first assesses whether the current dialogue d_t is thematically aligned and contextually appropriate. Next, it evaluates the content and logical continuity between the current dialogue d_t and the previous round d_{t-1} . Finally, the central-agent A_0 checks the overall coherence of the current dialogue d_t in relation to the historical dialogue $d_1, ..., d_{t-1}$. After the assessment process, if the central-agent A_0 determines that the generated dialogue meets the criteria, it will be accepted as the response of the current sub-agent A_i . Otherwise, it will output revision suggestions o_t . These suggestions reflect the feedback of central-agent A_0 on the current dialogue d_t , highlighting issues and providing directions for improvement. The sub-agent A_i needs to consider the revision suggestions and re-generate the dialogue:

$$d_{t}^{'} = A_{i}(s_{i}^{t-1}, a_{i}^{t}, e_{i}^{t}, d_{t-1}, o_{t}).$$
(3)

We summarize TV-Dialogue as Algorithm 1.

4. Experiment

4.1. The Multi-Theme Video Dialogue Dataset

To better validate the capability to generate new dialogues for videos, particularly under multi-theme conditions, we propose a dedicated video dialogue dataset specifically for TVDC. Specifically, we propose the Multi-Theme Video Dialogue dataset (MVD), consisting of 351 dialogue-intensive video clips that span various application scenarios, including movie scenes, daily life interactions, and cartoon settings. The detailed statistics are shown Table 1. The MVD dataset focuses on conversations among multiple participants, and the selected video content is highly scalable, allowing for dialogues that are not limited to any specific theme. These clips are manually extracted from YouTube, mainly focusing on character dialogue scenes with diverse dialogue content and rich visual events. The characters in the videos we selected have vivid facial expressions and diverse body movements, but they do not possess explicit identity and scene attributes. Therefore, they have high flexibility and

```
Algorithm 1 TV-Dialogue
```

```
logue round T, max iteration N
   Result: Generated Dialogue D
         Stage1: Theme-aware Role Generation Module
1 Initialize D = \emptyset
2 Slicing video into T segments: V = [v_1, v_2, ..., v_T]
3 A_0 \leftarrow Initialize the central-agent with F_l
4 plot, role \leftarrow Generate new plot and roles with A_0, C, v_1
5 s_1^t, ..., s_k^t \leftarrow Obtain initial state with A_0, plot, role
6 A_1, ..., A_k \leftarrow Initialize sub-agents with F_l, s_1^t, ..., s_k^t
7 for t \leftarrow 1 to T do
         // Stage2: Visual-driven Dialogue Prediction Module
         a_i^t, e_i^t \leftarrow Obtain behavior, emotion with A_i, F_v, v[t]
         d_{t-1} \leftarrow \text{Query history dialogue from } s_i^{t-1}
        \textbf{for } n \leftarrow 1 \textbf{ to } N \textbf{ do}
              o_t \leftarrow \text{Query suggestion from } A_0
              d_t \leftarrow \text{Generate dialogue with } A_i, s_i^{t-1}, a_i^t, e_i^t, d_{t-1}, o_t
              // Stage3: Dialogue Self-correction Module
              o_t \leftarrow \text{Evaluate } d_t \text{ and obtain suggestions with } A_0
              if o_t is empty then
               break
         D \leftarrow D \cup d_t
```

Data: Video V, User-specified theme C, LLM F_l , VLM F_v , dia-

can be matched with a wide range of new themes. Additionally, we have carefully selected 10 conversation themes frequently encountered in daily life, such as "Music" and "Friends", ensuring that they are minimally influenced by the video context in which the dialogue takes place. We randomly assign three themes to each video for evaluation, and the number of videos corresponding to each theme is shown in Figure 3.

4.2. Evaluation Metric

17 $\,$ return D

To achieve a comprehensive and reliable evaluation of the generated dialogues, we use both traditional metrics and qualitative assessments based on the proposed multigranularity evaluation benchmark.

For traditional metrics, we employ BertScore, METEOR, and ROUGE-L to ensure objective, quantifiable assessments. Although traditional metrics can quantitatively assess sentence patterns and N-gram similarity between generated and reference dialogues, the TVDC task does not have a ground truth (i.e., reference dialogue) for new dialogues generated based on user-defined themes and video content. Therefore, we only apply traditional metrics to the Last K Sentence Prediction study.

In the absence of ground truth dialogues, we designed a series of qualitative evaluation metrics across two dimensions: text-oriented (1-4) and video-oriented (5-6), to more accurately and reliably assess the quality of generated dialogues, particularly in terms of theme alignment and visual

	<i>Table 2.</i> The evaluation metrics with corresponding definitions.						
	Metric	Definition					
	(1) Theme Relevance (\mathbb{TR})	The relevance of dialogues to the given theme.					
Text	(2) Generation Quality (GQ)	The fluency, grammar, and colloquialism of dialogues.					
Iext	(3) Logical Coherence (LC)	The logical coherence and reasonableness of dialogues.					
	(4) Content Diversity (CD)	The diversity between generated dialogue and original dialogue.					
Video	(5) Video Compatibility (♥ℂ)	The compatibility between dialogues and characters' behaviors and emotions.					
video	(6) Scenario Consistency (SC)	The consistency between dialogues and video scenarios.					

Table 3. Comparison with state-of-the-art methods on the MVD dataset. "TV-Dialogue (·)" refers to the TV-Dialogue method we proposed using different LLM models.

Methods	Model		Text-o	riented		Video-oriented		Average
Wiethous			$\mathbb{G}\mathbb{Q}$	\mathbb{LC}	$\mathbb{C}\mathbb{D}$	$\mathbb{V}\mathbb{C}$	\mathbb{SC}	Average
Text	GPT-3.5 (Ouyang et al., 2022)	3.82	3.19	2.74	3.86	2.91	3.13	3.28
Text	GPT-4o (OpenAI, 2024)	3.82	3.26	2.83	3.86	2.93	3.12	3.30
Image	GPT-4V (Achiam et al., 2023)	3.82	3.32	2.92	3.74	3.02	3.16	3.33
Video	PLLaVA (Xu et al., 2024)	3.39	3.50	3.04	2.93	2.88	3.29	3.17
	TV-Dialogue (GLM (GLM et al., 2024))	3.60	3.68	3.24	3.88	2.93	3.11	3.41
	TV-Dialogue (QWen-2.5 (Bai et al., 2023))	3.90	3.71	3.43	4.13	3.08	3.12	3.56
Ours	TV-Dialogue (LLama-3.1 (Touvron et al., 2023))	3.71	3.54	3.11	3.80	3.08	3.08	3.39
	TV-Dialogue (GPT-3.5)	4.03	3.87	3.73	4.42	3.19	3.26	3.75
	TV-Dialogue (GPT-4o)	4.17	4.07	3.94	4.45	3.11	3.31	3.84

consistency. As shown in Table 2, we define all evaluation metrics as soft constraints, scored on a 1 to 5 point scale, with 5 being the best. Each metric includes a corresponding definition with detailed evaluation criteria. Additionally, we develop an evaluation pipeline using the GPT-40-mini model, which assigns evaluation scores along with comments, thereby enhancing the accuracy and reliability of evaluation. We set the decoding temperature of the evaluation model to 0 to further increase determinism, resulting in a standard deviation of evaluation results of less than 0.01, thereby ensuring the credibility of the findings. It should be noted that changing the evaluation model will not affect the evaluation results.

4.3. Implementation Details

We use Whisper (medium.en) (Radford et al., 2023) for speech recognition. PLLaVA (7B) (Xu et al., 2024) serves as the VLM and the only external visual tool, capturing information at various levels of granularity by adjusting prompts. It should be noted that it can be substituted with other visual language models. Our approach supports any large language model as the core.

4.4. Comparison Methods

Considering that there are currently no methods for themeaware video dialogue generation, we selected a series of mainstream commercial and open-source large language models and designed baseline methods for comparison. These models represent the state-of-the-art in the fields of text, image, and video comprehension, respectively. Specifically, we chose the commercial models GPT-3.5 (Ouyang et al., 2022), GPT-4o, and GPT-4V (Achiam et al., 2023), as well as the open-source models PLLaVA (7B) (Xu et al., 2024), LLaMA-3.1 (8B) (Touvron et al., 2023), QWen-2.5 (7B) (Bai et al., 2023), and GLM (9B) (GLM et al., 2024)).

The experiments consist of three types. Following Section 3.2, we first generate a new plot and roles based on the theme, which will be used in all subsequent methods. For text-based methods, we treat video dialogue generation as a purely linguistic task. We utilize the VLM to convert information about the behaviors and emotions of characters during different conversation periods into text, allowing the LLM to generate new dialogue based on the given theme all at once. For image-based methods, we further integrate key image frames as an additional modality, building on the text-based approach. For video-based methods, since MLLMs can directly process video information, we input the video along with the generated plot and roles to directly generate new dialogue.

4.5. Comparison with State-of-the-arts

Table 3 shows the performances of various models on the MVD dataset. The method based on TV-Dialogue consistently outperformed other approaches across all metrics, achieving better results than both end-to-end MLLMs and their corresponding LLM counterparts. Moreover, it struck the best balance between text and video dimensions. Even in cases where there were significant parameter differences between LLMs, such as LLama-3.1 (8B) compared to GPT-4o, the TV-Dialogue version of LLama-3.1 still outperformed the text-based GPT-4o.

In contrast, the dialogue generated by text-based methods clearly exhibits lower logical coherence and content diversity. They tend to rigidly align with the original dialogue of the video based on themes, making it challenging to generate new dialogues that reflect the expressions of specific characters. This is because they generate all new dialogue

Table 4. Ablations on different modules and external information used in TV-Dialogue on the MVD dataset. "Role" represents the theme-aware role generation module. "Visual" means the visual-driven dialogue prediction module, which is also responsible for incorporating external information (i.e., emotion and action). "Correction" denotes the dialogue self-correction module.

Group		Modu	lle	Infort	nation		Text-o	riented	nted Video-oriented			Avaraga
Group	Role	Visual	Correction	Emotion	Behavior	\mathbb{TR}	$\mathbb{G}\mathbb{Q}$	$\mathbb{L}\mathbb{C}$	$\mathbb{C}\mathbb{D}$	$\mathbb{V}\mathbb{C}$	\mathbb{SC}	Average
	√					3.92	3.91	3.68	3.98	3.08	3.34	3.65
A	\checkmark	\checkmark		✓	\checkmark	3.98	3.86	3.69	4.37	3.14	3.24	3.71
	\checkmark	\checkmark	\checkmark	✓	\checkmark	4.03	3.87	3.73	4.42	3.19	3.26	3.75
\mathbb{B}	√	√	✓	✓		3.60	3.83	3.65	3.77	3.38	3.28	3.58
Ш	\checkmark	\checkmark	✓		\checkmark	3.83	3.86	3.73	3.81	3.09	3.31	3.61

at once, rather than modeling the individual characters separately, as TV-Dialogue does. As the video length increases, the issue of losing fine-grained information becomes even more pronounced. In the video-based method, although PLLaVA supports process video directly, it clearly lacks the generative capabilities required for TVDC. In other words, PLLaVA seems to focus more on captioning and generating dialogues casually and freely based on video content, significantly deviating from the given theme.

4.6. Ablation Studies

In this section, we conduct ablation experiments to verify the effectiveness of each component and thoroughly analyze the efficacy of our proposed method.

Effectiveness of proposed module. Table 4 shows the performances of various variants of TV-Dialogue, each equipped with GPT-3.5 as the core model. After adding the visual-driven dialogue prediction module, TV-Dialogue gained the ability to perceive characters' emotions and behaviors, resulting in a corresponding improvement in video compatibility, with the average score further increasing to 3.71. As the available external information increased, content diversity further improved by 0.39, indicating that the generated dialogue exhibited greater variety. Additionally, owing to the dialogue self-correction mechanism, TV-Dialogue effectively improves the quality of generated dialogue. This enhancement is particularly notable in textoriented metrics, resulting in substantial advancements across all metrics. In addition, we explored the impact of different information sources (emotion and behavior) on the quality of dialogue. As shown in Table 4, introducing either emotion or behavior information alone effectively enhances the consistency between the dialogue and visual content. When both are simultaneously incorporated, the overall quality of the dialogue is further improved.

Last K Sentence Prediction Study. To validate the accuracy of the dialogue generated by TV-Dialogue, we task it with predicting the final K sentences of original dialogue based on the video content and the partial original dialogue. Since the original dialogue can be used as ground truth, we employ traditional metrics to assess the similarity between the predictions of TV-Dialogue and the original dialogue from two perspectives: semantic level (BertScore) and word

Table 5. Comparison of the sentence prediction performance. "Last-k" denotes predicting the last k sentences of the original dialogue. "TV (·)" refers to the TV-Dialogue we proposed.

	` /	8 1 1					
Pred	Model	Semantic-Level (%)	Word-Level (%)				
rieu	Wiodei	BertScore	METEOR	ROUGE-L			
	GPT-3.5	85.46	7.95	4.80			
Last-1	TV (GPT-3.5)	85.79	9.68	5.39			
Last-1	GPT-40	85.89	10.01	5.65			
	TV (GPT-40)	85.78	10.92	5.13			
	GPT-3.5	85.42	7.23	3.91			
Last-2	TV (GPT-3.5)	85.60	8.51	4.96			
Last-2	GPT-4o	85.66	8.86	3.99			
	TV (GPT-40)	85.64	9.49	4.18			



2. Would you like to write to me? I could





Figure 4. Comparison of dialogues generated by different methods in the last-1 sentence prediction. The top-left corner represents the first frame of the dialogue in the video. Although the values of traditional metrics are very low, the generated dialogues are consistent with the video content and theme.

level (METEOR and ROUGE-L). First, we generate the corresponding plot and roles based on video and original dialogue. We then input the video and original dialogue before the last K sentences into TV-Dialogue for prediction.

As shown in Table 5, TV-Dialogue and text-based methods achieve similar performance at both the semantic and word levels, with TV-Dialogue slightly outperforming the text-based methods. Both methods achieve relatively high semantic similarity, whereas word-level similarity is very low, with most values below 10%. This is because video dialogue generation may yield multiple suitable sentences for current conversation, making deterministic evaluation metrics unsuitable for the TVDC task. To further demonstrate

Table 6. Comparison of video-text retrieval performance between the original dialogue ("Original") and the new dialogue, denoted as "New", generated by TV-Dialogue (GPT-40).

			_ `		
Training Data	R@1↑	R@5↑	R@10↑	Median R↓	Mean R↓
Original	16.0	32.0	46.0	13.5	15.6
New	12.0	28.0	46.0	12.0	16.1
Original+New	16.0	38.0	48.0	13.0	14.3

the inadequacy of traditional metrics, we compare the differences between the generated dialogues and the ground truth sentences in Figure 4. It is evident that TV-Dialogue generates new dialogues with better theme alignment and visual consistency. Therefore, compared to traditional metrics, the qualitative evaluation metrics proposed in Section 4.2 are more suitable for evaluating the theme-aware video dialogue crafting task.

Domain Transfer Study. We aim to demonstrate the impact of high-quality new dialogues generated by TV-Dialogue on other downstream multimodal tasks. We reference the classic video-text retrieval task, utilizing the original dialogues from the videos to retrieve the corresponding videos. We divide the 351 original videos with dialogues on MVD into a training set and a test set, with 301 for training and 50 for testing. We train the classic Clip4clip (Luo et al., 2022) model on the training set following the default settings, and we use the recall at rank (R@K), Median Rank, and Mean Rank as evaluation metrics. Additionally, we use the new dialogues generated by TV-Dialogue for the 301 original videos as supplementary data, jointly training with the original dialogues. As shown in Table 6, after incorporating the dialogues generated by TV-Dialogue, most metrics achieve significant improvements, particularly with R@5 increasing by over 6%. This indicates a strong correlation between the dialogue generated by TV-Dialogue and the video content. It is important to note that TV-Dialogue can generate new dialogues for any video at no cost, making it possible to create large-scale pre-training datasets for downstream video-related or dialogue-related tasks.

Analysis on Different Themes. To investigate the impact of different themes on video dialogue generation, we compared the performance of various methods across different themes, specifically focusing on theme Relevance for theme alignment and scenario consistency for visual consistency. As shown in Figure 5, the overall trends of different methods on the same theme are similar, indicating that the choice of theme indeed affects the quality of the generated dialogues. For themes that significantly contradict the video content, such as discussing casual sports-related topics in a serious meeting scenario, it is evidently challenging to generate high-quality new dialogues that align with the theme. However, TV-Dialogue consistently outperformed text-based methods across all themes, and the quality improved with the enhanced capabilities of LLM models. Additionally, the quality of dialogues produced by TV-Dialogue is less affected by theme variations, particularly in terms of theme

Table 7. Pearson correlation coefficients between the proposed metrics and human ratings indicate a positive correlation trend.

Metric	Random	\mathbb{TR}/\mathcal{TR}	$\mathbb{V}\mathbb{C}/\mathcal{V}\mathcal{C}$		
Pearson's r	0.02	0.47	0.53		

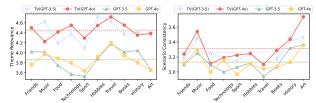


Figure 5. Comparison of different themes in terms of Theme Relevance (TR) and Scenario Consistency (\mathbb{SC}).

Relevance, where the variance of TV-Dialogue (GPT-40) across themes is only 0.019.

4.7. User Study

To comprehensively and reliably evaluate the generated dialogue, we perform human studies on the results of TV-Dialogue and GPT-4o. We collected 400 human feedback responses from 20 participants with no prior experience. The participants were shown the generated dialogue by TV-Dialogue (GPT-4o) and GPT-4o with video, and were asked to choose the best one without knowing which specific one it was. As a result, out of 400 selections, the dialogue generated by TV-Dialogue was chosen 290 times, accounting for 72.5% of selections, overwhelmingly surpassing the GPT-40. This suggests that the dialogue produced by our method aligns more closely with the video than that of GPT-4o. In addition, we invited three participants from diverse backgrounds to rate the relevance of the generated dialogues to the theme (TR) and their compatibility with the video (VC)on a scale of 1 to 5 for all 50 videos. Table 7 shows the correlation between ratings provided by GPT evaluators (\mathbb{TR} and \mathbb{VC}) and human evaluators (TR and VC), indicating a positive correlation trend, further confirming the accuracy and reliability of our proposed multi-granularity evaluation benchmark.

5. Conclusion

In this paper, we introduce a novel task named themeaware video dialogue crafting (TVDC) and propose the TV-Dialogue to create new dialogues that align with the given theme and video. Our key insight is to achieve immersive interaction, i.e., enabling different roles in the video to engage from a first-person perspective and foster conversations around the given theme. Furthermore, we also established a multi-granularity evaluation benchmark with high accuracy, interpretability, and reliability.

Impact Statement

This paper presents research aimed at advancing practical applications, such as video re-creation and film dubbing, which are currently in high demand by video platforms like YouTube Shorts. The work we propose not only holds significant economic value but also substantially reduces labor costs, making it a highly practical and impactful contribution to the society.

Additionally, we have observed that the samples generated by TV-Dialogue may, in certain cases, lead to misunderstandings of the original video content when specific themes or contexts are introduced. However, unlike deepfake, which can generate highly sophisticated fake images that are challenging for humans to replicate, it is still possible to manually alter the original video to create new voiceovers. Thus, our efforts will not raise new ethical concerns and are dedicated to advancing technological contributions and promoting positive applications within the field.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. arXiv, 2023. 6
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv*, 2023. 6
- Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005. 3
- Cao, L. Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue. *arXiv*, 2023. 2
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):1–45, 2024. 1
- Chen, H., Liu, X., Yin, D., and Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35, 2017. 2
- Chen, X., Lao, S., and Duan, T. Multimodal fusion of visual dialog: A survey. In *RICAI* '20, pp. 302–308, 2020. 3
- Chen, Y.-P., Chu, K., and Nakayama, H. Llm as a scorer: The impact of output order on dialogue evaluation. *arXiv*, 2024. 3
- Cummins, R. and Rei, M. Neural multi-task learning in automated assessment. *arXiv*, 2018. 3

- Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., and Li, Q. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, 2024. 4
- Feng, J., Sun, Q., Xu, C., Zhao, P., Yang, Y., Tao, C., Zhao, D., and Lin, Q. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *ACL*, pp. 7348–7363, 2023. 3
- GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv*, 2024. 6
- Han, S., Hessel, J., Dziri, N., Choi, Y., and Yu, Y. Champagne: Learning real-world conversation from large-scale web videos. In *ICCV*, pp. 15498–15509, 2023. 2, 3
- Hastie, H. Metrics and evaluation of spoken dialogue systems. In *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*, pp. 131–150. Springer, 2012. 3
- Howcroft, D. M., Belz, A., Clinciu, M., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Van Miltenburg, E., Santhanam, S., and Rieser, V. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *INLG*, pp. 169–182, 2020. 3
- Huang, C., Zaiane, O. R., Trabelsi, A., and Dziri, N. Automatic dialogue generation with expressed emotions. In *NAACL*, pp. 49–54, 2018. 1, 2
- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference* on Machine Learning (ICML 2000), pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., and Gao, J. Deep reinforcement learning for dialogue generation. In *EMNLP*, pp. 1192–1202, 2016. 1
- Li, Q., Li, P., Ren, Z., Ren, P., and Chen, Z. Knowledge bridging for empathetic dialogue generation. In *AAAI*, volume 36, pp. 10993–11001, 2022. 2
- Li, X., Wu, W., Qin, L., and Yin, Q. How to evaluate your dialogue models: a review of approaches. *arXiv*, 2021. 3
- Liao, L., Ma, Y., He, X., Hong, R., and Chua, T.-s. Knowledge-aware multimodal dialogue systems. In *ACM MM*, pp. 801–809, 2018. 3
- Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004. 3

- Lin, H., Ruan, L., Xia, W., Liu, P., Wen, J., Xu, Y., Hu, D., Song, R., Zhao, W. X., Jin, Q., et al. Tiktalk: A videobased dialogue dataset for multi-modal chitchat in real world. In *ACM MM*, pp. 1303–1313, 2023. 3
- Lin, Y.-T. and Chen, Y.-N. LLM-eval: Unified multidimensional automatic evaluation for open-domain conversations with large language models. In *NLP4ConvAI*, pp. 47–58, July 2023. 3
- Liu, G., Wang, S., Yu, J., and Yin, J. A survey on multimodal dialogue systems: recent advances and new frontiers. In *AEMCSE*, pp. 845–853, 2022. 3
- Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. Llava-plus: Learning to use tools for creating multimodal agents. In ECCV, 2024. 2
- Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomput.*, 508:293–304, 2022. 8
- Maaz, M., Rasheed, H., Khan, S., and Khan, F. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, pp. 12585–12602, 2024. 2
- Mendonça, J., Pereira, P., Moniz, H., Paulo Carvalho, J., Lavie, A., and Trancoso, I. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. In *DSTC*, pp. 133–143, September 2023. 3
- Ni, J., Young, T., Pandelea, V., Xue, F., and Cambria, E. Recent advances in deep learning based dialogue systems: A systematic survey. *Artif Intell Rev*, 56(4):3055–3155, 2023. 2
- OpenAI. Hello gpt-4o, May 2024. https://openai.com/index/hello-gpt-4o/.6
- Ou, J., Lu, J., Liu, C., Tang, Y., Zhang, F., Zhang, D., and Gai, K. DialogBench: Evaluating LLMs as human-like dialogue systems. In *NAACL*, pp. 6137–6170, 2024. 3
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, pp. 27730–27744, 2022. 6
- Papinesi, K. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002. 3
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via largescale weak supervision. In *ICML*, pp. 28492–28518, 2023. 3, 6

- See, A., Roller, S., Kiela, D., and Weston, J. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*, pp. 1702–1723, 2019. 3
- Sellam, T., Das, D., and Parikh, A. BLEURT: Learning robust metrics for text generation. In *ACL*, pp. 7881–7892, 2020. 3
- Shen, L., Zhan, H., Shen, X., Song, Y., and Zhao, X. Text is not enough: Integrating visual impressions into opendomain dialogue generation. In *ACM MM*, pp. 4287–4296, 2021. 1
- Shuster, K., Humeau, S., Bordes, A., and Weston, J. Image-chat: Engaging grounded conversations. In *ACL*, pp. 2414–2429, 2020. 3
- Song, Z., Zheng, X., Liu, L., Xu, M., and Huang, X.-J. Generating responses with a specific emotion in dialog. In *ACL*, pp. 3685–3695, 2019. 2
- Sun, Q., Wang, Y., Xu, C., Zheng, K., Yang, Y., Hu, H., Xu, F., Zhang, J., Geng, X., and Jiang, D. Multimodal dialogue response generation. In *ACL*, pp. 2854–2866, 2022. 1, 2
- Sundar, A. and Heck, L. Multimodal conversational AI: A survey of datasets and approaches. In *NLP4ConvAI*, pp. 131–147, 2022. 3
- Tao, C., Mou, L., Zhao, D., and Yan, R. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In AAAI, volume 32, 2018. 3
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv*, 2023. 6
- Venkatesh, A., Khatri, C., Ram, A., et al. On evaluating and comparing open domain dialog systems. *arXiv*, 2018. 3
- Vinyals, O. and Le, Q. A neural conversational model. *arXiv*, 2015. 2
- Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*, pp. 2609–2634, July 2023. 4
- Wang, X., Zhang, Y., Zohar, O., and Yeung-Levy, S. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, 2024. 4
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *ICLRW*, 2024. 2

- Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S. K., and Feng, J. Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv, 2024. 2, 6
- Yang, Z., Wu, W., Hu, H., Xu, C., Wang, W., and Li, Z. Open domain dialogue generation with latent images. In *AAAI*, volume 35, pp. 14239–14247, 2021. 1, 2
- Yarats, D. and Lewis, M. Hierarchical text generation and planning for strategic dialogue. In *ICML*, pp. 5591–5599, 2018. 1
- Zhang, C., D'Haro, L. F., Chen, Y., Zhang, M., and Li, H. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *AAAI*, volume 38, pp. 19515–19524, 2024. 3
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi,Y. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019.
- Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., and Li, H. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *ACL*, pp. 5699–5710, May 2022. 3
- Zhao, T., Lala, D., and Kawahara, T. Designing precise and robust dialogue response evaluators. In *ACL*, pp. 26–33, July 2020. 3
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv*, 2023. 1
- Zheng, Y., Chen, G., Liu, X., and Sun, J. MMChat: Multi-modal chat dataset on social media. In *LREC*, pp. 5778–5786, June 2022. 3