FROM NATURAL ALIGNMENT TO CONDITIONAL CONTROLLABILITY IN MULTIMODAL DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent advancement of Artificial Intelligence Generated Content (AIGC) has led to significant strides in modeling human interaction, particularly in the context of multimodal dialogue. While current methods impressively generate realistic dialogue in speech and vision modalities, challenges remain in multimodal conditional dialogue generation. This paper focuses on the natural alignment between speech, vision, and text, aiming at expressive dialogue generation through multimodal conditional control. Since existing datasets lack the richness and diversity in dialogue expressiveness, we introduce a novel multi-modal dialogue annotation pipeline to exploit meaningful dialogues from movies and TV series with fine-grained annotations across multi-modalities. The resultant dataset, MM-DIA, provides over 360 hours and 54,700 dialogues, facilitating the Multi-modal Dialogue Generation task through explicit control over style-controllable dialogue speech synthesis. While the proposed benchmark, MM-DIA-BENCH, containing 309 dialogues that are highly expressive with visible dual/single speaker scenes, supporting the evaluation of implicit cross-modal control through downstream multi-modal dialogue generation tasks to assess the audio-visual style consistency across modalities. Our experiments demonstrate the effectiveness of our data in enhancing style controllability and reveal limitations in current frameworks' ability to replicate human interaction expressiveness, providing new insights and challenges for multi-modal conditional dialogue generation. Code, demo and data will be released at: https://mmdiaiclr26.github.io/mmdiaiclr26/.

1 INTRODUCTION

Dialogue has long been considered one of the most natural forms of human interaction, involving multiple communication channels such as text, speech, vision, gestures, and etc. In the AIGC era, multimodal dialogue has become increasingly important for a wide range of applications in *human–computer interaction*, *social computing*, and *film-making*.

Existing research in multimodal dialogue primarily falls into two directions: (1) Semantic generation, which emphasizes producing coherent and contextually appropriate responses, as in large-scale dialogue systems, e.g., ChatGPT (OpenAI et al.). (2) Modality rendering, which projects the given semantics into output modalities such as speech (Zhu et al., 2025; Zhang et al., 2024) and motion (Kong et al., 2025b). However, both directions over-emphasize the transmission of dialogue content, while neglecting systematic modeling of interaction style controllability, resulting in limited expressiveness and controllability of the generated outputs.

To achieve expressive and controllable multimodal dialogue generation, several key challenges have been raised: (1) **Lack of high-quality native multimodal dialogue data.** Existing large-scale multimodal dialogue datasets, as shown in Tab. 1, face limitations in data source diversity and modality coverage, hindering their ability to capture the full complexity of multimodal interactions and offering limited expressiveness and generalizability. (2) **Lack of scalable annotation methods for interaction-level semantics.** Collecting naturally occurring dialogues with synchronized text, audio, and visual modalities is costly and complex. Existing datasets such as MELD (Poria et al., 2019) and MC-EIU (Liu et al., 2024b) provide human-labeled categorical emotion or intent annotations, but they are costly, limited in scope, and not easily extensible, failing to capture the nuanced, continuous nature of human interactions. (3) **Lack of systematic benchmarks and evaluation pro-**

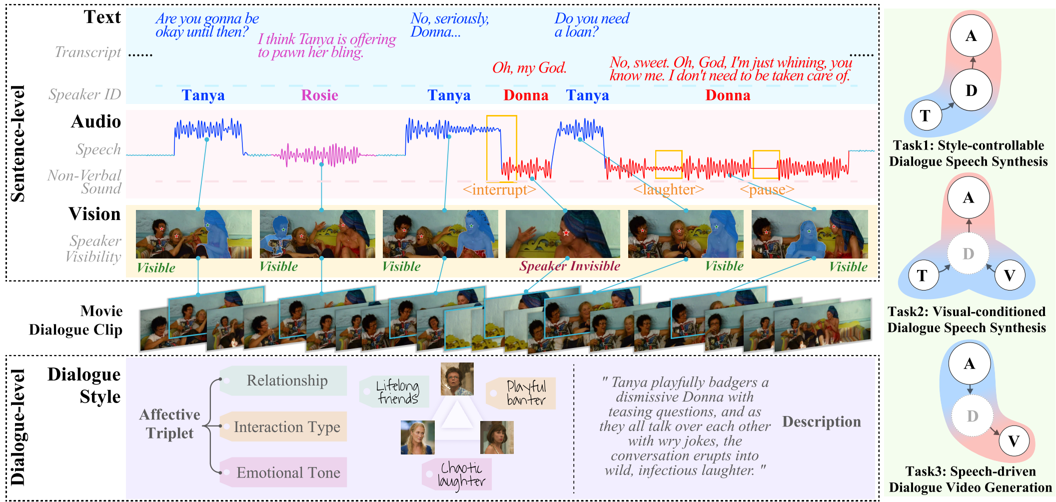


Figure 1: An example of a movie dialogue clip with sentence- and dialogue-level annotations in the MM-DIA and MM-DIA-BENCH datasets, highlighting multimodal dialogue interaction details. The right-hand side illustrates three examples dialogue-related cross-modal generation tasks involving text, audio, and vision, with both explicit (*Task 1*) and implicit control (*Task 2*, *Task 3*).

tocols. While existing tasks like semantic coherence and temporal alignment are well-established, new benchmarks and evaluation protocols tailored for emerging capabilities, such as dialogue-level controllability, are still lacking.

This paper seeks to address these gaps by constructing a large-scale expressive multimodal dialogue dataset, introducing new annotation paradigms, and establishing systematic benchmarks for controllable multimodal dialogue generation.

To compensate for the limited scale of high-quality multi-modal dialogue datasets, we develop an automatic data curation pipeline targeted for extracting dialogues with synchronized (text, audio, vision) streams and fine-grained interaction-level annotations, from in-the-wild movies and TV series. To resolve the challenges posed by complex scene transitions and audio-visual asynchrony,we devote special efforts to advancing dialogue boundary segmentation and multimodal speaker identification. To support controllability across diverse application scenarios, we define two complementary paradigms of "dialogue expressiveness": (1) **Affective Triplet**, consisting of *Relationship*, *Interaction Type*, and *Emotional State*, that jointly model role shaping, conversational dynamics, and emotional evolution; and (2) **Freestyle Description**, capturing per-speaker, turn-level style trajectories. Through extensive validation, we demonstrate that our pipeline achieves human-level quality in annotation consistency and reliability.

Applying the proposed data pipeline to over 700 hours of movies and TV series, we present a diverse, balanced, and interaction-rich multi-modal dialogue dataset, MM-DIA, which is characterized by 360.26 hours, 54,700 clips of highly expressive, contextually rich, and interaction-heavy dialogues. MM-DIA provides fine-grained annotation on various dialogue aspects, such as non-verbal sound, speaker identity and emotional dynamics at the individual and collective levels. To our best knowledge, MM-DIA is the first dataset to specifically center on dialogue expressiveness across multiple modalities.

Leveraging this dataset, we formally introduce Multimodal Dialogue Generation (MDG) as a conditional generation paradigm. Given multi-modal conversational context (text, audio, vision), generate multi-modal dialogue behaviors (one or more modalities) that not only ensure cross-modal alignment but also support conditional controllability with respect to interaction-level variables. To operationalize this controllability, we distinguish between two complementary forms: (1) *explicit control*, where style is specified through natural language prompts, and (2) *implicit control*, where conditions are conveyed through other modalities or structural cues.

For the explicit prompt control, we introduce the task of **Style-controllable Dialogue Speech Synthesis** (*Task 1*, as shown in Fig. 1), which directly supports generation of dialogue speech from the freestyle natural language description. With the supervised finetuning on MM-DIA data, current

Table 1: Comparison of the MM-DIA dataset with existing dialogue-related datasets across domain, scale, modality, annotation, and *open-source* (OS). Modality includes *text* ($\mathcal{T}$), *vision* ($\mathcal{V}$), and *audio* ($\mathcal{A}$), with audio-visual details on *speaker identity* (S-ID), *non-verbal annotations* (N-V), and *speaker visibility* (S-V).

Domain	Dataset	Scale			Modality			Audio-visual Details			Annotation		OS
		#Clip	#Utt.	#Dur.(h)	$\mathcal{T}$	$\mathcal{V}$	$\mathcal{A}$	S-ID	N-V	S-V	Granularity	Label	
Spoken Dia.	OpenDialogue (Zhu et al., 2025)	1M	6.5M	6.8K	✓	✗	✓	✓	✗	✗	Dialogue	None	✓
Textual Dia.	OpenViDial 2.0 (Wang et al., 2021)	-	5.6M	-	✓	✓	✗	✗	✗	✗	Dialogue	None	✓
	YTD-18M (Han et al., 2023)	18M	-	-	✓	✓	✗	✗	✗	✓	Dialogue	None	✓
Text-to-Video	OpenVid-1M (Nan et al., 2025)	1M	-	2.1K	✓	✓	✗	✗	✗	✗	Scene	Desc.	✓
	Captain Cinema (Xiao et al., 2025)	-	300K	500.0	✓	✓	✗	✓	✗	✗	Shot	Desc.	✗
MM Dia. Und.	MELD (Poria et al., 2019)	1.4K	14K	13.6	✓	✓	✓	✓	✗	✓	Sentence	Tag	✓
	MC-EIU (Liu et al., 2024b)	5.0K	56K	53.0	✓	✓	✓	✓	✗	✓	Sentence	Tag	✓
Movie Gen.	MovieBench (Wu et al., 2025)	16.0K	61K	69.2	✓	✓	✓	✓	✗	✗	Shot/Scene	Desc.	✓
MM Dia. Gen.	**MM-DIA (Ours)**	**54.7K**	**449K**	**360.3**	✓	✓	✓	✓	✓	✓	**Dia./Sent.**	**Desc./Tag**	✓
MM Dia. Gen.	**MM-DIA-BENCH (Ours)**	**309**	**1,851**	**1.7**	✓	✓	✓	✓	✓	✓	**Dia./Sent.**	**Desc./Tag**	✓

spoken dialogue models are able to generate high quality spoken dialogue with superior performance in intelligibility, speaker turn-taking accuracy, and emotional tone that adheres to the control of style instruction.

For the implicit cross-modal control, we introduce the following two tasks: (1) **Vision-conditioned Dialogue Speech** (*Task 2* in Fig. 1), which highlights the ability to generate coherent, contextually accurate speech aligned with turn-taking visual sequences. (2) **Speech-driven Dialogue Video Generation**, which focuses on generating videos capturing the essence of dialogue speech. Both of them require modeling implicit multimodal conditions and cross-modal generation, substantially increasing data demands and system complexity. This motivates us to introduce these tasks as open benchmarks for future research. Building on MM-DIA, we establish MM-DIA-BENCH, a diverse and balanced benchmark of 309 highly expressive dual-speaker dialogues with ensured speaker visibility. This benchmark is designed to evaluate style consistency in audio-visual communication throughout the dialogue turns, addressing a gap in traditional video evaluation, which often overlooks the assessment of cross-modal style consistency. Experiments reveal the limitations of current frameworks in audio-visual consistency when replicating the expressiveness of human interaction, offering new insights and challenges in cross-modal conditional dialogue generation.

2 RELATED WORKS

2.1 MULTIMODAL DIALOGUE DATASETS

In recent years, multimodal dialogue datasets have been pivotal for advancing research in multimodal AI systems. A significant number of existing dialogue datasets (Han et al., 2023; Zhu et al., 2025) provide valuable resources for training and evaluating dialogue systems. However, they primarily focus on single modality interactions, presenting challenges for further multimodal alignment and style control. In contrast, the web sourced video datasets (Ju et al., 2024; Wang et al., 2024; Nan et al., 2025) offer richer audio-visual data, but mainly feature casual chitchat or designated situational dialogues, limiting their diversity for flexible prompt control in multimodal dialogue. Similarly, movie-sourced video datasets (Han et al., 2024; Wu et al., 2025) offer a wealth of audiovisual content yet typically present unclear delineations of dialogue boundaries. To address these gaps, we introduces a novel multi-modal-based framework catering for dialogue-level style annotation, as shown in Tab. 2. Specifically, our approach focuses on synchronized audio-visual input instead of only key-frame image sequences ($\mathcal{I}$), contribute to the development of richer, more versatile datasets for advancing multimodal dialogue research.

Table 2: Comparison between MM-DIA and existing TV/Movie-sourced datasets in the annotation framework.

Dataset	Source	Segmentation	Anno. Input	Anno. Tool
MELD (Poria et al., 2019)	TV	Human	$\mathcal{V}+\mathcal{A}+\mathcal{T}$	Human
MC-EIU (Liu et al., 2024b)	TV	Human	$\mathcal{V}+\mathcal{A}+\mathcal{T}$	Human
MovieBench (Wu et al., 2025)	Movie	Vision-based	$\mathcal{I}+\mathcal{A}+\mathcal{T}$	GPT-4o
MM-DIA (Ours)	**TV/Mov.**	**Multi-modal**	$\mathcal{V}+\mathcal{I}+\mathcal{A}+\mathcal{T}$	**Gemini 2.5-pro**

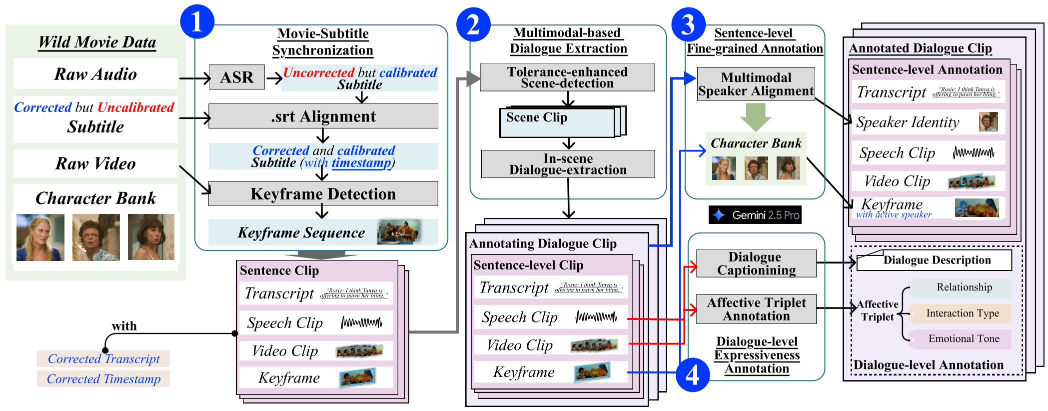


Figure 2: Framework of the Movie/TV-sourced in-the-wild data curation pipeline for multi-modal dialogue extraction with fine-grained interaction-level annotations.

2.2 Dialogue Generation from Multi-modality

Human interaction is fundamentally shaped by verbal exchanges, with dialogue serving as *the smallest and most structured unit* of social communication. In speech, recent advances in spoken dialogue generation (Labs, 2025; Ju et al., 2025; Zhu et al., 2025) capture realistic turn-taking and multi-speaker timbres (Boson AI, 2025), enabling more natural exchanges. In vision, progress in short movie generation (Xiao et al., 2025) supports high-fidelity multi-shot scenes, consistent character appearances (Liu et al., 2024a; Zhou et al., 2024), and immersive transitions (Blattmann et al., 2023; Zhang et al., 2023), producing coherent visual narratives. To enhance realism through synchronized facial movements, various talking video generation systems (Ki et al., 2025; Cui et al., 2025; Ji et al., 2025) are proposed. These advances establish the modality-specific foundations for conveying semantic information in dialogue. Yet, the challenge of flexibly controlling the interplay among speech, vision, and text for coherent and expressive multimodal multi-speaker dialogue remains largely unexplored. In this paper, we fill this gap by introducing a dataset and benchmarks for controllable, expressive multimodal dialogue.

3 MM-Dia: A Large-scale Expressive Multimodal Dialogue Dataset

Movies and TV series are two of the richest artistic forms that feature carefully crafted, context-sensitive performances. Dialogues from these sources exhibit stronger emotion, heightened tension, and greater resemblance to everyday interactions. However, the pursuit of strong cinematic sensory effects also poses challenges for data processing. Frequent background sounds, dramatic bursts, or ambiguous murmurs hinder the accuracy of automatic speech recognition (ASR), while artistic camera movements create complex audio-visual asynchrony, e.g., voiceovers or flashbacks, complicating dialogue boundary detection and speaker identification. As a result, a more cautious and comprehensive approach to the utilization of multi-modal information for various tasks is necessary, as introduced in following sections.

3.1 Pipeline Orientation with Data Preparation

In preparation of the dataset, we collect original movie & TV data from multiple public available sources, while the some of the official subtitle (SRT) files are unavailable. Although automatic speech recognition (ASR) can provide corrected time-stamped transcriptions of spoken content, the high word error rates associated with ASR, especially in complex movies and TV series, is unsatisfactory. To ensure high-quality subtitle due to the inherent trade-off between time correctness and content accuracy and further enlarge the dataset, we additionally crawled some multi-sourced uncalibrated subtitle files, combining them with ASR results to perform precise synchronization between the timestamps and the content. Selecting the matched specific ASR segments and subtitle entries as anchor points, we perform translation operations to adjust time and duration differences in the uncalibrated subtitle timestamps with minimal discrepancies. The qualified subtitle with low variance in discrepancy are double-checked by human to ensure usability. With the corrected timestamps from the calibrated subtitle, we extract the keyframe sequence from each subtitle line as representative for the upcoming dialogue boundary detection.

3.2 Multi-modal-based Dialogue Extraction

The automatic extraction of continuous dialogue from movies & TVs is challenging due to complex cinematic visuals. Dialogue boundaries often differ from shot or scene boundaries, as conversations may span multiple shots or shift within a single long scene. To address this, we introduce a tolerance-enhanced scene boundary detection method that first applies a Vision-Language Model (VLM) to identify scene continuity, followed by a Large Language Model (LLM) to refine in-scene dialogue boundaries.

Unlike traditional frame-to-frame matching methods (Wu et al., 2025; Xiao et al., 2025), our approach incorporates a buffer mechanism with a dynamic keyframe pool, allowing the model to bridge momentary visual disruptions such as rapid camera shifts, flashbacks, or perspective changes. This improves robustness in maintaining dialogue continuity across complex scenes. Based on the resulting scene-level segmentation, we further leverage subtitles and LLM-based semantic filtering to extract meaningful dialogue segments, particularly in long scenes exceeding 90 seconds. By combining visual and textual cues, the framework achieves coherent and accurate dialogue extraction, ensuring the integrity of multimodal context.

3.3 Sentence-level Fine-grained Annotation

Based on the dialogue boundaries determined by the previous two steps, we divide the movie into short dialogue segments. Next, we determine the attribution of the dialogue speech by assigning speaker identity to each line of dialogue. However, due to the unsatisfactory accuracy of speaker diarization in the audio modality, and the fact that movies and TV shows not always have visible speaker, visual modality-based active speaker detection is not very effective. Since it is difficult to accurately determine the speaker attribution for each line using only traditional automatic tools, we use Gemini-2.5-flash to assign the speaker based on the audio-visual synchronized video segments and dialogue subtitles. Geimini is prompted with the main character bank of the movie to recognize speakers, it will otherwise identify the speakers with their on-screen persona. Additionally, we label the non-verbal sounds or vocalizations during the dialogue process through this step to better capture the fine-grained details of dialogue expressiveness and context-related nuances. For downstream dialogue-related tasks like talking head generation, we further use the Insightface package to label the visibility of speakers that belong to the main characters in the corresponding keyframes.

3.4 Dialogue-level Expressiveness Annotation

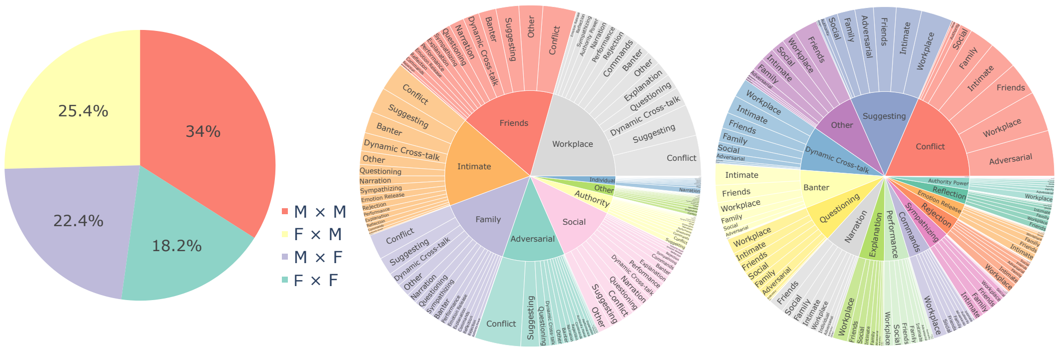
To enable systematic study of complex, interaction-level dialogue behaviors, we define the so-called "dialogue expressiveness" as what is consistent across modalities in a dialogue that makes it expressive beyond the semantic content. Two complementary paradigms of "dialogue expressiveness" are proposed:

(1) **Affective Triplet Control**, consisting of *Relationship*, *Interaction Type*, and *Emotional State*, that jointly model role shaping, conversational dynamics, and emotional evolution. It enables the precise control with the desired scenario of dialogue.

Table 3: Detailed statistics for MM-Dia and MM-Dia-Bench. Scored from Gemini/Human.

Statistic	MM-Dia	MM-Dia-Bench
Total Dialogues	54,700	309
Total Turns	449,138	1,851
Total Duration (h)	360.26	1.69
Avg. Spk. / Dia.	2.29	**2.00**
Avg. Dur. / Dia. (s)	23.71	19.69
Avg. Turns / Dia.	8.21	5.99
Avg. Dur. / Turn (s)	2.89	3.29
Avg. Turns / Spk. / Dia.	3.59	3.00
Avg. Rounds of Speaker Changes / Dia.	4.28	4.09
Speaker Visibility	Partial	**All**
Avg. Score on Emotion Intensity	6.76 / 5.22	**7.81 / 5.74**
Avg. Score on Volatility of Emotion Flow	5.32 / 4.36	**7.45 / 5.68**

(2) **Description Control**, capturing per-speaker, turn-level style trajectories. It enables the separate control over speakers, even the fine-grained emotion flow among the dialogue within the same speaker.



(a) (b) (c)

Figure 3: Distributions of (a) Dual-speaker Gender, (b) Relationship, and (c) Interaction Type among MM-DIA.

These paradigms cover common refined tag-based control as well as freestyle description-based natural language control forms. Given the speakers bank with the audio-visual synchronized video segments, we use Gemini-2.5-pro to annotate both paradigms of the dialogue expressiveness. To further quantify the abstract expressiveness, we also annotate the global emotional intensity of the dialogue as a whole and the local emotional volatility that occur at the level of individual speakers during the conversation. For instance, if a conversation remains consistently high-energy and intense throughout, the emotion intensity would be rated as high, while the emotion volatility would be low.

3.5 MM-DIA WITH MM-DIA-BENCH

Applying the data pipeline and annotation paradigms to over 700 hours data, including over 200 movies and 9 TV series, the resultant multi-modal dialogue dataset, MM-DIA, is characterized by 360.26 hours, 54,700 clips of highly expressive, contextually rich, and interaction-heavy dialogues, accompanied with fine-grained annotation on various dialogue aspects, such as non-verbal sound, speaker identity and emotional dynamics at the individual and collective levels. It is the first dataset to specifically center on dialogue-level expressiveness across multi-modalities. Fig. 3 further shows the balanced distribution of MM-DIA from multiple Affective triplet perspective. It is interesting to observe the corresponding connection between "Relationships" and "Interaction Type". For instance, the *Workplace* is the most common setting in happening *Commands* and *Questioning*, while people are more likely to engage in *Emotion Release* and *Banter* in *Intimate* relationships. The distribution further confirms the high consistency between the data and real-life distributions

Subsequently, we establish MM-DIA-BENCH, a diverse and balanced benchmark with carefully selected 309 instances of highly-expressive dual-speaker dialogues with assurance for speaker visibility. It meets the criteria of different kinds of downstream tasks in cross-modal dialogue generation. With the two invited annotators scoring on 100 random clips from each part of the data, as shown in Tab. 3, the results indicate that both Gemini and humans agree that MM-DIA-BENCH achieves a higher score in quantized dialogue expressiveness.

3.6 VALIDATION OF THE ANNOTATION SYSTEM AND DATASET

To validate the quality of the annotation system, we conduct a series of through evaluation (see Appendix. A.2) on each step component, demonstrating that our pipeline achieves human-level quality in annotation consistency and reliability.

4 MULTIMODAL DIALOGUE GENERATION TASKS

In this section, we first introduce a unified formulation of ***Multimodal Dialogue Generation* (MDG)**. Based on this framework, we then present three representative task definitions, each instantiating MDG under different control conditions and output modalities. These formulations establish the foundation for subsequent evaluation protocols and experiments.

4.1 Problem Formulation

To enable systematic study of multimodal dialogue behaviors, we formalize the task of MDG as a conditional generation problem. Given a multimodal conversational context $\mathcal{C} = \{c^{\text{text}}, c^{\text{audio}}, c^{\text{vision}}\}$, the goal is to generate multimodal dialogue behaviors $\mathcal{Y} = \{y^{\text{text}}, y^{\text{audio}}, y^{\text{vision}}\}$ that are (i) *semantically coherent* with the input context, (ii) *aligned across modalities*, and (iii) *controllable* with respect to interaction-level variables. Formally, MDG can be expressed as modeling a conditional distribution: $P(\mathcal{Y} \mid \mathcal{C}, \mathcal{Z})$, where $\mathcal{Z}$ denotes explicit/implicit control variables for dialogue style. This formulation unifies diverse downstream tasks such as style-controllable dialogue speech synthesis, keyframe-conditioned speech synthesis, and speech-driven dialogue video generation, providing a foundation for systematic benchmarking of controllable multimodal dialogue.

4.2 Task 1: Style-controllable dialogue speech synthesis

Definition. Given a dialogue transcript $T = \{c_1, c_2, \ldots, c_n\}$ and an explicit style condition $Z^{\text{exp}} \in (\{(c^{\mathcal{R}}, c^{\mathcal{I}}, c^{\mathcal{E}})\} \cup \mathcal{L}^*)$, i.e., either an Affective Triplet schema or a free-form natural language description, the goal is to synthesize a multi-speaker dialogue audio stream A: $A = f(Z^{\text{exp}}, T)$. Unlike conventional approaches that generate utterances turn by turn and concatenate them, we directly model A as a continuous dialogue speech sequence with embedded speaker changes but without explicit turn-taking boundaries, similar to *Zero-Shot Dialogue Generation* (ZSDG) (Zhang et al., 2024).

Challenges. Compared with conventional *Controllable Text-To-Speech* (CTTS) and ZSDG, our task presents several unique challenges: (i) generating a continuous single-pass end-to-end dialogue audio stream that naturally encodes rich multi-speaker interactions beyond turn-level concatenation; (ii) maintaining coherence and consistency across successive speakers, such as preserving role identity and interactional dynamics throughout the conversation; and (iii) supporting multi-level controllability, ranging from global conditions specified by structured triplets (e.g., relationship, interaction type, affective state) to fine-grained per-speaker expressive trajectories, such as emotional flow and intensity variation across dialogue turns.

4.3 Task 2: Vision-conditioned Dialogue Speech Synthesis

Definition. Let $I = \{I_1, \ldots, I_k\}$ be a temporally ordered sequence of keyframes that capture speakers' appearance, facial expressions, and scene context, together with temporal-aligned dialogue transcripts $T = \{T_1, \ldots, T_k\}$. The goal is to infer contextual style $S(I)$ from the visual sequence and generate multi-speaker dialogue speech A: $\hat{A} = g(S(I), T) = g(Z^{\text{imp}}, T)$, where $Z^{\text{imp}} = \psi(I)$ encodes implicit interaction-level conditions (e.g., relationship, interaction type, emotional state). This task instantiates MDG with $Y = \{\text{aud}\}$ under *implicit controllability*.

Challenges. Compared with explicit prompt-based control, this task requires the model to (i) reliably infer interactional variables from visual cues such as appearance, posture, and scene composition; (ii) capture temporal dependencies across the keyframe sequence to reflect evolving interactional dynamics in generated speech; and (iii) align inferred styles with textual content T so that the synthesized audio remains both semantically faithful and contextually expressive.

4.4 Task 3: Dialogue Video Generation

Definition. Given dialogue audio A and the corresponding transcript T, the objective is to synthesize a dialogue video $\hat{V}$ that is temporally synchronized with speech and affectively consistent with dialogue semantics: $\hat{V} = h(A, T, Z)$, where Z may include explicit style prompts or implicit cues inferred from prosody, turn-taking, and affective dynamics. This task instantiates MDG with $Y = \{\text{vis}\}$.

Challenges. Compared with text-to-video(T2V) tasks and single talking head generation, our task introduces three key challenges: (i) multi-speaker identity and scene continuity under rapid shot changes and partial visibility; (ii) multi-granularity audio–visual alignment—from lip–audio sync and utterance-level prosody/gesture to dialogue-level expressiveness (relationship, interaction type, affective state), often under weak/implicit control; and (iii) long-range cinematic reasoning to faithfully stage interactions (who, how, where), requiring shot planning and blocking beyond what standard quality or lip-sync metrics specify.

5 BENCHMARKING IN MULTIMODAL DIALOGUE GENERATION

In this section, we conduct several experiments to verify the effectiveness of MM-DIA and MM-DIA-BENCH on supporting the proposed multimodal generation tasks. Experiment results show that MM-DIA enables high-quality style-controllable spoken dialogue generation under explicit control, while MM-DIA-BENCH reveals key limitations of existing frameworks under implicit cross-modal control, offering new insights and challenges for future research.

5.1 EXPERIMENTS ON EXPLICIT CONTROL IN DIALOGUE SPEECH SYNTHESIS

A. Evaluation Settings.

1. Test sets: We prepared three test sets referred to as *Hard*, *Test*, and *Out-of-Domain* respectively. The *Hard* set is a superset of MM-DIA-BENCH containing 598 clips of highly-expressive data across MM-DIA. The remaining scope of MM-DIA is then randomly sampled into *Train*, *Valid* and *Test* by 90% : 5% : 5%. To further detect the generalizability, we curated another *Out-of-Domain* set with 60 clips of human-refined dialogue annotations. All experiment inference is conducted twice, taking the Description and Affective Triplet as style control for each.

2. Metrics: To evaluate the performance of the synthesized dialogue speech intrigued by MM-DIA, we established a dedicated evaluation from the speech, dialogue, and controllability-level.

Speech Quality: Word Error Rate (*WER*) and *UTMOS* (Takaaki et al., 2022) access the intelligibility and the overall quality of speech.

Dialogue Quality: Speaker Turn-Taking Accuracy (*cpWER*) and Speaker Aware Similarity (*saSIM*) respectively represent the intra-speaker similarity and inter-speaker timbre transition accuracy in spoken dialogue generation.

Expressiveness Controllability: Since there are no appropriate objective metrics to reflect the consistency between the text prompt and speech, we conduct subjective evaluation, including *Human-Mos Score* on the general quality and instruction-following capability. Inspired by MoonCast (Ju et al., 2025), we further involve *Gemini-as-Judge* for large quantities of nuance evaluation across Spontaneity, Coherence, Intelligibility, Quality, Timbre Similarity, and Instruction Following Capability. The Human Mos experiment Additionally, we calculate the mean recall accuracy on the label attributes of relationship and interaction type.

B. Baseline Models & Implementation Details.

To validate the effectiveness of MM-DIA, we perform supervised finetuning of pretrained backbones on our dataset with explicit style supervision, enabling controllability at both the global (triplet) and local (description) levels. We select two state-of-the-art pretrained backbones: Higgs-Audio-V2-Base (Boson AI, 2025) and Dia-1.6B (Labs, 2025). Both models support single-pass dialogue speech generation. Notably, Higgs-Audio-V2 allows flexible conditional inputs across multiple tasks, whereas Dia-1.6B is optimized for dialogue synthesis but does not natively support conditional inputs. To enable controllability, we introduce a lightweight adapter module that projects explicit style embeddings into Dia-1.6B's decoder.

C. Evaluation Results

Experimental results from Tab. 4 and Tab. 9 shows that spoken dialogue generation models outperform in generating high-quality style-controllable dialogue after supervised fine-tuning on MM-DIA while both tables exhibit a consistent trend. The supervised fine-tuning on Higgs-Audio-V2 successfully decreases the word error rate in the single-turn inference of multi-turn dialogue generation. The obvious reduction in cpWER indicates that the accuracy of dialogue tone conversion has significantly improved. In both subjective metrics and recall rate indicators, the models after SFT show notable advantages. These findings suggest that MM-DIA has helped the model generate more accurate and coherent dialogue while improving the ability to control styles effectively. We can also observe some slight reduction and sa-SIM metrics, suggesting the trade-off that while the

Table 4: Experimental results of Dialogue Speech Synthesis with **Description** as style prompt.

Model	Speech-Quality		Dialogue-Quality		Human-MOS		Gemini-as-Judge					
	WER↓	UTMOS↑	sa-SIM↑	cp-WER↓	Qual.↑	Instr. Follow.↑	Spont.↑	Coher.↑	Intellig.↑	Similar.↑	Qual.↑	Instr. Follow.↑
Dia-Base	19.991	2.272	0.389	51.713	$2.410_{\pm 0.940}$	$2.500_{\pm 0.890}$	3.993	4.335	4.446	3.738	4.248	3.807
Dia-SFT	29.071	1.974	0.447	57.813	$2.890_{\pm 0.690}$	$2.880_{\pm 0.710}$	3.626	4.071	4.171	3.590	3.971	3.598
Higgs-Audio-V2-Base	31.251	3.093	**0.475**	104.867	$3.580_{\pm 0.560}$	$3.110_{\pm 0.600}$	3.313	3.96	4.276	4.021	3.874	4.012
Higgs-Audio-V2-SFT	**4.450**	**3.280**	0.447	**33.765**	$\mathbf{4.440}_{\pm 0.290}$	$\mathbf{4.130}_{\pm 0.520}$	**4.277**	**4.881**	**4.965**	**4.640**	**4.851**	**4.707**

Table 5: Experimental results of Vision-conditioned Dialogue Speech Synthesis.

Model	Speech-Quality		Dialogue-Quality		Label-Recall	Gemini-as-Judge					
	WER↓	UTMOS↑	sa-SIM↑	cp-WER↓	Mean_acc ↑	Spont.↑	Coher.↑	Intellig.↑	Similar.↑	Qual.↑	Instr. Follow.↑
HarmoniVox	21.223	3.5704	0.62	30.981	40.47	1.790	3.390	4.238	1.657	1.895	2.410
Cascaded Gemini + Higgs	5.781	3.3245	0.499	16.267	42.33	3.081	4.129	4.927	2.605	3.21	3.347
Cascaded GPT + Higgs	5.793	3.4384	0.476	14.583	52.17	3.326	4.000	4.978	3.022	3.587	3.522

model has become better at generating dialogue with specific tones or styles, it might sacrifice some degree of generality or semantic accuracy in certain cases, since the domain shift in movie-sourced data brings challenges in preserving universal textual coherence and high-quality semantic fidelity.

5.2 EXPERIMENTS ON VISION-CONDITIONED DIALOGUE SPEECH SYNTHESIS

A. Evaluation Settings

1. Test sets: We use 132 clips from MM-DIA-BENCH, which guarantee single speaker visibility in each keyframe for the model to distinguish the utterance speaker.

2. Evaluation Metrics: Since Task 2 shares the same output paradigm as Task 1, we preserve most metrics while slightly modifying prompts for Gemini to compares the alignment in dialogue expressiveness between the speech and visual sequence.

B. Baseline Models & Implementation Details We implement several representative baseline models for comparison: (1) **HarmoniVox** (Zhou et al., 2025). This model implicitly infers the avatar's internal states from a visual image I, projects them into a talking style representation S, and then synthesizes speech audio A conditioned on S. We adopt sentence-level inference in our experiments and concatenate corresponding utterances into complete dialogue. (2) **Cascaded VLM + Higgs-Audio-SFT**. We employ a strong vision-language model (e.g., GPT-5, Gemini-2.5-pro) to first generate descriptive style prompts in human interaction from the visual dialogue context. These prompts are then cascaded into Higgs-Audio-V2-SFT for speech synthesis.

C. Evaluation Results

As shown in Tab. 5, although most data preserved stable performance in basic speech and dialogue metrics, the subjective score in Gemini-as-Judge appears to have a significant decline compared to the value. in Tab. 5. It mainly collapses into an uncontrollable spoken dialogue generation modal, but dismiss the style cue attached through the modality alignment. This initial experiment illustrates the limited capability of the existing frameworks in effectively interpreting the cross-modal human interaction style.

5.3 EXPERIMENTS ON DIALOGUE VIDEO GENERATION

Please refer to Appendix A.5 for detailed information.

6 CONCLUSION

In this paper, we propose MM-DIA, the first large-scale highly-expressive multi-modal dialogue dataset for the task of Multimodal Dialogue Generation, and the corresponding dual-speaker benchmark MM-DIA-BENCH for the evaluation of cross-modal conditional generation tasks. Experiments demonstrate that MM-DIA enhances the style controllability of dialogue generation model and MM-DIA-BENCH reveals the limitation in current cross-modal style consistency.

REFERENCES

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22563–22575, June 2023.

Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio, 2025. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In Asian Conference on Computer Vision, pp. 251–263. Springer, 2016.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21086–21095, 2025.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. Champagne: Learning real-world conversation from large-scale web videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15498–15509, October 2023.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel - back to the pixels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18164–18174, June 2024.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 193–203, 2025.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In Advances in Neural Information Processing Systems, volume 37, pp. 48955–48970. Curran Associates, Inc., 2024.

Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, Zhengtao Wang, Xu Tan, Xinyu Zhou, Tao Qin, and Xiangyang Li. MoonCast: High-quality zero-shot podcast generation, 2025. URL http://arxiv.org/abs/2503.14345.

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Float: Generative motion latent flow matching for audio-driven talking portrait, 2025. URL https://arxiv.org/abs/2412.01064.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025a. URL https://arxiv.org/abs/2412.03603.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation, 2025b. URL http://arxiv.org/abs/2505.22647.

N. Labs. Dia: A tts model capable of generating ultra-realistic dialogue in one pass. https://github.com/nari-labs/dia, 2025. Accessed: 2025-09-20.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6190–6200, June 2024a.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W. Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset, 2024b. URL http://arxiv.org/abs/2407.02751.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In Proceedings of the International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=j7kdXSrISM.

OpenAI, Josh Achiam, and Adler. GPT-4 technical report. URL http://arxiv.org/abs/2303.08774.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050.

Saeki Takaaki, Xin Detai, Nakata Wataru, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In Interspeech 2022, pp. 4521–4525, 2022.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges, 2019. URL https://arxiv.org/abs/1812.01717.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. OpenViDial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts, 2021. URL http://arxiv.org/abs/2109.12761.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In Proceedings of the International Conference on Learning Representations (ICLR), 2024. URL https://openreview.net/forum?id=MLBdiWu4Fw.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28984–28994, June 2025.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation, 2025. URL http://arxiv.org/abs/2507.18634.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, Sheng Zhao, and Michael Zeng. Covomix: Advancing zero-shot speech generation for human-like multi-talker conversations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 100291–100317. Curran Associates, Inc., 2024.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-XL: High-quality image-to-video synthesis via cascaded diffusion models, 2023. URL http://arxiv.org/abs/2311.04145.

Songtao Zhou, Xiaoyu Qin, Yixuan Zhou, Qixin Wang, Zeyu Jin, Zixuan Wang, Zhiyong Wu, and Jia Jia. Harmonivox: Painting voices to match the avatar's soul. In Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 110315–110340. Curran Associates, Inc., 2024.

Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Dong Zhang, Xin Zhang, Xingchen Song, Long Lin, and Daniel Povey. ZipVoice-dialog: Non-autoregressive spoken dialogue generation with flow matching, 2025. URL http://arxiv.org/abs/2507.09318.

Reproducibility Statement

We provide the MM-Dia dataset, a large-scale multimodal dialogue corpus, and the MM-Dia-Bench benchmark, both of which are integral to our research on style-controllable multimodal dialogue generation. Our experimental code and data curation pipeline will be made publicly available upon acceptance of the paper. The models and algorithms used in this paper can be reproduced using the provided dataset and benchmark, with all necessary details regarding model configurations, training procedures, and evaluation protocols included.

Ethics Statement

The MM-Dia and MM-Dia-Bench datasets include multimodal data sourced from movies and TV series, some of which may contain commercial content. We do not release the video or audio clips themselves; instead, we provide annotations (e.g., transcript, affective triplet, dialogue description, speaker identity, keyframe with active speaker, etc.d) and the methods used to generate them. Researchers are encouraged to obtain the corresponding media content independently and align it with the provided timestamps. For any further queries or information, readers are welcome to contact us.

We acknowledge the potential for biases inherent in the media content used and are committed to addressing these in future versions of the dataset by incorporating more diverse sources and refining our annotation methods.

LLM Usage Disclosure

We used GPT-5 for grammar checking and improving the clarity of sections 1 through 6 in this manuscript. All technical content, experimental design, and analysis are original human work. The LLM suggestions were manually reviewed and modified to ensure that they align with the paper's objectives and maintain technical accuracy.

A Appendix

A.1 Implementation Details for Dialogue Extraction

The automatic extraction of continuous dialogue in movies presents a challenge due to the inherent complexity of cinematic visuals. Notably, dialogue boundary is different from the shot or scene boundary. As a dialogue usually continues across multi-shot view, while multiple dialogues may happen sequentially in a long scene with the alternative in speaker composition or naturally change in topic. Therefore, we first introduce a tolerance-enhanced scene boundary detection method with Vision Language Model (VLM), following by Large Language Model (LLM) to determine the in-scene dialogue boundary.

Unlike static video content, movies often feature rapid camera shifts, inserted footage, and changes in perspectives. Traditional frame-to-frame scene continuation detection methods (Wu et al., 2025; Xiao et al., 2025), which are often based on direct visual comparison between adjacent frames, struggle to cope with these momentary disruptions, resulting in abrupt scene splits and false dialogue transitions. We introduce a buffer mechanism to dynamically update a keyframe pool of the current scene. Let $P = \{p_1, p_2, \dots, p_m\}$ represent the dynamic set of most representative keyframes from the current scene $S = \{s_{t-n}, \dots, s_{t-1}, s_t\}$, VLM uses the updated keyframes P to perform sparse comparisons of the similarity between the P and the frame s_{t+b} after a certain buffer interval b. Whenever the match fails, it falls back to the subsequent frame s_{t+1} through binary search. Once the match is successful, sparse comparisons start from the new end frame, recognizing the passed frames within the same scene. Meanwhile, the keyframe pool P is updated the by replacing a most similar frame within the pool with the new s_{t+b}.

$$S' = \{s_{t-n}, \dots, s_{t+b}\}, \quad P' = P \cup \{s_{t+b}\} \setminus \{p_{\text{most\_similar}}\}, \quad \text{if} \quad \text{VLM}(P, s_{t+b}) = \text{True}.$$

The buffer spanning multiple frames, together with the memory pool, enables the algorithm to ”bridge” temporary interruptions instead of triggering incorrect scene boundaries. This allows the algorithm to maintain the continuity of dialogue scenes over longer periods, providing greater resilience to the complex visual dynamics of movies.

With the resultant division from the vision modality at scene-level, we further extract relevant dialogue segments from the corresponding subtitle based on semantic meaning, especially to the long scene over 90 seconds. LLM is used to precisely extract meaningful dialogue within the correct scope. The framework effectively merges both visual and textual information to achieve robust dialogue extraction, ensuring the integrity and coherence of the dialogue context.

A.2 Validation results of the annotation system and dataset

1. Evaluation of the correctness in movie-subtitle synchronization.

With the *official* version of subtitle stands for the ground truth of content and human judgment on correctness of timestamps boundaries, the calibrated subtitle performs balanced in low word error with high time accuracy, successfully enlarge the dataset. Notably, both ASR and official subtitle tend to present the line slightly earlier than the actual time, while the start time is usually correct. As a result, we slightly extend the audio up to the next starting time in the subsequent training.

2. Evaluation on the buffer mechanism in boundary detection.

Firstly, we conduct human evaluation on a random sampled test set with six movies, with the reported boundary extraction accuracy to be 95.2%, comparing to 86.3% on the traditional frame-by-frame scene continuation detection methods.

As to the ablation study on the proposed buffer mechenism, inspired by the Intersection over Union (IoU) metric commonly used in *Object Detection*, we introduce a new metric called $F1_{Overlap}$ to represent the similarity between two continuous segmentation of a same sequence of clips, expressed as $\{A\}, \{B\}$:

Using A as the reference segmentation, for the n intervals in A, we take the corresponding interval in B that has the maximum overlap with it to calculate the percentage of the total overlapping duration of these n overlaps in A, denoted as $P(A, B)$. Formally, this can be written as: $P(A, B) =$

Table 6: TimeStamp accuracy and WER of different subtitle version.

Data Source	TimeStamp Accuracy	WER
ASR	**0.871**	0.34
SRT-Uncalibrated	0.179	0.43
SRT-Calibrated	0.857	0.03
SRT-Official	0.870	**0.00**

Table 7: Completeness and Hallucination of Dialogue Annotation from Qwen-72B, GPT-5 & Gemini-2.5-pro.

Annotation	Model	Comp. ↑	Hall. ↓
Non-verbal Sound	Qwen	1.25	2.12
	GPT	1.18	**1.00**
	Gemini	**4.66**	1.22
Affective Triplet	Qwen	3.45	2.56
	GPT	3.66	2.20
	Gemini	**4.76**	**1.38**
Description	Qwen	3.15	2.76
	GPT	3.60	2.16
	Gemini	**4.72**	**1.44**

Table 8: Ablation study on the buffer b with Qwen 7B and Qwen 72B model as VLM.

$F1\_Overlap$	b=1	2	3	4	5
Qwen 7B	0.771	0.866	0.841	0.839	0.836
Qwen 72B	0.947	0.975	0.977	0.978	**0.979**

$\frac{\sum_{i=1}^{n} \text{Overlap}(A_i, B_{\text{max}})}{\sum_{i=1}^{n} \text{Duration}(A_i)}$. Similarly, we reverse the roles of A and B to compute $P(B, A)$. The similarity between the two segmentations is then computed using the F1 score of $P(A, B)$ and $P(B, A)$: $F1_{Overlap} = 2 \times \frac{P(A,B) \times P(B,A)}{P(A,B)+P(B,A)}$. The $F1_{Overlap}$ metric prevents extreme segments, whether excessively dense or sparse, from receiving a high P score based on a single perspective. As shown in Tab. **??**, we leverage Qwen 72B with $b = 3$ to balance the time and performance.

3. Quality evaluation on the dialogue annotation.

Following MovieBench Wu et al. (2025), we invite two human annotators to evaluate the performance of Gemini annotation in the data curation pipeline, from the perspective of **Completeness** and **Hallucination**. Annotators are asked to score 1 to 5 for the three kinds of annotation of 100 randomly sampled movie/TV clips from MM-DIA. As indicated in Tab. 7, in comparision with Qwen 72B and GPT 5 (which instead takes sequential frames and audio as video input), Gemini outperforms in most aspects with the best interpretation of the movie style.

A.3 METRICS EXPLANATION IN TASK 1.

Speech Quality: ***WER, UTMOS***.

We used the official implmentation from Zhu et al. (2025) to compute Word Error Rate (WER) and UTMOS, accessing the intelligibility and the overall quality of speech.

Dialogue Quality: ***cpWER, saSIM***.

Speaker Turn-Taking Accuracy (cpWER) is computed by firstly concatenating all speech utterances by the same speaker after processing the speaker diarization to the generated spoken dialogue, then picking up the lowest WER among all the permutations of the generated transcripts with the concatenated ground truth.

Speaker Aware Similarity (saSIM) is acquired by computing mean speaker similarity among the permutations of each speaker's utterance after conducting the Montreal-Forced-Alignment.

A.4 EXPERIMENTAL RESULTS OF DIALOGUE SPEECH SYNTHESIS WITH **AFFECTIVE TRIPLET** AS STYLE PROMPT.

Table 9: Experimental results of Dialogue Speech Synthesis with **Affective Triplet** as style prompt.

Model	Speech-Quality		Dialogue-Quality		Label-Recall	Gemini-as-Judge					
	WER↓	UTMOS↑	sa-SIM↑	cp-WER↓	Mean_acc ↑	Spont.↑	Coher.↑	Intellig.↑	Similar.↑	Qual.↑	Instr. Follow.↑
Dia-Base	19.991	2.272	0.389	51.713	0.210	3.452	4.000	4.161	4.016	3.887	4.113
Dia-SFT	33.178	1.941	0.430	117.947	0.237	3.636	4.118	4.187	3.910	3.962	4.014
Higgs-Audio-V2-Base	39.684	3.066	**0.461**	75.847	0.352	3.169	3.816	4.075	3.843	3.704	3.850
Higgs-Audio-V2-SFT	**5.265**	**3.286**	0.459	**33.134**	**0.428**	**4.031**	**4.820**	**4.967**	**4.610**	**4.636**	**4.809**

A.5 Experiments on Dialogue video generation

A. Evaluation Settings

1. Test sets: We construct our evaluation splits from MM-Dia. We automatically screen dialogue clips and retain those with exactly one visible speaker in frame and an unobstructed face, yielding 133 dialogues. This set is curated to cover all annotated relationships and interaction types in MM-Dia, ensuring broad semantic coverage for cross-modal alignment assessment.

2. Evaluation Metrics:

We evaluate along three axes: *video quality* (Fréchet Video Distance, FVD (Unterthiner et al., 2019)), *lip-speech synchronization* (*LSE-C* and *LSE-D* (Chung & Zisserman, 2016)), and *cross-modal semantics/alignment*. We adopt the model-as-judge pipeline introduced in Sec. 5.1 to score *Spontaneity*, *Coherence*, *Intelligibility*, *Similarity*, *Overall Quality*, and *Instruction Following*, to quantify how well the generated dialogue videos align with the speech modality—from low-level timing (lip–speech sync) and utterance-level prosody/expressiveness to dialogue-level semantics (e.g., staging, flow, and instruction following). In addition, we report label accuracy/recall on *Relationship* and *Interaction Type* to test whether generated scenes faithfully reflect dialogue-level interpersonal semantics.

B. Baseline Models & Implementation Details

Because no system currently performs end-to-end dialogue-to-video generation, we evaluate two practical families:

- SI2V (Speaker-Image-to-Video). We split dialogue-level movie clips into sentence-level segments and drive the corresponding speaker images with each utterance, then concatenate per-sentence clips into dialogue videos. Given that SI2V models use reference keyframes, we do not evaluate relationship/scene accuracy here; we focus on lip sync and expressiveness alignment.
- T2V (Text-to-Video). Using sentence-level fine-grained and dialogue-level expressiveness annotations in MM-Dia, we construct rich text prompts to condition multi-speaker scene synthesis. Since audio is not explicitly input, we do not score lip sync for T2V; instead, we emphasize relationship/interaction and expressiveness alignment. During model-as-judge, we provide the corresponding audio for Gemini to evaluate the cross-modality alignment.

C. Evaluation Results

All experiments are conducted on MM-Dia-Bench dialogue clips with visible dyads and diverse expressiveness to ensure comparable shot complexity across systems.

Results in Tab. 10 show that no current system adequately solves dialogue video generation. Despite rich prompts, T2V models capture only a portion of high-level dialogue semantics; accurate staging of interaction scenes and who-interacts-with-whom remains unreliable. SI2V systems attain higher *Coherence*/*Intelligibility*/*Quality* on average, but *Instruction Following* and fine-grained *Spontaneity* alignment fluctuate across long dialogues.

To summarize, **SI2V** pipelines are complex and depend on keyframes; practical deployment will require coupling with keyframe generation to approach end-to-end usage. Additionally, small face extents and occlusions in natural dialogue shots make lip-sync brittle, often producing artifacts. Meanwhile, **T2V** systems lack explicit audio conditioning, making it difficult to synchronize with speech timing and match vocal expressiveness; they also underperform at faithfully reconstructing relationships and interaction patterns.

Table 10: Experimental results of Dialogue Video Synthesis.

Model	Visual-Quality	Lip-Sync		Label-Recall		Gemini-as-Judge					
	FVD↓	LSE-C↑	LSE-D↓	ACC-Rela. ↑	ACC-Interact.↑	Spont.↑	Coher.↑	Intellig.↑	Similar.↑	Qual.↑	Instr. Follow.↑
FLOAT (Ki et al., 2025)	572.187	4.805	9.502	-	-	2.703	2.405	3.050	3.339	2.248	3.050
MultiTalk (Kong et al., 2025b)	124.543	**5.305**	8.795	-	-	4.524	4.388	4.612	4.689	**4.922**	4.631
Sonic (Ji et al., 2025)	**117.096**	4.986	**8.503**	-	-	**4.592**	**4.583**	**4.750**	**4.800**	4.833	**4.750**
Wan-2.2 S2V (Wan et al., 2025)	154.261	4.288	9.873	-	-	4.205	4.116	4.357	4.589	4.652	4.384
HunyuanVideo (Kong et al., 2025a)	335.591	-	-	47.97%	13.82%	2.089	4.553	4.049	2.968	4.309	2.293
Wan-2.2 T2V (Wan et al., 2025)	300.092	-	-	**53.66%**	**18.70%**	3.114	4.634	4.602	3.732	4.423	3.268
Ground Truth	-	6.275	8.333	100.00%	100.00%	4.892	4.971	4.961	4.931	5.000	4.902

Overall, *neither* family is yet adequate for dialogue video generation. The results validate our benchmark design: quality and lip-sync alone are insufficient; cross-modal semantic alignment must be measured explicitly to drive progress. Future work should target: (1) End-to-end dialogue-to-video modeling that unifies keyframe planning, character visibility, lip/body sync, and scene continuity; (2) Multi-granularity alignment learning using sentence-level and dialogue-level expressiveness labels (relationship, interaction type, affect); (3) Cross-modal semantic discriminators that penalize misalignment during training; and (iv) Long-range dependence & shot planning for controllable staging in multi-speaker scenes, consistent with expressiveness schema in MM-DIA.

A.6 PSEUDO-CODE FOR MOVIE-SUBTITLE SYNCHRONIZATION

Algorithm 1 Subtitle Scene Segmentation with VLM

Require: Subtitle file srt, Video file video, Step size step, Buffer size buffer
Ensure: List of dialogue ranges

Load VLM model (Qwen2.5-VL-7B-Instruct) ParseScriptSrt
Extract subtitle blocks with index and timecode
return list of blocks ExtractFramevideo, timecode
Compute midpoint timestamp
Use ffmpeg to extract frame image
return image path IsContinuationframes
Prompt VLM with frames to check scene continuity
if last frame matches context **then**
    return True
else
    return False
end if
Initialize ranges list
for each block i **do**
    Try to extend range by comparing future blocks using VLM
    Allow up to step ahead, using up to buffer context frames
    if continuation fails **then**
        finalize current segment
    end if
end for
return ranges
Save ranges to JSON output

A.7 Explanation of Relationship and Interaction Type Categories

Table 11: Explanation of Relationship Categories with Typical Labels

Relationship	Explanation and Example
Workplace	Refers to professional relationships and environments, including people within a work setting. Example: *Colleague*, *Boss*, *Manager*, *Coworker*, *Client*.
Friends	A relationship between individuals characterized by mutual affection, trust, and companionship outside of family and work. Example: *Buddy*, *Pal*, *Companion*, *Mate*, *Peer*.
Intimate	Relationships of a more personal and romantic nature, typically involving emotional and physical closeness. Example: *Boyfriend*, *Girlfriend*, *Partner*, *Spouse*, *Fiancé*.
Family	Relationships defined by blood ties or marriage, including extended family members. Example: *Mother*, *Father*, *Sibling*, *Uncle*, *Cousin*.
Adversarial	Relationships characterized by opposition or conflict, often involving rivalry or animosity. Example: *Enemy*, *Opponent*, *Rival*, *Antagonist*, *Competitor*.
Individual	A relationship with oneself, or a solitary state where interaction with others is minimal or nonexistent. Example: *Solo*, *Loner*, *Isolated*, *Monologue*.
Social	Encompasses a wide range of social roles and interactions, from professional settings to casual encounters. Example: *Teacher*, *Doctor*, *Neighbor*, *Stranger*, *Host*, *Customer*.
Authority	Relationships based on power and control, typically involving leadership, governance, and decision-making. Example: *King*, *Judge*, *Mayor*, *President*, *General*.

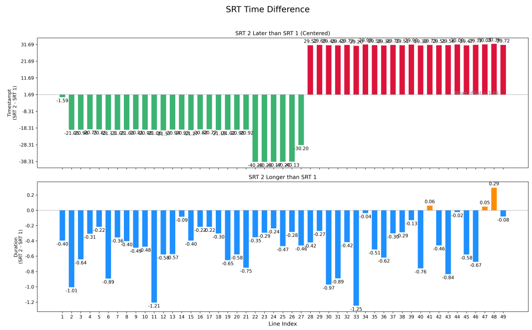
A.8 Explanation of Interaction Categories with Typical Labels

Table 12: Explanation of Interaction Categories with Typical Labels

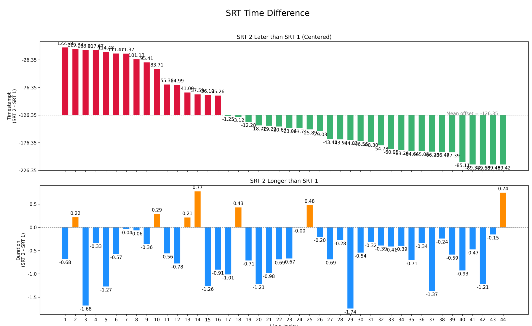
Interaction Type	Explanation and Example
Suggesting	The act of convincing someone to believe or do something through reasoning or emotional appeal. Example: *Persuasion*, *Convincing*, *Negotiation*.
Conflict	A state of disagreement or confrontation, often involving tension or hostility. Example: *Argument*, *Disagreement*, *Accusation*.
Questioning	Asking questions to gain information, clarify doubts, or provoke thought. Example: *Inquiry*, *Interrogation*, *Probing*.
Narration	The act of narrating a story or personal experience, often to entertain or inform. Example: *Storytelling*, *Flashback*, *Monologue*.
Explanation	Providing detailed information or clarification on a topic to ensure understanding. Example: *Justification*, *Diagnosis*, *Clarification*.
Commands	Issuing direct orders or instructions to prompt action. Example: *Orders*, *Demands*, *Instruction*.
Dynamic Cross-talk	A back-and-forth exchange of dynamic dialogue, often with interruptions or interjections. Example: *Interjection*, *Interruption*.
Sympathizing	Offering comfort or support to someone, often to alleviate concerns or anxiety. Example: *Comfort*, *Support*, *Encouragement*.
Rejection	Dismissing or refusing a request, idea, or proposal. Example: *Refusal*, *Dismissal*, *Avoidance*.
Banter	Playful, often teasing, interaction intended to entertain or create rapport. Example: *Teasing*, *Flirting*, *Joke*.
Authority Power	The use of authority or control to direct others' actions, often in a commanding or corrective manner. Example: *Domination*, *Criticism*, *Intervention*.
Performance	Delivering a structured or formal presentation, speech, or announcement to an audience. Example: *Presentation*, *Speech*, *Announcement*.
Reflection	Reflecting on one's thoughts, feelings, or experiences, often leading to a moment of realization. Example: *Introspection*, *Revelation*, *Discovery*.
Emotion Release	Expressing emotions, often related to frustration, anxiety, or relief. Example: *Venting*, *Confession*.
Invitation	Extending a request for someone to join an event or activity. Example: *Invitation*, *Offer*.

A.9 Typical Cases in Multi-sourced Movie Subtile Alignment

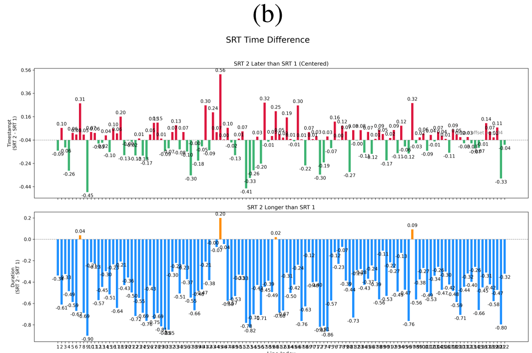
The unmatched subtitles are obvious through the time discrepancy sequences. The upper plot shows the start time discrepancy between anchor point start times in the subtitle and the ASR results. The lower plot shows the duration discrepancy.



(a)



(b)



(c)

Figure 4: (a) A bad case of Subtitle with edited movie segments. (b) A bad case of Subtitle with edited movie speed. and (c) A good case of Subtitle with potential usability with time translation.