LOLGORITHM: Funny Comment Generation Agent For Short Videos

Anonymous ACL submission

Abstract

This paper presents LOLGORITHM, a modular multi-agent system for stylized comment generation in short-form videos. The system employs video segmentation, emotion extraction, and stylized prompt construction to support six distinct comment styles: puns, rhymes, memes, sarcasm, humor, and content extraction. We construct a bilingual dataset covering both TikTok and YouTube platforms, encompassing five categories of popular videos. Experimental results demonstrate that LOLGORITHM outperforms existing methods in both automated metrics and human preference evaluations, achieving user preference rates exceeding 90% on TikTok and 87.55% on YouTube. This research provides a scalable framework for enhancing interactivity and cultural dissemination on short-video platforms.

1 Introduction

Short video platforms have become crucial channels for multimedia information dissemination, where comments play a vital role in content propagation, community interaction, and algorithmic feedback. With the development of generative AI, comments are transforming from auxiliary information to core driving forces in platform ecosystems.

Despite the capabilities of current multimodal large language models (MLLMs), they remain inadequate in generating stylistically diverse short video comments. Existing approaches focus either on video summarization, which lacks stylistic expressiveness, or live-streaming comment generation, which cannot adapt to the compact, information-dense nature of short videos. These methods generally lack support for multiple controllable styles and fail to align with platform-specific cultural and linguistic nuances.

We propose LOLGORITHM, a novel extensible modular multi-agent architecture for short video comment generation. The framework generates six controllable comment styles (puns, rhyming, memes, sarcasm, humor, and content extraction) using MLLMs as the generation backbone. The system comprises three main modules: video content extraction, classification, and comment generation. We construct a bilingual dataset from Douyin and YouTube, covering five high-engagement video types, to enhance contextual fit and community adaptability.

Our evaluation employs both automated metrics (originality, relevance, style conformity) and human preferences. Results show LOLGORITHM achieves leading scores on both platforms, with user preference rates exceeding 90% on Douyin and 87.55% on YouTube across 40 videos evaluated by 105 respondents.

The main contributions are: (1) the first modular multi-agent architecture specifically for short video comment generation; (2) a comprehensive bilingual dataset for training and evaluation; (3) empirical demonstration of effective style control and multimodal context modeling for generating platform-aligned comments. Code and data are provided as supplementary materials.

2 Related Work

2.1 Video Summarization

Apostolidis et al. (2021) surveyed over 40 deep learning-based video summarization methods, encompassing supervised, weakly supervised, unsupervised, and multimodal approaches. They proposed a unified classification framework, analyzing strategies in terms of extraction efficiency, contextual modeling, and scene adaptability. Otani et al. (2022) defined video summarization as extracting key information from long videos to improve viewing efficiency across domains, introducing a taxonomy based on Domains, Purposes, and Formats to guide standardization. Tonge and Thepade (2022) developed a static summarization method using

keyframe extraction and storyboard generation, improving narrative coherence and user engagement. Yao et al. (2022) proposed a multi-level spatiotemporal framework that jointly models frames, fragments, and shots, showing strong industrial applicability. Addressing data scarcity, Hua et al. (2024) introduced Instruct-V2Xum, a dataset of 30,000 YouTube videos (40–940s) with frame-referenced summaries and a 16.39% compression ratio, offering high semantic alignment for multimodal research.

2.2 Live-Streaming Danmaku Generation

Research on live-streaming danmaku generation has progressed from corpus construction to multimodal and personalized comment modeling. Ma et al. (2018) pioneered the field with a Bilibili corpus and the LiveBot framework. Wang et al. (2020) introduced the denser VideoIC dataset and MML-CG, a multimodal multitask model surpassing LiveBot in accuracy and diversity. Sun et al. (2023) incorporated "like" counts in ViCo-20k to reflect human preferences. Fang et al. (2020) broadened semantic scope via commonsense-grounded comment generation. Lalanne et al. (2023) proposed a Triple Transformer Encoder using visual, audio, and textual contexts, alongside the English LiveChat dataset. Luo et al. (2024) introduced socially engaging comment generation with a multimodal dataset annotated with "like" signals. Wu et al. (2024) integrated user preference modeling and video segmentation to enhance alignment and anthropomorphism. Lin et al. (2024) proposed PerVidCom for personalized video comment generation, emphasizing style migration and semantic alignment. Gao et al. (2023) focused on persona-driven interaction using a large-scale dialogue dataset with host speech, user comments, and profiles.

2.3 Research Gap

Despite advances in video summarization and live-streaming danmaku generation, directly applying these techniques to short video comment generation remains challenging. First, summarization methods focus on compressing core content, often overlooking user opinions, emotional cues, and social context—key elements that make comments engaging. Existing video summarization methods lack emotional naturalness and stylistic nuance, resulting in generated content that is overly monotonous, exhibits low style consistency, and lacks the interactivity that comments should possess. Although some research in live-streaming danmaku generation has explored personalization, current models heavily rely on historical user behavior within single-user, single-stream contexts, which limits generalization capability across users and content types, especially in cross-platform scenarios. Moreover, such frameworks exhibit weak robustness: when faced with short videos containing excessively large and complex amounts of information, they fail to generate comments and can only repeatedly output fixed replies from predefined templates. Furthermore, platform-specific designs hinder adaptability to varying content styles, linguistic norms, and interaction patterns across short video platforms. Finally, existing methods are mostly limited to a single language, lacking frameworks that can be used across multiple languages or on different language platforms. Additionally, there are no suitable datasets for short video comments, necessitating the creation of datasets required for training.

3 Methodology

LOLGORITHM is a novel, scalable, modular, multi-platform, and multilingual short video comment generation agent and framework. Currently, LOLGORITHM can generate Chinese and English comments for Douyin and YouTube platforms. This section provides a detailed description of LOLGORITHM through three components: the data preprocessing module, the dataset construction module, and the comment generation module. Figure 1 presents the complete workflow of the LOLGORITHM framework.

3.1 Data Preprocessing Module

The data preprocessing module provides structured semantic input for comment generation and serves a foundational role in dataset construction. This module consists of three sub-modules: information download, video processing, and video description. These components collaboratively complete the modal transformation from raw video to text description.

The workflow begins with creating a JSON file containing target video URLs, which serves as the entry point for data preprocessing. The information download module automatically downloads video content by calling official APIs from Douyin and YouTube. Videos are saved locally in MP4 format,

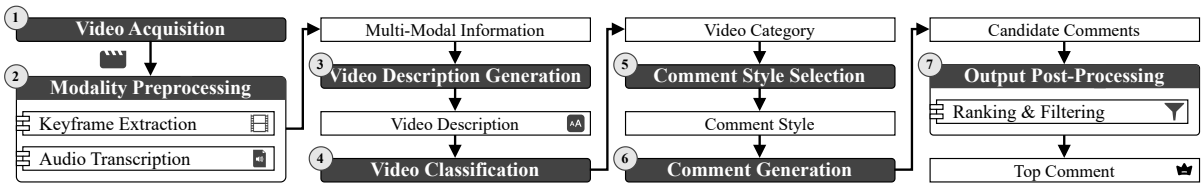

Figure 1: Workflow of the LOLGORITHM framework showing the three-stage pipeline: data preprocessing, dataset construction, and comment generation, with arrows indicating data flow between modules.

while video descriptions are stored as metadata in the JSON file, ensuring data integrity and providing semantic background for subsequent processing.

Downloaded videos are then processed by the video processing module, which transforms video content into analyzable textual information. The processing employs a dual detection mechanism using audio fluctuation and video light source variation to identify exciting segments. Let $A(t)$ represent the audio amplitude at time t and $L(t)$ represent the light intensity. The highlight score $H(t)$ at time t is computed as:

$$H(t) = \omega_a \cdot \left| \frac{dA(t)}{dt} \right| + \omega_l \cdot \left| \frac{dL(t)}{dt} \right| \quad (1)$$

where ω_a and ω_l are weights for audio and light variations respectively. A segment is identified as a highlight if $H(t) > \theta_h$, where θ_h is a predefined threshold.

After highlight detection, the system performs frame extraction with differential rates based on segment importance. The frame extraction rate $r(t)$ at time t is:

$$r(t) = \begin{cases} 10 \text{ fps} & \text{if } H(t) > \theta_h \text{ (highlight)} \\ 0.5 \text{ fps} & \text{otherwise (normal)} \end{cases} \quad (2)$$

Extracted images are saved and recorded in the JSON file along with audio transcription information from automatic speech recognition, forming complete timeline semantic data.

The video description module then processes the JSON file containing descriptions, transcriptions, and frame paths using a multimodal large language model. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ represent transcribed text segments and $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$ represent extracted frames. The video description D is generated by:

$$D = \text{LLM}_{\text{multi}}(\mathcal{T}, \mathcal{F}, \text{meta}) \quad (3)$$

where meta represents video metadata including title, description, and duration.

3.2 Dataset Construction Module

Current short video content generation research faces a critical bottleneck: the lack of high-quality, structured short video comment datasets, especially cross-platform Chinese-English bilingual datasets. This deficiency limits multilingual model training and evaluation while hindering comment generation system generalization and practicality globally.

We constructed a Chinese-English bilingual dataset for cross-platform short video comment research, with Chinese comments from Douyin and English comments from YouTube. These platforms represent the largest and most active short video platforms in their respective language systems, providing strong representativeness and coverage. This selection reflects user expression habits under different linguistic and cultural backgrounds while providing a solid corpus foundation for training multilingual comment generation models.

We selected 200 short videos as samples, with 100 each from Douyin and YouTube. To ensure content diversity and comment interactivity, we performed balanced sampling from five high-engagement video types: Talk Show (conversational content with hosts addressing audiences), Humorous Commentary (comedic narration on various topics), Funny Animal (amusing animal behaviors), Daily Life Jokes (humorous everyday situations), and Comedy Skits (short scripted performances). Each type has 20 videos selected per platform. The distribution is:

$$N_{p,c} = \frac{N_p}{|\mathcal{C}|} = \frac{100}{5} = 20 \quad (4)$$

where $N_{p,c}$ is the number of videos for platform p and category c, and $|\mathcal{C}| = 5$ is the number of categories.

Semantic descriptions of video content follow the data preprocessing module described above. We obtain videos and descriptions through the information download module, perform highlight detection, frame extraction, and audio transcription

through video processing, and generate structured content descriptions using a multimodal large language model. These descriptions provide semantic input for comment generation and constitute training data for video classification models, resulting in a semantic dataset of 200 video content descriptions.

Comments are obtained using platform APIs, downloading the top five highest-liked comments for each video. For video v_i with like count set $\mathcal{L}(v_i)$, we select the top-K comments where $K = 5$:

$$\mathcal{C}_s(v_i) = \{c_j \mid l_j \in \text{Top-K}(\mathcal{L}(v_i)), \quad j = 1, \ldots, K\} \tag{5}$$

where $\mathcal{C}_s$ denotes the selected comment set.

This strategy ensures comment representativeness and interactivity while maintaining legal compliance. The total comment count is:

$$N_{\text{comments}} = N_{\text{total}} \times K = 200 \times 5 = 1000 \tag{6}$$

We conducted manual annotation of all 1,000 comments, analyzing generation methods and language styles. The annotation was jointly completed by researchers with linguistics and short video analysis experience. Comments were categorized into six types: Puns/Homophones (creating humor using homophones), Rhyming (enhancing rhythm and readability), Meme Application (citing internet popular phrases), Sarcasm/Irony (satirical expressions), Plain Humor (direct humor without linguistic techniques), and Content Extraction (directly extracting video information). Let $\mathcal{S} = \{s_1, s_2, \ldots, s_6\}$ denote the style categories. Each comment c_i receives a style label:

$$\text{label}(c_i) \in \mathcal{S} \tag{7}$$

This systematic annotation provides language style labels and supervision signals for comment generation models. The resulting dataset possesses multilingual, multi-type, and multi-style characteristics, supporting various comment generation tasks including style transfer, semantic fit optimization, and cross-language generation.

3.3 Comment Generation Module

The comment generation module transforms video content into high-quality, style-matched original comments. This module consists of video classification and comment generation sub-modules, with tightly connected workflows relying on data preprocessing and semantic matching to ensure generated comments are relevant, stylistically consistent, and original.

The system first performs modal conversion on target videos through data preprocessing, transforming video content into structured text descriptions. This process extracts core semantic information including explicit elements (characters, scenes, actions) and implicit features (emotions, tone, rhythm), providing a rich semantic foundation for classification and generation.

The video content description enters the video classification module, which determines video category through semantic matching. The system preconstructs a dataset of video descriptions with clear categories. The classification module calculates semantic similarity between the target video and dataset samples using vectorized representations and cosine similarity. Let $\mathbf{v}_t$ denote the target video embedding and $\mathbf{v}_i$ the i-th sample embedding. Similarity is:

$$\text{sim}(\mathbf{v}_t, \mathbf{v}_i) = \frac{\mathbf{v}_t \cdot \mathbf{v}_i}{\|\mathbf{v}_t\| \|\mathbf{v}_i\|} \tag{8}$$

The category C is determined by:

$$C = \arg\max_{c \in \mathcal{C}} \sum_{i \in I_c} \text{sim}(\mathbf{v}_t, \mathbf{v}_i) \tag{9}$$

where $\mathcal{C}$ is the set of categories and I_c contains sample indices for category c.

Once categorized, the comment generation module screens style-matched comments from the corresponding category's comment dataset using a multi-layer voting mechanism. First, 100 comments are extracted and divided into 10 groups of 10 comments each. Let $\mathcal{C}_p$ denote the initial pool with $|\mathcal{C}_p| = 100$, partitioned as:

$$\mathcal{C}_p = \bigcup_{g=1}^{10} \mathcal{G}_g, \quad |\mathcal{G}_g| = 10 \tag{10}$$

The system evaluates style matching within each group, selecting the best-matched comment:

$$c_g^* = \arg\max_{c \in \mathcal{G}_g} S(c, v) \tag{11}$$

where $S(c, v)$ is the style matching score. Among the 10 candidates $\{c_1^*, c_2^*, \ldots, c_{10}^*\}$, the final style template is:

$$c^* = \arg\max_{c \in \{c_1^*, \ldots, c_{10}^*\}} S(c, v) \tag{12}$$

The style matching score combines structural, tonal, and length similarities. For comment c_j and video v:

$$S(c_j, v) = \alpha \cdot S_{st}(c_j, v) + \beta \cdot S_{tn}(c_j, v) + \gamma \cdot S_{ln}(c_j, v) \quad (13)$$

where S_{st}, S_{tn}, and S_{ln} denote structural, tonal, and length similarities respectively, with weights satisfying $\alpha + \beta + \gamma = 1$.

This style template serves as a reference structure rather than direct publication. The system inputs it with the video description to a large language model, instructing generation based on structure rather than semantic content. Let c_g denote the generated comment, D_v the video description, and $\mathcal{I}$ the instruction set:

$$c_g = \text{LLM}(D_v, c^*, \mathcal{I}) \quad (14)$$

This approach ensures consistent review style while creating novel and engaging content. Through collaborative stages including video understanding, semantic matching, style screening, and language generation, LOLGORITHM achieves automated transformation from video to high-quality comments, balancing style consistency with content originality.

4 Evaluation

We implemented LOLGORITHM with approximately 1,200 lines of Python code. The MLLM we used for the agents is GPT-4o. It is worth noting that in this paper, we do not focus on evaluating the performance of different MLLMs. Therefore, we evaluate only the performance of LOLGORITHM powered by GPT-4o. Any MLLMs with similar general capability as GPT-4o should allow LOLGORITHM to perform well.

Our evaluation aims to answer the following two research questions:

RQ1: Can LOLGORITHM generate comments similar to high-quality human-written comments?

RQ2: How does LOLGORITHM perform compared to existing approaches in terms of human preferences?

4.1 Evaluation Setup

4.1.1 Baseline Techniques

We compare LOLGORITHM against three representative baselines that span different approaches to comment generation:

Table 1: Details of the videos used for benchmarking. TS: Talk Show; HC: Humorous Commentary; FA: Funny Animal; DLJ: Daily Life Jokes; CS: Comedy Skits.

Platform	TS	HC	FA	DLJ	CS	Other
YouTube	3	3	3	4	3	4
Douyin	3	4	2	4	3	4

V2Xum-LLM (Hua et al., 2024) is a multimodal summarization model that focuses on extracting key information from video content to generate comments. This approach represents the content summarization paradigm in comment generation.

LiveChat (Gao et al., 2023) is a real-time interactive commentary system designed for live streaming scenarios. It represents the real-time interaction approach to comment generation.

GPT-4o (Direct) uses the prompt structure of LOLGORITHM without the multi-agent collaborative framework, serving as an ablation baseline to demonstrate the effectiveness of the multi-agent framework. This represents the direct large language model application approach.

These baselines represent different paradigms in comment generation: content summarization, real-time interaction, and direct large language model application.

4.1.2 Benchmark Videos

We constructed a dataset of 40 short videos crawled from Douyin (Chinese) and YouTube (English), with 20 videos from each platform. These short videos are different from those used in the tool's knowledge base. It is worth noting that we purposefully included several videos that do not fall into the categories of the tool's video classification to increase the challenge and simulate real-world scenarios, thereby evaluating how the tool performs on previously unseen types of videos. Table 1 shows the detailed information of the videos used for benchmarking. In addition to these videos, we also used the dataset construction module to collect five highly-rated comments for each video as reference standards for answering RQ1, and created a benchmark dataset for scoring in the automatic scoring framework.

4.1.3 Evaluation Metrics

To comprehensively assess the performance of LOLGORITHM, we adopt a dual evaluation strategy combining automatic scoring and human pref-

erence analysis. This approach ensures our system is benchmarked both computationally and socially, capturing both objective metrics and subjective user engagement potential.

Automatic Scoring Framework. We design an automatic scoring system that evaluates generated comments along three dimensions: originality (S_o), relevance (S_r), and style conformity (S_s). Each dimension has a maximum score of 10 points, and the final score is:

$$S_{\text{total}} = \frac{S_o + S_r + S_s}{3} \tag{15}$$

where $S_{\text{total}}, S_o, S_r, S_s \in [0, 10]$.

Originality (S_o): We evaluate originality by comparing the similarity between each generated comment c_i and existing comments. Let $\mathcal{D}_b$ and $\mathcal{D}_t$ denote benchmark and training datasets respectively, and $\mathcal{D} = \mathcal{D}_b \cup \mathcal{D}_t$. We compute:

$$m = \max\left(\max_{d \in \mathcal{D}} \text{sim}(c_i, d), \text{sim}(c_i, v_i)\right) \tag{16}$$

$$S_o = 10 \cdot (1 - m) \tag{17}$$

where m is the maximum similarity score and v_i is the target video content.

Relevance (S_r): Relevance measures appropriate connection with video content. We first compute a baseline b for human comments:

$$b = \frac{1}{|\mathcal{D}_b|} \sum_{(c_j, v_j) \in \mathcal{D}_b} \text{sim}(c_j, v_j) \tag{18}$$

For a generated comment c_i and video v_i:

$$S_r = 10 \cdot \exp\left(-\frac{|\text{sim}(c_i, v_i) - b|^2}{2\sigma^2}\right) \tag{19}$$

where σ is a scaling parameter.

Style Conformity (S_s): Style conformity combines length and sentiment matching. We found human-written English comments typically contain 63–72 words, while Chinese comments contain 25–35 characters. Using DistilBERT (Sanh et al., 2019) for English and BERT-Chinese (Devlin et al., 2019) for sentiment analysis:

$$S_s = S_l + S_{st} \tag{20}$$

where:

$$S_l = \begin{cases} 5 & L_{\min} \leq |c_i| \leq L_{\max} \\ 5e^{-\frac{(|c_i| - L_n)^2}{2\sigma_L^2}} & \text{else} \end{cases} \tag{21}$$

Table 2: Scoring results on Douyin dataset.

Model	Orig.	Rel.	Style	Total
V2Xum-LLM	1.58	8.24	1.10	3.64
LiveChat	1.56	7.49	5.44	4.83
LOLGORITHM	**2.11**	**8.48**	**6.07**	**5.55**
GPT-4o (Direct)	0.70	7.50	1.00	3.06

Table 3: Scoring results on YouTube dataset.

Model	Orig.	Rel.	Style	Total
V2Xum-LLM	3.12	**8.67**	1.70	4.50
LiveChat	3.08	4.90	**6.77**	4.92
LOLGORITHM	**3.26**	7.67	5.32	**5.42**
GPT-4o (Direct)	2.78	4.29	5.00	4.02

with $L_{\min}$, $L_{\max}$ as bounds and L_n as nearest bound, and:

$$S_{st} = \begin{cases} 5 & \text{sent}(c_i) = \text{sent}(v_i) \\ 0 & \text{else} \end{cases} \tag{22}$$

Human Preference Analysis. We conducted a manual evaluation using a multimodal anonymous questionnaire and this experiment has obtained ethical review approval. Each question consisted of a target video and four generated comments: one from each of the three baseline models and one from LOLGORITHM. Respondents watched the video and selected the comment they were most likely to engage with, simulating real-world platform interaction. Two questionnaires were administered to native speakers aged 20–50.

4.2 RQ1: Can LOLGORITHM generate comments similar to high-quality human-written comments?

To answer this research question, we conducted quantitative analysis of LOLGORITHM and baseline methods through our automatic evaluation framework. Table 2 and Table 3 present detailed scoring results on the Douyin and YouTube datasets, respectively.

As shown in Table 2, on the Douyin dataset, LOLGORITHM achieved the highest total score of 5.55, significantly outperforming all baseline methods. Specifically, LOLGORITHM excelled in the relevance dimension with a score of 8.48, surpassing other methods (V2Xum-LLM: 8.24, LiveChat: 7.49, GPT-4o: 7.50), demonstrating its superior ability to maintain appropriate topical connection with video content while balancing

the social interaction nature of short-video comments. In terms of style conformity, LOLGORITHM scored 6.07, substantially outperforming V2Xum-LLM (1.10) and GPT-4o (1.00), and exceeding LiveChat (5.44), indicating that its generated comments closely match the linguistic patterns, length requirements (25–35 characters for Chinese), and sentiment alignment expected on the Douyin platform. For originality, LOLGORITHM achieved 2.11, demonstrating moderate novelty while maintaining platform-appropriate style. The balanced performance across all three dimensions reflects LOLGORITHM's comprehensive approach to comment generation that neither sacrifices relevance for creativity nor conformity for originality.

On the YouTube dataset (Table 3), LOLGORITHM achieved the best overall performance with a total score of 5.42, outperforming LiveChat (4.92), V2Xum-LLM (4.50), and GPT-4o (4.02). LOLGORITHM demonstrated strong relevance with a score of 7.67, significantly exceeding LiveChat (4.90) and GPT-4o (4.29), though slightly lower than V2Xum-LLM's 8.67. This indicates LOLGORITHM's ability to generate contextually appropriate comments that connect meaningfully with video content. In originality, LOLGORITHM achieved the highest score of 3.26, surpassing V2Xum-LLM (3.12), LiveChat (3.08), and GPT-4o (2.78), showcasing its ability to generate fresh and distinctive comments. The style conformity score of 5.32 was competitive, positioned between LiveChat's leading 6.77 and V2Xum-LLM's lower 1.70, reflecting LOLGORITHM's strategy of balancing all three dimensions rather than optimizing for style alone. Notably, V2Xum-LLM's poor style conformity (1.70) despite high relevance (8.67) resulted in a lower total score (4.50), demonstrating the importance of balanced performance across all dimensions.

Overall, the automatic evaluation results demonstrate that LOLGORITHM has clear advantages in comment generation quality across both platforms, consistently achieving the highest total scores. Its multi-agent collaborative framework is particularly effective in maintaining strong relevance while generating original content and adapting to platform-specific style requirements. The results show that LOLGORITHM excels at balancing the three key dimensions—achieving high relevance scores (8.48 on Douyin, 7.67 on YouTube), the highest originality scores (2.11 on Douyin, 3.26 on YouTube), and strong style conformity (6.07 on Douyin, 5.32 on YouTube). These results validate that LOLGORITHM can generate comments similar to high-quality human-written comments, successfully balancing creativity, platform conformity, and topical connection.

4.3 RQ2: How does LOLGORITHM perform compared to existing approaches in terms of human preferences?

To evaluate LOLGORITHM's performance in terms of real user preferences, we conducted a large-scale human evaluation experiment. Figure 2 presents the human preference evaluation results on both Douyin and YouTube platforms.

The human evaluation results provide compelling evidence of our system's effectiveness. On the YouTube platform, with 51 valid responses collected across 20 video samples, LOLGORITHM was selected as the most preferred comment in 87.55% of cases, significantly outperforming other baseline methods: LiveChat received only 1.31% selection rate, V2Xum-LLM achieved 7.04%, and GPT-4o (Direct) obtained 4.10%. On the Douyin platform, the results were even more remarkable: through 56 valid responses across 20 video samples, LOLGORITHM achieved an overwhelming 92.41% selection rate, while all other methods performed below 5% (LiveChat: 1.96%, GPT-4o: 4.38%, V2Xum-LLM: 1.25%).

These results clearly demonstrate that LOLGORITHM has significant advantages in generating user-preferred comments, with its multi-agent collaborative framework better capturing the comment culture and user expectations of different platforms.

4.4 Case Study

Figure 3 further illustrates LOLGORITHM's advantages over baseline methods through specific YouTube and Douyin video cases.

In the YouTube case, LOLGORITHM generated concise and humorous content that perfectly matches YouTube users' interactive style. In contrast, GPT-4o (Direct) was overly lengthy and formal, V2Xum-LLM only provided simple video descriptions, while LiveChat lacked specificity.

In the Douyin case, LOLGORITHM cleverly used wordplay and internet slang, embodying the humorous style favored by Douyin users. GPT-4o (Direct), while rich in content, was overly formal and lengthy, not fitting Douyin's culture of concise and entertaining comments. V2Xum-LLM, though

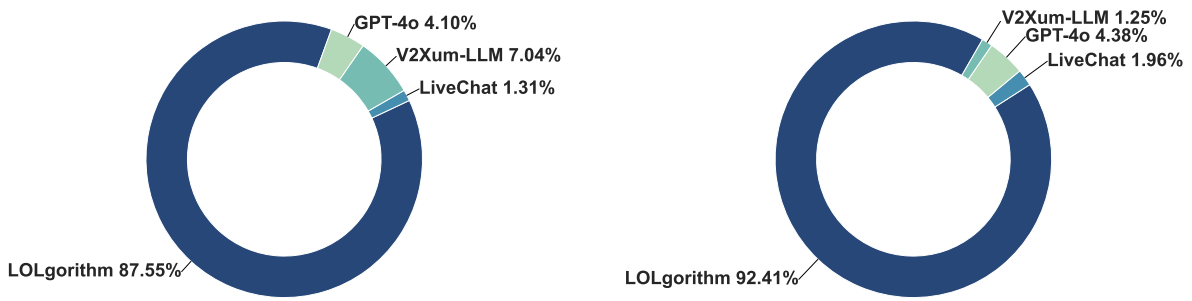

Figure 2: Human evaluation results for comment generation. (a) YouTube platform results. (b) Douyin platform results.

Figure 3: Example comments generated by different techniques. Top: YouTube example. Bottom: Douyin example.

containing emojis, lacked specific relevance to the video content, while LiveChat was completely irrelevant.

These cases clearly demonstrate LOLGORITHM's superior ability to understand platform-specific comment cultures and generate comments of appropriate length and style, explaining why it achieved such high selection rates in human preference evaluation.

5 Conclusion

LOLGORITHM is a modular multi-agent system for generating platform-compliant, stylistically diverse short-video comments, achieving over 90% user preference on Douyin and 87.55% on YouTube. It introduces a novel framework combining video understanding with style-aware generation, a bilingual dataset of 1,000 annotated samples, and balanced performance across originality, relevance, and style. This work lays the groundwork for scalable, culturally adaptive comment generation, with future plans for multilingual expansion, personalization, real-time deployment, and open-sourcing to support broader research.

6 Limitations

While LOLGORITHM demonstrates strong performance in generating stylistically diverse short-video comments, several limitations warrant discussion for future improvements.

Model Dependency and Cost. Our system relies heavily on GPT-4o as the core MLLM backbone, introducing substantial API costs and external service dependencies. We acknowledge that our evaluation focuses solely on GPT-4o without exploring alternative MLLMs, limiting our understanding of the framework's generalizability across different language models.

Language and Platform Coverage. LOLGORITHM currently supports only Chinese (Douyin) and English (YouTube), excluding other major short-video platforms such as TikTok International, Instagram Reels, and Kuaishou. This bilingual limitation prevents deployment in multilingual contexts where Spanish, Japanese, Korean, and other languages dominate user interactions.

Evaluation Scope. Our human evaluation involved 105 respondents and 40 videos, which may not adequately represent broader user demographics and preferences. Furthermore, we lack longitudinal studies on actual platform deployment, user engagement metrics, and long-term interaction pat-

terns that would validate real-world effectiveness.

Personalization and Real-time Constraints. The current architecture lacks user-level personalization capabilities and cannot adapt comment generation based on individual user histories or preferences. We have not addressed real-time deployment challenges, including latency requirements and computational efficiency for production environments.

Robustness and Adaptability. While our framework handles standard video content effectively, its performance on information-dense or highly complex videos remains underexplored. The system also lacks mechanisms for automatically adapting to emerging internet slang, cultural trends, and platform-specific linguistic evolution.

Cultural and Social Considerations. Although LOLGORITHM incorporates platform-specific cultural norms, it may not capture fine-grained regional variations and subcultural nuances. The handling of sensitive content, controversial topics, and potential harmful comments requires further investigation to ensure responsible deployment.

These limitations highlight important directions for future work, including expanding dataset scale, enhancing multilingual capabilities, implementing personalization mechanisms, and developing adaptive frameworks for evolving internet culture.

References

Evlampios Apostolidis, E. Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and I. Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109:1838–1863.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. In *Conference on Empirical Methods in Natural Language Processing*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Annual Meeting of the Association for Computational Linguistics*.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *AAAI Conference on Artificial Intelligence*.

Julien Lalanne, Raphael Bournet, and Yi Yu. 2023. Livechat: Video comment generation from audio-visual multimodal contexts. *ArXiv*, abs/2311.12826.

Xudong Lin, Ali Zare, Shiyuan Huang, Ming-Hsuan Yang, Shih-Fu Chang, and Li Zhang. 2024. Personalized video comment generation. In *Conference on Empirical Methods in Natural Language Processing*.

Ge Luo, Yuchen Ma, Manman Zhang, Junqiang Huang, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. 2024. Engaging live video comments generation. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2018. Livebot: Generating live video comments based on visual and textual contexts. In *AAAI Conference on Artificial Intelligence*.

Mayu Otani, Yale Song, and Yang Wang. 2022. Video summarization overview. *ArXiv*, abs/2210.11707.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yuchong Sun, Bei Liu, Xu Chen, Ruihua Song, and Jianlong Fu. 2023. Vico: Engaging video comment generation with human preference rewards. *Proceedings of the 6th ACM International Conference on Multimedia in Asia*.

Ashvini Tonge and Sudeep D. Thepade. 2022. A novel approach for static video content summarization using shot segmentation and k-means clustering. *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, pages 1–7.

Weiying Wang, Jieting Chen, and Qin Jin. 2020. Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. *Proceedings of the 28th ACM International Conference on Multimedia*.

Yihan Wu, Ruihua Song, Xu Chen, Hao Jiang, Zhao Cao, and Jin Yu. 2024. Understanding human preferences: Towards more personalized video to text generation. *Proceedings of the ACM Web Conference 2024*.

Mingyu Yao, Yu Bai, Wei Du, Xuejun Zhang, Heng Quan, Fuli Cai, and Hongwei Kang. 2022. Multi-level spatiotemporal network for video summarization. *Proceedings of the 30th ACM International Conference on Multimedia*.