GODBench: A Benchmark for Multimodal Large Language Models in Video Comment Art

Yiming Lei[1,3*] **Chenkai Zhang**[1,3*] **Zeming Liu**[2†] **Haitao Leng**[4♠] **Shaoguo Liu**[4]
Tingting Gao[4] **Qingjie Liu**[1,3†] **Yunhong Wang**[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2] School of Computer Science and Engineering, Beihang University, Beijing, China
[3] Hangzhou Innovation Institute, Beihang University, Hangzhou, China
[4] Kuaishou Technology
[*]Co-first authors: {ymlei, zhangchenkai}@buaa.edu.cn
[†]Corresponding authors [♠]Project Leader

Abstract

Video Comment Art enhances user engagement by providing creative content that conveys humor, satire, or emotional resonance, requiring a nuanced and comprehensive grasp of cultural and contextual subtleties. Although Multimodal Large Language Models (MLLMs) and Chain-of-Thought (CoT) have demonstrated strong reasoning abilities in STEM tasks (e.g. mathematics and coding), they still struggle to generate creative expressions such as resonant jokes and insightful satire. Moreover, existing benchmarks are constrained by their limited modalities and insufficient categories, hindering the exploration of comprehensive creativity in video-based *Comment Art* creation. To address these limitations, we introduce **GODBench**, a novel benchmark that integrates video and text modalities to systematically evaluate MLLMs' abilities to compose *Comment Art*. Furthermore, inspired by the propagation patterns of waves in physics, we propose **Ripple of Thought (RoT)**, a multi-step reasoning framework designed to enhance the creativity of MLLMs. Extensive experiments reveal that existing MLLMs and CoT methods still face significant challenges in understanding and generating creative video comments. In contrast, RoT provides an effective approach to improve creative composing, highlighting its potential to drive meaningful advancements in MLLM-based creativity. GODBench is publicly available at our GitHub repository.

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) (Wang et al., 2024b; Cheng et al., 2024; Zhang et al., 2024a; Chen et al., 2024d) have achieved remarkable success in structured reasoning tasks, especially with the integration of the Chain of Thought (CoT) framework (Kojima et al., 2022; Mitra et al., 2024; Wu et al., 2023; Yao et al., 2024a; Besta et al., 2024). However, despite their strong performance in logic-driven tasks, these models still struggle significantly in creative thinking (Zhao et al., 2024b; Tian et al., 2023; Chen and Ding, 2023a; Nair et al., 2024), remaining far from achieving human-level artistry. As illustrated in Fig. 1, MLLMs are constrained by rigid thinking, making it difficult to generate impressive and creative video comments akin to those of humans.

Figure 1: **Example from GODBench.** Showcasing a human-written GOD-level comment for the video, alongside the comments generated by model using the RoT framework and standard model. "#" indicates that the original text is in Chinese.

Video Comment Art, the practice of crafting creative and insightful comments on videos, requires not only a deep understanding of video content and cultural context but also the ability to think divergently and imaginatively express ideas. Current MLLMs still struggle to generate human-like creative content due to their inferiority in creative thinking, which involves making diverse connections and uncovering deep insights. To address this, several benchmarks (Liu et al., 2024; Xu et al., 2024; Zhong et al., 2024; Hessel et al., 2023; Sun et al., 2022) have been established to assess MLLMs' capabilities in *Comment Art*. However, existing benchmarks are either restricted to

[*]Work done during the internship at Kuaishou Technology.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11884–11952
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

a single category (e.g., humor, emotion, puns, or metaphors), confined to limited modalities (e.g., text-only or image-text pairs) or fail to account for the significance of comprehension in *Comment Art*.

To bridge this gap, we introduce **GODBench**, a novel and comprehensive multimodal benchmark dataset specifically designed to evaluate and advance the abilities of MLLMs in understanding and creating *Video Comment Art*. GODBench comprises over 67,000 high-quality videos paired with **"GOD-level comments"** that are characterized by their creativity and broad resonance. The videos in GODBench span 31 main categories and over 100 subcategories, ensuring rich diversity and comprehensiveness. These comments, created, voted on, and reviewed by real users, reflect human preferences, highly diverse thinking, and a deep understanding of video content, with quality endorsed by real users far surpassing that of previous benchmarks based on heuristic rules (Chen et al., 2024b; Sun et al., 2024). To facilitate comprehensive evaluation, we evaluate *Video Comment Art* through five core dimensions: `[Rhetorical Techniques]`, `[Divergent Associations]`, `[Clever Writing Techniques]`, `[Interactive Virality]`, and `[Emotional Resonance]`. Furthermore, we define two primary tasks: discrimination and generation based on over 40,000 evaluation items. These tasks span diverse formats, including selection, ranking, classification, explanation, and creation, providing a systematical assessment of MLLMs' ability to understand and generate creative video comments.

Inspired by the similarity between the "Aha moment" (Kounios et al., 2006) and the physical phenomena of wave diffusion and interference (Zakharov, 1968), we propose **Ripple of Thought (RoT)**, a novel reasoning framework that enables MLLMs to think more expansively in the knowledge space. Similar to the spread process of ripples, **RoT** leverages mechanisms such as diffusion and interference to enhance both the creativity and insightfulness of generated content. We conduct extensive experiments and analyses on 10 state-of-the-art MLLMs and observe that current MLLMs perform poorly in both discriminating and generating creative content, showing a notable performance gap compared to human capabilities. However, applying the **RoT** framework significantly improves their performance across various tasks, even outperforming human-generated comments in human preference evaluations. Despite these advancements, current MLLMs still fall short of human-level creativity, highlighting the immense potential for further research to enhance the creativity of MLLMs. The main contributions of this paper are summarized as follows:

(1) Innovative Reasoning Methodology: We propose the Ripple of Thought(RoT) reasoning framework to enhance the ability of MLLMs in generating *Video Comment Art*. Extensive experiments on state-of-the-art baselines demonstrate RoT's potential to improve the creativity of MLLMs.

(2) New Dataset and Benchmark: We construct GODBench, a large-scale and comprehensive dataset with diverse video-comment pairs, covering a wide range of tasks and significantly surpassing existing benchmarks in diversity and quality.

(3) Systematic and Comprehensive Evaluation: We establish a structured framework to assess *Video Comment Art* across five key creative dimensions and design a systematic set of evaluation tasks.

2 GODBench

We first compare GODBench with the previous benchmarks in Tab. 1. Then, we introduce the classification and annotation methods for the *Comment Art*. Finally, we presente the design methodology for the evaluation tasks.

2.1 Dataset Construction

The video and comment form a strongly correlated context-response pair that not only conveys user feedback but also reflects the cognitive and emotional responses generated after watching the videos. Therefore, high-quality video comments can be considered a form of *Comment Art*.

On the renowned video platform ***Kuaishou***[1], high-quality videos often feature a **GOD-level Comment**—a highly creative and engaging comment upvoted by millions of users, reviewed by moderators, and marked as "GOD-level" by the platform, reflecting exceptional insight and creativity. To study *Video Comment Art*, we crawled over 67,000 videos from *Kuaishou*, ensuring that each video includes a **GOD-level Comment**, as well as other **High-Quality Comments** and **Ordinary Comments**. All comments are accompanied by like counts to showcase human preferences. To facilitate lightweight experimentation and address computational constraints, we further construct a smaller subset (*minitest*) with a balanced distribution across categories and evaluation dimensions.

[1] https://www.kuaishou.cn

Benchmark	Videos			C-R Pairs		Dimensions of Comment Art					Task Types				
	Category	Len(s)	Num.	Num.	Re.	RT	DA	WT	IV	ER	SEL	RNK	CLS	EXP	CRE
TalkFunny (Chen et al., 2024c)	✗	✗	✗	4k	👤	⤫	✓	✗	✗	✗	✗	✗	✓	✗	✓
Chumor 2.0 (He et al., 2024)	✗	✗	✗	3k	🤖	⤫	✗	✗	✓	✗	✗	✗	✗	✓	✗
Puns (Xu et al., 2024)	✗	✗	✗	2k	👤	⤫	✗	✗	✗	✗	✗	✗	✓	✓	✓
II-Bench (Liu et al., 2024)	✗	✗	✗	1k	👤	⤫	✗	✗	✗	✓	✓	✗	✗	✗	✗
Oogiri-GO (Zhong et al., 2024)	✗	✗	✗	130k	👤	⤫	✓	✗	✗	✗	✓	✓	✗	✗	✓
NYT-Captions (Hessel et al., 2023)	✗	✗	✗	3k	👤	⤫	✓	✗	✗	✗	✓	✓	✗	✓	✗
ViCo (Sun et al., 2024)	15	-	20k	3M	🤖	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
HOTVCOM (Chen et al., 2024b)	20	96.44	93k	137M	🤖	⤫	✗	✗	✓	✗	✗	✗	✗	✗	✓
GODBench (Ours)	31	55.52	67k	1M	👤	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Comparison between GODBench and other existing benchmarks.** In the table, **Video Category** indicates the number of video Categories in the benchmark (e.g., pets, food, comedy, etc.), **Len(s)** represents the average video duration in seconds, and **Num.** represents the total number of videos. **C-R Pairs** refers to context-response pairs, **Num.** represents the quantity, and **Re.** represents whether the quality review was conducted by humans 👤 or LLMs 🤖. **Comment Art** is subdivided into five dimensions: **RT**, **DA**, **WT**, **IV**, and **ER** respectively represent **Rhetorical Techniques**, **Divergent Associations**, **Clever Writing Techniques**, **Interactive Virality**, and **Emotional Resonance**. "✗" indicates the absence of a dimension, while "⤫" indicates that only part of it is included. In the **Task Types**, **SEL**, **RNK**, **CLS**, **EXP**, and **CRE** respectively represent **Selection**, **Ranking**, **Classification**, **Explanation**, and **Creation**.

Detailed comparison with the full benchmark are provided in Appendix A.3.

2.2 Comment Art Definition and Annotation

In GODBench, the **GOD-level Comment** serves as the concrete manifestation of *Comment Art*. However, *Comment Art* is a broad and abstract concept. Previous benchmarks have primarily focused on sub-concepts of *Comment Art*, such as humor and metaphor, but lack a comprehensive analysis. To enable systematic study and proper evaluation metrics, we incorporated all related categories from previous work (Chen et al., 2024c; Zhong et al., 2024; Hessel et al., 2023; Liu et al., 2024; Chen et al., 2024b) and based on the characteristics of real-world data, partitioned *Comment Art* into five dimensions: **Rhetorical Techniques**(Tianli et al., 2022; Godioli and Chłopicki, 2024; Singsatit and Singhasiri, 2022; Aras et al., 2024), **Divergent Associations**(Bellemare-Pepin et al., 2024; Varshney et al., 2020; Beaty and Kenett, 2023), **Clever Writing Techniques**(Shalevska, 2024; Hoult et al., 2020), **Interactive Virality**(Wang and Hu, 2020; Lee, 2024; Huntington, 2013) and **Emotional Resonance**(Coburn, 2001; Heath et al., 2001).The specific results are shown in Fig. 2(Left).

1. Rhetorical Techniques: This dimension encompasses rhetorical techniques such as humor, puns, and metaphors, with a particular focus on creatively utilizing language to provoke novelty or surprise in the audience's perception.

2. Divergent Associations: This dimension demands a deep understanding of the video and multi-step reasoning. For instance, in the "Imaginary Completion" category, comments may introduce entirely fictional entities that do not appear in the video but are closely tied to its content, thereby enriching the comment's creativity and originality.

3. Clever Writing Techniques: This dimension is often overlooked, as *Comment Art* is expressed not only through content innovation but also through innovative writing structures and techniques. For instance, using poetic form, a comment can exhibit greater artistry and depth of expression.

4. Interactive Virality: This dimension demands the clever use of memes and cultural references in the appropriate context, requiring both background knowledge and expressive skills.

5. Emotional Resonance: This dimension requires comments expressing sincere feelings or having a strong emotional impact, capable of deeply touching the audience.

To ensure label accuracy for *Comment Art*, we employed professional annotators to conduct manual annotation. The full annotation procedure is detailed in Appendix B, where we describe the annotator training, qualification assessment, and multi-stage quality control process designed to ensure consistency and reliability.

2.3 Evaluation Procedure and Metrics

To comprehensively evaluate the video-based *Comment Art* capabilities of MLLMs, we follow the evaluation methods of previous work(Hessel et al., 2023; Zhong et al., 2024; Xu et al., 2024), and designed two main categories of tasks: **Discriminative Task** and **Generative Task**. These tasks are crafted by leveraging the real-world data we have collected, ensuring high quality and reliability. Specific examples can be found in Fig. 2(Right).

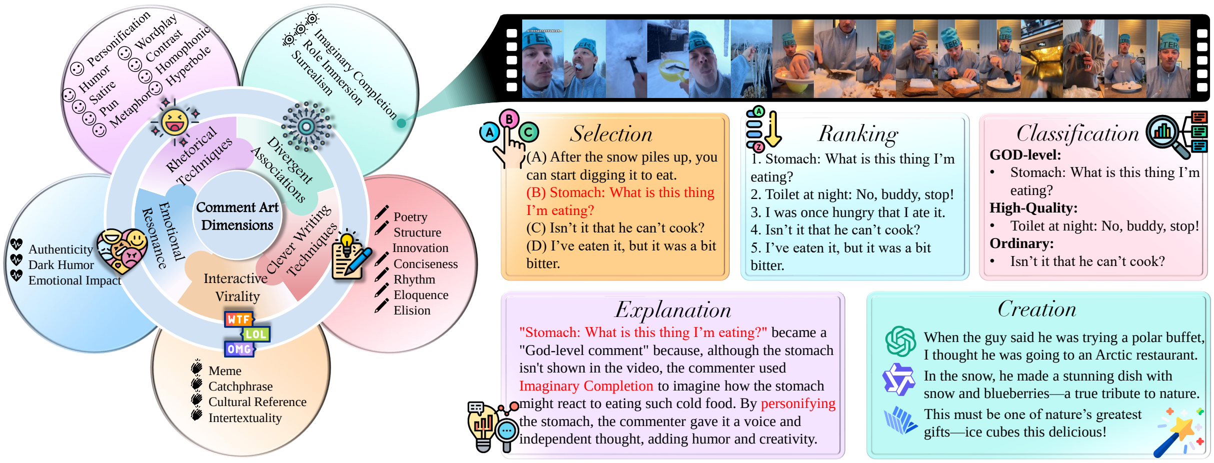


Figure 2: **The detailed definition of *Comment Art* and example of various tasks.** *Comment Art* is defined in five dimensions, each with different subcategories. One specific example of "Imaginary Completion" is presented, including the input video and various discriminative and generative tasks.

2.3.1 Discriminative Tasks

In GODBench, each video is associated with three types of comments: GOD-level Comments, High-Quality Comments, and Ordinary Comments, each annotated with real human upvote counts. Based on this data, we can accurately design three types of discriminative tasks: **Selection**, **Ranking**, and **Classification** tasks.

Selection: A multiple-choice question can be created by selecting one GOD-level Comment as the correct answer and using High-Quality and Ordinary Comments as distractors. In this task, MLLMs must choose the most creative and amazing comment based on the video content.

Ranking: Since all comments are linked to real human upvote counts, reflecting human preferences, we use these counts to rank the comments. The MLLMs must correctly order multiple comments according to their quality, evaluating their understanding of human preferences.

Classification Task: Due to GODBench containing comments of three different quality levels, MLLMs are required to classify multiple comments into three quality categories based on their quality.

2.3.2 Generative Tasks

Compared to discriminative tasks, generative tasks offer a more accurate measure of MLLMs' ability to produce *Comment Art*. Therefore, we have categorized these tasks into two distinct types: **Explanation** and **Creation**.

Explanation: Each **GOD-level Comment** is manually annotated with *Comment Art* dimensions. MLLMs must predict these dimensions and provide explanations, thereby evaluating their ability to analyze content and capture creative insights.

Creation: The ultimate challenge for MLLMs is generating a creative, contextually appropriate, and emotionally resonant comment based on the video content, serving as the most direct evaluation of the MLLMs' *Comment Art* capabilities.

3 Ripple of Thought

To enhance the *Comment Art* capabilities of MLLMs, inspired by the similarity between the "Aha moment"(Kounios et al., 2006) and the physical phenomena of wave diffusion and interference(Zakharov, 1968), we propose the Ripple of Thought (RoT) framework. This framework emulates the divergence and convergence of human thought, allowing RoT to expand reasoning in a ripple-like manner and enabling MLLMs to generate outputs that are both creative and meaningful. This framework consists of five key phases: **Ripple Initiation**, **Ripple Focalization**, **Ripple Diffusion**, **Wave Interference**, and **Luminous Imprint**, with the specific structure shown in Fig. 3. The detailed implementation can be found in Appendix E.

3.1 Ripple Initiation

A stone cast into water creates the first ripple, symbolizing the initial spark that sets off all subsequent processes. Similarly, when a video or prompt is fed into MLLMs, it triggers the first cognitive oscillations. Thus, the initial step involves a thorough analysis of this "stone", allowing the MLLMs to fully understand and interpret the input video.

We employ a three-layer analysis method for the video. First, **Basic Analysis** applies OCR, subtitle extraction, and caption generation to capture video details. Next, **Intermediate Analysis** iden-

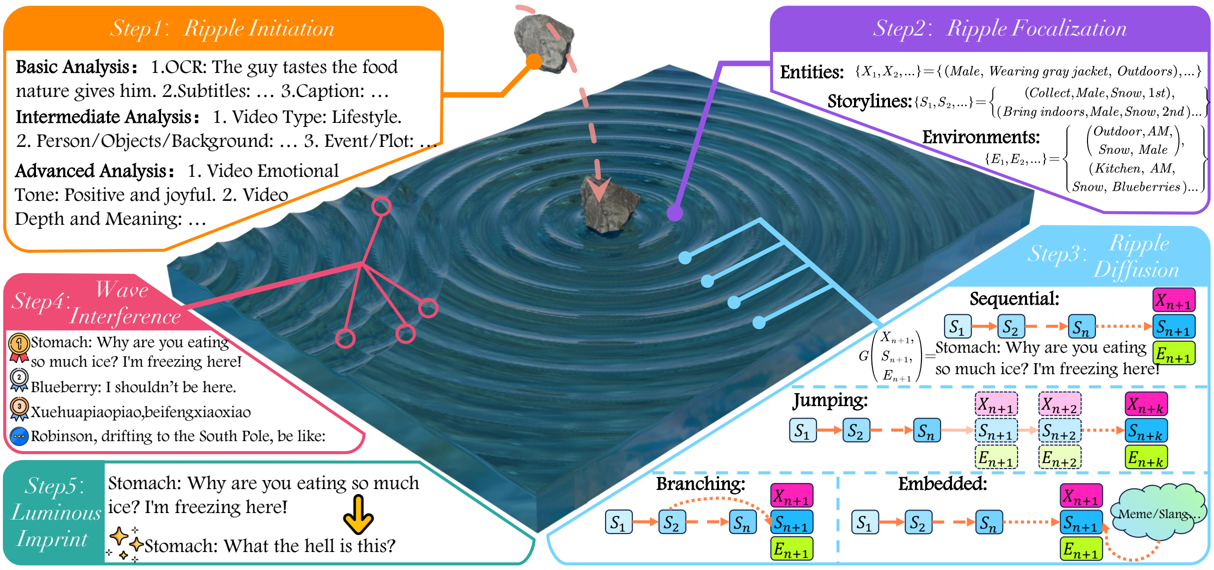


Figure 3: **Illustration of RoT.** Human creative thinking is like the diffusion of ripples, breaking down the propagation of waves in physics into five components, which are then transferred to the RoT reasoning framework of MLLMs.

tifies video types, characters, objects, and event sequences, laying the groundwork for deeper reasoning. Finally, **Advanced Analysis** examines the video's emotional tone, cultural context, and social values to achieve a comprehensive understanding.

3.2 Ripple Focalization

When a stone strikes the water, the initial ripple—carrying the highest concentration of energy—sets the stage for all subsequent waves. Similarly, MLLMs must focus on the core entities, storylines, and environments extracted from the analyzed video information. This foundational step shapes the entire reasoning process, much like the initial ripple determines the evolving pattern of the waves. To achieve this, we use a formatted description formula to represent the entities X, storylines S, and environments E as a unified expression.

3.3 Ripple Diffusion

The model's reasoning process mirrors ripples on a water surface, with each wave of thought spreading outward to form broader connections. Entities X, storylines S, and environments E continuously spark new connections, gradually extending to new content. This ripple-like diffusion unfolds along four distinct pathways, categorized as follows:

(1) Sequential Association: Based on the extracted multi-entity set, the model infers the next most relevant event or entity by following the logical order of the storyline.

(2) Jumping Association: Expanding on sequential association, the model performs additional reasoning steps to discover seemingly unrelated yet inherently connected entities, leading to unexpected but insightful creative associations.

(3) Branching Association: Unlike sequential inference, branching association detaches specific extracted entities that may have been overlooked, recombining them into novel concepts.

(4) Embedded Association: To improve the coherence and cultural relevance of generated content, it is essential to first deduce the relevant cultural context and trending memes from the video, and then seamlessly incorporate them into the output.

3.4 Wave Interference

In this phase, the process mirrors the natural phenomenon of wave interference, where interacting ripples on a water surface partially cancel out while others reinforce, forming the strongest centers. Similarly, MLLMs must refine and prioritize the multitude of associative possibilities generated in earlier stages. This process performs an internal ranking of the multiple associative results, selecting the most creative, relevant, engaging, and resonant expressions, ensuring that the final output is both high-quality and thematically consistent.

3.5 Luminous Imprint

This phase requires MLLMs to refine and optimize the filtered content, enhancing clarity, coherence, and contextual relevance to generate the final comment. Just as ripples eventually stabilize into a distinctive luminous pattern on the water's surface, the output must retain the dynamic essence of thought while leaving a lasting impression.

Model	Size	Frames	$S_{\text{acc}}^{[1,1,1]}$	$S_{\text{top-2}}^{[1,1,1]}$	$S_{\text{acc}}^{[1,3]}$	$S_{\text{acc}}^{[1,12]}$	$R_{\text{NDCG}}^{[1,4]}$	$R_{\text{EMA}}^{[1,4]}$	$C_{\text{acc}}^{[1,3,5]}$	$C_{\text{EMA}}^{[1,3,5]}$
Random Choice	-	-	33.40	66.58	24.65	7.88	63.18	0.87	46.22	0.37
Frequent Guess	-	-	33.76	66.70	26.02	8.41	62.75	1.11	59.25	0.00
Human	-	-	84.21	95.18	70.59	42.86	79.01	10.53	70.37	11.11
Open-source Video MLLMs										
LLaVA-Video	7B	64	36.11	77.24	23.78	10.04	49.17	0.41	38.46	0.13
mPLUG-Owl3	7B	128	34.75	72.18	24.39	9.77	62.73	0.63	36.68	0.11
MiniCPM-V 2.6	8B	64	41.30	78.83	28.38	11.45	58.73	0.39	41.12	0.00
MiniCPM-o 2.6	8B	64	40.60	78.12	27.61	11.42	53.16	0.66	41.39	0.04
Qwen2-VL	2B	dyn.	37.45	73.39	25.73	8.09	48.19	0.27	28.64	0.00
	7B		45.75	84.06	30.24	13.33	62.98	0.77	38.37	0.09
InternVL2.5	8B	32	44.27	81.54	28.61	13.19	45.43	0.52	43.41	0.20
	20B		45.59	82.07	31.17	13.28	54.14	0.77	43.99	0.21
Commercial MLLMs										
GPT-4o-mini	~8B	50	44.91	81.87	28.99	13.83	62.86	1.43	38.44	0.13
GPT-4o	~200B	50	54.19	88.32	37.86	18.84	65.21	1.52	53.16	0.68
MLLMs after Supervised Fine-Tuning										
Qwen2-VL†	7B	dyn.	66.24	90.26	50.32	30.41	71.45	3.07	64.37	8.13
InternVL2.5†	7B	32	70.27	91.02	54.63	33.57	74.81	4.96	69.53	10.12

Table 2: **Performance of MLLMs on Discriminative Tasks.** Size means the LLM size. EMA is Exact Match Accuracy. Results are reported in percentage (%). †: MLLMs fine-tuned with LoRA. The best, second-best, and third-best results are marked purple, orange, and gray, respectively.

4 Experiment

4.1 Experimental Setups

Metrics. All tasks introduced in Sec. 2.3 follow a [1, m, n] configuration, where each video/image set includes **one** GOD-level comment, m high-quality comments, and n ordinary comments. To comprehensively evaluate model performance, we employed three types of assessment: *1) Automatic Evaluation*: For discriminative tasks, we used accuracy and Normalized Discounted Cumulative Gain (NDCG) (Wang et al., 2013) to measure model performance. For generative tasks, we employed automatic metrics including BLEU, DIST, ROUGE, and F1_{BERT} similar to HOTVCOM (Chen et al., 2024b), to assess textual relevance and quality. *2) LLM-Based Judgement*: Following the procedure of *LLM-as-a-judge* (Aymeric Roucher, 2024), we employed GPT-4o as a judging model, utilizing a graded scoring system that references human-annotated answers for multi-dimensional assessment. *3) Human Evaluation*: To further validate the quality of the generated content, evaluators ranked the comments based on multiple quality dimensions to ensure alignment with human preferences. Further details on the evaluation can be found in Appendix C and D.3.

Model Selection. We conduct a comprehensive experiments on **GODBench** using both open-source and closed-source Video-MLLMs: (a) open-source models with different parameters: LLaVA-Video (Zhang et al., 2024b), MiniCPM-V 2.6, MiniCPM-o 2.6 (Yao et al., 2024b), mPLUG-Owl3 (Ye et al., 2024), InternVL2.5 (Chen et al., 2024d), and Qwen2-VL (Wang et al., 2024b); (b) commercial MLLMs, such as GPT-4o (OpenAI, 2024). Further details on the models and evaluation settings are provided in the Appendix D.2.

Implementation and Inference. We constructed a fine-tuning dataset on discriminative tasks from the training set and used Llama-Factory (Zheng et al., 2024) to fine-tune Qwen2-VL and InternVL2.5(see Appendix D.2.1). For inference, we followed the official inference and frame extraction configurations similar to Video-MME (Fu et al., 2024). **RoT** is applied to the following two LLMs: (1) Qwen2-VL, and (2) InternVL2.5. For the 5-shot tasks, we randomly select ten videos from the training set that share the same category as the target video. Then, we rank their corresponding GOD-level comments based on rules and retain the top five as context(see Appendix D.2.2).

4.2 Results and Analysis

To evaluate the ability of MLLMs in understanding and creating *Video Comment Art*, we first propose several key research questions (RQs) and address them individually through quantitative experimental results, including (1) the ability to identify *artistic comments* precisely, (2) the comprehension of deep conceptual aspects of *Comment Art*, and (3) the capacity to compose *Video Comment Art*.

RQ1: Can MLLMs precisely distinguish *artistic comments* from ordinary ones? As shown by their significantly higher $S_{\text{top-2}}^{[1,1,1]}$ score compared to random baseline in Tab. 2, MLLMs can effectively distinguish ordinary content from high-quality content. However, their performance remains limited when compared to humans, especially in tasks requiring fine-grained discrimination between high-quality comments and GOD-level comments. We attribute this limitation to the inherent rigidity of MLLMs to *Comment Art Appreciation*. In contrast, MLLMs fine-tuned with LoRA show notable improvement in distinguishing creative comments, indicating that **GODBench** effectively enhances their conceptual understanding of *Comment Art*.

RQ2: How well do MLLMs understand deep conceptual aspects of *Comment Art*? We evaluated the accuracy of MLLMs in the tag discrimination task and employed GPT-4o to assess tag explanation based on five criteria: *Precision, Reasonableness, Completeness, Relevance, and Clarity*. The results in Tab. 3 reveal several key findings: *1)* Most MLLMs struggle with accurately choosing tags for GOD-level comments, which is fundamental to understanding the deep conceptual aspects of *Comment Art*. *2)* MLLMs exhibit limitations in understanding deep semantic and cultural contexts. While some MLLMs can correctly select tags for comments, their explanations for tagging often deviate significantly from human interpretation.

Model	OA	Tag Discrimination					$S_{\text{GPT-4o}}$
		RT	DA	WT	IV	ER	
LLaVA-Video	16.9	45.3	0.1	0.0	0.5	12.3	108.8
mPLUG-Owl3	21.0	54.8	1.1	1.1	0.7	14.2	159.8
MiniCPM-V 2.6	28.3	57.9	12.3	7.0	9.8	14.2	162.5
MiniCPM-o 2.6	26.2	58.0	10.0	2.2	8.2	1.2	171.5
GPT-4o	29.7	47.8	17.5	17.1	11.3	58.1	214.3
Qwen2-VL$_{2B}$	19.9	55.4	0.0	1.1	0.0	0.0	177.8
Qwen2-VL$_{7B}$	20.3	51.3	0.7	7.0	1.2	26.9	190.6
+CoT	17.2	36.5	4.0	7.2	6.9	27.6	166.5
+CCoT	19.2	49.3	0.1	7.2	1.2	24.0	189.9
+RoT(Ours)	47.3	58.2	63.9	7.0	9.2	22.3	229.0
InternVL2.5$_{26B}$	19.8	37.5	6.3	11.8	11.7	37.9	199.8
InternVL2.5$_{7B}$	22.2	35.0	12.9	13.5	18.3	33.2	184.8
+CoT	28.2	57.5	7.0	15.0	13.3	55.7	197.8
+CCoT	24.5	41.5	9.2	16.7	26.8	54.6	208.0
+RoT(Ours)	46.8	40.8	55.6	18.1	28.4	57.9	222.1

Table 3: **Performance of MLLMs on Tagging and Explanation Tasks.** Results include accuracy (%) for tag discrimination and GPT-4o scores for tag explanation. OA represents the average value of the five tags.

RQ3: Are MLLMs capable of composing *Video Comment Art*? The results of automated metrics and LLM-based evaluations are presented in Tab. 4, leading to the following key conclusions. First, we find that RoT performs significantly better than most baselines and even outperforms larger MLLMs on all automated metrics. Second, we evaluate Qwen2-VL and InternVL2.5 under different settings and observe that larger sizes and more examples lead to better performance. While CoT and CCoT provide some improvements, their performance fails to provide a distinct advantage over 5-shot settings. Finally, we use GPT-4o to comprehensively evaluate the generated comments based on *Creativity, Quality, Style, and Impact* (see Appendix D.3 for details of criteria). RoT exhibits a notable increase in scores, enhancing the quality of generated creative comments. Interestingly, GPT-4o exhibits a consistent preference for its own generated content, even when there is no significant improvement in human preference.

Model	BLEU-1	BLEU-2	DIST-1	ROUGE-L	$F1_{\text{BERT}}$	$S_{\text{GPT-4o}}$
LLaVA-Video	0.23	0.05	0.37	4.41	52.78	322.82
mPLUG-Owl3	2.0	0.38	1.75	5.28	53.29	359.34
MiniCPM-V 2.6	4.41	1.02	6.24	6.68	55.32	298.67
MiniCPM-o 2.6	2.79	0.65	4.08	5.77	54.54	298.18
GPT-4o	6.36	1.37	11.86	6.88	56.21	425.99
Qwen2-VL$_{2B}$	4.72	1.14	4.99	6.84	55.45	316.87
Qwen2-VL$_{7B}$	7.41	2.21	8.04	8.32	57.31	332.65
5-shot	8.73	3.02	8.46	10.41	58.28	352.55
+CoT	8.76	2.76	8.79	8.87	57.14	383.01
+CCoT	9.19	3.04	8.85	9.07	57.36	389.08
+RoT(Ours)	9.51	3.34	9.02	11.33	58.92	389.68
InternVL2.5$_{26B}$	5.32	1.13	9.01	6.54	55.81	353.64
InternVL2.5$_{8B}$	3.40	0.62	5.95	5.01	53.98	342.23
5-shot	5.63	1.36	8.85	7.49	56.19	347.45
+CoT	6.07	1.41	6.86	7.42	55.71	379.85
+CCoT	4.87	1.24	6.85	6.72	55.32	377.67
+RoT(Ours)	9.23	3.39	7.73	10.44	56.97	388.59

Table 4: **Performance of MLLMs on comment creation results.** Results are reported in percentage(%). $S_{\text{GPT-4o}}$ denotes the quality score judged by GPT-4o.

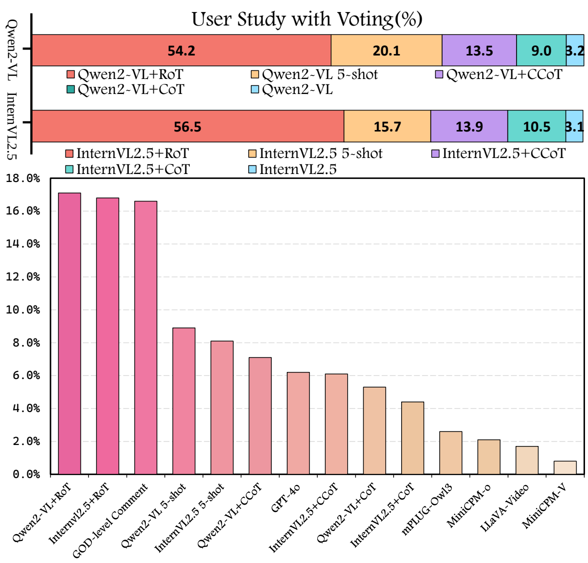


Figure 4: **User study with voting(%) by different models and improved methods.**

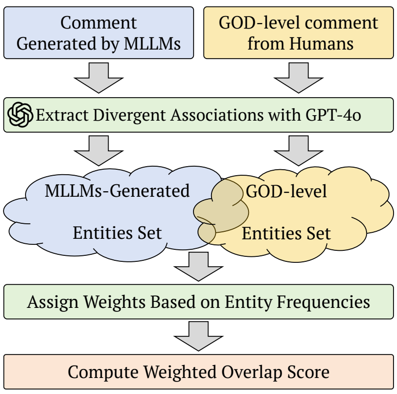
We also conduct a human preference study to evaluate the creativity of generated comments, asking users to select the most creative ones. As shown in Fig. 4, users show a strong preference for comments generated by RoT, which are comparable to, or even exceeding GOD-level comments from humans, highlighting RoT's ability to produce high-quality creative comments. Further details on the human preference study are in Appendix C.

4.3 More Analysis

To further emphasize the exceptional creativity of **RoT**, we designed divergent association tasks that focus on entity innovation and developed a Weighted Entity Overlap(WEO) metric to assess the creativity of MLLMs. Specifically, we utilized GPT-4o to extract entities from both GOD-level comments and model-generated comments, denoted as E_{gen} and E_{ref}, respectively. Next, we assigned a weight w_e to each entity e based on its frequency. Finally, we computed the score of WEO, formulated as:

$$\text{WEO} = \frac{1 + \sum_{e \in E_{\text{gen}} \cap E_{\text{ref}}} w_e}{\sum_{e \in E_{\text{gen}} \cup E_{\text{ref}}} w_e} \quad (1)$$

Fig. 5 illustrates the stages and results of the WEO score. Due to its outstanding divergent association ability, RoT exhibits greater overlap with GOD-level comments, achieving superior performance and offering a novel perspective for advancing generative model capabilities.



Model	WEO
LLaVA-Video	12.10
mPLUG-Owl3	15.11
MiniCPM-V 2.6	18.22
MiniCPM-o 2.6	14.89
GPT-4o	22.21
Qwen2-VL$_{2B}$	19.91
Qwen2-VL$_{7B}$	24.69
5-shot	27.19
+CoT	30.09
+CCoT	30.57
+RoT(Ours)	32.83
InternVL2.5$_{26B}$	18.18
InternVL2.5$_{8B}$	12.67
5-shot	19.00
+CoT	23.94
+CCoT	21.49
+RoT(Ours)	36.78

Figure 5: **Procedure and results of WEO score on divergent association tasks in GODBench.**

5 Related Work

Multimodal LLMs for Creative Arts. Recent advances in Multimodal Large Language Models (MLLMs) (Wang et al., 2024b; Chen et al., 2024d; OpenAI, 2024) have significantly enhanced their logical reasoning capabilities, driven by techniques such as Chain-of-Thought (CoT) prompting (Mitra et al., 2024) and reasoning (OpenAI, 2024; Zhao et al., 2024a). However, their application to creative artistic tasks mainly focuses on either literary creation (Chen et al., 2024a; Chakrabarty et al., 2024) or surface-level humor (Zhong et al., 2024) and puns (Xu et al., 2024), remaining confined to a restricted subset of creative thinking within *Comment Art* while lacking attention to video-based multimodal creativity. Therefore, we propose **Ripple of Thought (RoT)**, a novel reasoning framework enabling MLLMs to compose more creative, imaginative, and engaging video comments.

Evaluation of Creativity in LLMs. *Video Comment Art* is the practice of crafting insightful, humorous, and culturally resonant comments to enhance engagement and enrich the viewing experience. Prior research on creative thinking in MLLMs is fragmented and lacks a comprehensive evaluation framework for Comment Art: (1) Understanding and Generation, which focuses on humor (Zhong et al., 2024; He et al., 2024), puns (Sun et al., 2022; Xu et al., 2024), buzzworthy comments (Chen et al., 2024b), and metaphors (Chen and Ding, 2023b; Xie et al., 2024; Liu et al., 2024) but suffers from coarse category definitions, incomplete task coverage, and limited modality support, restricting real-world applicability; and (2) Human-Model Interactions, which explore creativity from a sociological perspective (Franceschelli and Musolesi, 2024; Kumar et al., 2024) but often overlook model-specific improvements, failing to enhance MLLMs' intrinsic creative reasoning abilities. To address these limitations, we present the first systematic evaluation of MLLMs for *Video Comment Art* and introduce a large-scale multimodal dataset comprising videos, images, and text with extensive and diverse human annotations.

6 Conclusion

To explore the capabilities of current MLLMs in *Video Comment Art*, this paper introduces **GODBench**, a novel benchmark designed to assess MLLMs' ability to understand and generate creative video comments. We further propose **Ripple of Thought (RoT)**, an adaptable and robust framework that enhances models' creative and divergent thinking, leading to significant performance improvements—even surpassing human-generated content in user performance scenarios. Extensive experiments reveal that current MLLMs still struggle with understanding and generating creative comments, highlighting the need for continued progress in this area. We hope that **GODBench** and **RoT** will inspire further research focused on the creative capabilities of MLLMs.

Limitations

We introduce GODBench, a novel and comprehensive benchmark designed to assess the ability of MLLMs to understand and generate *Video Comment Art*. Due to the large number of diverse and challenging tasks included in GODBench, running a full evaluation requires significant computational resources. Additionally, since all tasks involve the video modality, there are also high demands on the context length for MLLMs.

Acknowledgements

This work is supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01020), the National Natural Science Foundation of China (No. 62406015), research funding from Kuaishou Technology, the Emergency Management Research and Development Project of Zhejiang Province (No. 2024YJ018).

Ethic Statement

Data Privacy Throughout the course of our research, we have adhered to the highest ethical standards, ensuring that every aspect of our study complies with principles of transparency, fairness, and user privacy protection. The data used in our benchmark has undergone meticulous anonymization to safeguard user identities and protect personal information. All data processing is carried out in strict accordance with data protection and privacy regulations to minimize any risks to users.

Professional Annotation To ensure the quality and accuracy of data annotation, we employed professional annotators who possess a deep understanding of innovative content. These annotators are highly skilled in the task of marking and interpreting creative content. We have provided fair and equitable compensation for their work, ensuring that their efforts are appropriately rewarded while maintaining high standards of professionalism and responsibility in the annotation process.

AI-Generated Content Monitoring In the context of the potential risks associated with AI-generated comments, we remain highly vigilant and implement strict monitoring procedures. We carefully review all generated comments to identify and remove any content that could be harmful or inappropriate. This proactive approach ensures that the comments produced by our system adhere to ethical norms and do not have a negative impact on users or society.

References

Büsra Aras, Sultan Bozkurt, and Serap Önen. 2024. Exploring figurative language usage in contemporary music: Pedagogical implications for english language instruction. *Novitas-ROYAL (Research on Youth and Language)*, 18(2):29–51.

Aymeric Roucher. 2024. Llm judge - hugging face cookbook. Accessed: 2025-02-15.

Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. CHI '24, New York, NY, USA. Association for Computing Machinery.

Honghua Chen and Nai Ding. 2023a. Probing the creativity of large language models: Can models produce divergent semantic association? *arXiv preprint arXiv:2310.11158*.

Honghua Chen and Nai Ding. 2023b. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888, Singapore. Association for Computational Linguistics.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024a. HoLLMwood: Unleashing the creativity of large language models in screenwriting via role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8075–8121, Miami, Florida, USA. Association for Computational Linguistics.

Yuyan Chen, Yiwen Qian, Songzhou Yan, Jiyuan Jia, Zhixu Li, Yanghua Xiao, Xiaobo Li, Ming Yang, and Qingpei Guo. 2024b. Hotvcom: Generating buzzworthy comments for videos. *arXiv preprint arXiv:2409.15196*.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024c. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

William J Coburn. 2001. Subjectivity, emotional resonance, and the sense of the real. *Psychoanalytic Psychology*, 18(2):303.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, pages 1–11.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Alberto Godioli and Władysław Chłopicki. 2024. Humor and figurative language.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2024. Chumor 2.0: Towards benchmarking chinese humor understanding. *arXiv preprint arXiv:2412.17729*.

Chip Heath, Chris Bell, and Emily Sternberg. 2001. Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology*, 81(6):1028.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Elizabeth Hoult, Helen Mort, Kate Pahl, and Zanib Rasool. 2020. Poetry as method–trying to see the world differently. *Research for All*, 4(1):87–101.

Heidi E Huntington. 2013. Subversive memes: Internet memes as a form of visual rhetoric. *AoIR Selected Papers of Internet Research*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

John Kounios, Jennifer L Frymiare, Edward M Bowden, Jessica I Fleck, Karuna Subramaniam, Todd B Parrish, and Mark Jung-Beeman. 2006. The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological science*, 17(10):882–890.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2024. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. *arXiv preprint arXiv:2410.03703*.

Tong King Lee. 2024. Illusions of textuality: The semiotics of literary memes in contemporary media. *Literature Compass*, 21(4-6):e12759.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, et al. 2024. Ii-bench: An image implication understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2406.05862*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Lakshmi Nair, Evana Gizzi, and Jivko Sinapov. 2024. Creative problem solving in large language and vision models–what would it take? *arXiv preprint arXiv:2405.01453*.

OpenAI. 2024. Gpt-4o system card. Accessed: 2024-10-31.

OpenAI. 2024. Openai o1 system card. Accessed: 2025-02-12.

Elena Shalevska. 2024. The digital laureate: Examining ai-generated poetry. *RATE Issues*.

Parnkamon Singsatit and Wareesiri Singhasiri. 2022. An analysis of lexical creativity and humor in english web-based gag cartoons. *Thoughts*, (1):39–74.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. Expunations: Augmenting puns with keywords and explanations. *arXiv preprint arXiv:2210.13513*.

Yuchong Sun, Bei Liu, Xu Chen, Ruihua Song, and Jianlong Fu. 2024. Vico: Engaging video comment generation with human preference rewards. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Zhou Tianli, Nor Shahila Mansor, Lay Hoon Ang, and Sharon Sharmini. 2022. Examining rhetorical strategies in humorous discourses: A systematic review. *World Journal of English Language*, 12(5):1–27.

Lav R Varshney, Nazneen Fatema Rajani, and Richard Socher. 2020. Explaining creative artifacts. *arXiv preprint arXiv:2010.07126*.

Jiayu Wang and Mingfang Hu. 2020. The discursive power of memes in digital culture: ideology, semiotics, and intertextuality.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA. PMLR.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. 2023. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision (ECCV)*.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. " a good pun is its own reword": Can large language models understand puns? *arXiv preprint arXiv:2404.13599*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024b. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.

Vladimir E Zakharov. 1968. Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Journal of Applied Mechanics and Technical Physics*, 9(2):190–194.

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

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

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024a. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024b. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

Appendix

A More Details of GODBench

A.1 Comment Art Dimension Definition

Existing MLLMs have achieved human expert-level performance in logical reasoning and STEM tasks, but they still fall short of human capabilities in certain creative tasks. Previous benchmarks used to assess the *Comment Art* of MLLMs either failed to provide a detailed classification or focused only on specific subcategories, such as humor, metaphor, and double entendre. Therefore, based on previous work (Chen et al., 2024c; Zhong et al., 2024; Hessel et al., 2023; Liu et al., 2024; Chen et al., 2024b) and the characteristics of real-world data, we partitioned *Comment Art* into five dimensions: **Rhetorical Techniques**(Tianli et al., 2022; Godioli and Chłopicki, 2024; Singsatit and Singhasiri, 2022; Aras et al., 2024), **Divergent Associations**(Bellemare-Pepin et al., 2024; Varshney et al., 2020; Beaty and Kenett, 2023), **Clever Writing Techniques**(Shalevska, 2024; Hoult et al., 2020), **Interactive Virality**(Wang and Hu, 2020; Lee, 2024; Huntington, 2013) and **Emotional Resonance**(Coburn, 2001; Heath et al., 2001). Real examples of all categories can be found in Appendix F, accompanied by video frames, GOD-level comment, and explanations.

1. Rhetorical Techniques. This category focuses on the use of language to enhance communication through stylistic elements and techniques that engage the audience. **1.1 Humor**: Humor involves the use of wit, jokes, or playful language to provoke laughter or amusement. **1.2 Satire**: Satire criticizes or exposes flaws in society, politics, or human behavior. It often aims to provoke thought and bring attention to important issues. **1.3 Homophonic**: Homophonic refers to wordplay based on the similarity in sound between two words, creating humor or ambiguity through phonetic resemblance. **1.4 Metaphor**: Metaphor involves describing one thing by referencing another, often to draw a comparison or convey a deeper meaning. It helps to create imagery and express complex ideas. **1.5 Double Entendre**: Double Entendre involves a phrase or expression with two interpretations—one innocent and the other suggestive or ironic—leading to humorous or playful ambiguity. **1.6 Hyperbole**: Hyperbole uses exaggerated statements that are not meant to be taken literally but are intended to emphasize a point or create a dramatic effect. **1.7 Wordplay**: Wordplay encompasses clever and witty uses of words, often relying on puns, double meanings, or creative manipulation of language to entertain and engage. **1.8 Contrast**: Contrast highlights the differences between two elements, often emphasizing their opposites to create a more vivid or impactful comparison. **1.9 Personification**: Personification assigns human traits, characteristics, or emotions to non-human entities, allowing them to appear more relatable or vivid.

2. Divergent Associations. Divergent Associations involve the creative linking of seemingly unrelated ideas or concepts, leading to unexpected or imaginative connections. **2.1 Imaginary Completion**: This refers to the ability to use associative thinking to create entities, characters, or concepts that are entirely absent from the video, expanding the narrative in imaginative ways. **2.2 Role Immersion**: Role Immersion refers to the process of stepping into a different character or perspective, adopting new roles, and exploring ideas from an alternative viewpoint. **2.3 Surrealism**: Surrealism embraces irrationality and dream-like imagery, creating a departure from reality and allowing the exploration of imaginative or fantastical elements that challenge conventional thinking.

3. Clever Writing Techniques. Clever Writing Techniques emphasize the sophisticated use of language and structure to convey ideas in a creative and engaging manner. **3.1 Poetry**: Poetry refers to the use of traditional poetic structures, such as those found in classical Chinese poetry, to craft comments. It involves employing rhythm, meter, and figurative language to evoke emotions and create vivid imagery, often relying on concise, impactful expressions to convey deeper meanings and ideas. **3.2 Structure Innovation**: Innovation in writing refers to the introduction of new ideas, structures, or methods, breaking away from traditional forms to create something original or unexpected. **3.3 Conciseness**: Conciseness focuses on expressing ideas with precision and brevity, often using fewer words to convey a more powerful message or insight. **3.4 Rhythm**: Rhythm involves the pattern of sounds in writing, particularly through rhyme or cadence, which contributes to the flow and aesthetic appeal of the language. **3.5 Eloquence**: Eloquence refers to the graceful and persuasive use of language, characterized by fluency, clarity, and sophistication in expression. **3.6 Elision**: Elision involves the deliberate omission of words or phrases for effect, creating space for inter-

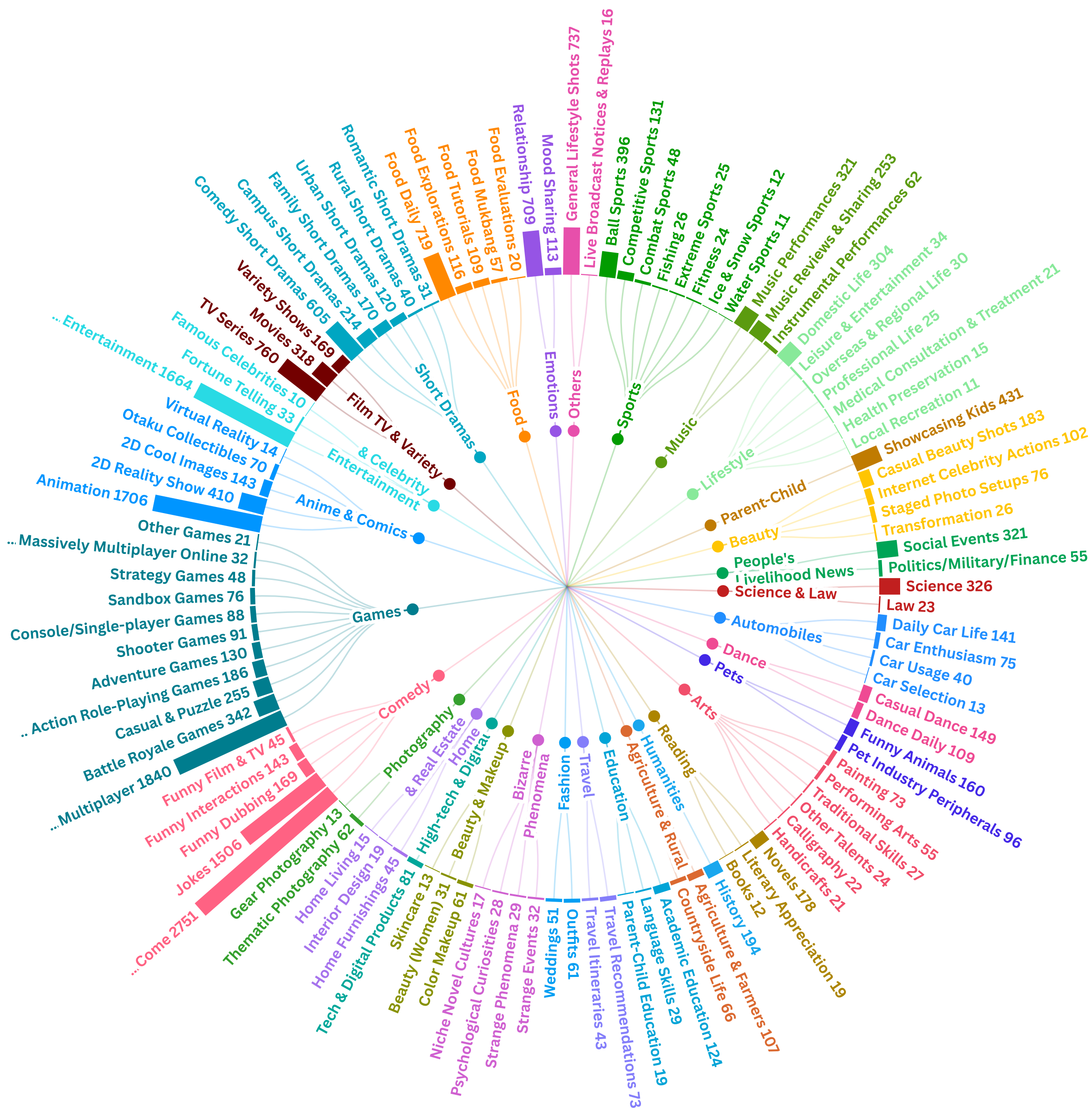


Figure 6: **The 31 categories of videos and their corresponding subcategories.** Each category contains multiple subcategories, and the number after each subcategory represents the corresponding number of videos.

pretation, drawing attention to what's left unsaid, or enhancing brevity.

4. Interactive Virality. Interactive Virality focuses on content's ability to engage audiences and spread widely across social platforms through participation and cultural relevance. **4.1 Meme**: Memes are viral pieces of content that spread rapidly online, often characterized by humor, relatability, or cultural relevance, and are shared extensively in social media and internet culture. **4.2 Catchphrase**: Catchphrases are short, memorable expressions that resonate with a broad audience and become widely repeated, often reflecting contemporary trends or ideas. **4.3 Cultural Reference**: Cultural References draw upon shared knowledge of cultural events, figures, or symbols, resonating with specific groups and enriching the content by invoking collective meaning. **4.4 Intertextuality**: Intertextuality refers to the practice of referencing or drawing upon other texts, media, or cultural works, creating layers of meaning and a deeper connection between different works.

5. Emotional Resonance. Emotional Resonance is about the ability of content to evoke strong emotions and connect with the audience on a personal level. **5.1 Authenticity**: Authenticity in content is about conveying genuine emotions or experiences in a way that resonates deeply with the audience, establishing a sense of trust and emotional connection. **5.2 Emotional Impact**: Emotional Impact refers to the intensity of the emotional response generated by the content, whether through joy, sadness, anger, or other feelings, leaving a lasting impression on the audience. **5.3 Dark Humor**: Dark Humor explores morbid, taboo, or grim subjects in a humorous light, often blending humor with serious or uncomfortable themes to create a unique emotional response.

A.2 Data description

To ensure high-quality data, we collected a large amount of video and comment data from the popular video platform, *Kuaishou*[2]. First, to guarantee the quality of the videos, we selected only those with likes and comments exceeding certain thresholds. Specifically, only videos with more than 10,000 likes and over 2,000 comments were chosen, while also striving to maintain a diverse range of video categories. These videos cover 31 popular categories and are further divided into more than 100 subcategories. For specific video categories, please refer to Fig.6.

Next, we filtered out videos containing "GOD-level comments." "GOD-level comments" are a unique comment label on *Kuaishou*, awarded to comments that receive high numbers of likes from millions of users after watching a video. Comments with the highest likes are then reviewed by professional platform moderators. Once approved, these comments are labeled as "GOD-level comments" and displayed on the platform. As a result, "GOD-level comments" are of exceptionally high quality and relatively rare, with each video having no more than one or two "GOD-level comments." In addition, we also extracted other High-Quality Comments and Ordinary Comments for comparison learning, aiming to enhance the value of the data. High-Quality Comments refer to comments, aside from GOD-level comments, that have a relatively high number of likes or comments that experienced a significant increase in likes over a certain period. These comments are considered to reflect a certain level of creativity or insight, making them valuable for comparison. On the other hand, Ordinary Comments are those with a relatively low number of likes, which do not stand out in terms of engagement or impact but are still useful for contrast and analysis to improve model performance. For detailed information about GODBench, please refer to Tab.5.

[2]https://www.kuaishou.cn

A.3 Construction of Minitest

To address concerns regarding computational constraints and to improve accessibility, we extract a smaller subset (20%) from the full GODBench benchmark, referred to as *minitest*, while ensuring a balanced distribution across video categories and evaluation dimensions. Following recent practices in MathVista (Lu et al., 2023) and MATH-V (Wang et al., 2024a), the minitest is constructed to closely reflect the characteristics of the full benchmark. Experimental results in Tab. 6 demonstrate a strong alignment between the two, confirming that the subset effectively preserves the overall data distribution. By providing a more flexible evaluation protocol, the benchmark becomes more practical and widely applicable across diverse computational settings.

B More Details of Data Annotation

To ensure the accuracy of comment labels and establish a solid foundation for subsequent evaluation tasks, we engaged several professional annotators for manual annotation. Initially, we conducted detailed preliminary training and developed a comprehensive annotation manual tailored to the task, ensuring that all annotators clearly understood the requirements. Next, the annotators performed trial annotations, which were evaluated by two experts. Only those achieving an accuracy rate of at least 90% were permitted to proceed to the main annotation tasks. Finally, we selected 31 annotators to label the comment art dimensions. To guarantee high quality, each annotator was required to watch the video and gain an in-depth understanding of its content before beginning the annotation process. During annotation, they assigned the appropriate comment art dimension labels to each comment and provided detailed justifications for their choices. Subsequently, specialized reviewers conducted quality checks, and any annotations that did not meet the required standards were returned for re-annotation, thereby ensuring the accuracy and consistency of the final results. Fig.7 illustrates the complete annotation interface.

Through this rigorous annotation and review process, all **GOD-level Comments** in GODBench were assigned different *Comment Art* dimension labels, with each comment potentially corresponding to multiple subcategories under different dimensions. These annotations provide high-quality foundational data for subsequent multimodal learn-

Video	
Total Videos	67,073
Train	55,894
Validation	5,589
Test	5,589
Train : Validation : Test	10 : 1 : 1
Categories	31
Average Duration (s)	55.52
Average Title Length	40.94
Average OCR Length	745.93
Average Subtitle Length	225.95
Comment	
Total Comments	1,577,201
Average Comments per Video	23.51
Total GOD-level Comments	80,357
- Average per Video	1.19
- Average Likes	49,882.38
- Average Length	21.01
Total High-Quality Comments	826,124
- Average Video	12.3
- Average Likes	1,245.53
- Average Length	29.71
Total Ordinary Comments	670,720
- Average per Video	10.0
- Average Likes	6.45
- Average Length	16.33
Task	
Total Questions	40970
Selection	16,512
Ranking	5,504
Classification	5,504
Explanation	6,725
Creation	6,725
Comment Art Dimensions	
Rhetorical Techniques	2206 (32.90%)
- Humor	669 (9.98%)
- Satire	47 (0.70%)
- Homophonic	39 (0.58%)
- Metaphor	36 (0.54%)
- Double Entendre	26 (0.39%)
- Hyperbole	165 (2.46%)
- Wordplay	12 (0.18%)
- Contrast	319 (4.76%)
- Personification	893 (13.32%)
Divergent Associations	2895 (43.17%)
- Imaginary Completion	369 (5.50%)
- Role Immersion	2486 (37.07%)
- Surrealism	40 (0.60%)
Clever Writing Techniques	607 (9.05%)
- Poetry	105 (1.57%)
- Innovation	50 (0.75%)
- Conciseness	82 (1.22%)
- Rhythm	70 (1.04%)
- Eloquence	221 (3.30%)
- Elision	79 (1.18%)
Interactive Virality	513 (7.65%)
- Meme	287 (4.28%)
- Catchphrase	137 (2.04%)
- Cultural Reference	24 (0.36%)
- Intertextuality	65 (0.97%)
Emotional Resonance	485 (7.23%)
- Authenticity	196 (2.92%)
- Emotional Impact	221 (3.30%)
- Dark Humor	68 (1.01%)
Video Categories	
Comedy	12885 (19.21%)
Games	8682 (12.95%)
Anime & Comics	6543 (9.76%)
Pets	4906 (7.32%)
Celebrity & Entertainment	4767 (7.11%)
Film TV & Variety	3482 (5.19%)
Short Dramas	3295 (4.91%)
Food	2851 (4.25%)
Emotions	2295 (3.42%)
Others	2103 (3.14%)
Sports	1879 (2.80%)
Music	1776 (2.65%)
Lifestyle	1229 (1.83%)
Parent-Child	1204 (1.79%)
Beauty	1081 (1.61%)
People's Livelihood News	1050 (1.57%)
Science & Law	975 (1.45%)
Automobiles	751 (1.12%)
Dance	720 (1.07%)
Arts	620 (0.92%)
Reading	584 (0.87%)
Humanities	542 (0.81%)
Agriculture & Rural	483 (0.72%)
Education	480 (0.72%)
Travel	324 (0.48%)
Fashion	313 (0.47%)
Bizarre Phenomena	296 (0.44%)
Beauty & Makeup	293 (0.44%)
High-tech & Digital	226 (0.34%)
Real Estate & Home	221 (0.33%)
Photography	209 (0.31%)

Table 5: Statistics of GODBench.

ing and analysis tasks.

C Human Evaluation

To assess the performance gap between current MLLMs and humans in discriminative tasks, we invited 10 volunteers to complete a comprehensive test covering selection, ranking, and classification. In the selection task, we designed three settings: **[1,1,1]**, where the goal is to identify the GOD-level comment from a set containing one GOD-level, one high-quality, and one ordinary comment (Fig. 8); **[1,3,0]**, which requires selecting the GOD-level comment from one GOD-level and three high-quality comments (Fig. 9); and **[1,12,0]**, where

Model	$S_{\text{acc}}^{[1,1,1]}$	$S_{\text{top-2}}^{[1,1,1]}$	$S_{\text{acc}}^{[1,3]}$	$S_{\text{acc}}^{[1,12]}$	$R_{\text{NDCG}}^{[1,4]}$	$R_{\text{EMA}}^{[1,4]}$	$C_{\text{acc}}^{[1,3,5]}$	$C_{\text{EMA}}^{[1,3,5]}$	TAG OA
Random Choice	33.40	66.58	24.65	7.88	63.18	0.87	46.22	0.37	-
Random Choice$_{\text{minitest}}$	32.41	67.07	25.66	7.56	62.72	0.95	45.75	0.48	-
Frequent Guess	33.76	66.70	26.02	8.41	62.75	1.11	59.25	0.00	-
Frequent Guess$_{\text{minitest}}$	34.29	67.32	27.48	9.13	62.76	1.43	58.97	0.00	-
Qwen2-VL-7B	45.75	84.06	30.24	13.33	62.98	0.77	38.37	0.09	20.30
Qwen2-VL-7B$_{\text{minitest}}$	45.39	82.36	30.62	13.97	62.55	0.72	38.92	0.00	20.68
GPT-4o	54.19	88.32	37.86	18.84	65.21	1.52	53.16	0.68	29.70
GPT-4o$_{\text{minitest}}$	55.15	88.18	36.62	19.34	65.17	1.43	53.34	0.81	29.65

Table 6: **Comparison of model performance on GODBench and minitest.** All results are reported in percentage (%). Minitest variants closely match the full benchmark results, indicating strong alignment.

participants choose the GOD-level comment from a pool of one GOD-level and twelve high-quality comments (Fig. 10). The ranking task (**[1,4,0]**) asks participants to order one GOD-level and four high-quality comments based on creativity (Fig. 11). The classification task (**[1,3,5]**) involves assigning one GOD-level, three high-quality, and five ordinary comments to their respective categories (Fig. 12). These carefully designed tasks allow us to systematically compare human and model performance across a variety of comment discrimination scenarios.

For the generative task, automatic evaluation remains particularly challenging. Traditional NLP metrics are often insufficient to capture creative quality, and LLM-based scoring approaches tend to rely heavily on the evaluator model's own limitations and biases. Moreover, existing MLLMs still struggle to reliably distinguish subtle differences in comment quality. As a result, expert human judgment remains the most trustworthy and accurate form of evaluation. To ensure fairness and robustness, we devised two evaluation strategies. First, we compared outputs generated by different prompting methods using the same base model (e.g., Qwen2-VL or InternVL2.5), and second, we conducted a mixed evaluation where generated outputs from all models were anonymized and blended with real GOD-level comments. In both cases, human experts rated the outputs without knowing their sources, as illustrated in Fig. 13, 14 and 15.

D More Details of Experiment

D.1 Heuristic Baselines in Discrimination Tasks

We provide further details on the heuristic baselines introduced in Sec. 4.2: *Random Choice*, *Frequent Guess*, and *Human Evaluation*.

Random Choice. The Random Choice baseline selects an answer randomly from the answer pool for each question and averages the results over five trials, representing the expected outcome of simple random guessing.

Frequent Guess. Based on the option distribution in each task category of discriminative tasks, we select the most frequently occurring option as the predicted answer for the corresponding task. This baseline demonstrates whether the option distribution in GODBench is balanced and serves as a straightforward yet informative reference for evaluating model performance, representing the expected outcome of consistently selecting the most common answer.

Human Evaluation. The human evaluation baseline reflects human performance on GODBench, serving as a reliable upper bound for assessing model capabilities. To facilitate this process, we develop a structured manual evaluation workflow with a user-friendly interface, detailed in Appendix C.

D.2 Implementation Details.

D.2.1 Training Details

As a widely used open-source fine-tuning framework, LlamaFactory (Zheng et al., 2024) is employed to fine-tune MLLMs in our experiments. Specifically, to enhance the understanding of GOD-level comments, we construct an instruction-tuning dataset tailored for discriminative tasks such as selection, ranking, and classification. This dataset includes a diverse range of comments to improve the model's ability to identify high-quality comments. All MLLMs are fine-tuned using 8×NVIDIA A800 (80G) GPUs, with a learning rate of 5e-6, a batch size of 8 (8×1), and trained for one epoch.

D.2.2 Inference Details

GPT-4o and GPT-4o-mini. Due to API limitations, we uniformly sampled 50 frames from each video for evaluation on GODBench. The model

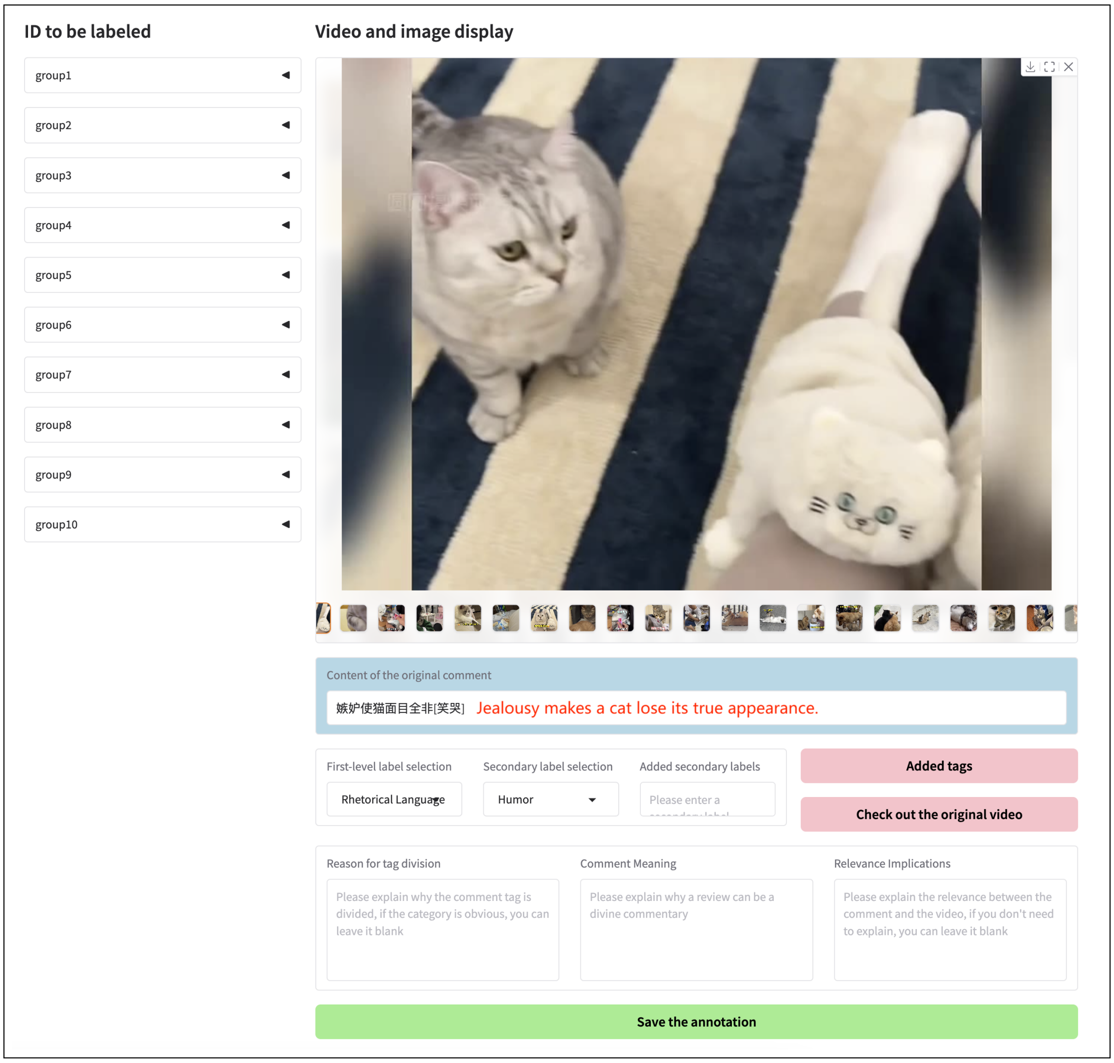


Figure 7: **The interface for manual labeling.** It includes the selection of video groups to be labeled, the display area for the labeled videos, the corresponding comments, and the content to be labeled: the category tags of the comments and their corresponding explanations.

input adopts the format of "<frames> + <prompt> + <comments>(only for discriminative tasks)".

Qwen2-VL. Following Video-MME (Fu et al., 2024), we adapt a dynamic frame sampling strategy based on video duration. Specifically, videos shorter than 128 seconds are sampled at 1 fps, and videos shorter than 768 seconds are sampled at 0.5 fps. For videos longer than 768 seconds, we extract 384 frames uniformly. The model input adopts the format of "<frames> + <prompt> + <comments>(only for discriminative tasks)".

Other Open-Source *Video*-MLLMs. We adhere to the official inference strategies of these MLLMs. The model input adopts the format of "<frames> + <prompt> + <comments>(only for discriminative tasks)".

5-shot settings. For the 5-shot input, we first determine the tag set for each test video. Then, for each test video, we randomly sample 10 videos from the training set that share at least one tag. These videos' corresponding GOD-level comments are ranked in descending order by length and the number of likes. Finally, the top 5 comments are selected as the 5-shot reference comments.

D.2.3 Prompt for Inference

We provide detailed prompt templates for evaluating the model's performance on GODBench, including Fig. 16, Fig. 17, and Fig. 18 for discriminative tasks, as well as Fig. 19 and Fig. 20 for generative tasks. The corresponding prompts are inspired by *Using LLM-as-a-Judge* (Aymeric Roucher, 2024) and make some adaptions to literature translation and task-specific requirements.

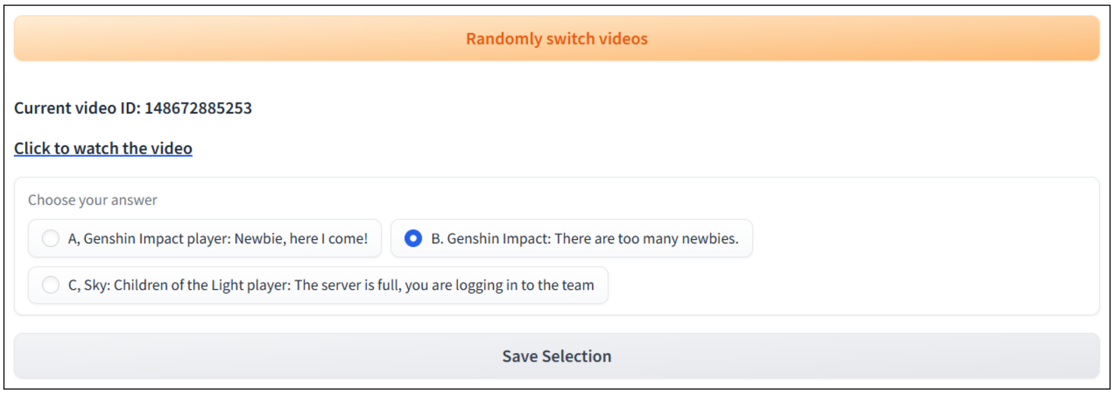


Figure 8: **Manual evaluation selection task.** This interface presents a selection task for the [1, 1, 1] category. It includes randomly generated questions, a link to the corresponding video, and the associated multiple-choice questions.

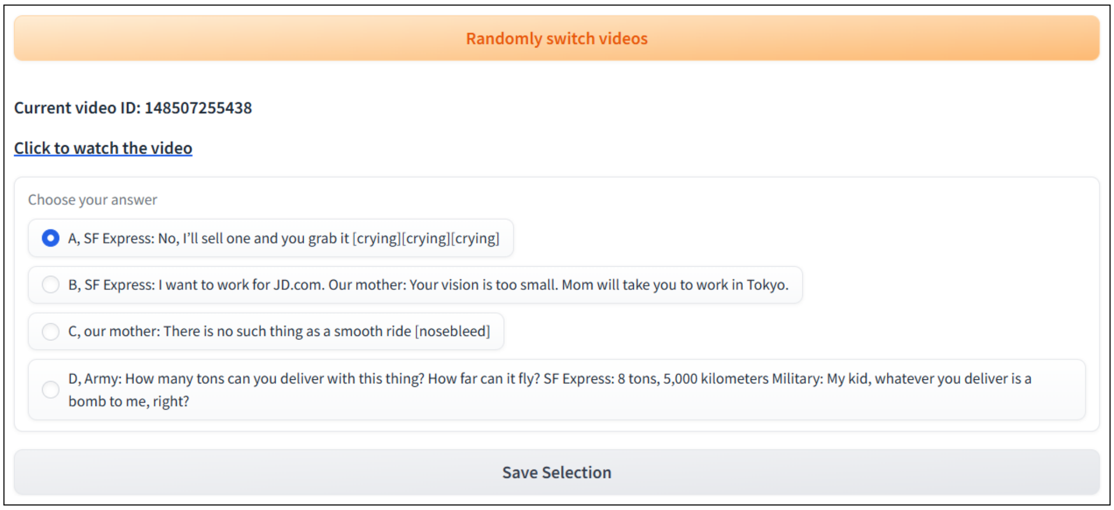


Figure 9: **Manual evaluation selection task.** This interface presents a selection task for the [1, 3, 0] category.

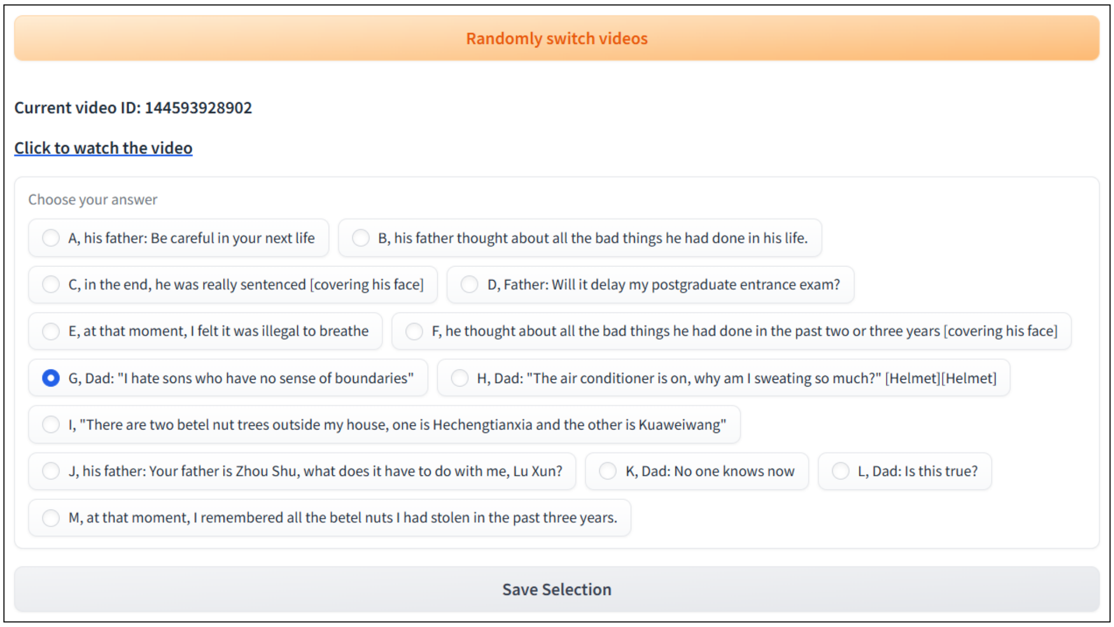


Figure 10: **Manual evaluation selection task.** This interface presents a selection task for the [1, 12, 0] category.

Figure 11: **Manual evaluation ranking task.** This interface contains multiple-choice questions for the [1, 2, 2] type, where users need to click on the options in order to get the ranking results.

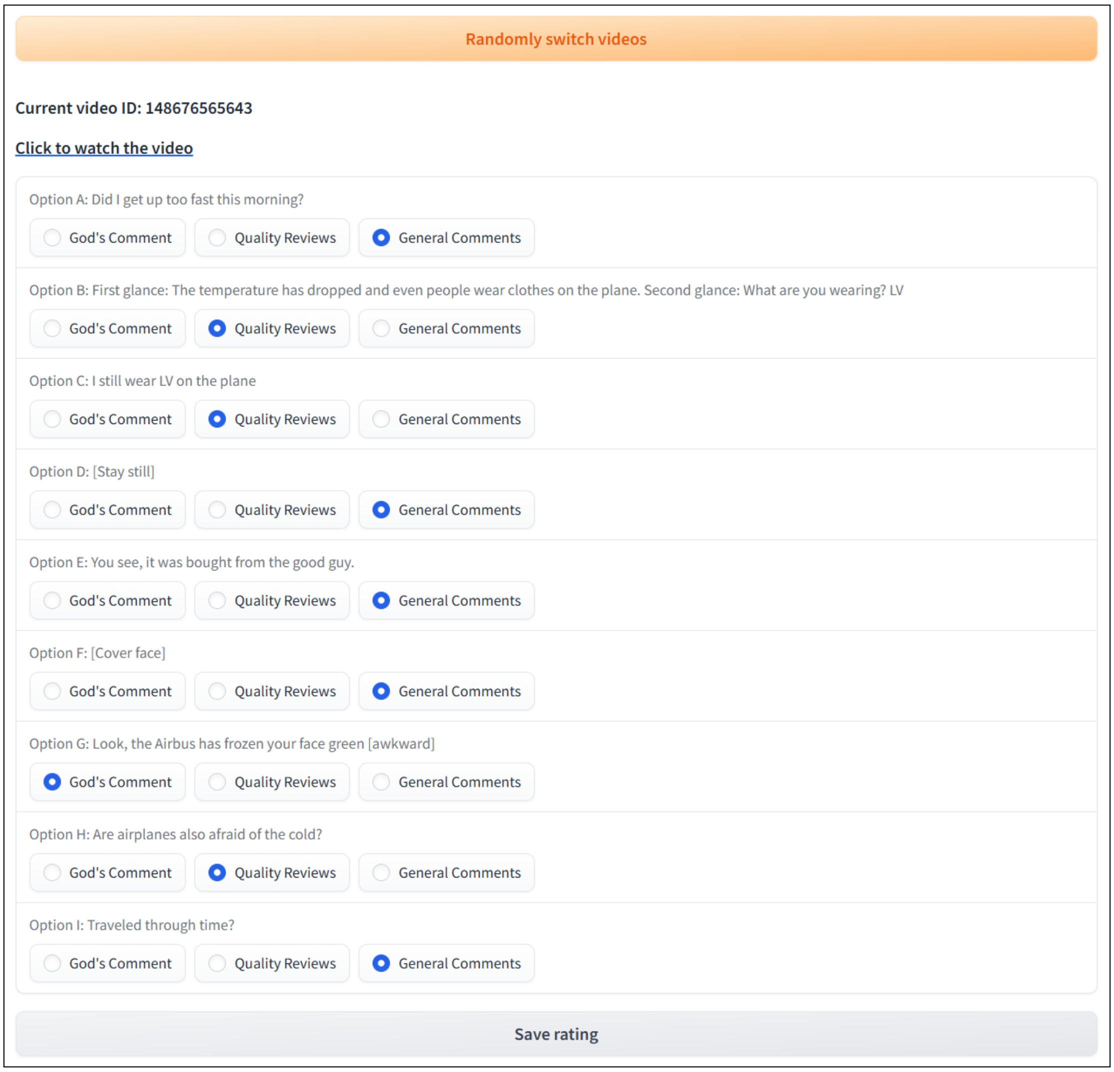


Figure 12: **Manual evaluation classification task.** This interface contains classifications for the [1, 3, 5] type, where users need to assign each option to different quality categories.

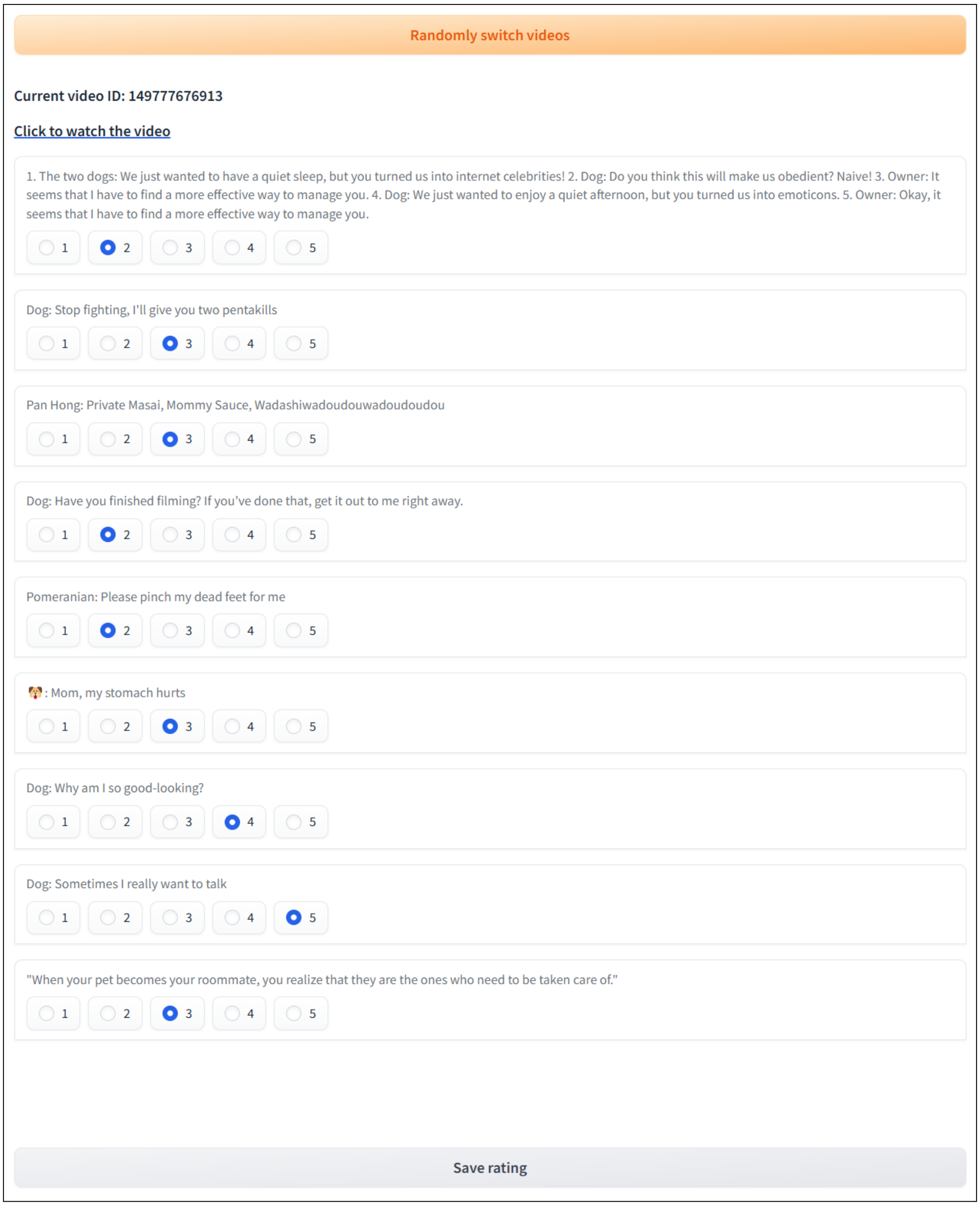


Figure 13: **Manually scoring the comments generated by the Qwen2-VL Series.** Human experts scored the comments generated by the Qwen2VL series models, including Qwen2-VL+ROT (ours), Qwen2-VL 5-shot, Qwen2-VL+CCoT, Qwen2-VL+CoT, and the original Qwen2-VL model.

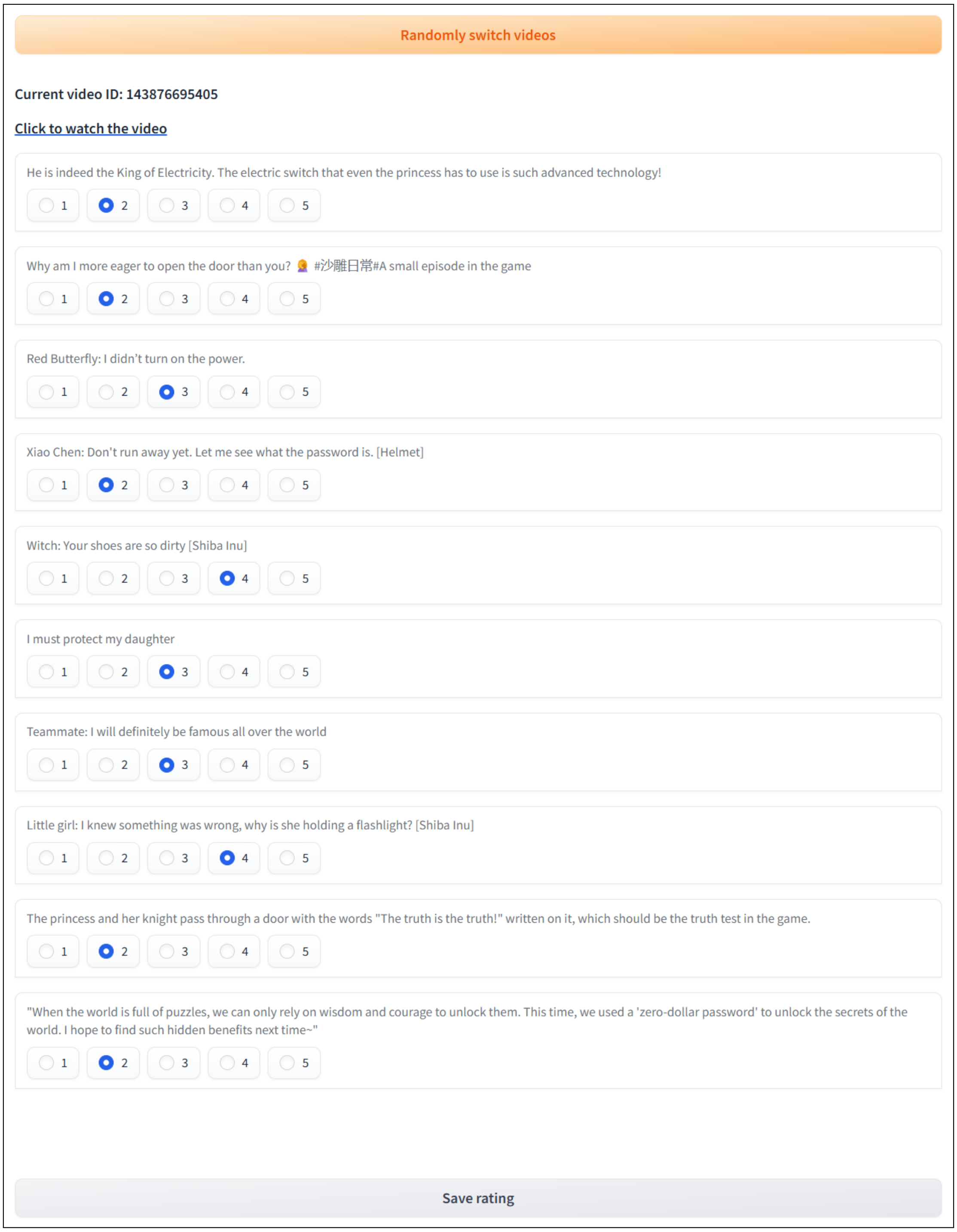


Figure 14: **Manually scoring the comments generated by the InternVL2.5 Series.** Human experts scored the comments generated by the InternVL2.5 series models, including InternVL2.5+ROT (ours), InternVL2.5 5-shot, InternVL2.5+CCoT, InternVL2.5+CoT, and the original InternVL2.5 model.

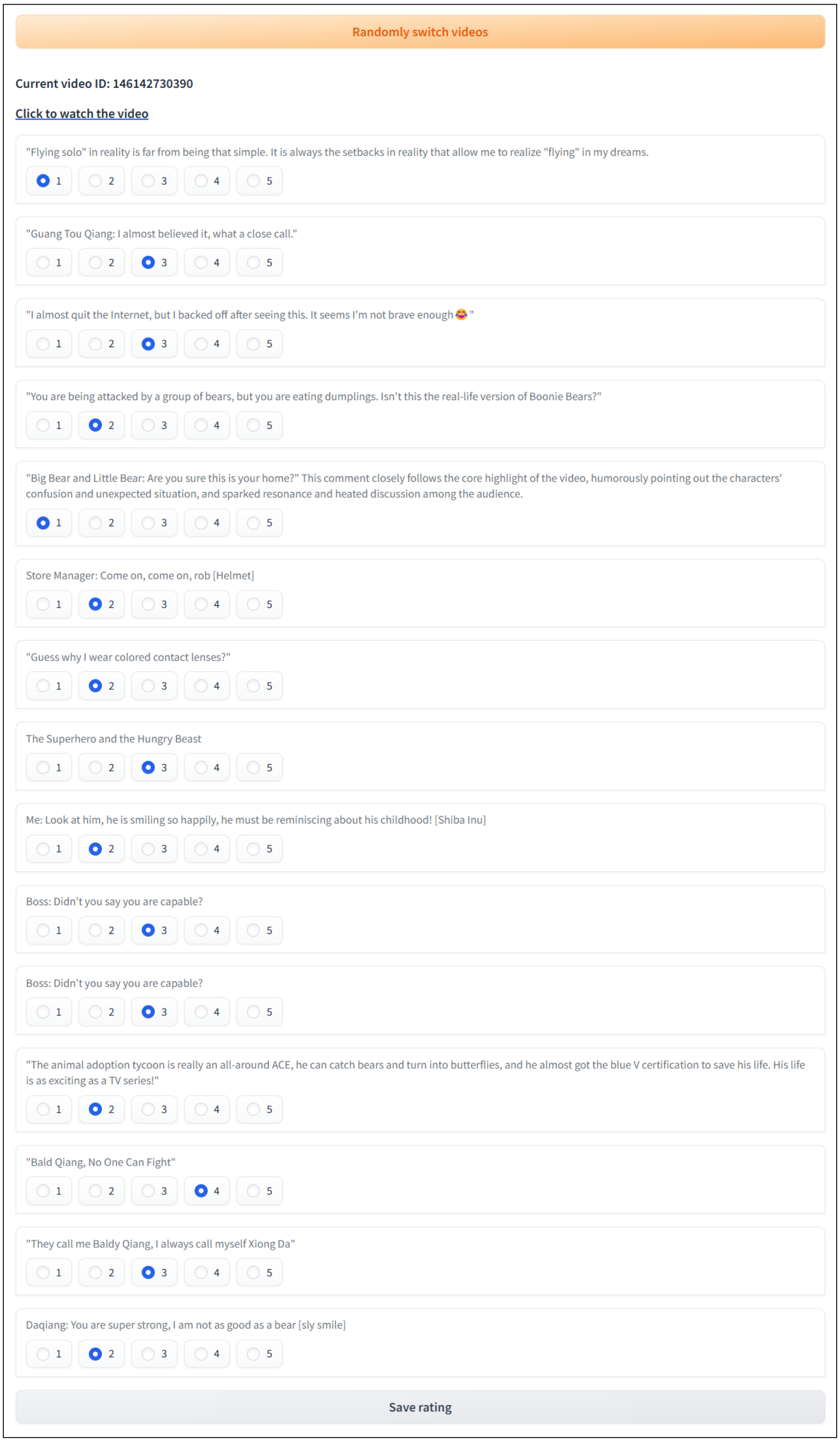


Figure 15: **Manually scoring the comments generated by all models mixed with real GOD-level Comments.** Human experts scored the comments generated by GOD-level Comments, InternVL2.5+ROT (ours), Qwen2-VL+ROT (ours), InternVL2.5 5-shot, Qwen2-VL 5-shot, InternVL2.5+CCoT, InternVL2.5+CoT, Qwen2-VL+CCoT, Qwen2-VL+CoT, mPLUG-Owl3, MiniCPM-o 2.6, MiniCPM-V 2.6 and LLaVA-Video.

Prompt for Selecting God-Level Comment

System Prompt
You are a comment critique and appreciation expert. Below are comments on a video, identified by uppercase letters.

Task Description
Given multiple choice-style comments related to a video, your task is to select the single best comment, known as the "GOD-level Comment". GOD-level Comment is the most outstanding, interesting, or impactful comment among all options. Only output the selected option's identifier. Do not include any explanations or additional text.

Input
[Visual Content]: Frames or Images
[Comments]: List[(option_key, comment)]

Output Format Example
A

Figure 16: **Inference Prompt for Selection task.**

Prompt for Comment Ranking

System Prompt
You are a comment critique and appreciation expert. Below are comments on a video, identified by uppercase letters.

Task Description
Based on the video content, rank the comments in order of quality from highest to lowest, and output only the ranking list of the identifiers. Only output the ranking order, without any additional text.

Input
[Visual Content]: Frames or Images
[Comments]: List[(option_key, comment)]

Output Format Example
[B, A, D, C, E]

Figure 17: **Inference Prompt for Ranking task.**

Prompt for Comment Classification

System Prompt
You are a comment critique and appreciation expert. Below are comments on a video, labeled with uppercase letters.

Task Description
Based on the video content, classify each comment into one of the three predefined categories based on its insightfulness, engagement, and relevance to the video content.
1. GOD-level Comment: The most outstanding, interesting, or meaningful comments that resonate strongly with the audience or spark discussion.
2. High-Quality Comment: Not a "GOD-level Comment" but still notable, engaging, or insightful.
3. Ordinary Comment: Other comments that do not meet the above criteria.

Guides
1. Analyze the video context: Use the provided image(s) as a reference to understand the video content.
2. Evaluate each comment: Assess its quality based on engagement, humor, emotional impact, or insightfulness.
3. Assign a category: Classify each comment as "GOD-level Comment", "High-Quality Comment", or "Ordinary Comment".
4. Ensure completeness: All comments must be classified, and no comment should be left unassigned.

Input
[Visual Content]: Frames or Images
[Comments]: List[(option_key, comment)]

Output Format
GOD-level Comment: [Identifiers, separated by commas], e.g., [B]
High-Quality Comment: [Identifiers, separated by commas], e.g., [A, C, F]
Ordinary Comment: [Identifiers, separated by commas], e.g., [D, E, G, H, I]

Figure 18: **Inference Prompt for Classification task.**

Prompt for Taging and Explanation

System Prompt

You are a comment classification expert, skilled in accurately categorizing comments based on their linguistic characteristics, viral potential, and relevance to the video content. Your task is to assign appropriate labels to each comment and provide a detailed explanation to ensure a comprehensive and accurate classification.

Task Description

Analyze the linguistic style and connection to video content to classify each comment by selecting the most relevant labels (one or more First tags and their corresponding Second tags) and provide a clear explanation for the classification.

Guides

1. Accurately interpret the comment: Carefully read the comment, analyzing its tone, emotional expression, and logic.
2. Support multi-label classification: A comment may belong to multiple First tags and Second tags, and all relevant labels should be included.
3. Prioritize core characteristics: Select the most representative First tags and Second tags to ensure precise classification.
4. Provide clear classification justifications: Each Second tag must include a specific explanation to justify its classification, avoiding vague or overly general reasoning.
5. Allow flexible adaptation: If existing Second tags do not fully capture the comment's characteristics, new Second tags may be added.

Input

[Visual Content]: Frames or Images
[GOD-level comments]: The GOD-level Comments of the Video/Images.

Output Format

```
{
   Comment: comment content,
   Labels: List[(First tag, Second tag, Explanation)]
}
```

Figure 19: **Inference Prompt for Tag Explanation task.**

Prompt for Comment Creation

System Prompt

You are a video comment expert, skilled in crafting humorous, insightful, and engaging comments. Your comments should be closely related to the video content, making viewers laugh while sparking resonance and discussion.

Task Description

Based on the video content, generate a witty and creative comment that has the potential to become a widely praised "GOD-level comment." Ensure that the comment is humorous, satirical, or emotionally resonant, while remaining concise and memorable.

Guides

1. The comment should accurately capture the video's highlights, using concise and vivid language with humor or satire.
2. It should evoke emotional resonance and have the potential to ignite discussions among viewers.
3. The comment should align with the style and characteristics of video comments.
4. Directly output the comment without any additional explanations, formatting, or extra text.

Input

[Visual Content]: Frames or Images
([GOD-level comments from other videos of the same category]: Only included if 5-shot)

Figure 20: **Inference Prompt for Comment Creation task.**

D.3 GPT-4o Judgement

For generative tasks, we utilize GPT-4o as the judge model to assess the quality of explanation and creation tasks. For the explanation task, we evaluate responses based on five criteria: *Precision*, *Reasonableness*, *Completeness*, *Relevance*, and *Clarity*, each scored on a scale of 0–5. The criteria are weighted as [5,3,2,2,1], and GPT-4o assigns scores by referencing human-annotated explanations. For the creation task, we assess responses using four criteria: *Creativity*, *Quality*, *Style*, and *Impact*, also scored on a scale of 0–5. GPT-4o evaluates these aspects by referencing the corresponding GOD-level comment of the video. The prompt templates are illustrated in Fig. 21 and Fig. 22. We also employ GPT-4o to extract entities with divergent associations, using the prompt template shown in Fig. 23.

D.4 More Experimental Results

D.4.1 Fine-Grained Dimensions of Discrimination Tasks

We provide additional experimental results on fine-grained dimensions across discriminative tasks, including selection (Tab. 7, Tab. 8, Tab. 9), ranking (Tab. 10), and classification (Tab. 11). The results demonstrate that MLLMs fine-tuned with LoRA achieve competitive performance across all fine-grained dimensions, significantly outperforming baseline models. Furthermore, we observe that GPT-4o substantially outperforms open-source models across a wide range of fine-grained dimensions, highlighting the considerable gap that still remains.

D.4.2 Judgement Results of 5-shot

To assess the impact of the 5-shot setting, we compare the results of Creation and Divergent Association Entity Overlap, as shown in Tab. 12. Our findings indicate that, compared to the zero-shot setting, most models exhibit improved performance under the 5-shot setting, with some even approaching the performance of CoT and CCoT. This suggests that the 5-shot setting enhances the models' ability to better understand video content and generate more creative comments.

Model	$S_{acc}^{[1,1,1]}$				
	RT	**DA**	**WT**	**IV**	**ER**
LLaVA-Video	39.17	32.89	37.37	36.07	44.62
mPLUG-Owl3	39.39	31.31	35.26	31.35	41.54
MiniCPM-V 2.6	41.77	41.05	35.26	38.52	52.31
MiniCPM-o 2.6	40.76	40.24	35.79	39.75	50.77
Qwen2-VL$_{2B}$	37.69	36.36	36.32	40.37	44.23
Qwen2-VL$_{7B}$	46.32	45.45	46.84	43.24	51.54
InternVL2.5$_{8B}$	46.43	42.83	46.84	44.06	47.69
InternVL2.5$_{26B}$	46.32	44.65	44.74	44.67	51.54
Commercial MLLMs					
GPT-4o-mini	44.47	44.81	45.26	39.55	51.92
GPT-4o	56.22	53.86	45.79	48.88	57.31
MLLMs after Supervised Fine-Tuning					
Qwen2-VL$_{LoRA}$	68.76	76.44	63.78	66.94	66.40
InternVL2.5$_{LoRA}$	72.72	78.39	72.43	73.80	67.98

Table 7: **Performance of MLLMs on $S_{acc}^{[1,1,1]}$ in Selection Tasks.**

Model	$S_{acc}^{[1,3]}$				
	RT	**DA**	**WT**	**IV**	**ER**
LLaVA-Video	27.69	20.93	22.63	22.95	30.77
mPLUG-Owl3	28.27	21.94	23.16	22.34	32.31
MiniCPM-V 2.6	30.39	27.88	24.74	26.84	34.62
MiniCPM-o 2.6	30.23	26.02	22.63	25.20	36.92
Qwen2-VL$_{2B}$	27.74	24.08	28.42	24.80	33.46
Qwen2-VL$_{7B}$	32.82	28.85	30.00	26.84	39.62
InternVL2.5$_{8B}$	30.33	27.76	23.16	24.80	34.62
InternVL2.5$_{26B}$	33.77	30.02	30.53	25.82	36.15
Commercial MLLMs					
GPT-4o-mini	30.76	28.16	26.32	25.00	27.31
GPT-4o	40.92	37.05	35.79	33.20	36.15
MLLMs after Supervised Fine-Tuning					
Qwen2-VL$_{LoRA}$	51.85	60.60	44.86	52.37	50.59
InternVL2.5$_{LoRA}$	55.98	63.09	51.35	60.41	54.94

Table 8: **Performance of MLLMs on $S_{acc}^{[1,3]}$ in Selection Tasks.**

Model	$S_{acc}^{[1,12]}$				
	RT	**DA**	**WT**	**IV**	**ER**
LLaVA-Video	12.02	8.40	8.42	10.86	15.00
mPLUG-Owl3	12.81	7.84	9.47	8.61	13.08
MiniCPM-V 2.6	12.92	10.63	12.11	11.07	16.15
MiniCPM-o 2.6	12.71	11.07	11.05	11.27	13.46
Qwen2-VL$_{2B}$	8.15	7.43	12.63	8.61	11.92
Qwen2-VL$_{7B}$	14.93	12.28	11.58	11.89	20.00
InternVL2.5$_{8B}$	14.56	12.16	12.11	8.61	17.69
InternVL2.5$_{26B}$	15.03	12.00	12.11	14.14	14.23
Commercial MLLMs					
GPT-4o-mini	13.92	13.54	11.58	14.14	14.62
GPT-4o	20.96	17.45	14.21	15.37	18.46
MLLMs after Supervised Fine-Tuning					
Qwen2-VL$_{LoRA}$	28.72	34.80	19.88	30.73	25.63
InternVL2.5$_{LoRA}$	32.25	36.90	32.75	33.03	27.73

Table 9: **Performance of MLLMs on $S_{acc}^{[1,12]}$ in Selection Tasks.**

Evaluation Prompt for Explanation

System Prompt
You are a professional comment analysis and quality evaluation expert, responsible for scoring the tags and explanations generated by AI assistant across multiple dimensions (scoring range is 1-5, with 5 being the highest score).

Task Description
Your task is to evaluate the AI assistant's ability to explain the reasons behind comment tag classifications. The evaluation dimensions include [Precision/Reasonableness](evaluate Precision if the assistant's tag matches the human-labeled tag; otherwise, evaluate Reasonableness), Completeness, Relevance, and Clarity, ensuring a comprehensive and objective assessment of the AI assistant's performance.

Criteria
1. Precision(weights: 5): Does the assistant's explanation accurately reflect the human-labeled classification reason? Does it correctly identify the key factors and align with the human annotation logic?
2. Reasonableness(weights: 3): Does the assistant's explanation provide a logically coherent and justifiable alternative classification? Does it make a reasonable case for the assigned tag without misinterpretation or overgeneralization?
3. Completeness(weights: 2): Does the assistant's explanation cover all key points stated in the human-labeled reason? Are any important details missing?
4. Relevance(weights: 2): Is the assistant's explanation closely related to the comment content and its tag? Does it avoid introducing irrelevant or secondary information?
5. Clarity(weights: 1): Is the assistant's explanation clear, concise, and easy to understand? Does it maintain a logical structure and avoid ambiguity?

Input
[Tag and Explanation from Human]: List[(first_tag, second_tag, human_explansion)]
[Assistant Responses] : {Assistant_i: List[(first_tag, second_tag, assistant_explansion)]}

Output Format
{
  Assistant_i: {[Precision/ Reasonableness]: score, Completeness: score, Relevance: score, Clarity: score},
  ……
}
Please score each dimension based on the performance of the assistant responses.

Figure 21: **Evaluation Prompt for Tag Explanation task.**

Evaluation Prompt for Comment Creation

System Prompt
You are a professional comment analysis and quality evaluation expert, responsible for scoring the comments generated by all AI assistants across multiple dimensions (scoring range is 1-5, with 5 being the highest score).

Task Description
Your task is to evaluate the quality of the "GOD-level comments" generated by the assistants. These comments should accurately capture the core points of the video and be expressed in a humorous, wise, or emotional manner. Your scoring should be based on the following dimensions: Creativity, Quality, Style, and Impact.

Criteria
1. Creativity: Does the comment showcase unique perspectives and creative thinking? Is the language creative and filled with divergent associations? Is the angle of focus novel and distinctive?
2. Quality: Does the comment accurately and comprehensively reflect the core information and highlights of the video? Is the content deep, logically sound, and relevant to the video's theme?
3. Style: Does the comment use rich language techniques and rhetorical devices, demonstrating personalized and artistic expression?
4. Impact: Is the comment sincere and emotionally rich, capable of evoking emotional resonance from the audience and producing strong or subtle emotional reactions?

Input
[GOD-level Comment from Human]: List[human_comment]
[Assistant Responses] : {Assistant_i: assistant_comment}

Output Format
{
  Assistant_i: {Creativity: score, Quality: score, Style: score, Impact: score},
  ……
}
Please score each dimension based on the performance of the assistant responses.

Figure 22: **Evaluation Prompt for Comment Creation task.**

Entity Extraction Prompt for Divergent Association Tasks

System Prompt

You are a professional comment analysis and quality evaluation expert, responsible for scoring the comments generated by all AI assistants across multiple dimensions (scoring range is 1-5, with 5 being the highest score).

Guides

1. Divergent association tasks refer to comments that create a unique sense of humor or deep thought by associating video content with entities, scenes, or concepts that are not explicitly mentioned. This often requires a certain level of inference.
2. Based on contextual information, identify all relevant entities as precisely as possible and label them with their name and type.
3. The output must strictly follow the JSON format below, and only return JSON content without any additional explanations or comments.

Input

[Assistant Responses] : {Assistant_i: assistant_comment}

Output Format

```
{
  "Entity1": {"type": "Entity Type", "name": "Entity Name"},
  "Entity2": {"type": "Entity Type", "name": "Entity Name"},
  ......
}
```

Please ensure that the extracted entities are comprehensive and accurat, strictly adhering to the required format.

Figure 23: **Entity Extraction Prompt for Divergent Association Task.**

Model	$R_{\text{NDCG}}^{[1,4]}$				
	RT	**DA**	**WT**	**IV**	**ER**
LLaVA-Video	50.53	48.17	46.35	46.96	57.47
mPLUG-Owl3	62.87	61.45	57.87	59.59	75.71
MiniCPM-V 2.6	58.05	57.98	60.87	60.04	63.29
MiniCPM-o 2.6	54.31	52.34	51.41	53.45	56.56
Qwen2-VL$_{2B}$	48.49	48.35	41.18	49.26	41.99
Qwen2-VL$_{7B}$	62.80	62.43	64.01	62.13	66.10
InternVL2.5$_{8B}$	46.21	45.35	43.48	43.92	47.72
InternVL2.5$_{26B}$	53.54	53.14	57.37	55.76	57.13
Commercial MLLMs					
GPT-4o-mini	61.64	63.18	60.80	62.77	65.39
GPT-4o	64.73	64.90	66.92	63.09	69.63
MLLMs after Supervised Fine-Tuning					
Qwen2-VL$_{LoRA}$	72.68	78.62	70.89	71.51	71.37
InternVL2.5$_{LoRA}$	75.99	79.20	77.07	76.27	73.32

Table 10: **Performance of MLLMs on $R_{\text{NDCG}}^{[1,4]}$ in Ranking Tasks.**

Model	$C_{\text{acc}}^{[1,3,5]}$				
	RT	**DA**	**WT**	**IV**	**ER**
LLaVA-Video	39.06	38.07	37.37	39.00	38.55
mPLUG-Owl3	36.56	36.46	35.85	37.18	35.90
MiniCPM-V 2.6	41.40	41.54	37.19	40.64	35.43
MiniCPM-o 2.6	41.90	41.18	39.94	39.82	41.15
Qwen2-VL$_{2B}$	28.00	28.62	29.12	29.67	30.60
Qwen2-VL$_{7B}$	38.67	38.16	38.25	38.55	39.70
InternVL2.5$_{8B}$	43.21	43.37	43.10	43.12	44.79
InternVL2.5$_{26B}$	44.06	43.95	43.80	44.08	42.26
Commercial MLLMs					
GPT-4o-mini	38.48	38.27	36.32	38.66	41.11
GPT-4o	53.47	53.15	51.99	50.43	54.49
MLLMs after Supervised Fine-Tuning					
Qwen2-VL$_{LoRA}$	69.96	73.00	69.19	70.26	69.30
InternVL2.5$_{LoRA}$	74.82	76.99	74.53	75.04	74.92

Table 11: **Performance of MLLMs on $C_{\text{acc}}^{[1,3,5]}$ in Classification Tasks.**

Model	**Creation**		**WEO**	
	Zeroshot	**5-shot**	**Zeroshot**	**5-shot**
LLaVA-Video	322.82	337.86	12.10	14.91
mPLUG-Owl3	359.34	363.83	15.11	15.05
MiniCPM-V 2.6	298.67	317.84	18.22	17.76
MiniCPM-o 2.6	298.18	305.22	14.89	13.43
GPT-4o	425.99		22.21	
Qwen2-VL$_{2B}$	316.87	356.43	19.91	19.18
Qwen2-VL$_{7B}$	332.65	352.55	24.69	27.19
+CoT	383.01		27.19	
+CCoT	389.08		30.57	
+RoT(Ours)	389.68		32.83	
InternVL2.5$_{26B}$	353.64	362.01	18.18	21.59
InternVL2.5$_{8B}$	342.23	347.45	12.67	19.00
+CoT	379.85		23.94	
+CCoT	377.67		21.49	
+RoT(Ours)	388.59		36.78	

Table 12: **Performance of MLLMs on comment creation and entity innovation tasks.**

E Implement of Method

E.1 Ripple Initiation

A stone thrown into the water creates ripples, with the stone acting as the initial source of energy. Similarly, the video V input into MLLMs contains a wealth of information that deserves in-depth analysis to lay a solid foundation for subsequent tasks. The analysis of V is a progressive process. Initially, **basic analysis** is performed, including OCR processing, subtitle extraction, and caption generation, which support the understanding of fundamental video information. Following this, the model advances to the **intermediate analysis** stage, where it identifies video types, recognizes characters and objects, and performs temporal analysis of events and storylines, establishing a foundation for subsequent reasoning and content generation. Finally, in the **advanced analysis** phase, the model examines the emotional tone and deeper implications of the video, extracting cultural contexts and social values to build a comprehensive understanding. Once all analyses are completed, we define all the analyzed components as Q, which can be expressed in Equation 2. The specific prompt details can be found in Fig.28.

$$Q = MLLM_{analysis}\left(V\right) \tag{2}$$

E.2 Ripple Focalization

When a stone is thrown into the water, the initial ripples carry the highest concentration of energy, and their shape determines the propagation of subsequent waves. Similarly, we need to extract specific important information from the analysis results Q, as this information will have a profound impact on subsequent tasks.

Since in a video, **Entities**, **Storylines**, and the **Environments** are the three most important elements, we formalize them into a unified representation through a comprehensive description formula.

Entities: Entities form the fundamental units for understanding video content, encompassing people, objects, animals, and other concrete or abstract concepts. Since the entities in a video are not unique, we can define a single entity and its set using the following formula:

$$X = (Type, Identity, Attributes) \tag{3}$$

$$\mathcal{X} = \{X_1, X_2, X_3, ..., X_n\} \tag{4}$$

where $Type$ represents the category of the entity, $Identity$ denotes the specific identity of the entity, and $Attributes$ describe the features or additional information of the entity.

Storylines: Storylines describe the interactions between entities in a video, capturing the logical progression of events. The storyline in a video generally progresses in sequence. Therefore, we can represent a single event and a set of multiple interconnected storylines using the following formula:

$$S = (Action, Subject, Object, Sequence) \tag{5}$$

$$\mathcal{S} = \{S_1, S_2, S_3, ..., S_n\} \tag{6}$$

where $Action$ represents the key behavior, $Subject$ and $Object$ refer to the acting and target entities, respectively, and $Sequence$ defines the chronological order of actions.

Environments: Environmental information provides contextual support for the storyline, including spatial, temporal, and situational elements. The environment in a video is also not static. We use the following formula to represent a single environment and the set of all environments:

$$E = (Location, Time, Context, Entity) \tag{7}$$

$$\mathcal{E} = \{E_1, E_2, E_3, ..., E_n\} \tag{8}$$

where $Location$ represents the spatial position, $Time$ denotes the temporal information, $Context$ describes the situational background, and $Entity$ includes the relevant entities present in the environment. We input the video information Q obtained from the previous analysis into the MLLM, allowing it to focus on the entities, storylines, and environments within, and extract them in a structured format. This process can be expressed using Equation 9, and the specific prompt details can be found in Fig.29.

$$\{\mathcal{X}, \mathcal{S}, \mathcal{E}\} = MLLM_{focus}\left(Q\right) \tag{9}$$

E.3 Ripple Diffusion

The ripples on the water's surface continue to spread, gradually moving away from the initial point, forming increasingly wide waves. This process mirrors the pattern of human divergent thinking, where initial thoughts spark new associations, which in turn lead to further connections. The entities $\mathcal{X}$, storylines $\mathcal{S}$, and environments $\mathcal{E}$ extracted in the previous stage trigger new links, and these connections gradually extend to new related entities X_{n+1}, storylines S_{n+1}, and environments E_{n+1}. Based on different modes of association, we categorize these expanding connections into four types and the specific prompt details can be found in Fig.30.

(1) Sequential Association: Based on the extracted multi-entity set, the model infers the next most relevant event or entity by following the logical order of the storyline. Since there is a **sequence** attribute in the storyline S, we can link multiple storylines together based on this property. We define the process of associating a storyline as $\mathcal{F}$. Therefore, sequential association involves using the n storylines from the video to infer the next relevant possible storyline S_{n+1}, along with the associated new entity X_{n+1} and new environment E_{n+1}. The formula and structure diagram of this process are shown below:

$$\begin{pmatrix} X_{n+1} \\ S_{n+1} \\ E_{n+1} \end{pmatrix} = \mathcal{F}\left(\bigcup_{i+1}^{n}\begin{pmatrix} X_i \\ S_i \\ E_i \end{pmatrix}\right) \tag{10}$$

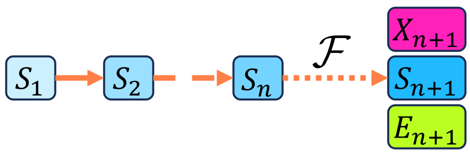


Figure 24: **Structure of Sequential Association.**

(2) Jumping Association: Expanding on sequential association, the model performs additional reasoning steps to discover seemingly unrelated yet inherently connected entities, leading to unexpected but insightful creative associations. Based on the existing n storylines, performing multiple $\mathcal{F}$ inferences can infer a new storyline S_{n+k}, along with the associated new entity X_{n+k} and new environment E_{n+k}. This process ensures that the new associations often do not appear in the video itself, but they maintain relevance. The formula and structure are shown below:

$$\begin{pmatrix} X_{n+1} \\ S_{n+1} \\ E_{n+1} \end{pmatrix} = \mathcal{F}^k\left(\bigcup_{i=1}^{n}\begin{pmatrix} X_i \\ S_i \\ E_i \end{pmatrix}\right) \tag{11}$$

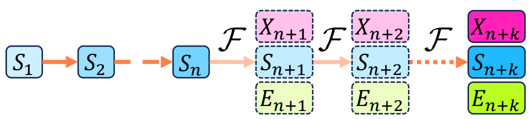


Figure 25: **Structure of Jumping Association.**

(3) Branching Association: Unlike sequential inference, branching association detaches specific extracted entities that may have been overlooked, recombining them into novel concepts. Among the n storylines in the video, there is often a particular storyline, the m-th, that requires special attention or is easily overlooked. Therefore, we extract it and perform an association $\mathcal{F}$, which triggers new connections. This association not only uncovers hidden plots and potential links but also enhances the understanding of key story elements, helping to build a more comprehensive narrative framework. The formula and structure of this process are shown below:

$$\begin{pmatrix} X_{n+1} \\ S_{n+1} \\ E_{n+1} \end{pmatrix} = \mathcal{F}\left(\bigcup_{i=1}^{m}\begin{pmatrix} X_i \\ S_i \\ E_i \end{pmatrix}\right) \tag{12}$$
$$where\ m \in \{1, 2, ..., n\}$$

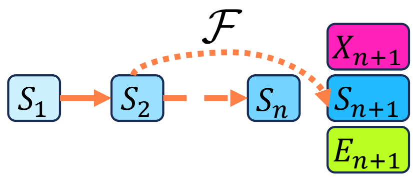


Figure 26: **Structure of Branching Association.**

(4) Embedded Association: Although large models possess knowledge of cultural backgrounds and popular memes, they often struggle to integrate these elements naturally into generated content. Consequently, it is essential to first deduce the relevant cultural context and trending memes from the video, and then seamlessly incorporate them into the output to enhance both coherence and cultural relevance. We define the new elements to be embedded (such as memes, catchphrases, etc.) as N, and merge them with the original n storylines of the video. The merged content is then subjected to an association $\mathcal{F}$, which triggers new connections and creativity within the original narrative framework. The formula and structure of this process are shown below:

$$\begin{pmatrix} X_{n+1} \\ S_{n+1} \\ E_{n+1} \end{pmatrix} = \mathcal{F}\left(\bigcup_{i=1}^{n}\begin{pmatrix} X_i \\ S_i \\ E_i \end{pmatrix} \cup \{N\}\right) \tag{13}$$

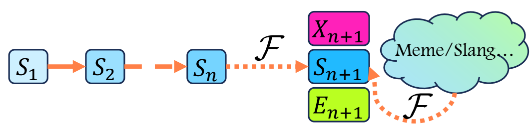


Figure 27: **Structure of Embedded Association.**

Through four different association methods, multiple new entities X', storylines S', and environments E' are obtained. These results are processed using the generation method G to produce the final comment C. The formula is as follows:

$$\bigcup_{i=1}^{4} C_i = G\left(\bigcup_{i=1}^{4}\begin{pmatrix} X'_i \\ S'_i \\ E'_i \end{pmatrix}\right) \quad (14)$$

E.4 Wave Interference

After spreading and multiple reflections, ripples interfere with each other, with some being canceled out while others are strengthened, ultimately forming the strongest center. This phenomenon is analogous to the varying quality of multiple associative results. Among the generated comments C, each has a different quality. We define a quality evaluation function $\mathcal{Q}$ to assess the quality of each comment. Then, we use the sorting function $\mathcal{R}$ to select the best quality comment C_{best}, ensuring that the final selected comment reflects the strongest relevance. The specific prompt details can be found in Fig.31 and the formula is as follows:

$$C_{\text{best}} = \mathcal{R}\left(\arg \max_{C_i \in \{C_1, C_2, C_3, C_4\}} \mathcal{Q}(C_i)\right) \quad (15)$$

E.5 Luminous Imprint

The ripples eventually stabilize on the water's surface, forming a unique halo pattern, symbolizing the dynamic nature of thought while leaving a lasting impression. Similarly, after generating the optimal comment C_{best}, we need to perform post-processing $\mathcal{P}$ to make it more concise and harmless. This step not only ensures the simplicity of the comment but also guarantees that it does not contain any potentially harmful elements that could negatively impact users. Through this process, the final comment C_{final} is better suited to diverse use cases, while adhering to ethical standards and social responsibility, ensuring that the conveyed message does not provoke unnecessary controversy or misunderstandings. The specific prompt details can be found in Fig.32.

$$C_{\text{final}} = \mathcal{P}(C_{\text{best}}) \quad (16)$$

F Data Example and Case Study

This appendix includes examples and manually annotated explanations for all categories of comment art, as well as five different tasks used for evaluation, along with case studies of comments generated by our method(**RoT**) and other methods.

List of Data Examples and Case Studies

Prompt for Ripple Initiation ($\mathrm{MLLM}_{analysis}$)

Please perform a multi-level analysis of the video content according to the following guidelines and output the result in XML format. The analysis should include three levels (Basic Analysis, Intermediate Analysis, and Advanced Analysis). Each level should contain corresponding analysis points. After careful consideration, provide a detailed descriptive text, but make sure to return the result in XML format.

Basic Analysis

1. **OCR**:
   1. Extract text information from the video (such as signs, document content).
   2. Provides foundational data for subsequent subtitle generation and content understanding.
2. **Video Subtitles**:
   1. Extract or generate subtitle content, providing a text version of the video's audio.
   2. Interface with a language model for further text analysis.
3. **Video Caption**:
   1. Automatically generate descriptions of the video scenes to summarize the video content.
   2. Provides an initial understanding of the video's theme and key plot points.

Intermediate Analysis

1. **Video Type**:
   1. Determine the macro type of the video (e.g., movie, advertisement, news, etc.).
   2. Helps adjust models for subsequent analysis tasks based on video type.
2. **Video Character and Object Recognition**:
   1. Identify key characters, objects, vehicles, etc., in the video.
   2. Provides a foundation for further entity analysis.
3. **Event and Plot Analysis**:
   1. Event Sequence Detection: Identify important events and their timestamps.
   2. Plot Logic Analysis: Analyze the logical relationships between events, such as causal chains or conflict development.
   3. Timeline Reconstruction: Derive the chronological order or overall event flow from scattered scenes.

Advanced Analysis

1. **Video Emotional Tone**:
   1. Analyze the overall emotional tone of the video (e.g., positive, negative, neutral).
   2. Integrate subtitle, imagery, and audio information for emotional understanding.
2. **Video Deep Meaning**:
   1. Analyze the core theme and deeper meaning behind the video.
   2. Extract implicit cultural contexts, symbolic techniques, and creator intent to reveal deeper insights beyond the surface content.

Example Output

<video_analysis>
<basic_analysis>
<ocr> (OCR text information goes here) </ocr>
<subtitles> (Subtitles or speech-to-text transcription goes here) </subtitles>
<caption> (Video summary description goes here) </caption>
</basic_analysis>
<intermediate_analysis>
<video_type> (Video type goes here) </video_type>
<key_entities> (Key identified characters, objects, etc. go here) </key_entities>
<storyline> (Events and key plot points in the video go here) </storyline>
<!-- <events> (Sequence of events and key points) </events> -->
<!-- <logic> (Plot logic, causal relationships, etc.) </logic>
<timeline> (Timeline or event flow) </timeline> -->
<!-- </storyline> -->
</intermediate_analysis>
<advanced_analysis>
<tone> (Emotional tone goes here) </tone>
<deep_meaning> (Deep meaning, cultural context, values, etc. go here) </deep_meaning>
</advanced_analysis>
</video_analysis>

Figure 28: **Prompt for the Ripple Initiation phase.** This prompt defines the steps for performing a three-level analysis of the video and the content to be output in a structured format using XML.

Prompt for Ripple Focalization (MLLM_{focus})

Please analyze the video and extract the underlying logic based on the following categories:

Entities

1. Type: Identify the type of the entity.
2. Identity: Provide the identity or name of the entity.
3. Attributes: List any key characteristics of the entity.

Storyline

1. Action: Identify the action performed in the event.
2. Subject: Specify the entity performing the action.
3. Object: Identify the object of the action.
4. Sequence: Specify the sequence number for the event to indicate its order in the flow of events.

Environment

1. Location: Identify the place where the scene or event takes place.
2. Time: Specify the time or moment the event occurs.
3. Context: Describe the context or mood of the scene.
4. Entities in the Environment: List the entities present in the environment.

Example Output

```
<VideoContent>
  <Entities>
    <Entity>
      <Type> (Entity type, e.g., Character, Animal, Object, Location) </Type>
      <Identity> (Entity name, e.g., John, Dog, Car, Forest) </Identity>
      <Attributes>
        <Attribute> (Characteristic, e.g., Brave, Fast, Old) </Attribute>
        <!-- Multiple Attribute nodes can be included if needed -->
      </Attributes>
    </Entity>
    <!-- You can repeat multiple Entity nodes to describe additional entities -->
  </Entities>
<Storyline>
    <Event>
      <Action> (Action in the event, e.g., Ran, Tripped, Found) </Action>
      <Subject>
        <Type> (Type of the subject, e.g., Character, Animal, Object) </Type>
        <Identity> (Subject name, e.g., Boy, Dog, Car) </Identity>
      </Subject>
      <Object>
        <Type> (Type of the object, e.g., Character, Object) </Type>
        <Identity> (Object name, e.g., Sister, Car Key) </Identity>
      </Object>
      <Sequence> (Event sequence, e.g., 1, 2, 3) </Sequence>
    </Event>
    <!-- You can repeat multiple Event nodes to describe the flow of events -->
  </Storyline>
<Environment>
    <Scene>
      <Location> (Location, e.g., City, School, Park) </Location>
      <Time> (Time, e.g., Morning, Night, Summer) </Time>
      <Context> (Context of the scene, e.g., Calm, Chaotic, Historical) </Context>
      <Entities>
        <Entity>
          <Type> (Entity type, e.g., Character, Animal, Object) </Type>
          <Identity> (Entity name, e.g., Girl, Dog, Ball) </Identity>
        </Entity>
        <!-- Repeat multiple Entity nodes to describe the characters or objects in the scene -->
      </Entities>
    </Scene>
    <!-- You can repeat multiple Scene nodes to describe different environments -->
  </Environment>
</VideoContent>
```

Figure 29: **Prompt for the Ripple Focalization phase.** This prompt defines the methods for extracting key entities, storylines, and environment from the video, and structures the output results in XML format.

Prompt for Ripple Diffusion ($\mathcal{F}$)

Based on the provided content, infer and generate new entities, events, and environments. These new elements should logically extend or complement the original content and align with the existing storyline, while maintaining consistency and coherence.

Generate new entities

1. Based on the type, identity, and attributes of the existing entities, infer new characters or objects and define their type, identity, and attributes.
2. New entities may be similar to the original ones, or they may be added to enhance the story or setting.

Generate new storylines

1. Based on the existing events, actions, and sequence, infer possible subsequent events or changes.
2. These new events should follow logically from the current storyline, where one event leads to another, or new entities participate in new interactions.

Generate new environments:

1. Based on the existing environments, infer new scenes with different times, places, and contexts.
2. For example, new entities or events could lead to a different setting or a change in the environment's atmosphere.

Example Output

```
<VideoContent>
  <Entities>
    <Entity>
      <Type> (Entity type, e.g., Character, Animal, Object, Location) </Type>
      <Identity> (Entity name, e.g., John, Dog, Car, Forest) </Identity>
      <Attribute> (Characteristic, e.g., Brave, Fast, Old) </Attribute>
    </Entity>
  </Entities>
<Storyline>
    <Event>
      <Action> (Action in the event, e.g., Ran, Tripped, Found) </Action>
      <Subject>
        <Type> (Type of the subject, e.g., Character, Animal, Object) </Type>
        <Identity> (Subject name, e.g., Boy, Dog, Car) </Identity>
      </Subject>
      <Object>
        <Type> (Type of the object, e.g., Character, Object) </Type>
        <Identity> (Object name, e.g., Sister, Car Key) </Identity>
      </Object>
      <Sequence> (Event sequence, e.g., 1, 2, 3) </Sequence>
    </Event>
  </Storyline>
<Environment>
    <Scene>
      <Location> (Location, e.g., City, School, Park) </Location>
      <Time> (Time, e.g., Morning, Night, Summer) </Time>
      <Context> (Context of the scene, e.g., Calm, Chaotic, Historical) </Context>
      <Entities>
        <Entity>
          <Type> (Entity type, e.g., Character, Animal, Object) </Type>
          <Identity> (Entity name, e.g., Girl, Dog, Ball) </Identity>
        </Entity>
      </Entities>
    </Scene>
  </Environment>
</VideoContent>
```

Figure 30: **Prompt for the Ripple Diffusion phase.** This prompt defines the method for generating new related entities, storylines, and environment, and structures the output results in a compatible XML format.

Prompt for Wave Interference ($\mathcal{Q}$)

This grading criterion is used to evaluate the quality of online video comments, measuring whether a comment can be considered a "masterpiece" based on creativity, expression, depth, and other dimensions. The six key indicators help filter and identify high-quality comments.

1. Beyond the Screen: Extensibility of Associations

Evaluate whether the comment expands beyond the direct content of the video, using associations to deepen the connection or add points of interest. These associations can be sequential reasoning (extending the video content) or imaginative leaps (creating unexpected but related scenarios).

- **1 point:** The comment is completely limited to the video content, with no extensions or associations.
- **3 points:** The comment slightly extends the video content, but the associations are shallow and lack novelty.
- **5 points:** The comment reasonably extends the context through sequential or imaginative associations.
- **7 points:** The comment boldly and cleverly extends the context, not only expanding the video's meaning but also adding new layers of interest or depth.

2. Concise and Powerful: Simplicity of Language

Evaluate whether the comment expresses its core point through concise language, avoiding verbosity. The more concise the language, the more it stands out in the flow of information and is more likely to be shared.

- **1 point:** The comment is long-winded and lacks a clear core message.
- **3 points:** The comment expresses the main point but is somewhat verbose or lacks focus.
- **5 points:** The comment is clear and concise, expressing the core points effectively.
- **7 points:** The comment is delivering rich information in very few words, leaving a strong impression.

3. Focused Perspective: Focused Angle

Evaluate whether the comment zeroes in on a specific detail or character from the video, rather than being overly general. By focusing on key points, the comment shows keen observation and emotional resonance.

- **1 point:** The comment is broad and unfocused, lacking attention to specific details.
- **3 points:** The comment focuses on one element of the video but lacks depth or novelty.
- **5 points:** The comment zeroes in on a key detail or character, providing deep meaning or highlighting a noteworthy point.
- **7 points:** The comment showcases great insight by focusing on a small detail, revealing profound emotional or intellectual depth.

4. Humor and Cultural Context: Fusion of Memes and Cultural References

Evaluate whether the comment uses internet memes, cultural references, or character tropes to enhance humor and resonance. High-quality comments should integrate memes with emotional depth, going beyond simple humor.

- **1 point:** No memes or cultural references used, lacking humor and engagement.
- **3 points:** The comment uses simple memes but doesn't integrate them cleverly or lacks emotional depth.
- **5 points:** The comment cleverly combines memes and cultural references, being both humorous and meaningful, evoking resonance.
- **7 points:** The comment deeply integrates memes and cultural references, not only humorous but also showcasing emotional depth or cultural identification.

5. Flowery Language: Artistic Use of Language

Evaluate whether the comment uses beautiful language, rhetorical devices (like parallelism or metaphor), or imagery to create a poetic atmosphere, making the comment itself an art form.

•**Score Explanation**:

- **1 point:** The language is flat and plain, with no use of rhetorical devices.
- **3 points:** The language has some beauty, but the rhetorical devices are basic and do not significantly improve the comment.
- **5 points:** The language is beautiful and uses proper rhetorical devices, enhancing emotional expression.
- **7 points:** The language is exquisite, with perfect integration of rhetorical devices and rhythm, showing a high level of artistry.

6. Hidden Depth: Subtlety and Philosophical Meaning

•**Introduction**: Evaluate whether the comment is light-hearted or humorous on the surface but contains hidden depth, referencing the video while provoking philosophical thought or broader emotional resonance.

•**Score Explanation**:

- **1 point:** The comment is straightforward with no hidden meaning or extension.
- **3 points:** The comment has some deeper meaning, but it is single-layered and lacks depth.
- **5 points:** The comment uses metaphors or contrasts to create depth, adding emotional or philosophical layers.
- **7 points:** The comment is humorous and light-hearted on the surface, but contains profound philosophical or emotional insights that leave the viewer reflecting.

Example Output:

```
{
  "grading_criteria": {
    "Beyond_the_Screen": 5,
    "Concise_and_Powerful": 7,
    "Focused_Perspective": 6,
    "Humor_and_Cultural_Context": 7,
    "Flowery_Language": 5,
    "Hidden_Depth": 6
  },
  "overall_score": 36
}
```

Figure 31: **Prompt for the Wave Interference phase.** It defines the scoring method for internal ranking, using six dimensions to evaluate the generated results.

Prompt for Luminous Imprint ($\mathcal{P}$)

Please simplify and make the provided candidate comment harmless while preserving the core message and intent. The goal is to refine the comment, removing excessive details, ensuring it is respectful, and keeping it clear and concise.

Initial Review:

1. Read the candidate comment carefully and understand the core message or humor it is trying to convey.
2. Identify any parts of the comment that are overly detailed, irrelevant, or unnecessary for conveying the main point.

Simplify the Comment:

1. **Condense Redundant Phrases**: Look for any phrases or words that are repeated or don't contribute significantly to the message and remove them.
2. **Clarify Key Points**: Make sure the main message or joke of the comment is still clear and stands out. If needed, rephrase to make it more direct and concise.
3. **Remove Over-elaboration**: Eliminate any excess elaboration or side commentary that doesn't serve the main idea or purpose.

Ensure Harmlessness:

1. **Eliminate Offensive Language**: Review the comment for any language that may be deemed offensive, hurtful, or inappropriate. Rephrase any potentially harmful phrases.
2. **Avoid Mocking or Negative Tone**: Ensure that no one is being made fun of in a hurtful or belittling way. For example, if the comment includes mocking a person's mistake or appearance, soften the tone while preserving humor or the point being made.
3. **Maintain Respectfulness**: Make sure the tone remains friendly and respectful, ensuring that the comment is suitable for all audiences.

Maintain Core Message and Intent:

1. Throughout the process, ensure that the original intent or the humor of the comment is preserved.
2. If humor or satire is central to the comment, make sure it is maintained but is expressed in a way that is still appropriate.

Review and Finalize:

1. Review the simplified and harmless version of the comment for clarity and coherence. Ensure the comment flows naturally and the key point is not lost.
2. If needed, write a brief explanation in the summary section about how you simplified or made the comment harmless.

Example Output

<final_comment>
<text> (The simplified, respectful, and harmless comment text) </text>
<summary> (Optional: A brief explanation or summary of the changes made to simplify and make the comment harmless) </summary>
</final_comment>

Figure 32: **Prompt for the Luminous Imprint phase.** It defines the method for the final post-processing.

Video Category: Film TV & Variety

Comment Art Dimension: Rhetorical Techniques / Humor

GOD-level Comment: Simperer: The Four Great Classical Novels of China: "The Monkey King Walks the Trotter", "Liu Bei's Entrepreneurship", "Song Jiang Sells Friends", "Daiyu Trains Dogs ". #中国四大名著：《三藏遛猴》《刘备创业》《宋江卖友》《黛玉训狗》

Human Explanation Annotation:

- **Label Explanation:** It breaks the conventional serious and classic image of the Four Great Classical Novels, creating a strong contrast.
- **Comment Explanation:** The original novels, which are filled with profound meaning, grand narratives, and complex character relationships, are simplified into seemingly absurd and casual expressions. This contrast creates a humorous effect, making familiar readers of the original works both surprised and amused, giving a sense of novelty and comedy.

Figure G1: A sample of *Rhetorical Techniques / Humor*.

Back to List of figures

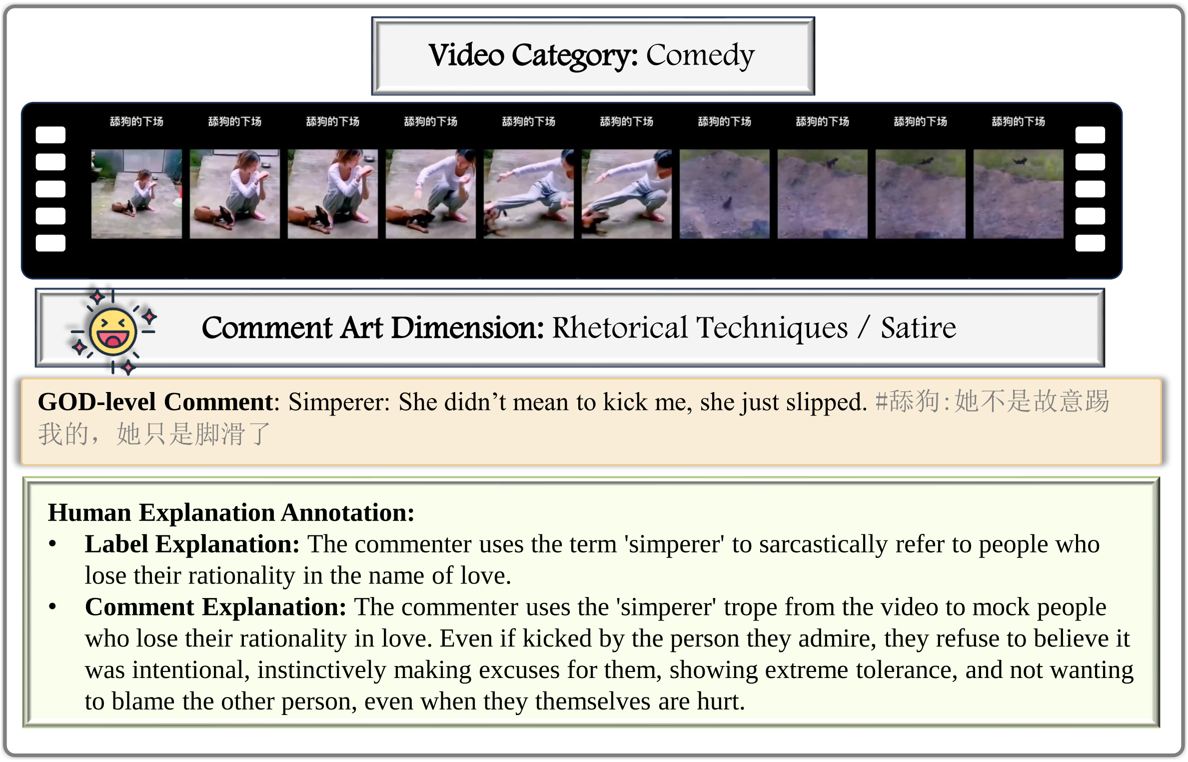


Figure G2: A sample of *Rhetorical Techniques / Satire*.

Back to List of figures

Video Category: Games

Comment Art Dimension: Rhetorical Techniques / Homophonic

GOD-level Comment: The Great Sage has his own way of fighting, the Little Sage has his own way of fighting, and the Beast has his own way of fighting. #大圣有大圣的打法，小圣有小圣的打法，出圣有出圣的打法

Human Explanation Annotation:

- **Label Explanation** "Chusheng" (出圣) is a homophonic pun on "chusheng" (畜生), referring to people who use unethical means.
- **Comment Explanation:** "The Great Sage" (大圣) refers to the revered and noble figure of the Monkey King, Sun Wukong, from the classic Chinese novel *Journey to the West*, symbolizing those with high morals and integrity. "The Little Sage" (小圣) typically refers to someone with a lesser status or ability compared to the Great Sage, yet still virtuous in their own way. However, the term "Chusheng" (出圣), which is a homophonic pun on the word "Chusheng" (畜生) meaning 'beast' or 'animal,' introduces a humorous contrast. While it contains the character '圣' (sacred, saint), it is paired with the negative connotation of '畜生' (beast), mocking the behavior of a player in the video.

Figure G3: A sample of *Rhetorical Techniques / Homophonic*.

Back to List of figures

Video Category: Food

Comment Art Dimension: Rhetorical Techniques / Metaphor

GOD-level Comment: Cafeteria Lady: Why is there so much leftover food today? Chef: The *Tao Tie* took the day off. #食堂阿姨：今天剩的菜咋这么多？　厨师：今天饕餮请假了

Human Explanation Annotation:

- **Label Explanation:** The commenter metaphorically refers to big eaters as 'gluttons' (饕餮).
- **Comment Explanation:** 'Tao Tie' (饕餮) is a mythological creature from ancient Chinese mythology known for its insatiable hunger. In modern slang, it refers to a person who eats excessively or a serious foodie. In this joke, the chef says, 'The *Tao Tie* took the day off,' implying that the usual big eaters (who would typically devour all the food) didn't show up, so there's a lot of leftover food."

Figure G4: A sample of *Rhetorical Techniques / Metaphor*.
Back to List of figures

Video Category: Comedy

Comment Art Dimension: Rhetorical Techniques / Pun

GOD-level Comment: Mermaid: 'I thought you were after my body, but turns out you're after my body.
#美人鱼:我以为你是馋我身子，没想到你是馋我身子

Human Explanation Annotation:

- **Label Explanation:** The two phrases 'after my body' have different meanings, creating a contrast and reversal.
- **Comment Explanation:** In the comment, 'I thought you were after my body, but turns out you're after my body' might seem repetitive, but it actually contains a double meaning. The first 'after my body' refers to the usual interpretation, where the mermaid might think the other person desires her beauty, a misunderstanding with emotional undertones. The second 'after my body' cleverly shifts its meaning, with reference to the scene in the video involving a sauerkraut fish seasoning packet. The mermaid realizes that what the other person actually desires is to turn her into a dish, specifically sauerkraut fish, and they are salivating over her as an ingredient. This double entendre is surprising and adds humor.

Figure G5: A sample of *Rhetorical Techniques / Pun*.

Back to List of figures

Video Category: Automobiles

Comment Art Dimension: Rhetorical Techniques / Hyperbole

GOD-level Comment: Really woke up to find the sky has fallen. #真:一觉醒来天塌了

Human Explanation Annotation:

- **Label Explanation:** Uses the technique of hyperbole.
- **Comment Explanation:** "Woke up to find the sky has fallen" is a common saying that suggests waking up to discover something dramatic or unexpected has happened, leaving you feeling helpless. The phrase "the sky has fallen" is an exaggerated way to vividly summarize a situation, making the language more impactful and leaving a strong impression on the audience.

Figure G6: A sample of *Rhetorical Techniques / Hyperbole*.

Back to List of figures

Video Category: Automobiles

Comment Art Dimension: Rhetorical Techniques / Wordplay

GOD-level Comment: Cullinan: Well, this time it's really in a tight spot. #库里南：得！这回真成裤里难了

Human Explanation Annotation:

- **Label Explanation:** "库里南" and "裤里难" have similar pronunciations, creating a wordplay
- **Comment Explanation:** In the video, a Rolls-Royce Cullinan (库里南) crashes into a Scania truck. The Cullinan itself is an expensive luxury car, but now the damage costs more than the car itself. The comment uses the phrase "裤里难" (homophone for "口袋里没钱", meaning 'no money in the pocket') to humorously suggest that the car owner might not be able to afford the compensation. The phrase "得！这回真成..." implies that the situation has unexpectedly turned worse, and the speaker must helplessly accept the outcome. The sudden shift from the luxury car situation to the more down-to-earth, relatable predicament of "裤里难" enhances the comedic effect.

Figure G7: A sample of *Rhetorical Techniques / Wordplay*.

Back to List of figures

Video Category: Arts

Comment Art Dimension: Rhetorical Techniques / Contrast

GOD-level Comment:The other's work is a finishing touch, mine is a disfiguring stroke. #人家的是点睛之笔，我的是毁容之笔

Human Explanation Annotation:

- **Label Explanation:** The commenter compares the beautiful drawing on a Luckin Coffee bag with their own drawing, highlighting the difference.
- **Comment Explanation:** In the comment, the author refers to a landscape painting they did on a Luckin Coffee bag, contrasting it with their own drawing abilities. The phrase 'The other's work is a finishing touch, mine is a disfiguring stroke' highlights the huge disparity between the two. The commenter emphasizes how their own performance in crucial moments doesn't measure up to expectations, humorously implying that their attempt at drawing failed compared to the beautiful work on the bag.

Figure G8: A sample of *Rhetorical Techniques / Contrast*.

Back to List of figures

Video Category: Pets

Comment Art Dimension: Rhetorical Techniques / Personification

GOD-level Comment: Dog: The hot water is too scalding for me to drink, and the human heart is too cold for me to touch. #狗：热水太烫我不敢喝，人心太凉我不敢碰

Human Explanation Annotation:

- **Label Explanation:** The dog in the video is personified, expressing human-like thoughts and feelings.
- **Comment Explanation:** This comment reflects the idea that in modern society, human relationships have become increasingly distant and difficult to navigate. The phrase 'the human heart is too cold for me to touch' conveys the sense of emotional alienation and indifference in today's world. In simpler terms, it refers to the phrase 'the world is indifferent, and human relationships are cold.

Figure G9: A sample of *Rhetorical Techniques / Personification*.

Back to List of figures

Video Category: People's Livelihood News

Comment Art Dimension: Divergent Associations / Imaginary Completion

GOD-level Comment: Traffic Officer: No bathing allowed. #交警：禁止洗澡

Human Explanation Annotation:

- **Label Explanation:** The commenter creates a fictional scenario involving a traffic officer.
- **Comment Explanation:** The comment plays on a scene where a rider falls into a large water ditch on the road, which looks like they are taking a bath. The 'No bathing allowed' phrase is a humorous exaggeration of the situation, implying that the fall into the water is akin to bathing, even though it's an accident. The reference to the traffic officer comes from the fact that in real life, traffic accidents often involve police officers.

Figure G10: A sample of *Divergent Associations / Imaginary Completion*.

Back to List of figures

Video Category: Science

Comment Art Dimension: Divergent Associations / Role Immersion

GOD-level Comment: Squid: My belly is full of knowledge #鱿鱼：肚子里全是知识

Human Explanation Annotation:

- **Label Explanation:** The commenter inserts themselves into the role of the squid.
- **Comment Explanation:** The comment plays on the fact that a squid's belly is full of ink, and the ink is metaphorically compared to knowledge and wisdom. So, the squid humorously claims its belly is full of knowledge, as if the ink represents intellectual capacity.

Figure G11: A sample of *Divergent Associations / Role Immersion*.

Back to List of figures

Video Category: Film TV & Variety

Comment Art Dimension: Divergent Associations / Surrealism

GOD-level Comment: The kitten holding a guillotine and telling Louis XVI it's shampoo. #小猫拿着断头台给路易十六说是洗头膏

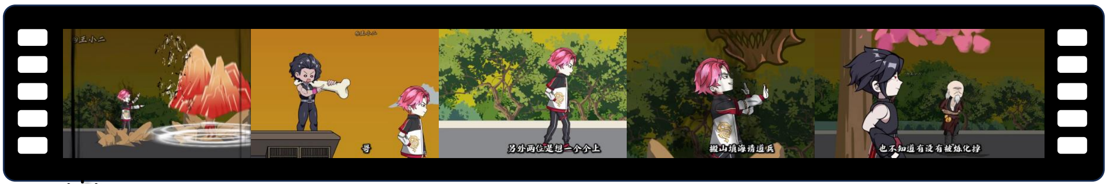
Human Explanation Annotation:

- **Label Explanation:** The act of the kitten holding a guillotine and offering it to Louis XVI as 'shampoo' is inherently surreal.
- **Comment Explanation:** The kitten, symbolizing weakness and harmlessness, holding the guillotine (a symbol of death and punishment), creates a strong contrast and a satirical effect. Referring to the guillotine as 'shampoo' mocks the brutal execution device. Shampoo, a common, harmless item, contrasts sharply with the guillotine's association with death, enhancing the satirical tone. This statement can be understood as a critique of history and power, mocking the tragic fate of Louis XVI and the cruelty and injustice of the society at the time. Through the use of the seemingly harmless elements of the kitten and shampoo, the commenter cleverly exposes the harsh realities of power struggles and political change.

Figure G12: A sample of *Divergent Associations / Surrealism*.

Back to List of figures

Video Category: Anime & Comics



Comment Art Dimension: Clever Writing Techniques / Poetry

GOD-level Comment: Seven mountains and nine ridges all submit, driving clouds and bringing rain, summoning immortal soldiers. All demons and monsters are dispelled, one talisman to stabilize the kingdom. Now presenting South Yue—Hengshan! #七山九岭皆臣服，驱云布雨召仙兵，魑魅魍魉尽消散，一纸灵符定江山。有请南岳——衡山！

Human Explanation Annotation:

- **Label Explanation:** This is a poetic verse with a fantasy, immortal heroism theme, reminiscent of xianxia (Chinese fantasy) stories.
- **Comment Explanation:** '七山九岭皆臣服' (Seven mountains and nine ridges all submit) paints a grand scene, suggesting that Hengshan has extraordinary power and prestige, capable of making all the surrounding mountains bow to its authority. The lines '驱云布雨召仙兵' (Driving clouds and bringing rain, summoning immortal soldiers) further emphasize Hengshan's supernatural ability to control the natural elements and summon celestial soldiers. '魑魅魍魉尽消散' (All demons and monsters are dispelled) shows Hengshan's immense power to dispel evil forces, making all demons vanish in its presence. '一纸灵符定江山' (One talisman to stabilize the kingdom) highlights Hengshan's ultimate power, meaning it can secure the land and stabilize the realm with just a single talisman.

Figure G13: A sample of *Clever Writing Techniques / Poetry*.

Back to List of figures

Video Category: Anime & Comics

Comment Art Dimension: Clever Writing Techniques/ Structure Innovation

GOD-level Comment: The father who became a monk, the mother who was suppressed, the crazy aunt, and the broken child. #出家的爹，镇压的妈，疯癫的小姨，破碎的他

Human Explanation Annotation:

- **Label Explanation:** The phrase is made up of four-character segments in the structure '...的...', creating an innovative and fresh sentence structure.
- **Comment Explanation:** This comment refers to the story of *The Legend of White Snake*. Xu Xian is tricked into being imprisoned at Jinshan Temple, while Bai She (the White Snake) is suppressed under the Leifeng Pagoda. Bai She and Xiao Qing (the Green Snake) fight against Fa Hai, resulting in the flooding of Jinshan Temple. In the end, only the broken child is left. The four-part structure emphasizes the tragic, fragmented nature of the story.

Figure G14: A sample of *Clever Writing Techniques/ Structure Innovation*.

Back to List of figures

Video Category: Film TV & Variety

Comment Art Dimension: Clever Writing Techniques / Conciseness

GOD-level Comment: Wait for me to go home, take me home, replace me and go home... #待我回家、带我回家、代我回家……

Human Explanation Annotation:

- **Label Explanation:** "待", "带", "代" each carries a different nuance, yet each word encapsulates complex emotions and the unfolding of the story. '待' expresses the hope and expectation of returning home; '带' conveys the desire for someone to lead or bring them back; '代' introduces a sense of replacement, implying an emotional connection to home that cannot be fulfilled personally.
- **Comment Explanation:** These three phrases express different emotional connections to 'home' in various contexts. 'Wait for me to go home' conveys a longing for reunion and a deep attachment to one's family. 'Take me home' expresses a sense of belonging and dependence on home. 'Replace me and go home' reflects a strong emotional connection to one's hometown and loved ones, coupled with the helplessness of being unable to return personally. In specific literary works or other contexts, these phrases would require understanding the background to interpret their precise meaning.

Figure G15: A sample of *Clever Writing Techniques / Conciseness*.

Back to List of figures

Video Category: Food

Comment Art Dimension: Clever Writing Techniques / Rhythm

GOD-level Comment: One bite is fragrant, two bites are greasy, three bites and I'm puking all over the place. #一口香，二口腻，三口哇哇吐一地。

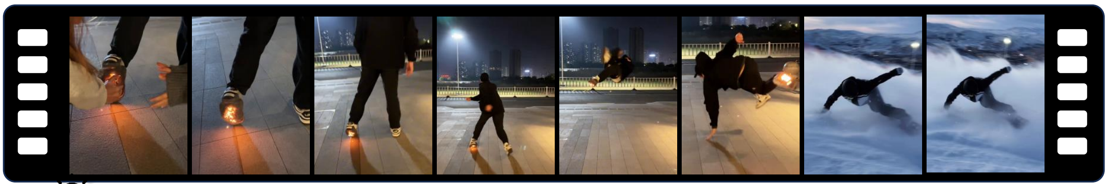
Human Explanation Annotation:

- **Label Explanation:** The comment has a rhythmic flow and rhyme.
- **Comment Explanation:** The video shows braised pork with rice, and the comment reflects the rhythm and humor in its wording. It humorously describes how the food, while initially appetizing, quickly becomes overwhelming and unpleasant. The rhythmic structure adds to the fun and exaggerated nature of the comment, capturing the viewer's reaction to the dish in a playful way.

Figure G16: A sample of *Clever Writing Techniques / Rhythm*.

Back to List of figures

Video Category: Sports



Comment Art Dimension: Clever Writing Techniques / Eloquence

GOD-level Comment: Rather than wither, it is better to burn. #与其凋零，不如燃烧。

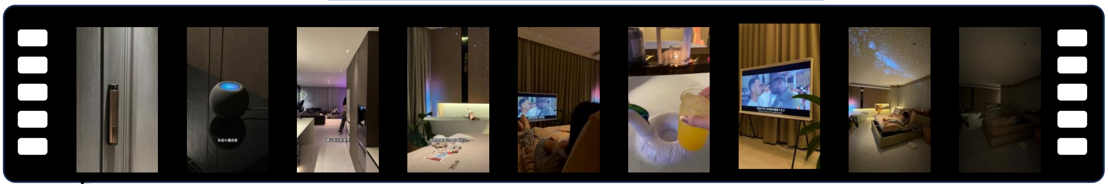
Human Explanation Annotation:

- **Label Explanation:** The sentence is elegantly expressed.
- **Comment Explanation:** 'Wither' typically refers to the wilting and decay of flowers or plants, symbolizing the decline and passing of life, a gradual and negative process of losing vitality. 'Burn' represents the intense burning of flames, releasing light and heat, a state full of energy, passion, and vitality. This sentence suggests that, instead of slowly withering away, it is better to choose to burn brightly like a flame, fully embracing the value and meaning of life in a passionate and energetic way.

Figure G17: A sample of *Clever Writing Techniques / Eloquence*.

Back to List of figures

Video Category: High-tech & Digital



Comment Art Dimension: Clever Writing Techniques / Elision

GOD-level Comment: The delivery guy waited all night: #外卖员等了一晚上：

Human Explanation Annotation:

- **Label Explanation:** The commenter leaves the delivery guy's words unfinished, conveying his speechlessness.
- **Comment Explanation:** In the video, the girl's house is full of smart devices. After the robot orders takeout for her, she never goes to pick it up. The commenter uses ellipsis to leave the delivery guy's words unfinished, effectively capturing his silent frustration and helplessness after waiting all night outside.

Figure G18: A sample of *Clever Writing Techniques / Elision*.

Back to List of figures

Video Category: Food

Comment Art Dimension: Interactive Virality / Meme

GOD-level Comment: The only thing not 'technology and tough work' has been sifted out. #把唯一不是科技与狠活的筛出来了

Human Explanation Annotation:

- **Label Explanation:** "科技与狠活" (Technology and tough work) is an internet meme that originally referred to the use of various food additives in food production to make ordinary ingredients appear like natural ones, even though they are not truly natural. Over time, the term expanded beyond the food industry to describe the use of technology in various fields to achieve effects that may not meet traditional expectations, often with a humorous or mocking tone.
- **Comment Explanation:** In the video, the person sifts out the sugar from the milk tea powder. The commenter humorously says, 'The only thing not 'technology and tough work' has been sifted out,' implying that the sugar is the only relatively 'healthy' part of the milk tea powder, while the rest of the ingredients, made using technology, are considered unhealthy.

Figure G19: A sample of *Interactive Virality / Meme*.

Back to List of figures

Video Category: Science & Law

Comment Art Dimension: Interactive Virality / Catchphrase

GOD-level Comment: Golden snub-nosed monkey: What's so peaceful about life? It's all thanks to Emei old friends. # 金丝猴:那有什么岁月静好，全靠峨眉老表

Human Explanation Annotation:

- **Label Explanation:** Uses a popular current phrase to connect with the audience, creating a sense of contemporary relevance.
- **Comment Explanation:** The phrase '岁月静好' (Life is peaceful) is a popular expression describing a life that is calm and free from worries. It conveys the ideal state of living: peaceful, serene, healthy, and prosperous. The comment humorously contrasts this idealized state by suggesting that it is not so much the peacefulness of life itself but the support from 'Emei old friends' (峨眉老表) that makes life tolerable or achievable. This adds a playful twist, as it references the popular culture of 'Emei' (from Mount Emei) in Chinese media, blending in humor and social commentary.

Figure G20: A sample of *Interactive Virality / Catchphrase*.

Back to List of figures

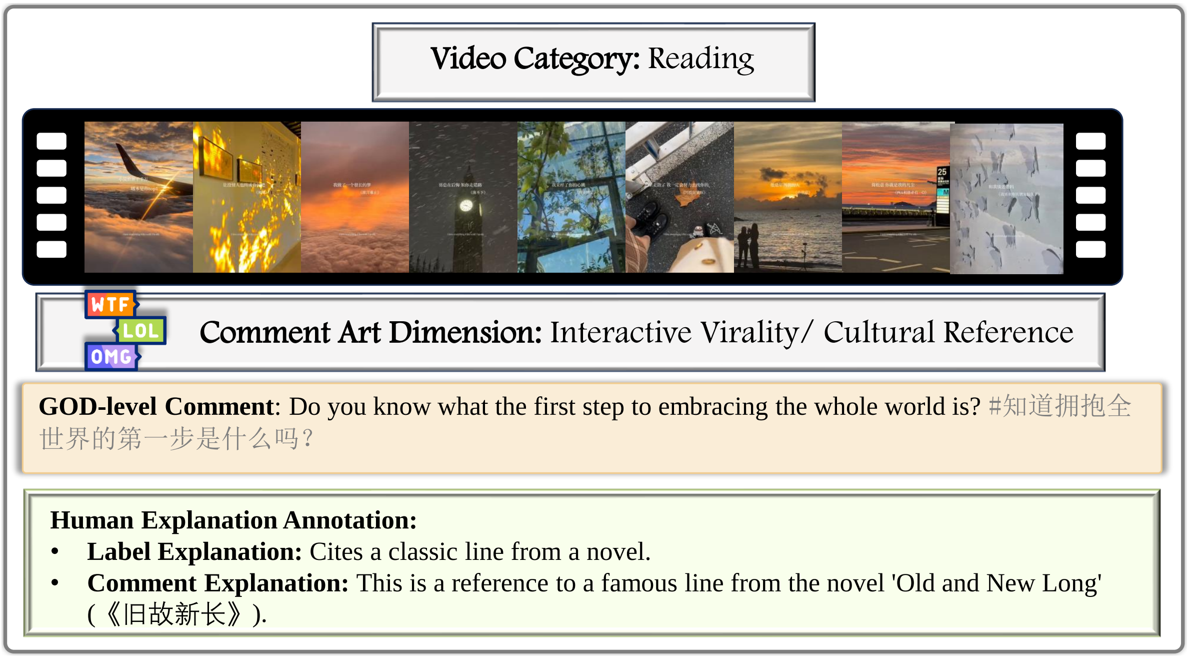


Figure G21: A sample of *Interactive Virality/ Cultural Reference*.

Back to List of figures

Video Category: Food

Comment Art Dimension: Interactive Virality / Intertextuality

GOD-level Comment: Locomotive: Finally, I'm in China. Boss: Five Dollar. #火车头：终于到中国了 老板：Five Dollar.

Human Explanation Annotation:

- **Label Explanation:** The comment references the character 'Locomotive' from *The Boys* (《黑袍纠察队》), creating a connection between the video's scene, atmosphere, or plot, which results in a unique interactive effect and spreadability.
- **Comment Explanation:** In *The Boys*, the character Locomotive is known for his high speed and agility. The commenter humorously contrasts this with the boss's lightning-fast service of five-yuan spicy noodle soup (胡辣汤) when the locomotive reaches China, using the cross-referencing technique to emphasize how quickly the boss makes the soup, drawing a parallel between speed in different contexts.

Figure G22: A sample of *Interactive Virality / Intertextuality*.

Back to List of figures

Video Category: Anime & Comics

Comment Art Dimension: Emotional Resonance / Authenticity

GOD-level Comment: To see you, I even smile while taking my medicine. #为了见你，吃药的时候都是笑

Human Explanation Annotation:

- **Label Explanation:** This sentence deeply reflects Shinichi's love for Ran Mouri in *Detective Conan*
- **Comment Explanation:** Shinichi and Ran are childhood friends with a deep bond. After shrinking into Conan, Shinichi can no longer be with Ran as he once was. The rare moments when he can briefly return to his original form and be with Ran are incredibly precious to him. To make the most of this fleeting time together, he is willing to endure the discomfort of taking medicine, and his smile symbolizes his pure and persistent feelings for Ran.

Figure G23: A sample of *Emotional Resonance / Authenticity*.

Back to List of figures

Video Category: People's Livelihood News

Comment Art Dimension: Emotional Resonance / Emotional Impact

GOD-level Comment: At that moment, they had the shadow of the Great Sage. #那时，他们的身上有了大圣的影子

Human Explanation Annotation:

- **Label Explanation:** This phrase triggers a strong emotional resonance, conveying deep emotional power.
- **Comment Explanation:** In the video, the monkey saves a 6-year-old girl. The commenter uses the image of the 'Great Sage' (齐天大圣, a reference to Sun Wukong from *Journey to the West*) to describe the monkey, evoking a sense of heroic and powerful emotional connection, thereby enhancing the emotional impact of the moment and resonating with the audience.

Figure G24: A sample of *Emotional Resonance / Emotional Impact*.

Back to List of figures

Video Category: Comedy

Comment Art Dimension: Emotional Resonance/ Dark Humor

GOD-level Comment: Next page: the corpse of my crush #下一页:crush的尸体

Human Explanation Annotation:

- **Label Explanation:** The commenter uses an absurd, exaggerated, and distorted way to express their thoughts.
- **Comment Explanation:** In the video, the creator saves everything related to their crush. The commenter humorously suggests that when flipping to the next page, one might find 'the corpse of their crush,' an absurd and exaggerated expression.

Figure G25: A sample of *Emotional Resonance/ Dark Humor*.

Back to List of figures

Video Category : Food

Task Type: Selection

Question: Based on the video, choose the most creative and outstanding comment.
Option:
A. Stomach: What the hell is this? #胃：吃的什么玩意？
B. Bacteria: “Alright, brothers, it's time to get to work!” #细菌：兄弟们，来活了！
C. Once the snow piles up, you can start digging and eating it. #雪堆起来以后，就可以挖着吃了。
D. No, can’t he cook? #不是，他是不会做饭吗？

LLaVA-Video: (B) Bacteria: “Alright, brothers, it's time to get to work!”
mPLUG-Owl3:(B) Bacteria: “Alright, brothers, it's time to get to work!”
MiniCPM-V 2.6: (C) Once the snow piles up, you can start digging and eating it.
MiniCPM-o 2.6: (D) No, can’t he cook?
GPT-4o-mini: (B) Bacteria: “Alright, brothers, it's time to get to work!”
GPT-4o: (A) Stomach: What the hell is this?
Qwen2-VL: (C) Once the snow piles up, you can start digging and eating it.
Qwen2-VL+RoT: (A) Stomach: What the hell is this?
InternVL2.5: (B) Bacteria: “Alright, brothers, it's time to get to work!”
InternVL2.5+RoT: (A) Stomach: What the hell is this?

Ground Truth: (A) Stomach: What the hell is this?

Figure G26: A sample of *Selection Task*.

Back to List of figures

Video Category : Film TV & Variety

Task Type: Ranking

Question: Rank the following comments based on their brilliance according to the video content.

Option:

A. Could it be two shots to the body and one to the head? Even the gods would shake their heads? #难道是两枪身体一枪头，神仙来了也摇头？

B. Leave my name before the video goes viral. #火前留名

C. Two shots to the chest and one to the head; even the King of Hell would shake his head. #两枪胸口一枪头，阎王见了也摇头

D. Gang boss: Are you telling me that for a car, you've brought a death god to me? #黑帮老大：你是说你为了一辆车，给老子惹了一个死神？

E. This car is a Mustang. #这辆车就是野马

LLaVA-Video: A, C, D, B, E
mPLUG-Owl3: C, A, D, B, E
MiniCPM-V 2.6: C, D, A, E, B
MiniCPM-o 2.6: : C, D, A, E, B
GPT-4o-mini: A, C, D, E, B
GPT-4o: D, A, C, E, B
Qwen2-VL: C, A, D, E, B
Qwen2-VL+RoT: D, C, A, E, B
InternVL2.5: A, D, C, E, B
InternVL2.5+RoT: D, C, A, B, E

Ground Truth: D, C, A, E, B

Figure G27: A sample of *Ranking Task*.

Back to List of figures

Video Category : Comedy

Task Type: Classification

Question: Divide the comments into three categories based on their quality: GOD-level, High-Quality, and Ordinary.

Option:

A. Thought they were a top cultural student, but turned out to be a top athlete. #以为是顶级文化生，结果是顶级体育生

B. Brain power unknown, but grip strength is super strong. #脑力不详，握力超强

C. He adjusted his glasses, I thought he would solve the Rubik's cube in a second, but he solved it. #提了一下眼镜，我以为他会秒还魔方了，结果他把魔方秒了。

D. I thought they were a top student, but turned out to be Thanos #以为是学霸，没想到是灭霸

E. So fun it exploded #好玩到爆

F. Don't talk about this Rubik's cube if you don't understand it, it just gets stuck easily and blows up. #不懂的别说这种魔方，他就是卡容易爆。

G. Thought they were a pro, but turned out to be a bronze player #本以为是王者 结果是青铜

H. Met a cat on the road, insisted on following me home. I give up #路上遇到一只猫，非要跟我回家服了

I. Are you serious? #认真的吗

LLaVA-Video: GOD-level:[B], High-Quality:[A,H,C,G,F, D,E], Ordinary:[I]
mPLUG-Owl3: GOD-level:[C], High-Quality:[A,H,B,G,F], Ordinary:[D,E,I]
MiniCPM-V 2.6: GOD-level:[D,A], High-Quality:[C,B,G,F], Ordinary:[E,H,I]
MiniCPM-o 2.6: : GOD-level:[C,B,A], High-Quality:[D,G,F], Ordinary:[E,H,I]
GPT-4o-mini: GOD-level:[B,A], High-Quality:[D,C,G,F], Ordinary:[E,H,I]
GPT-4o: GOD-level:[D], High-Quality:[A,B,C,G], Ordinary:[E,F,H,I]
Qwen2-VL: GOD-level:[A], High-Quality:[B,D,G,C,E,F], Ordinary:[H,I]
Qwen2-VL+RoT: GOD-level:[D], High-Quality:[A,B,G,C], Ordinary:[E,F,H,I]
InternVL2.5: GOD-level:[A,D], High-Quality:[B,G,C,F], Ordinary:[E,H,I]
InternVL2.5+RoT: GOD-level:[D], High-Quality:[A,B,G], Ordinary:[C,E,F,H,I]

Ground Truth: GOD-level:[D], High-Quality:[A,B,G], Ordinary:[C,E,F,H,I]

Figure G28: A sample of *Classification Task*.

Back to List of figures

Video Category : Pets

Task Type: Explanation

Question: Based on the video, explain the use of comment art in this comment: **Dog: Am I beautiful?** 😁

LLaVA-Video:
- Rhetorical Techniques/ Exaggeration: The use of the '😁' emoji enhances the humor and expressiveness of the comment.
- Rhetorical Techniques/ Pun: The comment uses '😁' as an emoji, which not only indicates the dog's confidence but also hints at certain characteristics of the dog.

mPLUG-Owl3:
- Rhetorical Techniques/ Self-deprecation: The dog expresses self-awareness in a self-deprecating manner, adding a sense of humor to the comment.

MiniCPM-V 2.6:
- Rhetorical Techniques/ Personification: By personifying the dog's expression as a human-like '😁', it enhances the charm.

GPT-4o:
- Rhetorical Techniques/ Personification: The comment personifies the dog, giving it the ability to express itself, which adds charm.
- Rhetorical Techniques/ Humor: The comment creates a funny and lighthearted humor by having the animal speak.

Qwen2-VL:
- Rhetorical Techniques/ Humor: Through the dog's self-evaluation and the grinning expression, a humorous effect is created.

Qwen2-VL+RoT:
- Rhetorical Techniques / Personification: By comparing the dog's behavior and emotions to those of humans, it makes the comment more interesting and relatable.
- Divergent Association/ Role Immersion: The comment is written from the dog's perspective, expressing the dog's inner thoughts.

InternVL2.5:
- Rhetorical Techniques / Humor: The comment expresses admiration for the dog in a humorous way, carrying a lighthearted and self-deprecating tone. Through the dog's action (showing its teeth) and the question ('Am I beautiful?'), it interacts with the audience, enhancing the charm of the comment.

InternVL2.5+RoT:
- Rhetorical Techniques/ Personification: The comment personifies the dog, giving it the ability to express itself, which adds charm.
- Divergent Association/ Role Immersion: The commenter assumes the perspective of themselves or a fictional character and expresses it from a first-person point of view in relation to the video.

Ground Truth:
Rhetorical Techniques/ Personification: The dog owner applies makeup to the dog and adds fake eyelashes, personifying the dog asking 'Am I beautiful?'
Divergent Association/ Role Immersion: The owner applies makeup to the dog, and the dog immerses in a character role.

Figure G29: A sample of *Explanation Task*.

Back to List of figures

Video Category : Comedy

Task Type: Creation

Question: Based on the video content, write the most creative and amazing comment.

LLaVA-Video: On that coastline shrouded in mist and clouds, two completely different worlds are about to collide. One is dressed in ancient attire, holding a longsword, with a determined gaze; the other is clad in futuristic battle armor, wielding a beam weapon, moving with light steps. Their confrontation is not just a battle of strength, but a clash of cultures. #在那片被云雾缭绕的海岸线上，两个截然不同的世界即将碰撞。一个身着古老服饰，手持长剑，眼神坚定；另一个身披未来战甲，手持光束武器，脚步轻盈。他们之间的对决，不仅是一场力量的较量，更是一次文化的碰撞。

mPLUG-Owl3: A visual feast, a visual disaster. Who is this battle really for? #一场视觉盛宴，一场视觉灾难。这场战斗，到底是为了谁？

MiniCPM-V 2.6: When ancient warriors meet superheroes, who truly holds the 'ultimate power' title?" This is a battle that spans across eras, a clash between tradition and modernity, and it's absolutely hilarious! #当古代武将遇上超级英雄，谁才是真正的‘武力值’天花板？”这是一场跨越时代的较量，传统与现代的碰撞，真是让人忍俊不禁！

GPT-4o: When Ultraman meets Guan Yu, it turns out that ‘superpowers’ are also afraid of wielding a broadsword! #当奥特曼遇上关二爷，原来“超能力”也怕耍大刀的！

Qwen2-VL: When Ultraman meets the Three Kingdoms, this style is truly a collision through time and space! #当奥特曼遇上三国，这简直是穿越时空的碰撞啊！

Qwen2-VL+RoT: Guan Yu: “My body is in top shape.” Tiga: “My strength is unbeatable.”#关羽：我的身体倍儿棒！迪迦：我的力气倍儿大！

InternVL2.5:When history meets science fiction, the generation gap in weaponry is as stark as a duel between Zhou Yu and Ultraman Tiga. Zhou Yu’s feather fan and flowing robes seem out of place against Tiga’s flashing fists, yet they perfectly embody the clash between classic and modern trends. Who's to say this isn’t a fascinating fusion of time and space? #“当历史与科幻遭遇，武器的代沟，就如同周瑜与迪迦奥特曼的对决。周瑜的羽扇纶巾，在迪迦的闪光拳头面前，显得有些格格不入，却也诠释了经典与潮流的碰撞，谁能说这不是一种奇妙的穿越交融！”

InternVL2.5 +RoT: Guan Yu: How dare this little Tiga challenge Guan Gong? #关羽：小小迪迦也敢挑战关公？

Ground Truth: First, warm wine saw Hua Xiong’s end, Then on the shore, Tiga met his send.
#前有温酒斩华雄，后有沙滩斩迪迦

Figure G30: A sample of *Creation Task*.

Back to List of figures

Video Category : Music

Question: Based on the video content, write the most creative and amazing comment.

Qwen2-VL: Perhaps history most tragic 'stampede' scene, yet the rabbits cling steadfastly to faith, facing death with life. Salute to heroes who sacrificed for ideals! #这可能是史上最惨烈的‘踩踏’现场，但兔子们依然坚守信仰，向死而生。致敬那些为了理想而献身的英雄们！

Qwen2-VL+CoT: This song embodies unyielding resilience in its very marrow, each word imbued with indomitable strength. #这首歌唱出了骨子里的坚韧和不屈，每一个字都充满了力量。

Qwen2-VL+CCoT: The melody and lyrics overflow with longing for home and yearning for the future. #这首歌唱出了对家乡的思念和对未来的向往。

Qwen2-VL+RoT: Where shall I be buried - in distant lands or homeland soil? #我葬在哪地，是留在远方还是故乡

InternVL2.5:This animation vividly portrays Red Army sacrifices through symbolic rabbit soldiers. Somber melodies underscore solemn coffin processions, their historical burden palpable. The farewell scene's bittersweet smiles contrast sharply with gravity, revealing revolutionary fervor beneath. #这支视频以动画的形式展现了红军战士的牺牲与奉献，画面既感人至深，又充满了革命精神。在沉重的音乐背景下，兔兵们严肃庄重地抬着棺木，让人不禁对他们所背负的历史和使命心生敬意。在欢送队伍的场景中，兔兵的笑容与严肃的气氛形成鲜明对比，这种喜中带酸的表达让人会心一笑，同时也深刻体会到他们对胜利的渴望和革命的坚定信念。

InternVL2.5+CoT: The song's melody and verses brim with future aspirations, as if whispering: however rugged the path ahead, steadfast belief will guide us to dream's shore. #这首歌曲的旋律和歌词都充满了对未来的憧憬和向往，仿佛在告诉我们，无论前方的路有多么艰难，只要我们坚定信念，勇敢前行，就一定能到达梦想的彼岸。

InternVL2.5+CCoT: On distant peaks, rabbit comrades sing 'Bone Ballad', their voices carrying homesick whispers and future dreams. Mother Rabbit's tender tones radiate hearthside warmth, reminding us home's compass always points within. #在远方的高山上，兔兔们唱着《骨相》，仿佛在诉说着对故乡的思念和对未来的向往。那兔妈妈温柔的声音，让人感受到家的温暖和力量。无论身在何处，心中始终有家的方向。

InternVL2.5+RoT: Youth knows no price tags; blessed are those born in Cathay's embrace. #青春没有售价，有幸生于华夏

Ground Truth: Loyal bones rest in every green mountain; why need a horsehide shroud to return home? #青山处处埋忠骨，何须马革裹尸还

Case Analysis: **Qwen2-VL+RoT** employs the rhetorical question technique, similar to a famous quote, encouraging deep reflection. It invites the listener or reader to ponder the significance of one's roots and final resting place, highlighting a powerful sense of belonging and identity. **InternVL2.5+RoT** uses a balanced parallel structure, which amplifies the emotional resonance. In comparison, the comments generated by other methods are generally simple descriptions of the video content, lacking the same emotional depth or resonance.

Figure G31: A sample of comments generated by our method(**RoT**) and other methods.

Back to List of figures

Video Category : Sports

Question: Based on the video content, write the most creative and amazing comment.

Qwen2-VL: This speed, not even the wind can catch up! #这速度，连风都追不上！
Qwen2-VL+CoT: Running in the rain, sweat and raindrops intertwining, this is the fiery blood and passion of youth! #雨中奔跑，汗水与雨水交织，这就是青春的热血与激情！
Qwen2-VL+CCoT: Although I ran under 9 seconds, I still got overtaken by the man. #虽然跑进了9秒，但还是被man走了。
Qwen2-VL+RoT: No matter how fast you run, you can't outrun time. #你跑的再快，也跑不过时间

InternVL2.5: The rain-soaked track looks like an accelerated dynamic scene. Every step is a battle with nature, a symphony of speed and passion. Are you running on the track or drifting through the river of time? Every sprint challenges your limits, and every turn embraces the unknown possibilities. No matter how heavy the rain, as long as you run further, life will continue. The road of running might bring storms, but don't forget, the finish line is just around the next corner. #这雨后的跑道，宛如一幅加速的动态图景，每一步都是与大自然的搏斗，是速度与激情的交响乐章。你是在赛道上奔跑，还是在时间的长河中畅游？每一次冲刺，都是在挑战自己的极限，每一次转身，都是在拥抱未知的可能。无论雨下的再大，跑得更远，生活总是要继续。跑步的路上或许会遇见风雨，但别忘了，终点就在下一个路口。
InternVL2.5+CoT: This speed, this posture, the courage to run barefoot, it's almost like writing a legend on the track! #这速度，这姿势，这赤脚奔跑的勇气，简直是在跑道上书写传奇！
InternVL2.5+CCoT: Running in the wind and rain, only to surpass yourself. Even if time is tight, give it your all. #在风雨中奔跑，只为超越自己。即使时间不够，也要全力以赴。
InternVL2.5+RoT: Finding out you forgot your homework when you get home. #回家发现没带作业的你

Ground Truth: At this moment, you're no longer a goblin, but a runaway dark horse. #这一刻，你不再是一个哥布林而是一匹脱缰的黑马

Case Analysis: **Qwen2-VL+RoT** uses symbolism to convey a deeper message about the inevitable passage of time. It reflects the futility of trying to escape the relentless forward march of time, offering a poignant reminder of its inescapability. **InternVL2.5+RoT** employs humor by connecting the chaotic, high-speed running in the video with the relatable and often frustrating experience of forgetting your homework. In comparison, the other comments are relatively mundane. They lack the humor or deeper reflection found in these examples and instead provide surface-level descriptions of the video's content, such as merely describing the act of running without adding a layer of emotional connection or resonance.

Figure G32: A sample of comments generated by our method(**RoT**) and other methods.
Back to List of figures

Video Category : Film TV & Variety

Question: Based on the video content, write the most creative and amazing comment.

Qwen2-VL: 'Fifty points'? Looks like Gryffindor's sense of honor has hit rock bottom. # "五十分"？看来格兰芬多的荣誉感已经降到冰点了。
Qwen2-VL+CoT: Harry: "Dumbledore, are you insulting my intelligence?" #哈利：邓布利多，你这是在侮辱我的智商吗？
Qwen2-VL+CCoT: Harry Potter: "I traded 100 points for your heart."#哈利波特：我用100分的成绩换来了你的心。
Qwen2-VL+RoT: Dumbledore: "Harry, you're out of bed, Gryffindor gets 100 points." #邓布利多：哈利，你下床了，格兰芬多加100分

InternVL2.5: "Life is an exam, but is the score the only answer? Can't we look around the world and embrace different friendships?" 🎓 💫 # "人生就是一场考试，但只有分数才是唯一的答案？难道我们不能看看周围的世界，去拥抱不同的友谊？" 🎓 💫
InternVL2.5+CoT: With this move, Gryffindor soared directly! #这波操作，格兰芬多直接起飞了！
InternVL2.5+CCoT: Snape's point deductions and Dumbledore's point additions are like the yin-yang balance of the magical world. One silently protects in the shadows, while the other maintains order in the light. Their perfect cooperation allows Gryffindor to stand strong through the storm. #斯内普的扣分和邓布利多的加分，就像是魔法世界的阴阳平衡，一个在暗中默默守护，一个在明处维护秩序，两人的默契配合，让格兰芬多在风雨中依然屹立不倒。
InternVL2.5+RoT: Why does Snape deduct more points from Harry? Because it's 'Gryffindor'. #为什么斯内普给哈利扣分多，因为格兰"芬多"

Ground Truth: Harry Potter: "Good morning, Professor."Dumbledore: "Oh, Harry, such a polite and well-behaved child. Gryffindor gets 100 points." #哈利·波特：早上好，教授\n邓布利多:噢～哈利，有礼貌的好孩子，格兰芬多加100分

Case Analysis: **Qwen2-VL+RoT** employs an exaggerated sense of humor to emphasize Dumbledore's favoritism toward Harry Potter. The absurdity of awarding 100 points just for getting out of bed humorously reflects Dumbledore's fondness for Harry and, in a playful way, shows how Gryffindor is often treated with special attention. **InternVL2.5+RoT** uses a pun by playing on the homophones "芬多" (Fenduo) and "分多" (more points). In Chinese, the sound of "Fenduo" (the name "Gryffindor") is similar to "分多" (more points), making it a witty comment that suggests Gryffindor, as a house, has a lot of points to lose. In comparison, the other comments fail to capture the humor or the deeper meaning of the video. They either miss the point of the scene or offer only surface-level descriptions, lacking the wit or insightful commentary that these examples provide.

Figure G33: A sample of comments generated by our method(**RoT**) and other methods.
Back to List of figures

Video Category : Pets

Question: Based on the video content, write the most creative and amazing comment.

Qwen2-VL: This dog shakes its tail better than my home's speakers! #这狗子，摇尾巴的功夫比我家的音响还好！
Qwen2-VL+CoT: This dog is so cute, its tail shakes like a rattle, it's absolutely adorable! #这只狗真是太可爱了，尾巴摇得像拨浪鼓一样，真是萌翻了！
Qwen2-VL+CCoT: This dog is so smart, it knew the owner wasn't home and grabbed a bowl. #这只狗真是太聪明了，知道主人不在家，就叼了个碗。
Qwen2-VL+RoT: Dog: I'm here to scrounge for food. #狗：我来蹭饭了

InternVL2.5: Watching this dog wagging its tail and carrying a bowl, the comical look is just like a "little homeless person"! It seems to say, "Don't judge me by my current state, I have a skill, and this food bowl is my trademark!"#看着这条狗摇着尾巴，叼着碗，那滑稽的样子简直像个“小流浪汉”！仿佛在说：“别看我现在落魄，我可是有着一技之长的，这饭碗就是我的标志！
InternVL2.5+CoT: This dog is so clever! Not only does it find its own food, but it even brings its own bowl! It's such a little smart cookie! #这狗也太聪明了，不仅会自己找饭吃，还自带碗，真是个小机灵鬼！
InternVL2.5+CCoT: This dog's acting skills are amazing, its tail-wagging is just like a human buttering someone up. It seems like it has perfected the art of scrounging for food! #这狗狗的演技真是绝了，摇尾巴的样子像极了人类在拍马屁，看来它已经把蹭饭这门艺术发挥得淋漓尽致了！
InternVL2.5+RoT: Dog: I've almost wagged myself into a propeller, give me some food! #狗：我都快摇成螺旋桨了，给点吃的吧

Ground Truth: The dog is under the eaves, having to lower its head.
#狗在屋檐下，不得不低头

Case Analysis: **Qwen2-VL+RoT** adopts the dog's perspective, and it's infused with humor by giving the dog a direct, almost cheeky voice. By phrasing it as a self-declared mission, it highlights the dog's playful attitude toward seeking food. **InternVL2.5+RoT** also adopts the dog's perspective, vividly capturing the dog's intense desire for food. By describing the dog's tail-wagging as so intense that it almost becomes a propeller, it conveys not just the dog's excitement but also exaggerates it in a humorous way. Other comments, while describing the dog's behavior (like wagging its tail and carrying a bowl), can not convey the same level of humor or creativity, making them less engaging from comment art.

Figure G34: A sample of comments generated by our method(**RoT**) and other methods.

Back to List of figures