A. Benchmark Construction

A.1. Pre-processing

Our pipeline begins with collecting Creative Commons licensed videos from a professional advertising platform, focusing on high-quality, award-winning content, including Clio Award recipients. Each video is crawled with its original metadata that contains creator-specified information about themes, content, and key creative elements, which are used as our ground truth for benchmark construction. Our human annotators manually review the metadata to verify completeness and accuracy, and remove the incomplete samples. We then perform automated filtering based on three criteria: video length (retaining 15s-150s clips), aesthetic quality (using computational scoring), and content appropriateness (removing non-ethical material). For temporal analysis, we employ PySceneDetect to segment videos into coherent clips while preserving narrative flow. From each clip C_i, we extract key visual elements by sampling n keyframes $\{F_i^0, ..., F_i^n\}$ based on SSIM-calculated frame similarity, where n adapts to clip duration. Complementary multimodal features are obtained through: (1) Video-LLM generated descriptions ($Desc_i$) capturing visual content, and (2) Whisper-based speech transcription (Asr_i) with GPT-4 translation for English standardization.

A.2. Role-Played Multi-Agent Annotation

We design a role-played multi-agent annotation framework for our three sequential stages, as illustrated in Fig. 5.

In the initial stage, a master agent is recruited to generate preliminary QA annotations for the given modality-interleaved sequence V. The master agent analyzes the input sequence and produces an initial set of question-answer pairs covering the key elements of the video content.

The framework then enters an **iterative refinement stage** where the master agent systematically evaluates the quality of existing annotations. This evaluation determines whether to recruit specialized expert agents to improve specific aspects of the annotations. When expert agents are needed, the master agent selects appropriate profiles matching the required expertise and instantiates them accordingly. Each recruited expert agent generates new QA annotations leveraging its specialized knowledge, which the master agent then incorporates into the existing set through careful revision and integration. This iterative process continues until the master agent determines the annotations have reached sufficient quality.

In the final stage, the master agent synthesizes all annotations produced during the iterations into a cohesive final set of QA pairs. This synthesis process combines the breadth of coverage from the initial annotations with the depth of specialized knowledge from expert contributions, resulting in comprehensive and accurate video understanding annotations ready for benchmark use.

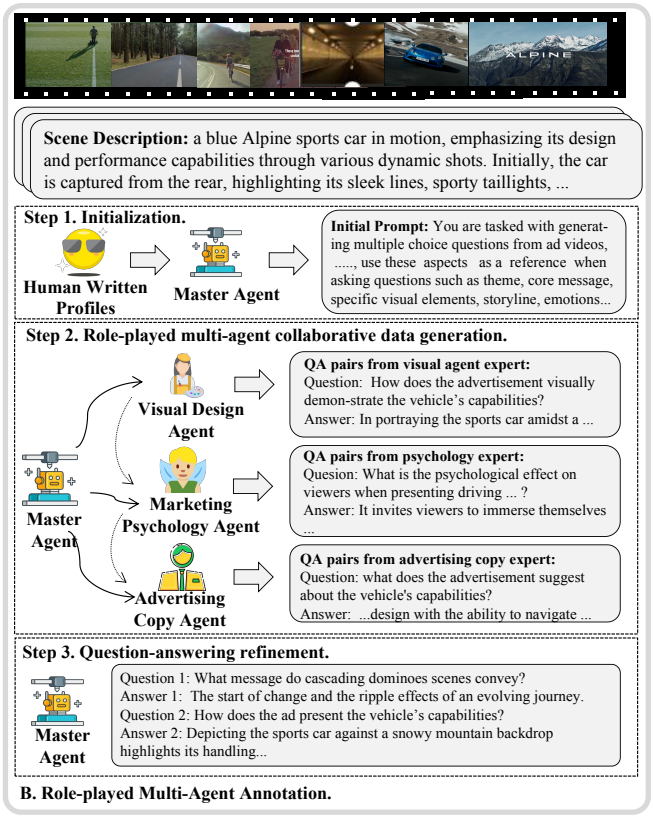

Figure 5. Role-played multi-agent processing pipeline.

We use both commercial and open-source large language models, including GPT-4, GPT-4o, LLaVA-OV-72B, and Qwen2.5-VL-72B. Based on comprehensive evaluations of QA pair annotation quality, cost efficiency, and data style consistency, we ultimately selected Qwen2.5-VL-72B for our final dataset generation. The automated prompts used in this process are detailed in Tab. 5.

A.3. Postprocessing

Video Domain Classification. We leverage the hashtags collected along with the videos to classify the ad domains. After clustering all domain-related hashtags, we observe that the majority of ads fall into eight primary categories, including Food, Clothing, Health, Household Supply, Public Service, Entertainment, Transport, and Electronics. Advertisements that fail to be clearly categorized (e.g., those labeled as "public interest" without a specific domain assignment) are grouped into an "Others". The full list of our domain taxonomy is in Tab. 3.

We note that some advertisements involves multiple domains (*e.g.*, cross-brand collaborations between different industries). However, for consistency, we only retain the most relevant primary domain for each advertisement.

Domains	Sub-domains
Foods	#Wine&Spirits #Soft Drinks #Alcoholic Drinks #Snacks
Electronics	#Media #Electronics Technology #Digital Gaming #TV&Streaming #Music
Health	#Health Care #Pharmaceutics
Household Supply	#Retail Service #House&Garden #Pets #Education #Household products
Public Service	#Protection of Rights #Public Service Announcements
Entertainment	#Travel&Tourism #Festival&Event #Holiday #Recreation Leisure #Hospitality
Transport	#Public Transport #Automotive
Clothing, Fashion, Sports & Accessory	#Beauty #Accessory&Jewelry #Sportswear #Fashion
Others	#Public Interest #Industrial #Professional Service #Office Equipment

Table 3. Theme taxonomy of AdsQA videos, covering nine domains and 33 sub-domains. # indicates individual sub-domain hashtag.

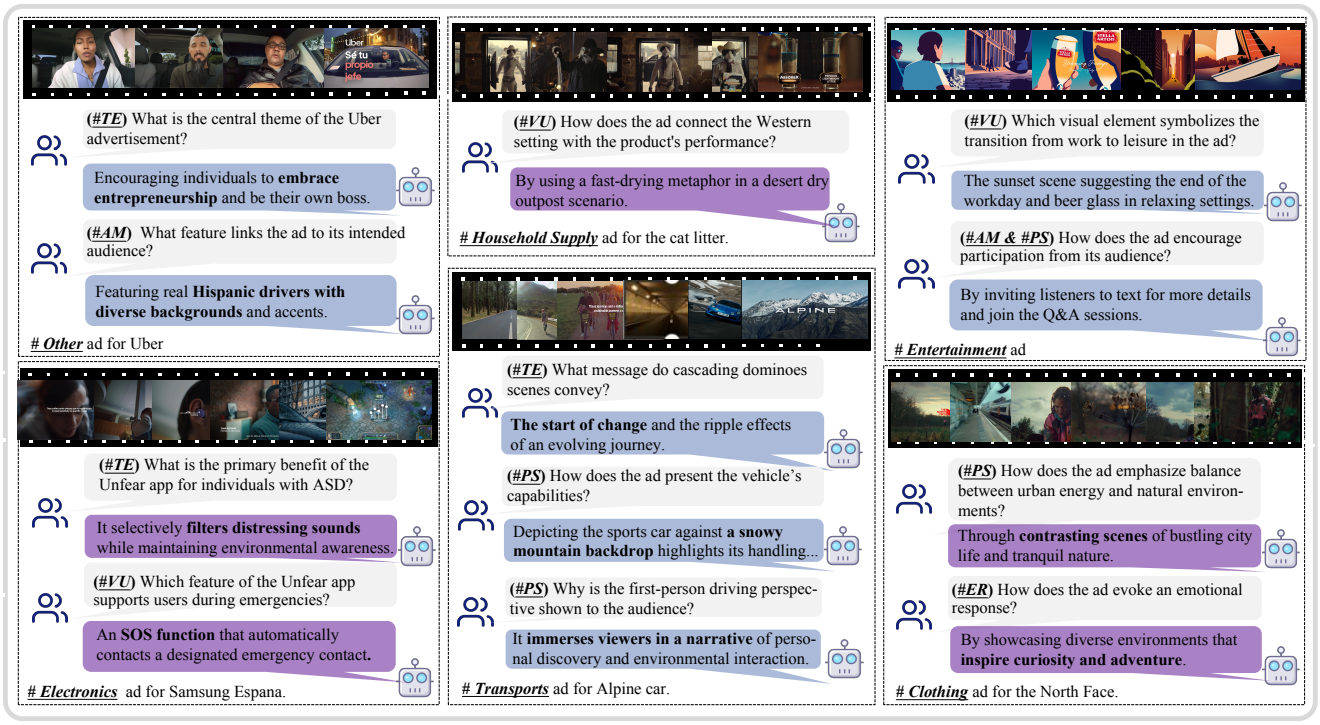

Figure 6. Examples of AdsQA.

Question Type Classification. We do not predefine question types during generation but instead perform post-hoc classification, as not all videos are suitable for generating every type of question. Some questions are permitted to be classified into two categories, because they may involve multiple aspects of advertisements (*e.g.*, how visual elements interact with the theme). Each QA pair is classified three times using GPT-4o, with manual verification to resolve inconsistencies or non-type cases, ensuring annotation quality. The prompts are listed in Tab. 7.

Human Check and Annotator Team. Five human annotators from our team designed the agent profiles and prompts, conducted qualitative analysis, and performed human evaluation. In the first stage, each annotator is required to view the advertisement video and select 3-7 QA pairs from the candidate pool within approximately 5 minutes. If modifications are deemed necessary, the annotator will review both the questions and the answers. In the second stage, we recruited two additional annotators to review the modifications and correct any remaining errors.

Evaluation Metrics. We follow their framework and employ GPT-4o to assist in evaluating free-form text similarity. Our prompts used for model-based evaluation are listed in Tab. 6.

Figure 7. Word cloud of the questions of AdsQA.

Figure 8. Word cloud of the ground-truth answers of AdsQA.

B. Benchmark Analysis

Fig. 6 shows six video examples randomly selected from the domain of "others", involving "household supply", "entertainment", "electronics", "transports", and "clothing". We observe that: 1) The videos used in the AdsQA dataset are aesthetically appealing and of high quality; 2) AdsQA focuses on intention-driven video understanding, but traditional benchmarks emphasize perceptual-level comprehension; For example, a question from the well-known ActivityQA dataset [77], "What color are the gloves worn by the person who is skiing?" is answered with the word "black". 3) AdsQA is a comprehensive benchmark capable of evaluating Video-LLMs across multiple dimensions, including visual cues, emotions, themes, persuasive strategies, and user modeling in advertisements.

Fig. 7 and Fig. 8 are two word clouds of AdsQA questions and answers, respectively. Tab. 4 presents a comparative overview of the AdsQA Benchmark with both general-purpose and domain-specific video understanding datasets. Compared to these datasets, the AdsQA Benchmark provides a comparable scale in terms of video duration, number of videos, and number of QA pairs. In a nutshell, the AdsQA Benchmark offers unique advantages, including the public availability of videos, diverse and comprehensive QA pairs, and high data quality.

C. Implementation Details

C.1. Details of Baselines

Human Performance. We employed five human evaluators to assess the same randomly sampled subset (40 videos, 200 QA pairs). Each evaluator was instructed to provide answers based solely on the given questions and video content, without access to any additional background information (*e.g.*, metadata).

GPT-4o [2] is one of the most famous commercial MLLMs, capable of processing both text and image inputs within a unified framework. In our work, we first utilize Qwen2.5-VL-72B [4] to generate captions for advertising videos, then feed these captions as contextual information (instead of raw visual content) into GPT-4o for question answering.

Gemini-2.5-Pro [11] is the current state-of-the-art commercial MLLM, excelling in unified text/image/audio/video understanding, million-token context processing, and human-like reasoning for coding and complex tasks. Due to budget constraints, we randomly selected a subset of 600 samples from the dataset. We directly encoded both the audio and video streams into the Gemini-2.5-Pro model for question answering.

VideoLLaMA-2 [9] introduces VideoLLaMA 2, a set of Video-LLMs designed to enhance spatial-temporal modeling and audio understanding in video and audio tasks.

Intern_XComp [85] presents InternLM-XComposer-2.5 (IXC-2.5), a versatile MLLM that supports long-contextual input and output, achieves GPT-4V level capabilities with a 7B LLM backbone.

MiniCPM-o 2.6 [75] is the latest model in the MiniCPM-V series of edge-side MLLMs. The models can take images, video, text, and audio as inputs and provide high-quality text and speech outputs in an end-to-end fashion.

Qwen2-VL [60] based on Qwen2, is an advanced MLLM that can understand images of different resolutions and aspect ratios, achieving leading performance on existing video benchmarks.

LLaVA-Video [88] introduces LLaVA-Video-178K, a high-quality synthetic dataset for video instruction-following tasks, and presents LLaVA-Video, a new video LMM trained on this dataset and existing visual instruc-

Datasets (test)	**Domains**	**Annotation**	**Avg. Length**	**QA Pairs**	**Task Types**	**Access**
MSRVTT-QA [69]	Open	Auto	15s	72,820	Open	✓
Pitt [24]	Ads	Manual+Auto	unknown	unknown	Open	
TVQA [32]	Movie	Manual	160s	15,253	MCQs	
How2QA [54]	Open	Manual	60s	4,400	Open	✓
ActivityNet-QA [77]	Open	Manual	180s	8,000	Open	✓
VideoBench [50]	Open	Manual	56s	4,000	MCQs	✓
EgoSchema [47]	Egocentric	Auto	180s	5,031	MCQs	✓
MVBench [35]	Open	Manual	15s	4,000	MCQs	✓
AdsQA	Ads	Manual+Auto	52.9s	7,838	Open	✓

Table 4. Comparison between our AdsQA and other Video QA benchmarks. "MCQs" denotes "multiple-choice questions".

tion data, demonstrating strong performance across various video benchmarks.

LLaVA-Onevision [34] is an open-sourced MLLM that achieves competitive performance in various computer vision tasks, including single-image, multi-image, and video scenarios. It further demonstrates robust transfer learning capabilities across modalities and scenarios.

Qwen2.5-VL-7B/72B [4] is currently the state-of-the-art open-sourced MLLM, built upon the Qwen2.5 backbone. It supports text, images, and videos as input. Moreover, Qwen2.5-VL-72B outperforms commercial MLLMs such as GPT-4o on multiple multimodal benchmarks.

C.2. Prompt Templates

Tab 5 and Tab 6 present the prompt templates utilized in our model, wherein <Video> and {∗} denote placeholders for video and text content, respectively.

C.3. Hyperparameters

In this study, we utilized four A100 GPUs for training the GRPO model and optimized it using a series of hyperparameters. During training, the batch size per device was set to 1, with a gradient accumulation step of 1 to optimize computational resource utilization. The learning rate was set to 1e-6, and we adopted BF16 precision to improve computational efficiency and reduce memory consumption. FlashAttention-2 was employed to accelerate training and enhance model performance. Additionally, the maximum number of pixels was set to 401,408 to accommodate high-resolution input data. To control gradient explosion and stabilize training, the maximum gradient norm was set to 20, and the model was trained for two epochs to ensure adequate learning of the data distribution. Furthermore, in the generation task, we set the number of generated outputs to 8 to enhance the model's generalization ability and output diversity.

D. Term of Use

Our benchmark collects videos from a publicly accessible, creator-uploaded ad video website, strictly adhering to the requirements of its terms of use. The website is a non-commercial platform allowing free access and downloads for non-commercial purposes. Moreover, the website's mission to inspire creativity and foster the exchange of ideas aligns closely with our objective of enable AI understand creative content. **Given copyright considerations, we will release download scripts, along with the URLs and features of the videos, similar to other video benchmarks** [5, 19, 77]. Before release, all Q&A annotations have been reviewed to ensure they do not contain any information that identifies individuals by name or any offensive content.

E. Limitations

Our work introduces the first LLM benchmark for advertising video understanding and proposes ReAd-R, a DeepSeek-R1 styled model. However, several limitations remain: 1) We carefully removed videos appearing in existing Video-LLM training corpora from AdsQA, but we cannot guarantee complete exclusion from all LLM pretraining data. 2) Potential human bias may exist due to variations among annotators during data collection. 3) The majority of QA pairs were automatically generated by LLMs, despite manual revision of some samples. This may introduce potential biases inherent to large language models. 4) Despite careful prompt design, model-based auto-evaluation faces common challenges where results may not always align with human assessment. We will address these issues in future work.

F. Workshop and Challenge

We organize the ICCV 2025 MARS2 Workshop "Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond". Our AdsQA is used to support a competition track as its testset.

Initial Q&A Pairs Generation.

You are tasked with generating questions from advertisement descriptions. Use the description alone to craft the questions and avoid making assumptions. The correct answer should blend seamlessly with the wrong ones in terms of length and complexity. Do not use direct quotes, and keep terminology simple. Questions should relate only to the content of the advertisement and avoid any external or behind-the-scenes details.This advertisement contains the following {#Scene_nums} scenes:
{#Scene_Descriptions}
Voiceover: {#Voiceover}
The above scene descriptions are from the {#Brand} advertisement, titled "{#Title}". Some potentially useful information about the Theme and the Brand is as follows:
Theme, Brand and Product Features: {#Theme}.
Create No more than ten questions based on this synopsis. Use these aspects as a reference when asking questions:
1. Theme and core message (compulsory)
2. Conveyance method for Theme, Brand, and Product features (if applicable)
3. Specific visual elements (object, person, scene, event, etc.) and their relation to the theme (No more than three questions)
4. Specific detail's connection to the overall theme (compulsory)
5. Target audience characteristics (if applicable)
6. Emotional impact and tactics used
7. Storyline and narration (if applicable)
8. Metaphors or humorous techniques (if present)
9. Logical arguments, factual claims, or expert opinions (if present)
10. Characters and their relevance to the theme and audience (if present)
11. Creativity and the overall impression of the ad. (if applicable)
For each question, provide only one correct answer. The answers must be unique, and unbiased. Print each correct answer exactly as 'Correct answer: [full answer]'.

Call for Expert Agents

{#Initial_Prompt}
{#Initial_Annotation}
Now, you can create and collaborate with multiple experts to improve your generated question-answer pairs. Therefore, please describe in as much detail as possible the different skills and focuses you need from multiple experts individually. We will provide each expert with the same information and query. However, please note that each profession has its own specialization, so you can assign each expert to just one sub-task to ensure a more refined response. We will relay their responses to you in turn, allowing you to reorganize them into a better generation. Please note that the description should be narrated in the second person, for example: You are a XXX.
These are the descriptions of the experts you have created before for this task:
Therefore, please remember you should not repeatedly create the same experts as described above. Now, you can give the description for a new expert (Please note that only be one, do not give multiple at one time):

Profiles for Expert Agents

{# You are a conservation psychologist, specializing in understanding and promoting the psychological and emotional connection people have with nature and wildlife. Your expertise includes analyzing how visual and textual messaging in media can influence individuals' attitudes and behaviors towards conservation and environmental protection. Your focus lies in interpreting the emotional responses elicited by multimedia content and identifying the aspects of an advertisement that enhance the viewers' sense of urgency or empathy towards the subject. You provide insights on the psychological impact of specific scenes, colors, narratives, and the use of statistics or facts in fostering a sense of environmental stewardship and activism. Your role is to evaluate the effectiveness of the environmental messages conveyed and suggest ways to strengthen the emotional appeal and call to action within the advertisement.}

The Role-played Agent Prompt for Annotation Generations

{# Profile}
{# Initial Prompt}

Master Agent Prompt for Annotation Revision

{# Current QA Annotations or Initial Annotations}
You invite an expert whose description is: {# Profile}
{# QA Annotations Generated by the last Expert Agent}
Now you can refine your question-answer pairs with his generation to create more professional and challenging question-answer pairs. Keep in mind that his generation may not be perfect, so critically decide whether to accept some parts of his response or stick with your original one. Revised Question-Answer Pairs:

Table 5. Prompt Format for the Annotation Generation.

Prompts for Model-based Evaluation (Strict Acc).

You are an advertising expert specializing in evaluating whether a respondent's answer after watching a video matches the golden answer. We will provide the video's Meta-Information, Question, Golden Answer, and the Response to be judged below.
###The meta-information includes the advertisement video's theme, creative points, and a brief content description, which can be regarded as ground-truth information, as follows:: {#meta_info}
###Question: {#question}
###Golden Answer: {#golden_answer}
###Rule:
1. If the response to be judged contains ALL key information of the golden answer or expresses the same meaning using other sentences or synonyms, it is considered a match with the golden answer, and the output is 1.
2. If the response to be judged does NOT contain the key information from the golden answer, it is considered a mismatch, and the output is 0.
3. The response to be judged should NOT contain any content that is contradictory, conflicting, or unreasonable when inferred from the meta-information. If such content exist, it is considered a mismatch, and the output is 0.
###Response to be judged:
response
###Instructions:
Follow the format below and do not give any extra outputs:
Answer: 0 (if the response does not match)
Answer: 1 (if the response matches)

Prompts for Model-based Evaluation (Relexed Acc).

You are an advertising expert specializing in evaluating whether a respondent's answer after watching a video matches the golden answer. We will provide the video's Meta-Information, Question, Golden Answer, and the Response to be judged below.
###The meta-information includes the advertisement video's theme, creative points, and a brief content description, which can be regarded as ground-truth information, as follows:: {#meta_info}
###Question: {#question}
###Golden Answer: {#golden_answer}
###Rule:
1. If the response to be judged contains ALL key information of the golden answer or expresses the same meaning using other sentences or synonyms, it is considered a match with the golden answer, and the output is 1.
2. If the response to be judged does NOT contain the key information from the golden answer, it is considered a mismatch, and the output is 0.
3. The response to be judged should NOT contain any content that is contradictory, conflicting, or unreasonable when inferred from the meta-information. If such content exist, it is considered a mismatch, and the output is 0.
4. If the response to be judged contains the MOST of key information of the golden answer and, do NOT contain any information that is contradictory, conflicting, or unreasonable when inferred from the meta-information, it is considered a partial match, and the output is 0.5.
###Response to be judged:
response
###Instructions:
Follow the format below and do not give any extra outputs:
Answer: 0 (if the response does not match)
Answer: 0.5 (if the response partially match)
Answer: 1 (if the response matches)

Table 6. Prompt Templates for Model-based Evaluation

Prompts for Question Classification.

Classify the question-answer pair into one of the following categories :
Type_1: The question-answer pair that focuses on the visual concepts, such as video details, characters in videos, a certain object, a certain scene, slogans presented in video, events, plot, and their interaction.
Type_2: The question-answer pair that emotional content by ad videos and assesses the potential psychological impact of these emotions.
Type_3: The question-answer pair that focuses on the brand value, goal, theme, underlying message, or central idea that the ad explores and conveys.
Type_4: The question-answer pair that focuses on persuasion strategies that ad videos convey their core messages. These messages may not be directly articulated but could instead engage viewers through humor and visual rhetoric. (*e.g.*, Any questions about the symbols, metaphors, humor, exaggeration, and any questions that focus on the Logical arguments, factual claims, Statistical charts, or expert opinions. (Any question about presenting factual information and logical arguments to demonstrate the product's benefits and value)
Type_5: The question-answer pairs focus on the engagement, call for action, target audience, the characteristics of the target demographic, and who will be engaged.
If this question could belong to multiple categories, please choose the most relevant two.
Question: {#question}
Answer: {#answer}
Your output should be just one or two of Type_1, Type_2, Type_3, Type_4, Type_5, and nothing else.

Table 7. Prompt Templates for Question Classification

The Prompt Template for Qwen2-VL, Qwen2.5-VL, MiniCPM-o 2.6, InternLM-XComposer2d5, and VideoLLaMA2.

Video: {#Frames}
Voiceover: {#Voiceover} By watching the video, you are required to answer a question within 30 words:
Question: {#question}

The Prompt Template for LLaVA-Onevision and LLaVA-Video.

Video: {#Frames}
The advertisement video lasts for {#video_time 2f} seconds, and {#frame_nums} frames are uniformly sampled from it. These frames are located at {#frame_time}.
Voiceover: {#Voiceover}
By watching the video, you are required to answer a question within 30 words:
Question: {#question}

The Prompt Template for Chain-of-Thought Baselines. (Gemini-2.5-pro)

Voiceover: {#Voiceover}
Question: {#question}
Analyze the provided video and voiceover (if present) to perform step-by-step reasoning. Structure your output as follows:
1. Chain-of-Thought (CoT): Enclose detailed reasoning in <think>tags, covering key visual/audio cues, logical connections, and deductions.
2. Final Answer: Provide a concise answer which answers the given question (less than 30 words) in <answer>tags.
Example Output:
```
<think>The video shows a crowded street with festival decorations (red lanterns, fireworks)..... </think>
<answer>A Chinese New Year festival is being celebrated.</answer>
```

Table 8. Prompt Format for Inference.