AdsQA: Towards Advertisement Video Understanding

Xinwei Long[1] Kai Tian[1*] Peng Xu[1✉] Guoli Jia[1] Jingxuan Li[2] Sa Yang[3]
Yihua Shao[4] Kaiyan Zhang[1] Che Jiang[1] Hao Xu[5] Yang Liu[2]
Jiaheng Ma[2] Bowen Zhou[1,6✉]
[1] Tsinghua University, [2] Independent Researcher, [3] Peking University, [4] CASIA
[5] Harvard University,[6] Shanghai Artificial Intelligence Lab
{longxw22, tk23}@mails.tsinghua.edu.cn; {peng_xu, zhoubowen}@tsinghua.edu.cn

Project page: https://github.com/TsinghuaC3I/AdsQA

MARS2 workshop & challenge: https://lens4mllms.github.io/mars2-workshop-iccv2025/

Abstract

Large language models (LLMs) have taken a great step towards AGI. Meanwhile, an increasing number of domain-specific problems such as math and programming boost these general-purpose models to continuously evolve via learning deeper expertise. Now is thus the time further to extend the diversity of specialized applications for knowledgeable LLMs, though collecting high quality data with unexpected and informative tasks is challenging. In this paper, we propose to use advertisement (ad) videos as a challenging test-bed to probe the ability of LLMs in perceiving beyond the objective physical content of common visual domain. Our motivation is to take full advantage of the clue-rich and information-dense ad videos' traits, e.g., marketing logic, persuasive strategies, and audience engagement. Our contribution is three-fold: ***(1)*** *To our knowledge, this is the first attempt to use ad videos with well-designed tasks to evaluate LLMs. We contribute* `AdsQA`*, a challenging ad Video QA benchmark derived from 1,544 ad videos with 10,962 clips, totaling 22.7 hours, providing 5 challenging tasks.* ***(2)*** *We propose* `ReAd-R`*, a Deepseek-R1 styled RL model that reflects on questions, and generates answers via reward-driven optimization.* ***(3)*** *We benchmark 14 top-tier LLMs on* `AdsQA`*, and our* `ReAd-R` *achieves the state-of-the-art outperforming strong competitors equipped with long-chain reasoning capabilities by a clear margin.*

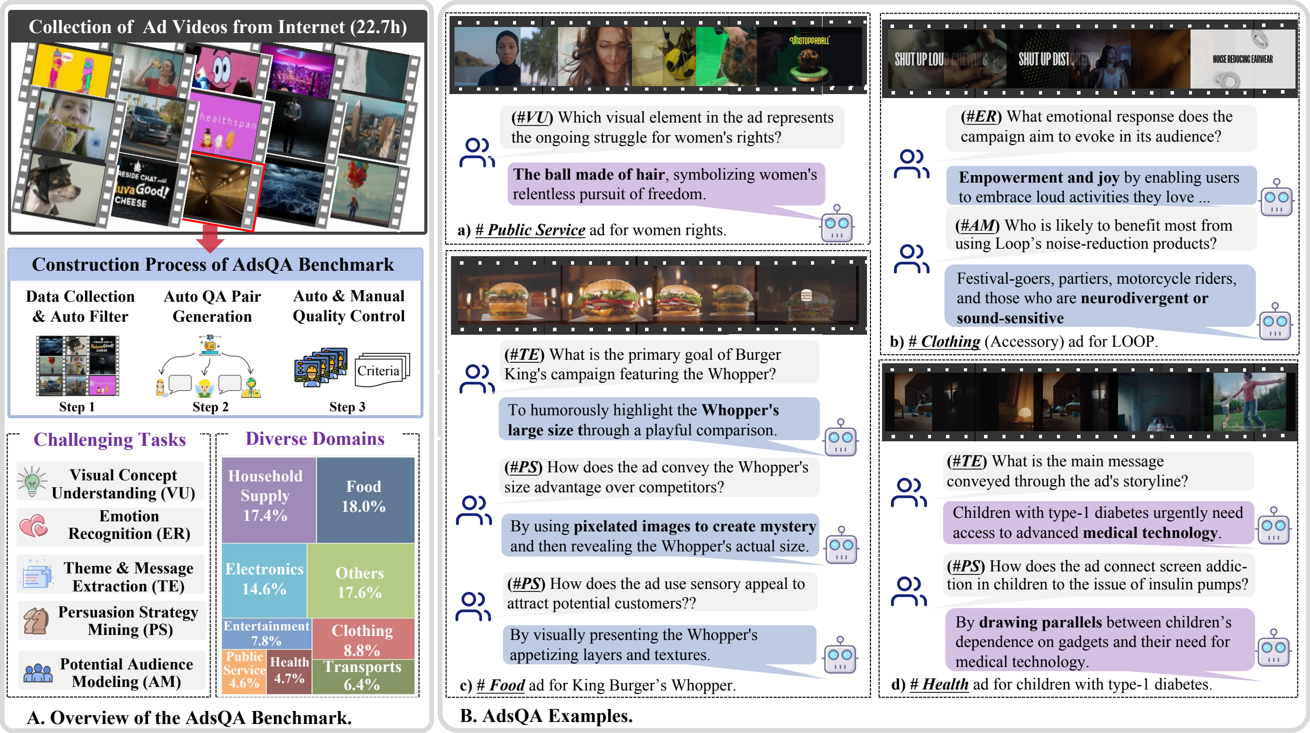
1. Introduction

A recent milestone flagged by OpenAI o1 [25] and DeepSeek-R1 [21] has been in the spotlight, and already opened up an era of large reasoning models. Meanwhile, going beyond general domains, the specialized domains [13, 93] are increasingly studied to boost these general-purpose LLMs to evolve continuously. The specialized domains contribute significantly in optimizing and evaluating generalist models towards specialized reasoning like domain experts [18, 52]. Recently, a clear phenomenon [20, 84] has been observed that LLMs are already good at reasoning a kind of step-wise problems represented by math and programming. As has been widely discussed [20, 84], the step-wise problems based on explicit theorems and programming syntax are compatible with the chain-style reasoning approach `''if A, then B''`. To further extend the diversity of specialized reasoning for LLMs, it would be better to exploit other novel domains where the domain-specific reasoning is unmanageable for "if A, then B".

Therefore, we, for the first time, propose to use advertisement videos to probe the reasoning boundary of LLMs in perceiving implicit multimodal reasoning. The domain-unique features of clue-rich and information-dense ad videos can be summarized as key words implicit, non-physical, mental, heuristic, *etc*. Different from user-uploaded videos on social media, ads are typically meticulously crafted by commercial or non-commercial organizations, making them entertaining, creative, aesthetically appealing, and capable of offering enjoyment and attracting viewer engagement.

Based on these domain-specific advantages of ad videos, we introduce `AdsQA`, a comprehensive and carefully curated VideoQA benchmark derived from 1,544 ad videos containing 10,962 clips, spanning a total of 22.7 hours of video content. The `AdsQA` introduces five tasks, each requiring different types and levels of reasoning: (1) Visual Concept Understanding: Identify and analyze visual elements in ads. (2) Emotion Recognition: Detect emotions and infer their roles in ads. (3) Theme and Core Message Extraction: Summarize the central theme and key messages

[*] Kai Tian has equal contribution with Xinwei Long.



(***#VU***) Which visual element in the ad represents the ongoing struggle for women's rights?

The ball made of hair, symbolizing women's relentless pursuit of freedom.

a) ***# Public Service*** ad for women rights.

(***#ER***) What emotional response does the campaign aim to evoke in its audience?

Empowerment and joy by enabling users to embrace loud activities they love ...

(***#AM***) Who is likely to benefit most from using Loop's noise-reduction products?

Festival-goers, partiers, motorcycle riders, and those who are **neurodivergent or sound-sensitive**

b) ***# Clothing*** (Accessory) ad for LOOP.

(***#TE***) What is the primary goal of Burger King's campaign featuring the Whopper?

To humorously highlight the **Whopper's large size** through a playful comparison.

(***#PS***) How does the ad convey the Whopper's size advantage over competitors?

By using **pixelated images to create mystery** and then revealing the Whopper's actual size.

(***#PS***) How does the ad use sensory appeal to attract potential customers??

By visually presenting the Whopper's appetizing layers and textures.

c) ***# Food*** ad for King Burger's Whopper.

(***#TE***) What is the main message conveyed through the ad's storyline?

Children with type-1 diabetes urgently need access to advanced **medical technology**.

(***#PS***) How does the ad connect screen addiction in children to the issue of insulin pumps?

By **drawing parallels** between children's dependence on gadgets and their need for medical technology.

d) ***# Health*** ad for children with type-1 diabetes.

Figure 1. Overview of `AdsQA` benchmark. Subfigs A - B: statistics & diversity, and examples.

of the ad. (4) Persuasion Strategy Mining: Analyze the strategies used to persuade the audience. (5) Potential Audience Modeling: Identify and characterize the target audience. The five tasks are formalized as the open-ended QA format. During the construction process, we propose an innovative Role-Played Multi-Agent Annotation framework, which simulates the role of advertising experts to generate specialized data and significantly reduces the workload of human annotators. To ensure data quality, we conduct multiple rounds of automated data cleaning and manual data revision. Moreover, based on our `AdsQA`, we benchmark 14 well-known LLMs, including the GPT-4 level models, recently-released reasoning-based models, *etc*.

Humans process visual stimuli from ads and naturally form mental impressions, rather than relying on rigid logical reasoning templates, *e.g.*, "if A, then B". Inspired by this heuristic process, we propose `ReAd-R`, a Deepseek-R1 styled Reinforced Ad Reasoner, which learns to gather effective visual stimuli from ad videos, then reflects on the questions, and finally answers the given questions in a human-like manner. Specifically, `ReAd-R` utilizes a reward function to evaluate the model's responses and adjusts its parameters based on correctness, facilitating learning through trial and error. As a result, `ReAd-R` enhances reasoning abilities through outcome-reward-driven optimization, eliminating the need for costly step-wise supervision or chain-of-thought (COT) training data. Experiments demonstrate that our model brings reasoning ability an obvious gain in comprehending implicit logic of ad videos. In contrast, other reasoning-based methods obtain limited improvement due to fixed COT templates and sub-optimal process reward models.

To our knowledge, our main contribution can be stated as:
(1) `AdsQA` is the first video QA benchmark for advertisement domain, which is also the first ad benchmark for LLMs, with domain-unique features implicit, non-physical, mental, heuristic, *etc*. It presents a new challenge to the current mainstream `''if A, then B''` LLM thinking approach, hence further extend the domain specialization reasoning scenarios for LLMs. Moreover, it advances ad video understanding beyond physical-content dominated shallow perception towards deeper cognitive reasoning.
(2) We propose `ReAd-R`, a DeepSeek-R1 styled RL reasoning model, which enhances the specialized reasoning capability to understand implicit logic in ad videos like the human thinking process from perception to cognition. `ReAd-R` also can be regarded as one of the earliest attempts evolving R1-technique to vision research.
(3) We benchmark 14 flagship LLMs on `AdsQA`, and our `ReAd-R` achieves the state-of-the-art outperforming strong competitors equipped with long-chain reasoning (*e.g.*, variants of GPT4, LLaVA, and Qwen) by a clear margin.

2. Related Work

Video Question Answering Benchmarks. Video question answering (Video QA) [7, 48, 55, 58, 68] aims to answer user questions based on a given video, requiring a detailed understanding of both the spatio-temporal information and the relationships between objects and events [5, 12, 66, 74]. The Video QA community has developed a variety of benchmarks. From the perspective of video content, most benchmarks focus on human-centric videos, such as NextQA [65], ActivityQA [77], MVBench [36], FunQA [67], and VideoMME [16], which are derived from movies [19, 59], TV shows [32], social media [67], *etc*. Additionally, there are benchmarks centered on object-centric videos, such as EgoSchema [47] , which are sourced from instructional and operational videos [38, 92]. Despite their advancements, these benchmarks still have several limitations. MovieQA [59] overly relies on dialogue understanding, which restricts the in-depth comprehension of visual content. Although TGIF-QA [26] requires some complex reasoning, the GIFs used are typically no longer than 3 seconds. ActivityQA [77] and Next-QA [65] have introduced open-ended QA tasks, but the annotated answers are often overly simplistic, usually consisting of just a few words. To sum up, current video QA benchmarks face limitations in video diversity and the setups of QA pairs.

Video QA Methods and Video-LLMs. Early video QA methods employed graph structures [29, 56, 90] or transformers [17, 33, 37] to model the spatiotemporal relationships in videos and capture interactions between humans and objects. Recently, Video-LLMs [6, 45, 46] have demonstrated impressive capabilities in video understanding. Representative Video-LLMs include the VideoLLaMA [9, 47] series, LLaVA series [34, 39, 41, 70, 86], InternLM series [62, 85], and Qwen series [4, 60]. Video-LLMs take both the question and the video as input and directly generate the answer, implicitly performing the reasoning and thinking process in an efficient manner. Moreover, Some methods attempt to explicitly model the reasoning process through step-by-step thinking or multi-round debates. For example, some studies [15, 22, 57, 63, 81] leverages chain-of-thought to simulate the thinking process, while VideoAgent [14, 61, 73] improves reasoning capabilities through agent collaboration.

Advertisement Understanding. Internet companies generate substantial profits by automatically distributing advertisements to target users, making the understanding of ad content highly important [28, 79]. The Pit dataset [24], one of the earliest efforts in automatic ad understanding, formalized this task as a visual question-answering (VQA) problem [43]. Subsequent research has built upon the Pit dataset or further developed it [53], or utilized their own private datasets [72]. These studies have mainly focused on a single aspect of advertisement understanding, such as persuasive strategies in image ads [31, 76], image ad search [89], intent understanding [27], and visual metaphor comprehension [3, 53, 71, 82, 83]. Although the Pit dataset primarily focuses on image advertisements, it also provided a subset of video advertisements they collected. However, the Pit dataset suffers from issues such as data inaccessibility, lack of diversity, and limited Q&A formats. Therefore, there is currently no comprehensive Video QA benchmark available in the advertising domain.

Reinforcement Learning for Reasoning. Recent research efforts [25, 30, 44, 94] have attempted to improve the reasoning capabilities of LLMs through reinforcement learning. Recently, DeepSeek-R1 [21] achieved significant improvements in reasoning based on reinforcement learning, eliminating the need for intermediate reasoning signals. S1 [49] demonstrated that even with only a few hundred data points, the model can effectively perform reasoning in specialized tasks. However, current research on RL-based reasoning primarily focuses on math and code problems, with few works [42] attempting to apply it to multimodal domains, particularly open-ended video QA tasks.

3. The `AdsQA` Benchmark

3.1. Task Definition

To comprehensively evaluate the model's ability to understand ad videos, we divided the questions into five types, with each one targeting a specific angle of ad analysis.

Visual Concept Understanding (VU) task evaluates the model's ability to comprehend specific visual concepts, such as characters, objects, scenes, slogans, and other details, as well as their interrelationships.

Emotion Recognition (ER) assesses if the models understand **what** emotion the ad evokes, and **how** ads establish connections with audiences through emotions.

Theme and Core Message Extraction (TE) drives the models to extract the underlying message or central idea that the ad explores. This task requires deep reasoning to gather visual information from the full video.

Persuasion Strategy Mining (PS) task evaluates the model's ability to uncover the strategies used to convey core messages and persuade audiences, such as humor, exaggeration, and visual rhetoric. The task includes questions, such as: **how** an ad conveys its central message, **why** the ad is appealing, and **what** strategies the ad employs.

Potential Audience Modeling (AM) probes the model's performance in identifying potential audience groups and their profiles. It sets key questions, such as **who** the ad targets and **what** the characteristics of the audience are. This task directly reflects the ad's influence and value.

These five tasks are defined as the open-ended video QA and each question is annotated with a ground-truth answer.

3.2. Dataset Construction Pipeline

AdsQA benchmark construction pipeline is in three stages:
Pre-Processing. To serve the ad video understanding, we consider the creativity, aesthetic quality, and availability of videos during our collection process. We do not use any private data; instead, we collect creative ad videos that adhere to the Creative Commons License from a creative community platform [1]. This platform offers publicly accessible, high-quality ad videos uploaded by creators. Specifically, we first crawl videos along with their metadata. The metadata is typically written by the uploader (*i.e.*, usually the creator of the ad) and includes information such as the theme, content, and key creative elements of the ad. It can be considered the "ground-truth" information of the ad video and serves as an important reference for our benchmark construction. Then, human experts are asked to carefully review the completeness and accuracy of the metadata and manually remove incomplete samples. Additionally, we conduct automated filtering based on video duration, aesthetic score, and content to ensure video quality and exclude non-ethical and negative content.

Considering efficiency and accuracy, we employ PySceneDetect to segment ad videos into fine-grained video clips to avoid missing detailed information in subsequent steps. For each clip C_i, we use Video-LLMs to generate its description $Desc_i$. To preserve key visual elements, we sample n keyframes $\{F_i^0, ..., F_i^{N_{max}}\}$ from each clip based on frame similarity calculated using SSIM [64], where n is determined by the duration of each clip. To retain speech information, we extract speech information Asr_i for each clip using the Whisper model, and translate them into English using GPT-4 [23]. Therefore, an ad video with N clips can be represented as a modality-interleaved sequence, as Eq. 1,

$$V = \{M, \{F_i^0, ..., F_i^{n_i}, Desc_i, Asr_i\}_{i=1}^{N}\}, \tag{1}$$

where M denotes the meta-information of the video. The modality-interleaved sequence will be used for automated annotation generation.
Role-Played Multi-Agent Annotation. To balance data quality and cost, ensure data diversity, and avoid template-driven outputs, we propose a Role-Played Multi-Agent Annotation framework to automatically generate video QA pairs. Inspired by previous research [78], we found that agents can develop specialized capabilities in the ad domain by configuring specific skill sets, such as marketing, visual design, and consumer psychology. Therefore, we can leverage AI agents to act as ad experts, analyzing diverse facets of ad videos and designing more specialized and challenging questions. We ask human experts to create profiles for advertising expert skills, enabling AI agents to adaptively select profiles (or autonomously generate new profiles) to accomplish role-playing tasks. Additionally, human experts provide examples of QA pairs as in-context demonstration to guide the agents in generating appropriate QA pairs.

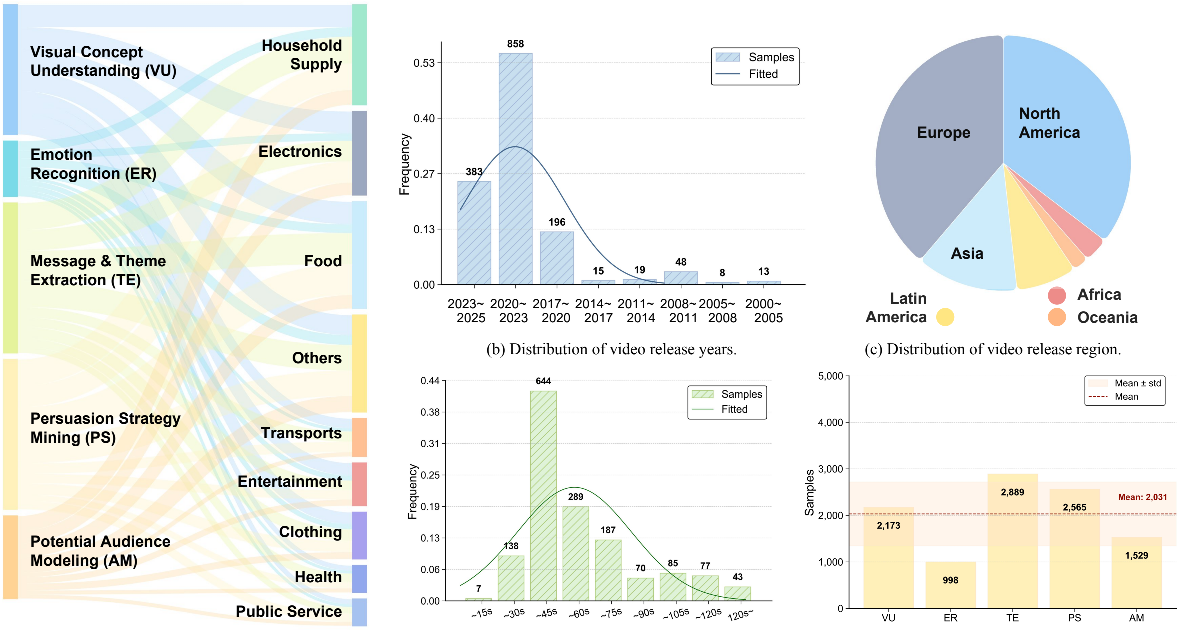
The Role-Played Multi-Agent Annotation is formulated into a three-stage pipeline. (1) In the initial stage, a master agent is recruited to generate a preliminary QA annotation for the given modality-interleaved sequence V. (2) In the iterative stage, the master agent checks the quality of the current QA annotations and decides whether to recruit a specialized expert agent. If necessary, the master agent selects a profile for the expert agent and instantiates it accordingly. The expert agent then generates new QA annotations leveraging its specialized expertise. Finally, the master agent revises the original annotations based on the expert agent's provided annotations and assesses whether to terminate the iteration or continue to recruit a new expert agent. (3) After terminating the iteration, the master agent synthesizes outputs from previous iterations to produce the final QA annotation.
Automated Data Cleaning. After automated annotation through multi-agent collaboration, we employ the IXC-2.5-Reward model [80] to automatically verify the correctness of the answers. Given the video, question, and meta information, the reward model scores the answer. QA pairs with scores below a certain threshold are flagged and carefully reviewed by annotators.

3.3. Manual Check and Quality Control

Following the automatic annotation generation, we conduct a careful manual check and revision on all Q&A pairs, focusing on the question suitability, annotation correctness, video content, and task difficulty. To ensure the quality of Q&A pairs, we first remove questions that are homogeneous, unrelated to the ads' theme, or cannot be answered based on the ad video (*e.g.*, those requiring external knowledge). Then, we incorporate the meta information as ground truth to verify the accuracy of the reference answers. We eliminated pairs containing inaccurate, ambiguous, or biased answers. After two rounds of rigorous selection, only 37% of the QA pairs remain. Additionally, we manually revise the wording of some QA pairs to improve their clarity and accuracy. Any modifications to a sample are reviewed by other annotators to ensure consistency and reliability.

3.4. Dataset Statistics

Videos. Our AdsQA Benchmark includes a total of 1,544 unique advertisements, comprising 10,962 video clips for evaluation. As shown in Fig. 2a, these ad videos span 9 primary domains. The video durations range from 15 to 120 seconds, with an average length of 52.9 seconds, amounting to a total of 22.7 hours of content. Each ad video is accompanied by metadata, including a title, tags, automatic speech recognition (ASR) results, and descriptions. This metadata is utilized for Q&A generation, as described in Sec. 3.1.



(a) Alluvial diagram of task and domain distribution.

(b) Distribution of video release years.

(c) Distribution of video release region.

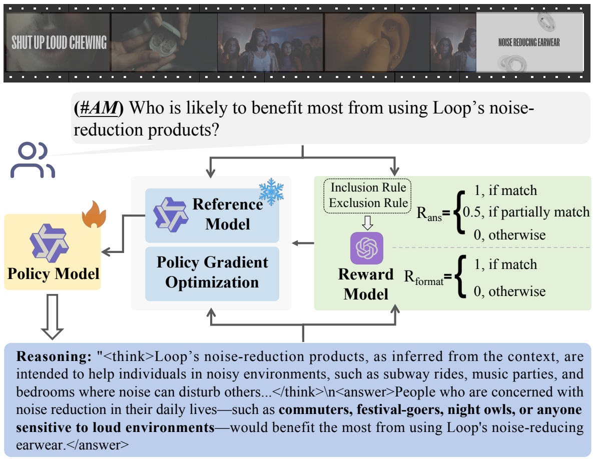
(d) The distribution of the video duration in the dataset.

(e) Number of Q&A entries for different tasks.

Figure 2. Statistics of Our `AdsQA` Bench.

Most of the ad videos are in English and primarily originate from North America and Europe.

Q&A Pairs. Our `AdsQA` Benchmark includes 7,859 QA pairs (among which 29.2% of the questions can be categorized into two types, resulting in a total of 10,154 QA pairs for evaluation). Among these: Visual Concept Understanding accounts for 21.4% of the QA pairs, Theme and Message Extraction accounts for 28.5%, Persuasion Strategy Mining accounts for 25.3%, Potential Audience Modeling accounts for 15.1%, and Emotion Recognition accounts for 9.8%. Typical persuasion strategies include metaphor, symbolism, humor, expert opinion, and others. Some questions may fit into two categories, such as "What is the role of a particular visual concept in relation to the theme?". On average, each question consists of 17 words and each answer contains approximately 12 words.


SHUT UP LOUD CHEWING
NOISE REDUCING EARWEAR
Reference Model
Policy Model
Policy Gradient Optimization
Inclusion Rule
Exclusion Rule
Reward Model
$R_{ans} = \begin{cases} 1, \text{ if match} \\ 0.5, \text{ if partially match} \\ 0, \text{ otherwise} \end{cases}$
$R_{format} = \begin{cases} 1, \text{ if match} \\ 0, \text{ otherwise} \end{cases}$

(***<u>#AM</u>***) Who is likely to benefit most from using Loop's noise-reduction products?

Reasoning: "<think>Loop's noise-reduction products, as inferred from the context, are intended to help individuals in noisy environments, such as subway rides, music parties, and bedrooms where noise can disturb others...</think>\n<answer>People who are concerned with noise reduction in their daily lives—such as **commuters, festival-goers, night owls, or anyone sensitive to loud environments**—would benefit the most from using Loop's noise-reducing earwear.</answer>

Figure 3. Framework of `ReAd-R`. Given a question and video, the policy model generates multiple responses. The reward model evaluates and scores them, and the rewards are used to update the policy model via policy gradient optimization.

4. Methodology

As illustrated in Fig. 3, we propose `ReAd-R`, a DeepSeek-R1 styled Reinforced Ad Reasoner, to simulate human heuristic thinking and learn from trial and error. The model takes an ad video and a question as input. The policy model generates a set of reasoning processes, including thoughts and answers, based on the given input. Each reasoning process is evaluated using a reward function to compute its reward value. After calculating the reward values for all outputs, each reasoning process is assessed and used to update the policy model. Additionally, `ReAd-R` employs KL divergence to control the difference between the policy model and the reference model, ensuring stable training.

Data Preparation. Inspired by previous work [21, 49], RL-based reasoners can improve reasoning capabilities with limited high-quality data. We additionally crawled a total of 1,053 ad videos along with their metadata and automatically generated annotations using the same method. We utilized metadata tags collected during data crawling (*e.g.*, topic, domain) to ensure diversity, adopted Video-LLM evaluation results to assess difficulty, and performed manual quality assessment based on the same criteria as in Sec. 3.3. Finally, we selected 100 videos and 500 question-answer pairs. We designed a prompt format to guide the model to output its reasoning process before providing the final answer.

Reward Modeling. The reward model guides the model to find the optimal direction through trial and error. Ideally, ReAd-R's responses should (1) include as many elements of the standard answer as possible and (2) avoid containing content not mentioned in the ad video (*e.g.*, hallucinations). To achieve this, we propose a rule-guided LLM evaluator as the reward model to assess the quality of each response. We introduce two sets of rules: inclusion rules and exclusion rules. The inclusion rule requires that the generated answer incorporate as many elements of the standard answer as possible. If satisfied, the reward value is 1.0; otherwise, it is 0. The exclusion rule states that if the generated content includes elements not mentioned in the standard answer and these elements cannot be inferred from the meta-information, they should be judged as incorrect. To avoid sparse reward values in the early stages, we relaxed the inclusion rule. If the generated response partially includes factual elements, it is assigned a reward value of 0.5. Additionally, the format reward is used to enforce the model's predictions to adhere to the required format of <think> and <answer>, as $R(\cdot) = R_{ans}(\cdot) + R_{format}(\cdot)$.

Reinforced Fine-tuning. Following DeepSeek-R1, we employ the GRPO algorithm instead of PPO to optimize our model. ReAd-R first generates n distinct responses $O = \{o_1, o_2, ..., o_n\}$ from the current policy model $\pi_{\theta_{old}}$. Then, the reward model $R(\cdot)$ evaluates these responses O to obtain their reward value as $\{r_1, r_2, ..., r_n\}$, and the advantage A_i is defined as

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_n\})}{\text{std}(\{r_1, r_2, ..., r_n\})}. \tag{2}$$

A_i denotes relative quality of the i^{th} answer when compared to the group of rewards $\{r_1, r_2, ..., r_n\}$. This strategy encourages the model to generate better answers with higher reward values within the group. We use GRPO algorithm [20] to optimize our policy model based on A_i.

5. Experiments

5.1. Experimental Settings

Evaluated Models. We select 14 top-tier LLMs as strong competitors covering diverse categories: (1) GPT-4 Level Baselines: including GPT-4o [23], GPT-4V [2] and Qwen2.5-VL-72B [4]. (2) Open-sourced Video LLMs: We present at least one representative model released in the past six months from each series of Video-LLMs, such as Qwen2.5-VL-7B [4], LLaVA-Onevision-7B [34], and others. (3) Reasoning Based Methods: We re-implemented several reasoning-based methods and applied them to the AdsQA task. These methods include the Video Chain-of-Thought (VOT) [15], the Role-Played Agent (EvoAgent) [78], and the Monte Carlo Tree Search (MCTS) algorithms [84]. Detailed descriptions and implementation methods are provided in the appendix.

Evaluation Metric. All five tasks of AdsQA are evaluated as open-ended QA, allowing for the free-text evaluation methods. Traditional free-text metrics [40, 51] are sensitive to lexical variations, which may introduce bias when evaluating different LLMs. Recent studies [8, 10, 19, 67] have shown promising results in using large generalist models to evaluate generated text. Therefore, we follow their framework and employ GPT-4o to assist in evaluating free-text similarity. Specifically, we use the same inclusion and exclusion rules mentioned in Sec. 4 to guide the scoring of text generated by the large model. We provided each sample with a ground-truth answer and meta-information; each meta-information includes relevant information such as the creative elements, content, storyline, and themes of the advertisement video. We propose both relaxed and strict scoring modes. Our **strict accuracy** Acc_{strict} is defined as:

$$\text{Acc}_{\text{strict}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = y_i\}, \tag{3}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function that returns 1 if the model's response x_i semantically matches all elements of the standard answer y_i, and 0 otherwise.

In the relaxed mode, the **relaxed accuracy** Acc_{relaxed} is defined as:

$$\text{Acc}_{\text{relaxed}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}\{x_i = y_i\} + \lambda \cdot \mathbb{1}\{x_i \approx y_i\}), \tag{4}$$

where $\mathbb{1}\{x_i \approx y_i\}$ returns 1 if the response x_i partially addresses the elements of y_i, and 0 otherwise. It is observed $\lambda = 0.5$ works well in experiments. Our prompt template for model-based evaluation are in the appendix.

Implementation Details. Our ReAd-R is model-agnostic. In our main experiments, we utilize Qwen2-VL-7B as the base model. We froze the parameters of the visual backbone and only fine-tuned the parameters of the language model. We conducted the experiments with a batch size of 4 and a learning rate of 1e-6 on a server with eight A100 GPUs, with a training duration of 12 hours. Additional details and hyperparameters are provided in the appendix.

5.2. Results and Observations

Although GPT-4 achieves over 85% accuracy on other datasets such as Next-QA, its strict accuracy on our dataset is only 29.4%, and its relaxed accuracy is just 56.6%. This shows that our benchmark exceeds the capabilities of some

Model	Strict Accuracy						Relaxed Accuracy					
	VU	ER	TE	PS	AM	Overall	VU	ER	TE	PS	AM	Overall
Human Results∗	55.1	39.4	57.1	57.5	45.8	51.3	74.5	65.9	75.7	76.2	64.6	71.4
Commercial Large Muitimodal Model												
GPT-4o [23]	24.9	26.5	32.6	32.3	31.0	29.4	50.1	57.8	62.4	55.2	54.8	56.6
Gemini-2.5-Pro∗ [11]	36.0	44.0	40.4	29.3	33.3	35.9	59.2	69.0	64.0	54.4	62.4	60.7
Open-sourced Video-LLMs												
VideoLLaMA2-7B [9]	4.56	7.75	7.28	6.06	7.38	6.48	21.6	29.0	28.4	22.6	27.2	25.2
InternLM-XComp.2.5-7B [85]	11.2	10.7	16.3	11.9	12.1	12.6	34.7	35.9	41.5	34.1	37.3	36.5
LLaVA-OneVision-7B [34]	11.8	11.6	16.2	13.7	15.1	14.0	35.8	39.5	43.2	36.8	41.7	39.1
LLaVA-Video-7B [87]	14.0	14.4	18.7	15.2	17.3	16.1	37.8	41.6	45.1	38.1	43.6	41.0
MiniCPM-o_2.6-7B [75]	13.0	15.3	17.3	14.9	18.9	16.3	38.1	43.4	45.7	39.1	45.0	42.5
Qwen2-VL-7B [60]	**13.7**	**15.4**	**20.2**	**16.0**	**20.0**	**17.2**	**36.8**	**41.3**	**46.0**	**37.7**	**44.6**	**41.0**
Qwen2.5-VL-7B [4]	20.2	23.2	24.6	20.5	24.3	23.0	45.8	50.4	51.8	45.3	50.9	48.9
Qwen2.5-VL-72B [4]	26.7	30.2	34.8	29.8	34.7	31.0	51.8	57.4	59.5	53.8	59.2	55.8
Reasoning-based Models												
VOT (Qwen2-VL-7B) [15]	14.3	16.4	19.5	12.6	22.0	17.0	36.5	43.3	42.5	35.2	45.1	40.2
EvolAgent (Qwen2-VL-7B) [78]	15.0	16.1	21.6	16.5	20.4	18.3	38.6	43.4	46.4	40.4	45.3	42.6
MCTSr (Qwen2-VL-7B) [84]	14.8	10.2	19.3	16.1	16.4	17.1	40.6	41.3	46.5	41.1	42.6	43.4
Qwen2-VL-7B (SFT) [60]	10.5	12.5	17.9	13.8	15.9	14.1	34.8	42.1	42.7	34.9	41.4	38.8
ReAd-R (Qwen2-VL-7B) (Ours)	**15.8**	**19.3**	**21.7**	**18.5**	**18.8**	**18.6**	**42.3**	**46.4**	**50.0**	**43.4**	**47.8**	**44.9**
ReAd-R (Qwen2.5-VL-7B) (Ours)	20.4	27.9	27.2	22.1	25.5	25.0	46.2	56.2	54.6	48.2	52.6	51.5

Table 1. Experimental Results.

multi-modal large models. We observe that the state-of-the-art multi-modal models show strong visual perception, often generating partial answers by describing video segments. However, fully understanding the underlying meaning of ads is difficult for all baseline models. This requires not only identifying specific visual elements but also interpreting their implicit logic.

Different tasks vary in difficulty, but their performance is comparable. Though tasks focus on different aspects, these aspects are closely linked. For example, ads may use emotions or specific visuals to convey themes. Most models perform best on Theme and Message Extraction and Audience Modeling tasks. This is because ad videos aim to deliver clear messages to specific audiences, making it relatively easier to identify themes and target users. Even without fully understanding creative elements or marketing strategies, models can infer these from characters, scenes, or slogans in the ad. The Persuasion Strategy Mining task is more challenging. It requires models to explain how and why ads use certain designs or elements. These questions are often not directly expressed in the video and may even require interpreting counterfactual or unexpected visual information, which increases the difficulty of this task. Models also perform worse on the Visual Concept Understanding (VU) task. While other video QA benchmarks [77] focus on surface-level questions like "What color are the gloves?", our VU task requires understanding global ad information, such as "Which scene represents the ad's theme?" Answering this question requires understanding the theme, identifying related scenes, and describing them briefly.

On the AdsQA task, reasoning models that excel in code and math tasks achieved only marginal results, even with ten times the computational cost. This shows that AdsQA reasoning differs significantly from tasks like math and code. Math and code rely on structured "if A, then B" reasoning, while ads require associative reasoning, connecting concrete visuals to abstract concepts. Prior work [91] also suggests chain-of-thought reasoning can hinder abstract reasoning performance. Though MCSTr improved results by 2.4% over Qwen2-VL, its search process is inefficient and highly dependent on the reward model. In some cases, Qwen2-VL answers correctly, but MCSTr fails due to incorrect search directions from the reward model. Moreover, training an effective reward model is difficult, especially in the data-scarce ad domain.

Our method, ReAd-R, achieves a 3.9% improvement using only 500 video-question-answer triplets, outperforming Qwen2-VL and other reasoning models. During inference, our model reasons about video content and questions by generating free-text explanations, eliminating the need for

	Strict Acc.	Relaxed Acc.
ReAd−R (Our Model)	**18.6**	**44.9**
w. Uncurated Data	13.2	36.2
w. Strict Reward	15.1	41.7
w.o. Prompt Constraints	16.8	42.9
Qwen2-VL (SFT)	14.1	38.8
w. Uncurated Data	16.0	40.9
VOT (Chain length: 3)	17.0	40.2
w. (Chain length: 5)	13.6	36.2

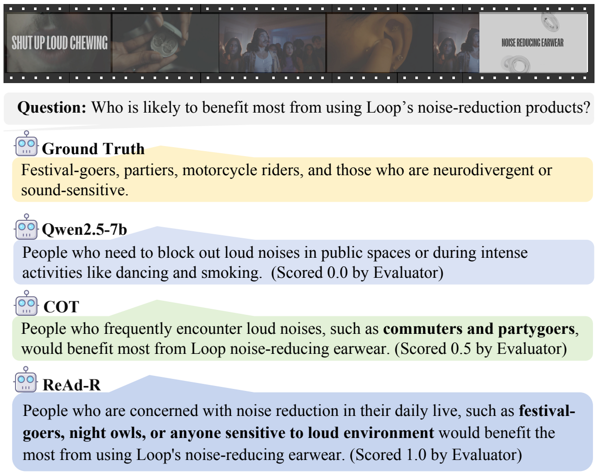
Table 2. Ablation study on ReAd−R, Qwen2-VL SFT, and VOT.

fixed chain-of-thought (COT) templates (*e.g.*, VOT), additional reward functions (*e.g.*, MCTSr), or iterative search processes (*e.g.*, MCTSr and EvolAgent). Interestingly, we observe that supervised fine-tuning (SFT) on the same data caused a sharp performance drop of 2.2% for Qwen2-VL. Using limited but high-quality data does not improve the performance of SFT models. SFT may overfit to the features of the training data, leading to poor generalization on diverse test sets. In contrast, our ReAd−R model improves reasoning capabilities through RL on limited data, achieving generalization across diverse advertisement videos.

5.3. Ablation Study

We conduct ablation studies for our ReAd−R model in Tab 2. We first show results of RL fine-tuning on 1,053 uncurated videos and 5,159 QA pairs. Our model performs worse than Qwen2-VL. This is due to the GRPO algorithm's sensitivity to data quality. Early in fine-tuning, low-quality data can mislead optimization, causing suboptimal results. Next, we test a strict reward rule (only fully correct answers receive a R_{ans} reward of 1; otherwise, the reward is 0). Our method does not achieve improvement through RL fine-tuning. During training, we observe that the strict reward rule makes it hard for the model to earn rewards, limiting gains. We also find ReAd−R is sensitive to prompt templates. Properly constraining the model's thinking and answers (*e.g.*, avoiding repetitive content in <think> </think> and <answer> </answer> tags) stabilizes training and improves performance.

We also conduct ablation studies on our implemented baselines (*e.g.*, VOT, Qwen2-VL SFT). We present the results of Qwen2-VL fine-tuned using uncurated data. Interestingly, the model fine-tuned on uncurated data improves by 2.1% over limited-data fine-tuning but still lags behind the non-SFT model. This shows SFT does not enhance reasoning on ad data. Additionally, we experiment with VOT using a COT length of 5 (as set in its original paper) and observe a significant performance drop of 4.0%. The model does not benefit from step-by-step reasoning; instead, accumulated errors degrade its performance. This highlights the difference between abstract reasoning in the ad domain and chain-style logical reasoning.



Question: Who is likely to benefit most from using Loop's noise-reduction products?

Ground Truth
Festival-goers, partiers, motorcycle riders, and those who are neurodivergent or sound-sensitive.

Qwen2.5-7b
People who need to block out loud noises in public spaces or during intense activities like dancing and smoking. (Scored 0.0 by Evaluator)

COT
People who frequently encounter loud noises, such as **commuters and partygoers**, would benefit most from Loop noise-reducing earwear. (Scored 0.5 by Evaluator)

ReAd-R
People who are concerned with noise reduction in their daily live, such as **festival-goers, night owls, or anyone sensitive to loud environment** would benefit the most from using Loop's noise-reducing earwear. (Scored 1.0 by Evaluator)

Figure 4. Case Study. Due to the limited page space, we omit the model's thinking process and show only final answer.

5.4. Case Study

As shown in Fig. 4, Qwen2.5-VL generates vague and overly general answers (scored 0 by GPT-4). The COT baseline produces better answers (scored 0.5 by GPT-4) by gathering information step-by-step through a reasoning process, but it still misses some important clues in the video. Our ReAd−R learns to perform video-grounded reasoning through reinforcement fine-tuning, deriving the final answer by summarizing its reasoning process. The reinforced thinking-answer mechanism helps ReAd−R generate more complete and precise answers.

6. Conclusion

We contibute AdsQA, the first advertisement benchmark for LLMs, with domain-unique features implicit, non-physical, mental, heuristic, *etc*. It presents a new challenge to the current "if A, then B" LLM thinking approach, hence further extend the domain specialization reasoning scenarios for LLMs. We propose ReAd−R, a DeepSeek-R1 styled RL reasoning model, no need of costly supervision, which enhances the specialized reasoning capability to understand implicit logic in ad videos. We benchmark 14 flagship LLMs on AdsQA, and our ReAd−R achieves the state-of-the-art outperforming strong competitors equipped with long-chain reasoning capabilities (*e.g.*, variants of GPT4, LLaVA, and Qwen).

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2022ZD0160603), and NSFC No.62306162.

References

[1] https://www.adsoftheworld.com/. 4

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 15

[3] Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211, 2023. 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 7, 15, 16

[5] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *CoRR*, abs/2410.10818, 2024. 3, 16

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[7] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *arXiv preprint arXiv:2408.14469*, 2024. 3

[8] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*, 2023. 6

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024. 3, 7, 15

[10] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023. 6

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7, 15

[12] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2024. 3

[13] Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, et al. Automating exploratory proteomics research via language models. *arXiv preprint arXiv:2411.03743*, 2024. 1

[14] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 3

[15] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024. 3, 6, 7

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 3

[17] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14773–14783, 2023. 3

[18] Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23923–23932, 2025. 1

[19] Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. Short film dataset (SFD): A benchmark for story-level video understanding. *CoRR*, abs/2406.10221, 2024. 3, 6, 16

[20] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. 1, 6

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 3, 5

[22] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025. 3

[23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4, 6, 7

[24] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *CVPR*, pages 1100–1110. IEEE Computer Society, 2017. 3, 16

[25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander

Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 1, 3

[26] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 3

[27] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996, 2021. 3

[28] Zhiwei Jia, Pradyumna Narayana, Arjun R Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. Kafa: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. *arXiv preprint arXiv:2305.18373*, 2023. 3

[29] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11109–11116, 2020. 3

[30] Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F Chen, Shafiq Joty, and Furu Wei. Preference optimization for reasoning with pseudo feedback. *arXiv preprint arXiv:2411.16345*, 2024. 3

[31] Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. In *Proceedings of the AAAI conference on artificial intelligence*, pages 57–66, 2023. 3

[32] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, pages 1369–1379. Association for Computational Linguistics, 2018. 3, 16

[33] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 3

[34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 6, 7, 16

[35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multimodal video understanding benchmark. In *CVPR*, pages 22195–22206. IEEE, 2024. 16

[36] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 3

[37] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8658–8665, 2019. 3

[38] Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-sentence grounding for long-term instructional video. In *European Conference on Computer Vision*, pages 200–216. Springer, 2024. 3

[39] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122, 2023. 3

[40] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[42] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3

[43] Xinwei Long, Zhiyuan Ma, Ermo Hua, Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Retrieval-augmented visual question answering via built-in autoregressive search engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 24723–24731, 2025. 3

[44] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024. 3

[45] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. 2023. 3

[46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arxiv*, 2024. 3

[47] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023. 3, 16

[48] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 3

[49] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. 3, 5

[50] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *CoRR*, abs/2311.16103, 2023. 16

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[52] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: a comprehensive evaluation. *arXiv preprint arXiv:2407.08940*, 2024. 1

[53] Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, and Sumit Shekhar. Seeing the unseen: Visual metaphor captioning for videos. *arXiv e-prints*, pages arXiv–2406, 2024. 3

[54] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 16

[55] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 3

[56] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. *arXiv preprint arXiv:2504.13092*, 2025. 3

[57] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Unlocking video-llm via agent-of-thoughts distillation. *arXiv preprint arXiv:2412.01694*, 2024. 3

[58] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 3

[59] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640. IEEE Computer Society, 2016. 3

[60] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 7, 15

[61] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 3

[62] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 3

[63] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024. 3

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[65] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE, 2021. 3

[66] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 3

[67] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024. 3, 6

[68] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025. 3

[69] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 16

[70] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 3

[71] Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. Exploring chain-of-thought for multi-modal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, 2024. 3

[72] Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. Synchronized video storytelling: Generating video narrations with structured storyline. *arXiv preprint arXiv:2405.14040*, 2024. 3

[73] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). *arXiv preprint arXiv:2401.08392*, 2024. 3

[74] Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Yufei Wu, Shichao Su, Yifan Xu, Wenxi Zeng, et al. Lens: Multi-level evaluation of multimodal reasoning with large language models. *arXiv preprint arXiv:2505.15616*, 2025. 3

[75] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7, 15

[76] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1308–1323, 2019. 3

[77] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134. AAAI Press, 2019. 3, 7, 15, 16

[78] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. *arXiv preprint arXiv:2406.14228*, 2024. 4, 6, 7

[79] Majid Zahmati, Seyed Morteza Azimzadeh, Mohammad Saber Sotoodeh, and Omid Asgari. An eye-tracking study on how the popularity and gender of the endorsers affected the audience's attention on the advertisement. *Electronic Commerce Research*, 23(3):1665–1676, 2023. 3

[80] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025. 4

[81] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 3

[82] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, 2021. 3

[83] Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154, 2023. 3

[84] Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*, 2024. 1, 6, 7

[85] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 3, 7, 15

[86] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3

[87] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7

[88] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 15

[89] Kang Zhao, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wen Tao, Cong Han, Shuanglong Li, and Lin Liu. Enhancing baidu multimodal advertisement with chinese text-to-image generation via bilingual alignment and caption synthesis. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2855–2859, 2024. 3

[90] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, page 8, 2018. 3

[91] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 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, 2024. 7

[92] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[93] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025. 1

[94] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. 3