

# ARC-Hunyuan-Video-7B: Structured Video Comprehension of Real-World Shorts

Yuying Ge<sup>1,\*</sup>, Yixiao Ge<sup>1,\*†</sup>, Chen Li<sup>1,\*</sup>, Teng Wang<sup>1,\*</sup>, Junfu Pu<sup>1,\*</sup>, Yizhuo Li<sup>1,\*</sup>, Lu Qiu<sup>1,\*</sup>, Jin Ma<sup>2</sup>, Lisheng Duan<sup>2</sup>, Xinyu Zuo<sup>2</sup>, Jinwen Luo<sup>2</sup>, Weibo Gu<sup>3</sup>, Zexuan Li<sup>4</sup>, Xiaojing Zhang<sup>2</sup>, Yangyu Tao<sup>3</sup>, Han Hu<sup>3</sup>, Di Wang<sup>3</sup>, Ying Shan<sup>1</sup>

<sup>1</sup>ARC Lab, Tencent PCG, <sup>2</sup>Search Application Department, Tencent CSIG, <sup>3</sup>Tencent Hunyuan, <sup>4</sup>Big Data Platform Department, Tencent PCG

\*Core contributors, †Project lead

Real-world user-generated short videos, especially those distributed on platforms such as WeChat Channel and TikTok, dominate the mobile internet. However, current large multimodal models lack essential temporally-structured, detailed, and in-depth video comprehension capabilities, which are the cornerstone of effective video search and recommendation, as well as emerging video applications. Understanding real-world shorts is actually challenging due to their complex visual elements, high information density in both visuals and audio, and fast pacing that focuses on emotional expression and viewpoint delivery. This requires advanced reasoning to effectively integrate multimodal information, including visual, audio, and text. In this work, we introduce ARC-Hunyuan-Video<sup>a</sup>, a multimodal model that processes visual, audio, and textual signals from raw video inputs end-to-end for structured comprehension. The model is capable of multi-granularity timestamped video captioning and summarization, open-ended video question answering, temporal video grounding, and video reasoning. Leveraging high-quality data from an automated annotation pipeline, our compact 7B-parameter model is trained through a comprehensive regimen: pre-training, instruction fine-tuning, cold start, reinforcement learning (RL) post-training, and final instruction fine-tuning. Quantitative evaluations on our introduced benchmark ShortVid-Bench and qualitative comparisons demonstrate its strong performance in real-world video comprehension, and it supports zero-shot or fine-tuning with a few samples for diverse downstream applications. The real-world production deployment of our model has yielded tangible and measurable improvements in user engagement and satisfaction, a success supported by its remarkable efficiency, with stress tests indicating an inference time of just 10 seconds for a one-minute video on H20 GPU<sup>b</sup>.

**Date:** July 28, 2025

**Code:** <https://github.com/TencentARC/ARC-Hunyuan-Video-7B>

<sup>a</sup>The version supports Chinese and English videos and particularly excels at Chinese.

<sup>b</sup>The reported inference time is based on a deployment accelerated with vLLM.

## 1 Introduction

The explosive growth of user-generated short videos on platforms such as WeChat Channel and TikTok has fundamentally reshaped mobile internet consumption. These short videos, characterized by their brevity, diversity, and high engagement, have become a dominant medium for information sharing, entertainment, and social interaction. This shift has created an urgent need for structured, temporally-aware, and in-depth comprehension of real-world video content, which is essential for enabling a wide range of video-centric applications, including search, recommendation, and emerging intelligent video services.

Yet, the very characteristics that make these videos so appealing also present significant challenges for automated understanding. These videos typically contain dense visual elements (e.g., visual effects, text overlays), high-information audios (including speech and sound effects), and rapid pacing that emphasizes emotional expression and viewpoint delivery. Effectively comprehending such complex content requires not only multimodal integration of visual, audio, and textual information, but also advanced reasoning capabilities

**real-world shorts**



0:00 0:02 0:07 0:11 0:15 0:27 0:36 0:42 0:58

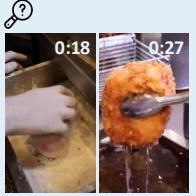
<b>Event-level caption</b> The video begins with a man excitedly announcing making tonkatsu...	He then displays his method: pork belly is coated in a thick batter...	He uses a tape measure to show thickness, expresses disappointment at its ...	The video shows the man running into restaurant ...	Inside the restaurant, a chef in a uniform is shown preparing a large, thick pork belly, cutting it, breading it, deep-frying it...	The chef then pours a large amount of beaten egg into a pan and cooks...	The chef then places perfectly fried tonkatsu on top of the egg ...	Finally, the man is shown eating the meal, measuring the thickness of the tonkatsu, and taking a bite, expressing his satisfaction with the final product...
---	--	---	---	---	--	---	--

0:00 0:11 0:42 0:58

<b>Chapter-level caption</b> A man shows how to make tonkatsu at home, but is <b>disappointed with the result</b> because it is too thin.	After the man runs into a restaurant, a chef in the restaurant demonstrates <b>the professional preparation of tonkatsu</b> , including frying a thick pork belly, scrambling eggs, and assembling the dish.	The man enjoys the professional tonkatsu, <b>showing the perfect thickness</b> and eating it with satisfaction.
--	--	---

**Summary**  
This video presents a **humorous comparison between homemade and professional tonkatsu**. It starts with a man attempting to make the dish at home, ... This is immediately contrasted with a professional chef in a restaurant kitchen, ... The creator's initial disappointment at his homemade version transforms into awe and satisfaction upon seeing the professional's creation. The video **concludes with him enjoying the professional's delicious meal, confirming that the professional's method is far superior**.

**Temporal Grounding**  
Question: When does the chef in the restaurant frying the tonkatsu in the video?  
**<think>**The user wants me to identify the time range when a chef is frying tonkatsu in a restaurant. I will scan the video to find this specific action. At 00:00:18, a chef in a white coat starts the process by dipping a large piece of raw pork into an egg mixture. This is followed by coating it in breadcrumbs. The actual frying happens **between 00:00:23 and 00:00:27**, where the tonkatsu is submerged in hot oil. The action concludes at 00:00:27 when the fried tonkatsu is lifted out of the oil. Therefore, the most comprehensive time range for the entire frying process shown in the restaurant is **from when the preparation starts to when it finishes****</think>**  
**<answer>**00:00:18 - 00:00:27**</answer>**



**Open-ended QA**  
Question: Is this man satisfied with the tonkatsu he made by himself?  
**<think>**The video starts with the man enthusiastically preparing his own tonkatsu, but his reaction is one of disappointment. At 00:00:08, he **uses a tape measure to check the thickness of the cutlet**, and at 00:00:09, he **holds up a piece and says 'So thin'** with a look of dismay. This indicates that he is not satisfied with the quality of his homemade version. The question asks if he is satisfied with his homemade version, and the answer is clearly no.**</think>**  
**<answer>**The man is not satisfied with the tonkatsu he made himself. He is visibly disappointed when he measures the thickness of the cutlet at 00:00:08, commenting on how thin it is, and his facial expression conveys disappointment.**</answer>**



ARC-Hunyuan-Video-7B

**Figure 1** Model capabilities of ARC-Hunyuan-Video-7B, which supports multi-granular timestamped captioning (output time span and corresponding description), summarization, temporal grounding, and open-ended question answering through integrating and reasoning over both visual and audio cues in the user-generated short videos.

to grasp the core intent of the content. However, current multimodal models, which are primarily designed for general video understanding, struggle to address these challenges and fall short of delivering the level of comprehension required for real-world applications.

To bridge this gap, we introduce **ARC-Hunyuan-Video**, a multimodal model for comprehensive understanding of real-world short videos. Our model processes visual, audio, and textual inputs to achieve what we term **Structured Video Comprehension**: the ability to decompose a video into its constituent events and narrative elements with temporal precision. This includes generating multi-granularity timestamped video caption and summary, answering open-ended questions through video reasoning, and performing accurate temporal grounding of events as illustrated in Fig. 1. This structured understanding is crucial for real-world scenarios, as it allows the model to move beyond surface-level analysis and truly understand what happens in user-generated content, when it happens, why it matters, and what intentions the creator wanted to convey.

Specifically, **ARC-Hunyuan-Video** is built on top of the Hunyuan-7B vision-language model and has two key incremental designs to meet the requirements of effective structured video comprehension: (1) an extra audio encoder with fine-grained visual-audio synchronization for temporally aligned visual-audio inputs, and (2) a timestamp overlay mechanism on visual frames that explicitly provides the model with temporal awareness for accurate event localization. Additionally, we collect millions of in-house real-world videos and develop a totally automated bootstrapped annotation pipeline that generates high-quality data, enabling a comprehensive training regimen. This includes (i) pre-training for fundamental knowledge and atomic capability acquisition, (ii) instruction fine-tuning for task alignment, (iii) cold-start initialization, (iv) reinforcement learning (RL) post-training, and (v) final instruction fine-tuning using high-quality human-annotated data and trajectories obtained through rejection sampling. A key aspect of our RL strategy is the design of objective questions such as multiple-choice questions and temporal grounding to enhance the model’s holistic comprehension of the video. This design is grounded in our pilot experiments, which demonstrate that verifiable tasks with RL significantly benefits the learning of high-quality subjective data (e.g., video summary).

To rigorously evaluate our model’s ability to understand real-world short videos, we construct a specialized, human-annotated benchmark named **ShortVid-Bench** with multiple-choice questions. Empirical evaluations demonstrate that our compact 7B-parameter model not only achieves exceptional performance in real-world video understanding on our proposed benchmark, but also excels in temporal video grounding benchmarks. Furthermore, **ARC-Hunyuan-Video** exhibits strong versatility, supporting zero-shot inference for a range of tasks and adapting to various downstream applications, such as video abstract for search and tagging for recommendation, with minimal fine-tuning data required. The deployment of our fine-tuned model in real-world product scenarios has resulted in significant and measurable improvements in user engagement and satisfaction. This success is underpinned by the model’s remarkable efficiency: stress test reports show an inference time of just 10 seconds for a one-minute video on NVIDIA H20 GPU, yielding an average of 500 tokens, with inference accelerated by the vLLM ([Kwon et al., 2023](#)) framework.

To facilitate further research and application, we have open-sourced both the model checkpoint, API, and inference code. We hope that **ARC-Hunyuan-Video** will contribute to advancing the field of structured video comprehension and inspire new developments in the comprehension of real-world short videos.

## 2 Related Work

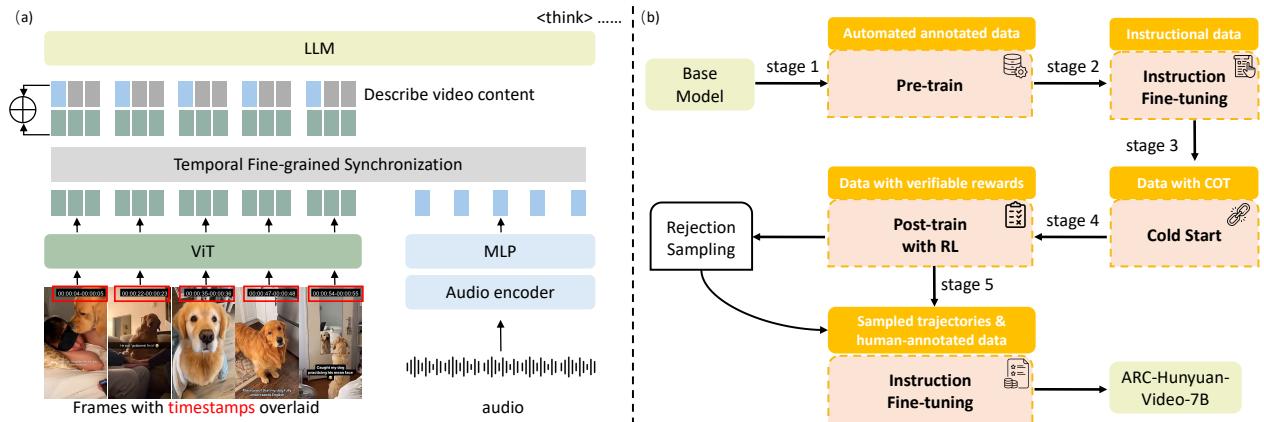
Short-form video has emerged as a dominant medium for communication, entertainment, and information dissemination across social media platforms. The ability to automatically understand such content is crucial for a wide range of downstream applications, including content retrieval, personalized recommendation, automated video tagging, and content moderation. Despite its importance, comprehending real-world short videos presents unique challenges. These videos are characterized by dense visual elements (e.g., dynamic effects, text overlays), rich audio streams (speech, music, sound effects), and rapid narrative pacing that emphasizes emotional expression and viewpoint delivery. These characteristics necessitate joint modeling of visual, audio, and textual modalities with advanced reasoning to comprehend key events, temporal relationships, and decipher underlying intentions.

Recent efforts have sought to address these challenges. Our concurrent work Keye-VL-8B ([Team et al., 2025](#)) introduces a multimodal foundation model designed specifically for short-video understanding. However, Keye-VL-8B does not directly integrate raw audio signals; instead, it relies on transcripts generated by automatic speech recognition (ASR). This approach discards important non-speech audio cues, such as emotional tone and environmental sounds, and can lead to temporal misalignment between audio and visual

content. Meanwhile, audio-visual LLMs (Shu et al., 2023; Zhang et al., 2023; Sun et al., 2024, 2025; Xu et al., 2025; Tang et al., 2025) have been developed to jointly process video, audio, and text input, but they focus on video understanding of general scenarios, which feature slower pacing, and lower information density. As a result, they often struggle to capture the dynamic, fast-paced, and information-rich nature of user-generated short videos.

In this work, we propose **ARC-Hunyuan-Video**, a compact 7B-parameter multimodal model that achieves structured video comprehension of real-world shorts through synchronized audio-visual-text processing and sophisticated reasoning.

### 3 Method



**Figure 2** (a) Model architecture. Built upon the Hunyuan-7B VLM, we incorporate an audio encoder with fine-grained visual-audio synchronization to obtain temporally aligned multimodal inputs. Timestamps are overlaid on visual frames to provide the model with temporal awareness.(b) Training stages including pre-training, instruction fine-tuning, cold start initialization, RL post-training and final instruction fine-tuning using high-quality human-annotated data and trajectories selected via rejection sampling.

#### 3.1 Model Architecture

As shown in Fig. 2, ARC-Hunyuan-Video is built upon the Hunyuan-7B vision-language model (VLM), with an additional audio encoder and fine-grained visual-audio synchronization mechanism to obtain temporally aligned multimodal features as input to the LLM. To explicitly provide the model with temporal awareness, we directly overlay the timestamp of each sampled frame onto its corresponding visual frame.

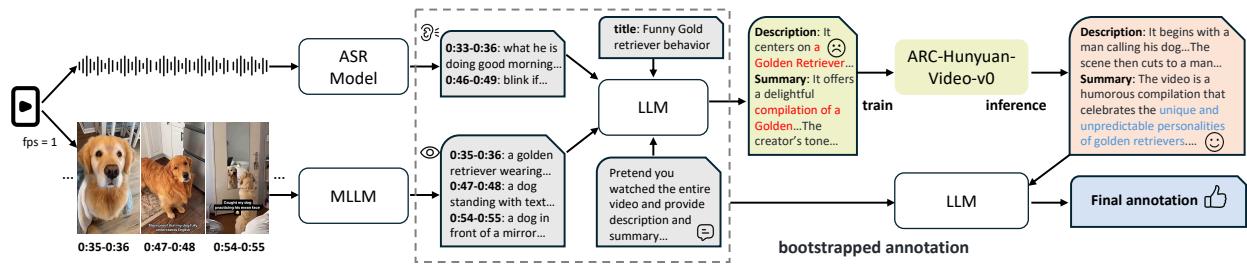
**Visual Encoding.** To process the visual input, we first sample frames at a rate of one frame per second (1 fps). For videos exceeding 150 seconds, we uniformly sample a total of 150 frames to maintain a manageable sequence length. To enhance model’s temporal awareness, we explicitly render the corresponding timestamp of each frame in an HH:MM:SS format onto its top-right corner. This provides the model with a direct, explicit signal for temporal localization. These timestamped frames are then fed into a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2020) encoder. While the Hunyuan ViT architecture inherently supports dynamic input resolutions, we resize all frames to a fixed resolution of  $640 \times 640$  for enabling the visual-audio synchronization detailed below. For each input frame, the ViT encoder outputs a sequence of 112 visual tokens.

**Audio Encoding.** For the audio modality, we leverage the pre-trained audio encoder from Whisper (Radford et al., 2023). The raw audio waveform is first segmented into 30-second chunks. The encoder processes each chunk to produce a sequence of 1500 feature tokens. For videos longer than 300 seconds, we split the audio into exactly 150 segments and truncate each to the initial 2 seconds before encoding—a design choice that optimizes temporal synchronization with visual frames. Finally, the audio features extracted by the audio

encoder are passed through a multi-layer perceptron (MLP). This projection layer aligns the dimensionality of the audio features with that of the visual tokens, preparing them for fusion.

**Visual-audio Synchronization.** The fine-grained synchronization aims to obtain multimodal representation where visual and audio tokens that correspond to the same time interval are fused, ensuring that the LLM receives temporally aligned multimodal signals. We adopt an adaptive and parameter-free synchronization strategy based on video duration. Specifically, for each sampled video frame, we align and fuse the corresponding audio segment by zero-padding the audio tokens to match the number of visual tokens, then adding them to form synchronized multimodal embeddings. This approach ensures that, regardless of video length, each fused embedding consistently represents the same temporal interval. The resulting sequence of synchronized embeddings, with positional encodings added, is then input to the LLM.

### 3.2 Pre-training



**Figure 3** Our automated bootstrapped annotation pipeline for pre-training. It extracts timestamped speech via ASR model and frame-level descriptions via MLLM; these, along with meta information (e.g., title), are input to an LLM for initial video annotation. The annotated data is used to train an initial version of the model, whose inference results are further integrated to produce the final annotations.

#### 3.2.1 Annotation pipeline

To generate high-quality video descriptions and summaries that capture the essence of real-world short videos (i.e., “truly understanding” the content), we designed an automated bootstrapped annotation pipeline as shown in Fig. 3. This pipeline is structured to iteratively refine annotations through a self-improving process, ensuring robust and accurate multimodal integration.

We begin by extracting and integrating multimodal information from videos using a series of specialized models. Specifically, we employ Whisper-v3 (Radford et al., 2023) to transcribe speech with precise timestamps, providing synchronized audio-text data. Concurrently, we use InternVL-2.5-8B (Chen et al., 2024) to generate detailed captions and detect textual overlays (e.g., on-screen text) for sampled video frames. These outputs, combined with video metadata such as titles, are fed into a closed-source large language model (LLM) for comprehensive synthesis. To ensure the model “truly understands” the video, we implement a Chain-of-Thought (COT) prompting strategy. This guides the LLM to first output intermediate elements: a step-by-step description of events, the creator’s attitude, and potential audience-interest tags. These elements are then integrated into a final summary that encapsulates the video’s core intent, emotional expression, and viewpoint delivery.

Leveraging the initial annotations, we train a preliminary version of our model. This model is then deployed to generate new descriptions and summaries, which are combined with the outputs from the initial pipeline (e.g., speech transcriptions, frame captions, and metadata). The aggregated data is reprocessed through the same closed-source LLM for polishing, using the COT approach to resolve inconsistencies and enrich detail. This iterative loop, where the model’s own outputs refine the annotations, yields high-quality, final descriptions and summaries that are used for pre-training.

### 3.2.2 Pre-training data

**Video description and summary.** Using our automated bootstrapped annotation pipeline, we annotate 4.5M short-form videos with detailed descriptions and summaries. Additionally, to ensure general video understanding capabilities, we apply the same pipeline to annotate 0.2M publicly available academic videos.

**Image caption and OCR.** Frame-level image understanding, including captioning and OCR, is one of the fundamental capabilities for short-form video comprehension. Therefore, we leverage the frame-level captions and OCR results obtained during the video annotation process as training data, resulting in a total of 4.7M image-text pairs.

**ASR.** Automatic speech recognition (ASR) is another fundamental capability for understanding short-form videos. Therefore, we utilize the ASR results obtained during the video annotation process as training data. We further employ an LLM to filter out transcribed speech samples without meaningful semantics (retaining only a small portion and labeling them as “no speech detected”), resulting in a total of 3.2M audio-text pairs.

**Video temporal grounding.** Precise temporal video grounding, aligning textual queries with specific temporal segments within a video, is essential for structured comprehension, which improves the model’s temporal awareness. To develop this capability, we leverage a diverse collection of 0.5M temporally grounded instances sourced from multiple public datasets. These instances provide paired textual queries (comprising natural language descriptions or questions) and their corresponding temporal intervals within the videos.

**Video multi-granular caption.** To support multi-granular video captioning, our dataset includes both event-level and chapter-level captions, each paired with their corresponding time spans and descriptive texts. For event-level captions, we leverage timestamped annotations from public datasets, applying filters to remove videos with excessive segment overlap or insufficient caption coverage relative to video duration, resulting in 50K high-quality samples. For chapter-level captions, we automatically generate captions for the time spans in 80K in-house videos, where the time spans themselves were annotated by humans.

### 3.2.3 Training Recipe

Building upon the pre-trained Hunyuan-7B VLM with established visual understanding capabilities, our training adopts a progressive two-stage strategy to integrate audio modality while preserving core competencies.

In the first stage, we conduct warm-up training using Automatic Speech Recognition (ASR) data to adapt the model to audio feature inputs as shown in Fig. 2. To prevent degradation of existing visual understanding, this stage simultaneously incorporates image-text pair data. When a modality is missing, we feed an all-zero input into the corresponding modality encoder. The dual-task design ensures the model develops initial audio-text alignment while retaining its foundational visual understanding abilities.

In the second stage, we perform full multimodal pre-training (video/audio/text) via next-token prediction. We freeze parameters of both the ViT and audio encoder to preserve their feature extraction capabilities. Only the MLP adapter layers and the full LLM backbone are updated. This phase employs a learning rate of 2e-5, leverages DeepSpeed Zero Stage 1 optimization, and operates with an extended context length of 20K tokens.

## 3.3 Post-training

The post-training stage aims to further enhance the structured comprehension capability of ARC-Hunyuan-Video for real-world shorts. Since our automated annotation pipeline inevitably introduces noise in the generated data, we collect a small set of high-quality human-annotated data. We conduct pilot experiments to investigate how to fully leverage this high-quality data to enhance the model’s ability for truly understanding videos.

### 3.3.1 Pilot Experiments

To quantitatively evaluate our model’s capability for video summarization, we curate 140 real-world shorts with human-annotated summaries and employ LLM-as-a-judge scoring (scale: 1-10) comparing model outputs

against human annotations. Initial experiments revealed that supervised fine-tuning the pre-trained model directly on this human-annotated summary data yielded no significant performance gains (pretrained 6.42 vs. fine-tuned 6.67). We further explored using Direct Preference Optimization (DPO) (Rafailov et al., 2023), treating the human annotations as positive samples and model outputs as negative samples, but similarly observed no improvement (pretrained 6.42 vs. DPO-tuned 6.50). We hypothesize that this stems from a potential distribution mismatch between the human annotations and the model’s learned representations.

Inspired by the success of approaches like DeepSeek-R1 (Guo et al., 2025), which utilized rule-based rewards on tasks with verifiable outputs (e.g., mathematics, coding) with Generalized Reinforcement Policy Optimization (GRPO) algorithm, we design two objective tasks for structured video understanding: (1) Multi-dimensional Multiple-Choice QA covering five critical aspects: (a) spatial fine-grained understanding, (b) temporal fine-grained understanding, (c) timeline analysis, (d) intent comprehension (creator’s attitude/purpose), and (e) event reasoning; (2) Temporal Video Grounding. We found that conducting GRPO-based post-training on these verifiable tasks, followed by fine-tuning on the human-annotated summaries, led to substantial gains in comprehension performance (pretrained 6.42 vs. GRPO-sft 6.99). The MCQ task enhances the model’s understanding across multiple respects by explicitly targeting diverse dimensions of video comprehension, while the grounding task increases the model’s temporal awareness by requiring precise localization of events within the video timeline. This combination enables the model to develop a more holistic and temporally sensitive understanding of video content, thereby better preparing the enhanced model to effectively learn from high-quality data.

After verifying that GRPO post-training on verifiable tasks followed by fine-tuning strengthens model’s comprehension ability, we design a comprehensive post-training regimen including an initial instruction fine-tuning for instruction alignment, a cold-start phase to initialize the model for reinforcement learning, a targeted reinforcement learning phase using GRPO, and a final instruction fine-tuning stage using high-quality human-annotated data.

### 3.3.2 Stage 1: Initial Instruction Fine-tuning

The Instruction Fine-tuning stage is adopted to equip ARC-Hunyuan-Video with robust instruction-following capabilities. To achieve this, we construct a comprehensive and high-quality supervised dataset that covers a wide range of tasks. Specifically, our data includes 460K open-ended question answering (QA) samples and 70K multiple-choice QA from publicly available academic datasets, as well as 20K QA samples collected from real-world short videos, ensuring both general coverage and domain-specific relevance. For temporal video grounding, we incorporate 10K samples from academic datasets and 5K samples from real-world short videos, enabling precise event localization within diverse video content. Additionally, the dataset contains 45K video description and summarization samples from real-world videos, along with 12K multi-granular captioning samples. Similar to pre-training, only the MLP adapter layers and the full LLM backbone are updated during this stage. We use a learning rate of 1e-5, leverage DeepSpeed ZeRO Stage 1 optimization, and operate with an extended context length of 20K tokens.

### 3.3.3 Stage 2: Cold Start Initialization for Reinforcement Learning

To prepare a strong initial policy for the subsequent reinforcement learning phase, we fine-tune the model on a curated dataset of 146K samples featuring Chain-of-Thought (CoT) reasoning. This “cold start” stage aims to teach the model how to perform step-by-step reasoning across a broad spectrum of tasks, thereby building a versatile reasoning foundation. For 90K multiple-choice QA samples covering both general video understanding and real-world shorts scenarios, we use a powerful MLLM to generate CoT rationales, retaining only instances where the final answer was correct. Similarly, for 18K temporal grounding tasks again spanning general videos and real-world shorts, we generate CoT rationales for timestamp prediction and filtered for samples with an Intersection-over-Union (IoU) above 0.6 between the predicted and the ground truth time span. This process was extended to 20K open-ended QA samples, where an LLM judge verifies the correctness of the final answer, and to 15K video summarization and 3K chapter-level captioning tasks, where event-level captions serve as the intermediate reasoning steps. This model, trained with the same recipe as the initial instruction fine-tuning stage, serves as the starting point for reinforcement learning.

### 3.3.4 Stage 3: Reinforcement Learning with GRPO

Our pilot experiments demonstrate that effective reinforcement learning on verifiable tasks benefits the learning of high-quality subjective data. Therefore, the phase narrows its focus to the two tasks with objective, verifiable reward signals: 100K multiple-choice questions and 35K temporal video grounding instances. For multiple-choice questions, a binary reward of 1.0 is assigned for correct answers and 0.0 for incorrect ones. For temporal grounding tasks, the reward is determined by the Intersection over Union (IoU) between the predicted time span and the ground truth. Using the GRPO (Guo et al., 2025) algorithm, we exclusively fine-tune the parameters of the large language model (LLM). Training employs a learning rate of 2e-5, leverages DeepSpeed ZeRO Stage 3 optimization, and operates with an extended context length of 20K tokens. The KL divergence coefficient within the GRPO algorithm is set to 0.1.

### 3.3.5 Stage 4: Final Instruction Fine-tuning

In the final stage, we return to the high-quality, human-annotated data. Having undergone the targeted reasoning enhancement of GRPO, the model is now capable of effectively learning from this nuanced data. We use a dataset of 25K human-annotated subjective questions for instruction fine-tuning including open-ended QA, video summarization and chapter-level captioning. We further leverage the enhanced capabilities of the GRPO-tuned model to generate 100K high-quality multiple-choice questions with CoT rationales and 50K temporal grounding instances with detailed reasoning traces through rejection sampling. By combining the high-quality human annotations with accurate self-generated trajectories, this final instruction fine-tuning stage polishes the model’s capabilities, aligning it closely with human-level comprehension of real-world short videos. The training employs a learning rate of 1e-5 and utilizes DeepSpeed ZeRO Stage 1 optimization.

## 4 Experiments

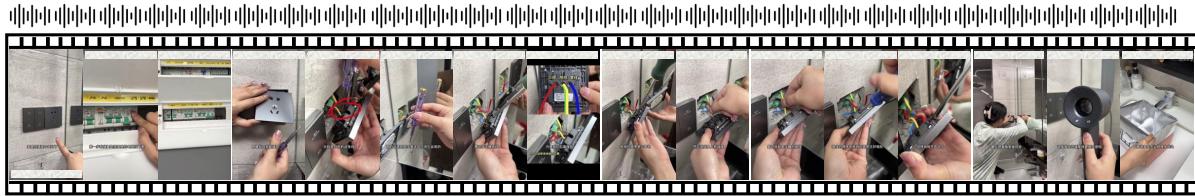
### 4.1 Qualitative Evaluation

#### 4.1.1 Model Capability

To intuitively showcase the advanced capabilities of ARC-Hunyuan-Video-7B, we present a series of qualitative evaluations on diverse real-world short videos as shown in Fig. 4, Fig. 5 and Fig. 6. These examples highlight our model’s proficiency in leveraging joint audio-visual reasoning and temporal awareness to achieve deep, structured comprehension, demonstrating its applicability in various real-world scenarios.

**Joint audio-visual reasoning for complex queries.** Our model excels at integrating information from multiple modalities to answer complex questions that are unanswerable from a single modality. It is crucial to note that although many short videos feature subtitles, these visual texts can be easily missed or only partially captured at low frame sampling rates. This makes processing the complete audio stream essential for a reliable comprehension of the spoken content, which might otherwise be lost. In the outlet replacement tutorial, when asked how to verify the absence of electricity (Fig. 4), the model correctly synthesizes the visual action of using a voltage tester with the narrator’s spoken instructions to provide a precise and safe procedure. Furthermore, when tasked with summarizing purchase advice from a product review video (Fig. 6), the model effectively extracts and organizes detailed specifications, prices, and target user profiles for different models, demonstrating its utility in structured information extraction from content where information is distributed across visuals, on-screen text, and narration.

**Fine-Grained temporal grounding and summarization.** A core strength of our model is its ability to understand the chronological flow of events. For instance, when analyzing a tutorial video on replacing an electrical outlet (Fig. 4), ARC-Hunyuan-Video-7B accurately segments the entire process into time-stamped, coherent steps, from turning off the breaker to testing the new installation. This demonstrates its capacity for fine-grained event localization and generating structured, step-by-step summaries. Similarly, for a fast-paced video montage of different morning routines (Fig. 6), the model successfully identifies and describes each distinct scene with its corresponding time range, showcasing robust scene segmentation. This temporal awareness is crucial for applications like video highlight generation and structured data extraction.



Localize video events with temporal boundaries and the corresponding sentence description.



<think>

The video provides a step-by-step tutorial on how to safely replace a home electrical outlet. The creator begins by showing a faulty outlet and then proceeds to demonstrate the entire replacement process. First, she shows how to turn off the main electrical breaker to ensure safety (00:00:00 - 00:00:11). Next, she demonstrates how to remove the old outlet by using a screwdriver to pry off the cover and unscrew the internal screws, emphasizing the importance of checking for residual power with a voltage tester (00:00:11 - 00:00:31). Then, she explains the function of the lettering on the outlet's back, identifying the 'L' for live (red) wire, 'N' for neutral (blue) wire, and 'D' for earth (yellow/green) wire, advising viewers to take a photo for reference (00:00:31 - 00:00:44). Following this, she carefully removes the old wires from the outlet, following the correct sequence of 'L', 'N', 'D' (00:00:44 - 00:00:53). She then proceeds to install the new outlet, connecting the wires in the reverse order (D, N, L) and securing them with screws, ensuring the copper parts are fully inserted into the socket for safety (00:00:53 - 00:01:10). After the new outlet is assembled, she puts the cover back on, screws it in, and then tests the new outlet by plugging in a hairdryer (00:01:10 - 00:01:19). The video concludes with a preview of the next tutorial on how to fix a water faucet and a call to follow her channel for more self-sufficiency tips (00:01:19 - 00:01:27).

</think>

<answer>

**00:00:00 - 00:00:11** The video begins by showing a faulty outlet and then demonstrates the first step: **turning off the main electrical breaker for safety**.

**00:00:11 - 00:00:31** The creator demonstrates **how to remove the old outlet** by using a screwdriver and then emphasizes the importance of using a voltage tester to check for any remaining power.

**00:00:31 - 00:00:53** The video explains the function of the lettering on the outlet's back, identifying the live, neutral, and earth wires, and then proceeds to **remove the old wires**.

**00:00:53 - 00:01:10** The creator **installs the new outlet** by connecting the wires in the correct sequence (D, N, L) and securing them with screws, ensuring the copper parts are fully inserted.

**00:01:10 - 00:01:27** The final steps involve **putting the cover back on, screwing it in, and testing the new outlet**. The video concludes with a preview of the next tutorial on fixing a faucet.

</answer>



After removing the outlet, how can we tell if there is no electricity inside?



<think>

The video shows a specific safety step after the outlet is taken apart. At **00:00:21**, the narrator says 'After removing the outlet, you must definitely use a penlight to check if there is no electricity'. A penlight is a type of multimeter or electrical test pen. The demonstration follows: the penlight is inserted into the outlet's live (red) terminal. At **00:00:26**, the penlight does not light up, indicating no electricity. The narrator confirms this by saying the penlight is safe and explaining that a different penlight (with a visible light) would light up if there was power. This is a crucial safety check before proceeding to disconnect the wires.



</think>

<answer>

After removing the outlet, you can test for the presence of electricity by inserting a penlight into the outlet's live (red) terminal. **If the penlight does not light up, it indicates that the outlet is safe and there is no electricity**. This step is crucial to prevent electrical shock before proceeding to disconnect the wires.

</answer>

**Figure 4** An example of ARC-Hunyuan-Video-7B. Given an instructional short video, our model can accurately identify and summarize the content of each step along with the corresponding time spans. For specific questions, the model is also able to locate the relevant time segments within the video, thereby providing precise answers.



列出这个视频值得借鉴的创意。



<think>

视频的核心信息是关于环境保护和人类责任。为了强化这一主题，创作者采用了多种高级的视听语言。**一个非常典型的创意是“分屏”技术**，它将截然不同的景象（如干旱沙漠与碧绿海洋、工业污染与自然生机）并列，形成强烈对比，直观地揭示了问题的严重性(00:00:49 - 00:00:51)。另一个创意是**视觉叙事的递进**，视频通过一系列对比鲜明的场景（如火山爆发与人类探索、森林火灾与生态修复、工业废水与清洁能源）来构建一个从“质疑”到“选择”的逻辑链。此外，视频还**运用了象征手法，如用“选择向前”对应登山者的前行，用“选择敬畏”对应城市的夜景，用“选择坚持”对应在沙地中栽树的孤独身影，用“选择长久”对应工业与自然和谐共存的愿景，来赋予“选择”这一概念以深刻的内涵(00:00:25 - 00:00:52)**。最后，**视频的结尾将宏大的主题落脚于具体的行动号召上**，通过“6.5世界环境日”的字样和“未来在我们手中”的口号，将抽象的环保理念转化为具体的、可操作的号召，极具感染力和号召力(00:01:14 - 00:01:46)。同时，**视频的配乐、剪辑节奏和画面切换都服务于其主题**，营造出一种既忧虑又充满希望的氛围，增强了视频的说服力和感染力。</think>

<answer>

该宣传片在创意上主要体现在以下几个方面：

1. \*\*强烈的视觉对比与并置\*\*：视频大量使用分屏技术，将同一场景下的两种极端状态（如干旱与生机、污染与纯净）并列，直观地展现了人类活动对环境造成的破坏与挑战，引发观众对问题的深刻反思。
2. \*\*叙事性视觉语言的递进\*\*：视频构建了一个从“质疑”到“选择”的叙事逻辑。通过“在质疑中选择向前”、“在欢呼中选择敬畏”等一系列递进的字幕，将抽象的环保理念与具体的行动号召相结合，赋予了“选择”这一概念以强大的情感和行动意义。
3. \*\*象征意义的运用\*\*：视频中的许多场景都带有强烈的象征意义，例如将“选择向前”与登山者的前行联系起来，象征着人类在面对未来时的责任与决心；将“选择敬畏”与城市的繁华夜景并置，引发观众对人与城市关系的思考。
4. \*\*最终的情感升华与行动号召\*\*：视频的结尾部分，通过展示人类在环保事业中的积极行动（如风力发电、生态修复），最终落脚于“未来在我们手中”这一强有力的口号，将宏大的主题与具体的、可操作的环保行动紧密联系起来，极大地增强了视频的感染力和号召力。
5. \*\*卓越的视听语言运用\*\*：宣传片在配乐、剪辑和画面选择上都非常出色，其激昂与舒缓交织的音乐、富有冲击力的画面切换、以及将自然景观与人类活动并置的叙事手法，共同营造出一种既忧虑又充满希望的氛围，极大地增强了视频的说服力和感染力。

</answer>



你觉得最能体现主题的视频片段是哪一个？



<think>

这个问题要求理解视频的主旨，并从中找出最能概括或升华这一主旨的画面。视频的核心信息是关于人类面临的环境挑战以及我们应如何应对。我需要寻找一个画面，它能够象征性地总结出视频所传达的关于希望、选择与未来的信息。视频结尾(**00:01:46 - 00:01:49**)的标题“未来在我们手中”直接点明了主题。然而，在此之前的画面中，**有一幕极具象征意义：在广阔的绿色山脉背景下，出现了“未来在我们手中”这几个大字**。这不仅是对主题的总结，也像是一种宣言，将希望和责任交给了观众。因此，这最后一帧是整个视频情感和信息的集中体现。

</think>



<answer>

最能体现视频主题的一幕是视频结尾处(**00:01:46 - 00:01:49**)的标题画面。**该画面在壮丽的绿色山脉背景上，清晰地展示了“未来在我们手中”这句核心主旨**，将视频所探讨的关于人类、选择与环境的关系最终落脚于人类自身的责任与希望上，起到了点睛和升华主题的作用。</answer>

**Figure 5** An example of ARC-Hunyuan-Video-7B. Given a real-world video with excellent audiovisual quality, our model can analyze the video from visual, auditory, and thematic perspectives, and through reasoning, provide fine-grained segment recommendations.



List the detailed purchase advice for the iPhone 16 in bullet point.



<think>The user wants a comprehensive list of purchase advice. I need to identify all the models mentioned and their specific features and price points to create a complete guide. The video presents three main models with different price points and features. I will list them in the order they are introduced: the budget-friendly 16e, the standard 16, and the high-end 16 Pro. For each model, I will state its price range, its main selling point (e.g., battery, display, camera), and any key features (e.g., chip, lighting). The 16e is for \$600, the 16 is for \$800, and the 16 Pro is for \$1000 and up. I will format this information into a clear, ordered list. </think>

<answer>The video provides the following detailed purchase advice for the iPhone 16 models:

1. \*\*iPhone 16e (00:00:02 - 00:00:12):\*\* This is the cheapest option, priced at \$600. Its main selling point is a long-lasting battery and a reliable chip set designed for users who are not tech-savvy. It features a 6.1-inch display.
2. \*\*iPhone 16 (00:00:13 - 00:00:34):\*\* This model starts at \$800. Its key upgrades include a brighter display with a dynamic island, a more powerful camera system with an ultra-wide and macro capability, and a smaller enhancement like the MagSafe charging ring. It comes in two sizes: a 6.1-inch model and a 6.7-inch model.
3. \*\*iPhone 16 Pro (00:00:35 - 00:00:55):\*\* These are the most advanced models, starting at \$1000. They feature 120Hz displays, a significantly improved camera system with a telephoto lens, LiDAR for 3D mapping, better speakers and microphones, a brighter flashlight, and a faster A18 Pro chip. They are available in two sizes: a 6.3-inch model and a 6.9-inch Pro Max model.</answer>



How many scenes appear in the video? Describe each scene in detail with the time range.



<think>The video is a compilation of different morning routines. I need to identify the distinct scenes and their corresponding timestamps. The first scene, labeled 'Italy', shows a man being served a cappuccino and croissant from a waiter (00:00:00 - 00:00:11). The second scene, labeled 'UK', features a man pouring tea from a teapot into a cup and then drinking it (00:00:11 - 00:00:20). The third scene, labeled 'Switzerland', shows a man in a shirt dancing energetically on the street in front of a restaurant (00:00:20 - 00:00:23). The final scene, labeled 'USA', depicts a man with a paper as a bib, eating pancakes while wearing sunglasses (00:00:23 - 00:00:31). </think>

<answer>The video presents four distinct scenes, each representing a different morning routine. The scenes are structured as follows:

1. \*\*Italy (00:00:00 - 00:00:11):\*\* A man in a black polo shirt is seated at an outdoor cafe. A waiter brings him a cappuccino and a croissant on a plate. He takes a bite of the croissant, then picks up his cup and takes a sip of the coffee.
2. \*\*UK (00:00:11 - 00:00:20):\*\* A man in a brown shirt is at a cafe. He pours hot tea from a white teapot into a cup, smiles, and then lifts the cup to take a drink.
3. \*\*Switzerland (00:00:20 - 00:00:23):\*\* The scene opens with a view of a pink and grey building. A man in a shirt and dark pants suddenly runs and dances energetically on the street in front of the restaurant window.
4. \*\*USA (00:00:23 - 00:00:31):\*\* A man in a grey shirt and sunglasses is seated at a cafe table with a paper draped over his shoulder like a bib. He smiles while looking at the camera.</answer>

**Figure 6** Examples of ARC-Hunyuan-Video-7B. Given a review-style short video, the model can extract the characteristics of different products based on both visuals and speech. Given a short video consisting of multiple distinct scenes, our model is able to analyze the transitions between these scenes and accurately discern the main theme.

**High-level thematic and creative understanding.** Beyond literal descriptions, ARC-Hunyuan-Video-7B demonstrates a remarkable ability for thematic reasoning, which is vital for understanding content focused on emotional expression and viewpoint delivery. When analyzing a real-world promotional video about environmental protection, it can identify sophisticated creative strategies (Fig. 5), such as the use of “strong visual contrast”, “symbolism” and “narrative progression” to convey the message. Moreover, it can pinpoint the single most thematically resonant moment in the video (Fig. 5), the final shot with the slogan “The future is in our hands”, showcasing its capacity to grasp the core intent and emotional weight of the content.

In conclusion, these qualitative results validate that ARC-Hunyuan-Video-7B moves beyond surface-level perception to truly understand what happens in a video, when it happens, and why it matters. This deep, structured comprehension makes it a powerful and versatile tool for a wide range of real-world applications.

#### 4.1.2 Model Comparison

To qualitatively assess the capabilities of ARC-Hunyuan-Video-7B, we conduct a comparative analysis against several baseline models, including Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025), and Keye-VL-8B-8B (Team et al., 2025). The results, summarized across four representative cases as shown in Fig. 7, Fig. 8 and Fig. 9, highlight our model’s superior performance in structured video comprehension, particularly in leveraging joint audio-visual reasoning and precise temporal awareness.

**Superior thematic understanding through audio-visual fusion.** A primary limitation of video-only models is their inability to process audio, which is often crucial for understanding the context and intent of short videos. In Fig. 7, a comedic skit titled “POV: Parent Logic”, the humor and narrative are driven by the audio narration explaining the parent’s illogical assumptions. Video-only models like Qwen2.5-VL-7B-Instruct and Keye-VL-8B, deprived of this audio context, misinterpret the visual cues. They describe the physical actions (e.g., a child peeking, a parent checking) but fail to grasp the core comedic premise. Keye-VL-8B even hallucinates non-existent dialogue like “the entropy theme”. In contrast, ARC-Hunyuan-Video-7B correctly identifies the skit’s satirical nature by integrating the audio narration with the visual scenes, accurately summarizing the central theme of “parent’s tendency to assume the worst in his child’s activities”. While the audio-visual model Qwen2.5-Omni-7B captures the basic events, its summary remains a literal play-by-play, lacking the deeper thematic insight that our model provides.

**Deeper nuance comprehension in real-world scenarios.** Beyond just understanding the plot, grasping the nuance and emotional tone is key to short video comprehension. In Fig. 8, a video contrasting the “imagination vs. reality” of holding an umbrella for a partner, all models identify the basic visual contrast. However, ARC-Hunyuan-Video excels in capturing the video’s intended purpose and emotional impact. Its summary describes the “excellent comedic effect” and how the video “resonates with the audience” by showing a “relatable and humorous” side of love. This demonstrates a more profound level of reasoning compared to the baselines, which offer more superficial, descriptive summaries. This ability to understand why a video is engaging is a critical advantage for real-world applications.

**Enhanced temporal awareness for accurate event grounding.** The fast-paced nature of short videos makes temporal localization a significant challenge. This weakness is evident in the temporal grounding tasks (Fig. 9). For the first example, which asks for the time range of a woman cooking, the baseline models produce wildly inaccurate predictions, both completely missing the event. Similarly, in the second example, their predictions for when a woman comments on food are imprecise. ARC-Hunyuan-Video-7B, however, leverages its explicit temporal awareness, achieved through our timestamp overlay mechanism, and joint visual-audio reasoning, to pinpoint the events with remarkable accuracy. It correctly identifies the cooking scene and the food commentary. This precision demonstrates that our model does not just see what happens, but understands precisely when it happens, a cornerstone of structured video comprehension.

In summary, these qualitative comparisons underscore the effectiveness of ARC-Hunyuan-Video-7B. By robustly fusing audio-visual information and maintaining a strong sense of temporality, our model overcomes the limitations of video-only and general-purpose multimodal models, delivering a more accurate, nuanced, and structured understanding of real-world short videos.



Describe the video content.

**Qwen2.5-VL-7B-Instruct:** The video humorously portrays a scenario from the perspective of "Parent Logic," where a parent is trying to monitor their child's activities. The child, dressed casually in a green t-shirt and jeans, is seen peeking through doorways while holding a phone, suggesting they are trying to avoid being caught. The parent, wearing a plaid shirt, is shown checking on the child multiple times, first by looking through a doorway, then by sitting at a desk with papers, and finally by peeking around a corner. The child alternates between pretending to do homework and playing games on the phone, all while the parent tries to catch them in the act. The setting appears to be a home environment, with simple decor and furniture visible in the background. The overall tone of the video is comedic, highlighting the often humorous and sometimes futile attempts parents make to keep track of their children's activities.

**Qwen2.5-Omni-7B:** The video is from the perspective of a parent. It shows a boy in a green shirt with a towel around his neck. He's in a room, and there's a guy in a plaid shirt sitting on a bed, looking at his phone. The parent leaves and comes back, seeing the boy on his phone, so he assumes the boy has been playing on his phone the whole time and gets angry. Then, the parent leaves again and comes back to find the boy doing homework. He assumes the boy only started homework when he came back and gets angry. When the parent leaves and comes back to find the boy doing homework, he assumes the boy stopped doing homework the moment he left and only started it when he came back, and he gets angry again. The parent is really frustrated with the boy's behavior.

**Keye-VL-8B:** <think>First, I need to carefully look at the video's timeline and dialogue. First, the video starts with an open door, then a person in a plaid shirt (possibly a child) holding a phone, ..... Then the other person (parent) wearing a light green T-shirt and carrying a towel appears, saying "the entropy theme" and "very angry." .....Overall, the video uses comedic parental logic to depict the common scenario of parents trying to manage children's behavior, such as doing homework and using phones, with the parent switching from angry to eventually giving in, reflecting the contradictory yet loving nature of parental logic.</think>

<answer>The video, titled "POV: Parent Logic," comically depicts a common scenario of parents managing children's behavior .....The child tries to explain he was "playing on his phone," but the parent, upon seeing the child holding a homework paper with "started homework," feigns understanding yet subtly criticizes "long term consequences." .....<boxed>The video "POV: Parent Logic" humorously depicts a common parental scenario where a parent interacts with a child about their behavior, such as using a phone and doing homework. The parent initially appears angry, mentioning "the entropy theme" and "very angry," but later offers help with homework, eventually allowing the child to go outside despite their earlier grievances. The video comically showcases the contradictory yet affectionate nature of parental logic in everyday interactions with children.</boxed></answer>

**ARC-Hunyuan-Video-7B:** <think>The video, titled 'POV: Parent Logic', presents a comedic skit about a parent's perception of their child's activities. A young man, playing both the child and the parent, begins by explaining his 'logic' (00:00:00 - 00:00:08). He states that when he leaves and returns, seeing his child on the phone, he will assume the child has been playing all the time, which will make him very angry. The scene then shifts to illustrate a different scenario (00:00:08 - 00:00:16). The parent leaves and returns to find the child doing homework. In this case, the parent's logic is that the child only started homework the moment they came back, causing the parent to be angry. Next, the scenario is reversed (00:00:16 - 00:00:24). The parent leaves and then returns to find the child playing on the phone. The parent's assumption is that the child started playing the instant they left, leading to anger. In the final scenario, the parent leaves and returns to find the child doing homework (00:00:24 - 00:00:35). The parent's logic is that the child stopped doing homework the moment they left and only started when they returned, again causing the parent to be angry.</think>

<answer>This video is a humorous and relatable skit that satirizes the 'parent logic' phenomenon. The creator portrays a parent who, upon returning from leaving the house, experiences a series of negative thoughts based on what he sees. The core of the comedy lies in the parent's tendency to assume the worst in his child's activities. The video effectively uses this exaggerated scenario to create a funny and shareable piece of content about the unique perspective and often illogical assumptions parents make about their children's behavior. </answer>

**Figure 7** A qualitative comparison between baseline models and our model in understanding short videos where one person plays multiple roles. Our model can accurately identify the events in each scene and provide precise understanding of the main video theme.



描述视频的内容和主题

**Qwen2.5-VL-7B-Instruct:** 这段视频展示了一对情侣在雨天的互动场景。视频开始时，画面中出现了一个男人背对着镜头，他正在为一个女人打伞。伞上布满了雨滴，显示出天气的阴沉和湿润。接着，镜头切换到女人的特写，她面带微笑，看起来非常开心。她的脸上有“美丽”、“浪漫”、“幸福”、“端庄”的文字标注，这些词汇突显了她此时的心情和状态。随后，视频切换到男人的视角，他戴着一副眼镜，穿着灰色的衣服，并且头上戴着透明的塑料袋作为临时的雨具。他的表情显得有些无奈或困惑，背景是模糊的城市街道，显示出下雨天的环境。整个视频通过对比“想象中”与“现实中”的场景，展示了情侣之间在雨天打伞时的真实感受。虽然想象中的场景可能更加浪漫和幸福，但现实中的情况可能更为实际和有趣。这种对比增加了视频的趣味性和真实感。

**Qwen2.5-Omni-7B:** 视频里有两个人在雨中。一开始，一个男人背对着镜头，他穿着灰色上衣，手里拿着透明雨伞，伞上有很多水珠。然后镜头切换到一个女人，她穿着白色上衣，长发披肩，面带微笑，看起来很美丽，浪漫，幸福，端庄。接着又切回男人，他戴着眼镜，头发被雨水打湿，看起来有点狼狈。视频的主题就是展现给老公打伞的样子，从想象中的美丽浪漫幸福端庄，到现实中的狼狈。你要是还有啥想法，咱们可以再聊聊。

**Keye-VL-8B:** <think>首先，我需要仔细看每个片段的信息。首先，第一个片段是一个人撑着伞在雨中，文字是“**你以为给老公打伞的样子**”。然后第二个片段是一个**穿着白色外套的女性**，手里也拿着伞，文字是“想象中”，周围还有标签比如“美丽”“浪漫”“幸福”“端庄”，这应该是展示人们想象中给老婆伞浪漫的画面。接下来第三个片段是一个戴眼镜的男性，同样撑着伞在雨中，文字是“现实中”，这里的场景可能是在街道上，背景有建筑物和车辆，看起来比较普通，不是特别浪漫，可能是实际生活中给老婆打伞的样子，比较真实或有点狼狈。所以视频的内容应该是对比了人们想象中给老婆打伞的浪漫样子和现实中的真实情况。主题就是这种生活中的反差萌，或者对现实与想象对比的幽默呈现。现在需要把这些内容整理一下。视频通过两个场景对比：首先展示“想象中”给老婆打伞的样子，是浪漫、美丽的画面；然后展示“现实中”的样子，是比较普通甚至有点狼狈的真实情况，形成反差，主题是现实与想象的对比，幽默地呈现生活中的小真相。</think>

<answer>视频首先呈现了“想象中”给老婆打伞的样子，画面中一位穿着白色外套的女性手持伞，周围伴有“美丽”“浪漫”“幸福”“端庄”等标签，营造出浪漫美好的氛围；随后转为“现实中”的场景，**一位男性在雨中撑着伞**，背景是街道和建筑物，画面更贴近真实生活的普通状态。通过对比，视频展现了人们想象与现实之间的反差，幽默地呈现了生活中的小真相。</answer>

**ARC-Hunyuan-Video-7B:** <think>视频开头，镜头从后方拍摄一位男士在雨中撑着透明雨伞行走的背影，配文“**你以为给老公打伞的样子**”，营造出一种浪漫温馨的氛围(00:00:00 - 00:00:02)。紧接着，镜头切换到一位女士的自拍视角，她面带微笑，手持雨伞，**背景音乐轻柔**，画面上浮现出“想象中”、“美丽”、“浪漫”、“幸福”、“端庄”等字样，进一步强化了理想中的场景(00:00:02 - 00:00:06)。最后，视频画风突变，切换到“现实中”的场景，镜头对准了那位男士，**他正顶着一把被风吹得完全变形的雨伞，雨水顺着他的脸和眼镜不断滑落**，他本人则一脸狼狈和无奈，与之前的浪漫想象形成强烈反差(00:00:06 - 00:00:12)。</think>

<answer>该视频通过一个简单而生动的“想象 vs 现实”的对比，幽默地记录了情侣间一个充满反差感的日常瞬间。视频前半部分**通过浪漫的配乐**和“你以为”的文字，构建了一个温馨、体贴的理想化场景：一位女士在雨中为男士撑伞，自己则显得美丽、幸福。然而，视频的后半部分迅速反转，揭示了“现实”中的情景——**男士不仅没能被妥善保护，反而被一把被大风吹得东倒西歪的雨伞“欺负”**，狼狈不堪。这种强烈的反差制造了出色的喜剧效果，让观众在会心一笑中感受到真实、接地气的爱情。视频的核心在于其真实性和幽默感，它打破了人们对浪漫的刻板印象，展现了爱情中平凡、甚至有些笨拙但依然充满乐趣的一面，成功地用一个生活小插曲引发了观众的共鸣。</answer>

**Figure 8** A qualitative comparison between baseline models and our model in understanding short videos with rich visual information. Our model can accurately describe the visual content, analyze the background music, and identify the main theme of the video.



视频中一个女士在厨房做饭发生在什么时间段?请给出时间范围



**Qwen2.5-VL-7B-Instruct:** 该事件发生的时间区间为[80.9, 123.4]

**Qwen2.5-Omni-7B:** 视频片段从183.45秒开始, 200.05秒结束, 描述了一个女士在厨房做饭的场景。

**Keye-VL-8B:** <think>从提供的描述来看, 大部分片段是关于人物在户外的场景, 比如在绿色环境中说话、乡村场景、围坐吃饭等。比如第0.0到4.1秒是女士在户外, 然后5.1秒提到“家里又来客人了”, 接着6.2秒女士和他人说话, 8.3秒到14.5秒是户外吃饭的场景, 直到15.6秒开始提到“厨房”。具体看15.6秒的描述: 有女性端着食物进入室内”, 这直接对应了.....<think><answer>通过分析视频描述, 关键场景出现在15.6 - 16.6秒: 该片段明确提到“厨房”, 并描述“有女性端着食物”, 符合“在厨房做饭”的核心事件。<boxed{开始时间: 15.6秒, 结束时间: 16.6秒}></answer>

**ARC-Hunyuan-Video-7B:<think>**

1. 我浏览了整个视频, 寻找符合描述的场景。
  2. 在视频的中间部分, 我注意到一个场景转换到了室内厨房。
  3. 从00:01:17开始, 镜头切换到厨房内部, 一个中年女性正在案板上用刀切菜。
  4. 接着, 她开始炒菜, 整个做饭的过程被完整地记录下来。
  5. **这个做饭的场景一直持续到00:01:44, 此时菜已经做好, 她端着菜从门口出来, 准备上桌。**
  6. 可以确定完整时间范围是从她开始切菜到她把菜端出厨房为止, 即00:01:17到00:01:44。</think>
- <answer> 00:01:17 - 00:01:44 </answer>



When can we observe that the woman in the video comment on the food she made?



**Qwen2.5-VL-7B-Instruct:** The event occurs from 18.4 to 24.3 seconds.

**Qwen2.5-Omni-7B:** The event happens in 26.9 - 31.8 seconds.

**Keye-VL-8B:** <think>Initially, the video begins with someone pouring chocolate into a glass, .....The video shows her holding the food and speaking from 8 to 22 seconds. For example, at 11 seconds, she shows the food to the camera, then says “OMG,” at 13 seconds “how old is this thing??”These are all her comments on the food.....<think><answer>During this period, the woman holds the food, expressing her feelings through lines such as “OMG,” “how old is this thing??” and provides comments on the food she made.<boxed{8.0 to 22.0}></answer>

**ARC-Hunyuan-Video-7B:<think>** I need to locate the part of the video where she eats the treat and gives her opinion. At 00:00:27, she takes a bite of the frozen treat. She then comments on its texture and taste, saying 'As always, mine isn't pretty... that's really soft and so delicious'. This comment segment ends at 00:00:33 when she gives her final score. Therefore, the relevant time range is from when she starts eating and commenting until she finishes her sentence.</think>

<answer> 00:00:27 - 00:00:34 </answer>

**Figure 9** A qualitative comparison between baseline models and our model in the ability of temporal video grounding on real-world videos. Our model can effectively analyze visual and audio cues to accurately determine the start and end times of events.

## 4.2 Quantitative Evaluation

### 4.2.1 Evaluation Benchmark

**Real-world shorts understanding.** Existing benchmarks often fall short in capturing the nuanced complexities of user-generated content. To rigorously evaluate our model’s ability to understand real-world short videos, we construct a specialized benchmark named **ShortVid-Bench**. Specifically, we develop an automated pipeline to generate multi-dimensional questions for each video, targeting capabilities that signify a deep, holistic comprehension through integrating both visual and audio cues. These dimensions include: (1) Temporal Reasoning and Localization, (2) Affective Intent Classification, (3) Creator Intent Taxonomy, (4) Narrative Comprehension, (5) Humor & Meme Deconstruction, (6) Creative Innovation Analysis as shown in Fig. 10. For objective assessment, we employ a multiple-choice question (MCQ) format following previous work (Li et al., 2023; Chen et al., 2023; Qiu et al., 2024). Each question is carefully curated by human annotators who provide the ground-truth answer and design challenging, plausible distractors. Collectively, these dimensions push the evaluation beyond mere descriptive captioning, demanding a genuine comprehension of the video’s context, intent, and narrative.

**Video temporal grounding.** We further evaluate our model on temporal video grounding tasks including Charades-STA (Gao et al., 2017), which contains 3,720 long videos capturing indoor human activities for testing, and ActivityNet (Caba Heilbron et al., 2015), which comprises 17,031 test samples.

**General video understanding and reasoning.** While our primary focus is the structured comprehension of real-world short videos, we also evaluated ARC-Hunyuan-Video-7B on several established general-purpose benchmarks. Specifically, we report performance on (1) MVBench (Li et al., 2024), (2) the multiple-choice task of VCR-Bench (Qi et al., 2025), and (3) Video-Holmes (Cheng et al., 2025). The first two benchmarks encompass a mixture of perception and reasoning tasks across various video types, while Video-Holmes is specifically designed to test complex video reasoning, with a focus on suspenseful short films.

### 4.2.2 Evaluation Results

As shown in Tab. 1, ARC-Hunyuan-Video-7B achieves the highest accuracy on our proposed ShortVid-Bench, which demonstrates its superior ability to comprehend real-world short videos by integrating visual and audio signals with advanced reasoning capabilities. Furthermore, our model outperforms all baselines in temporal video grounding, which is largely attributed to our strategy of directly overlaying timestamps onto the video frames for enhancing temporal awareness. With limited general-purpose video training data, our model also shows promising results on general video understanding and reasoning benchmarks.

**Table 1** Quantitative evaluation on different benchmarks, which use accuracy as the evaluation metric, except for the grounding tasks, which use mIoU.

Model	fps	#frames	think	Real-world Shorts Und		Temporal Video Grounding		General Video Und & Reasoning		
				ShortVid-Bench	Charades-STA	ActivityNet	MVBench	VCR-Bench	Video-Holmes	
Qwen2.5-VL-7B-Instruct	1.0	150	✗	67.8	46.9	25.1	62.9	<b>53.7</b>	41.6	
Qwen2.5-Omni-7B	1.0	150	✗	68.3	30.5	13.0	<b>64.8</b>	51.0	<b>43.9</b>	
Keye-VL-8B	1.0	150	✓	53.5	25.1	14.9	35.7	34.9	35.7	
ARC-Hunyuan-Video-7B	1.0	150	✓	<b>74.3</b>	<b>54.8</b>	<b>41.7</b>	62.6	50.5	40.9	

## 4.3 Downstream Application

To demonstrate the practical utility and adaptability of ARC-Hunyuan-Video-7B, we conduct supervised fine-tuning on a set of downstream tasks with minimal task-specific data for real-world application scenarios. We map specific supervised fine-tuning tasks to their corresponding real-world applications: (a) Brief Summary for Video Retrieval, (b) Detailed Summary for comprehensive Video Tagging, and (c) Extended Browsing Words for Video Recommendation.



At which moment did the cleaner show the joy in his work?

- A. 00:00-00:11 The cleaner starts cleaning the truck's back door...
- B. 00:11-00:21 The cleaner washes the side of the truck...
- C. 00:21-00:26 The cleaner signs on the dirt and then...**
- D. 00:00-00:11 The cleaner signs on the dirt and then
- E. 00:11-00:21 The cleaner signs on the truck as a completion mark...
- F. 00:21-00:26 The cleaner finishes all the work and checks the quality...

#### Temporal Reasoning and Localization



What is the ultimate goal of the creator in making this video?

- A. The common social phenomenon where gardening enthusiasts...
- B. Promote a wearable kneeling device to alleviate knee discomfort...**
- C. Educate the audience about the scientific knowledge and preventive...
- D. Record and share the testing process of new gardening tools...
- E. Test the durability and performance of the kneeling equipment...
- F. Solicit audience opinions on equipment design improvements for...

#### Creator Intent Taxonomy



What is the core humorous point in the video?

- A. The humor arising from the awkward mistake of a man entering...
- B. The sense of loneliness and loss shown by a woman when she walks...
- C. The two people imitate the actions of the stick figures on the sign...**
- D. The unexpected caused by the sudden change of the heart emoji...
- E. The everyday behavior of couples holding hands and walking to...
- F. The humor in this video comes from the editing of the footage...

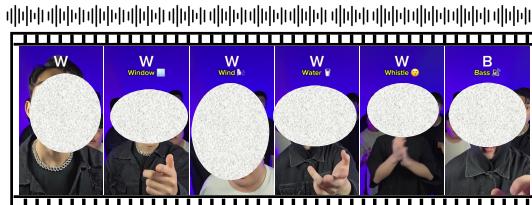
#### Humor & Meme Deconstruction



What kind of emotion do video creators mainly want to convey?

- A. A profound concern of the proliferation of brands in the fashion...
- B. Anger and condemnation over the poor quality of counterfeit good...
- C. By exaggerating misspelling to satirize the absurdity of counterfeit...**
- D. The pride and satisfaction of collecting authentic sneakers...
- E. Sympathy and concern for consumers being easily deceived ...
- F. Excitement and appreciation for innovative design...

#### Affective Intent Classification



How do the participants organize and carry out the activity?

- A. Randomly select words to imitate, and participants improvise...
- B. Imitate the sound of a word starting with the randomly appearing...**
- C. Group by sound type with priority to imitating animal sounds...
- D. After imitating the sound, you must guess the word...
- E. Only the central character is responsible for imitation, while others...
- F. Use props to imitate, such as using a bottle to simulate the sound...

#### Narrative Comprehension



What are the core creative concepts and techniques in the video?

- A. Dance clips are seamlessly compiled into the same cheerful music...**
- B. Using mascot characters to add fun and appeal, in order to...
- C. Highlight the details of the dance through slow-motion effects...
- D. Create narrative connections between different segments, such...
- E. Combining strong audio rhythm and visual synchronization...
- F. The video is stitched together from dance performances by multiple...

#### Creative Innovation Analysis

**Figure 10** Examples from ShortVid-Bench. The questions, spanning six distinct dimensions, require integrating both visual and audio information for a genuine comprehension of the real-world short videos.



**Figure 11** Demonstration of ARC-Hunyuan-Video-7B's versatility through minimal fine-tuning for various downstream applications. Specific supervised fine-tuning tasks are mapped to their corresponding real-world scenarios: (a) Brief Summary for Video Retrieval, (b) Detailed Summary for comprehensive Video Tagging, and (c) Extended Browsing Words for Video Recommendation.

#### 4.3.1 Experimental Setup

Based on common business scenarios, we consider three typical ones as examples. The specific task definitions are as follows:

- **Brief Summary.** The concise summary of the core content such as people, location and event, condensing the key information in simple language, facilitating a quick understanding of the overall picture. Brief summaries can effectively simplify the functional requirements in scenarios such as video retrieval and

video aggregation, and transform traditional cross-modal or pure visual analysis into more mature plain text operations.

- **Detailed Summary.** The detailed summary of the entire content of the video, which should include a complete description of the video’s content, as well as information about the shooting techniques, background music, subtitles, etc., and also a brief analysis of the video’s meaning. Similar to brief summary, detailed summary is also a compromise made to maximize utilizing the efficiency and high quality of plain text retrieval. The difference is that detailed summary pay more attention to the content details of the video and can make more accurate matches to the video content details during retrieval.
- **Extended Browsing Words.** Extended browsing words refer to search terms that users may be interested in and search for after browsing a given video, and then extend their browsing. This is a typical recommendation scenario. Traditional strategies may implement recommendations based on video similarity (Wray et al., 2021; Fang et al., 2021) or collaborative filtering (Wu et al., 2022) or association rules (Liao et al., 2021; Qin et al., 2021) based on user behavior, but extended browsing words given by content-based reasoning can effectively expand the scope of recommendations and have better prospects in terms of cold start and preference prediction.

For each of the three tasks, we obtain 1,100 samples by manual annotation, of which 1,000 are used for supervised fine-tuning and the remaining 100 are used for evaluation. To ensure the quality of data annotation, we randomly select 10% of the labeled data for cross-validation, and the pass rate was greater than 95%.

For evaluating the performance of the fine-tuned ARC-Hunyuan-Video from both qualitative and quantitative perspectives, we utilize the following two indicators.

**Pass Rate (PR).** We manually define a score scale for evaluating model outputs, which is divided into three grades from low to high, i.e., 0-2 points. Specifically, a score of 0 indicates that the model output contains obvious errors, which are clearly inconsistent with the original video content or task definition; a score of 1 means that the model output has acceptable minor issues that do not affect the understanding of the original video content and do not excessively violate the rules; a score of 2 represents that the model output has no problems at all and fully conforms to the task definition. In actual business operations, we generally believe that a score of 1 or above can be considered a pass.

**Good vs. Same vs. Bad (GSB).** (Zou et al., 2021; Zhao et al., 2022; Ye et al., 2023; Li et al., 2025) GSB metric is widely adopted in industry, and it is evaluated by experts judging the superiority or inferiority of results from different sources. Specifically, judges are given two results for a single input: one generated by System A, and the other by its competitor, System B. Crucially, annotators are unaware of which system each result corresponds to. Their task is to determine which result is of higher quality based on an assessment of the overall quality of the output results.

#### 4.3.2 Implementation Details

We use the ARC-Hunyuan-Video-7B as the base model and optimize the parameters of the MLP adapter layers and the full LLM backbone with a learning rate of 1e-5 for 3 epochs. The prompt settings in the training data consist of rule descriptions in natural language. In particular, in the reasoning of extended browsing words, we design a simple thought of chain structure, which outputs the understood video content and then infers possible extended browsing words.

#### 4.3.3 Experimental Analysis

**Main Results.** As shown in Table 2, compared with the current online baseline, the supervised fine-tuned ARC-Hunyuan-Video-7B performs significantly better in the three tasks. Specifically, by comparing the model scores, it can be found that supervised fine-tuned ARC-Hunyuan-Video-7B received fewer 0 points and more 2 points, so the PR is also significantly higher than the baseline, proving that it can better and more correctly understand the video content according to the instructions. At the same time, the relatively large difference in the GSB score based on manual evaluation further proves that even one instance can get a score of 2 points. The supervised fine-tuned ARC-Hunyuan-Video-7B can better meet the preferences of the

**Table 2** Evaluation results of three tasks, where baseline is the original online model (different for three businesses) and Ours is ARC-Hunyuan-Video-7B after supervised fine-tuning for different tasks. Our model shows marked improvements, with a significantly higher Pass Rate (PR) and a dominant win rate in the GSB (Good vs. Same vs. Bad) human preference comparisons.

Model	Brief Summary					Detailed Summary					Extended Browsing Words				
	0	1	2	PR	GSB	0	1	2	PR	GSB	0	1	2	PR	GSB
Baseline (Method A)	29	22	49	0.71	16:4:80	37	44	19	0.63	8:15:77	18	36	46	0.82	14:44:42
Ours (Method B)	18	4	78	0.82		26	35	39	0.74		12	34	54	0.88	

reviewers and provide a better experience. Going deeper, since both are summary tasks, the main difference between brief summary and detailed summary is the level of detail in the description of the video content. By comparing the PR of the two, it can be found that the detailed summary is more difficult, which is also in line with intuition. After all, the rules of detailed summary are more difficult, and the description of video details is more prone to errors and omissions. Further observation of the distribution of 0, 1, and 2 scores of these two tasks also shows that the number of 0 points in the detailed summary is larger, and the amount of 2 points in the detailed summary is also significantly lower than that of the concise summary, which further indicates that the current model still has room for improvement.

**Real-world Production Benefits.** The above three types of supervised fine-tuning tasks are not actually directly output to users for interaction, but can provide indirect support for different user services as intermediate products. For example, the brief summary, as a description of the core content of the video, can directly serve video retrieval services. Therefore, we apply it to the video retrieval of our products as a retrieval target with user query, which significantly improves the user’s retrieval experience. Specifically, our retrieval CTR increased by 5.88%, the landing page consumption time increased by 5.11%, the video floating layer click CTR increased by 7.26%, and the long click rate increased by 3.34%. Similarly, we also applied it to the video aggregation application. After the function was launched, the number of goals per capita increased by 0.63%, the average QV per capita increased by 0.55%, and the proportion of satisfied QV increased by 1.77%.

**Case Study.** To more clearly compare the performance differences after supervised fine-tuning, we provide specific examples for the brief summary in Figure 12. By observing these examples, we can find that after supervised fine-tuning, ARC-Hunyuan-Video-7B is able to better incorporate task rules and achieve rule-guided video understanding.

## 5 Conclusion

This paper introduces ARC-Hunyuan-Video, a powerful multimodal model designed to tackle the challenges of understanding real-world short videos. Faced with the complexity of user-generated content, characterized by dense information, multimodal integration, and rapid pacing, we propose the concept of **Structured Video Comprehension**, which focuses on fine-grained, temporally-precise understanding of a video’s narrative, events, and underlying intent. Built upon Hunyuan-7B VLM, we adopt an audio encoder for fine-grained audiovisual synchronization and a timestamp overlay mechanism for explicit temporal awareness. This model is trained using a multi-stage strategy on a large-scale dataset of millions of real-world videos, annotated via an automated bootstrapped pipeline. A core finding of our work is that grounding the model in objective tasks with RL is key to unlocking high-quality, subjective understanding. Extensive experiments demonstrate that ARC-Hunyuan-Video achieves state-of-the-art performance on short video comprehension benchmarks and exhibits strong versatility for downstream applications. We believe ARC-Hunyuan-Video represents a significant step towards enabling more sophisticated, in-depth, and practical video-centric AI services, paving the way for a new generation of intelligent video applications.



Baseline: 大哥午饭留鸡腿给继女。(The eldest brother saved the chicken leg for his stepdaughter at lunch.)

ARC-Hunyuan-Video (Brief Summary): 视频中, 继父午饭只吃青菜馒头, 却把鸡腿让给放学回家的继女, 继女感动不已。(In the video, **the stepfather** only ate vegetable buns for lunch, but gave the chicken leg to his stepdaughter who **came home from school**, and **the stepdaughter was very touched**.)



Baseline: 视频讲述媳妇为参加同学聚会努力减肥, 展现搞笑家庭日常。(The video tells the story of a wife trying hard to lose weight in order to attend a class reunion, and shows a funny family routine.)

ARC-Hunyuan-Video (Brief Summary): 视频中, 丈夫为即将参加同学聚会、想要拥有好身材的妻子, **用胶带缠绕其腰部和臀部, 帮助她快速瘦腰**, 充满趣味和幽默感。(In the video, **the husband wraps tape around the waist and hips of his wife**, who is about to attend a class reunion and wants to have a good figure, to help her lose weight quickly. The video is full of fun and humor.)

**Figure 12** Examples of brief summary. In the first case, we observe that the fine-tuned ARC-Hunyuan-Video-7B can correctly infer the relationship between the characters in the video, and infer that the person is going home from school through the behavior (carrying a schoolbag into the door). In the second example, the fine-tuned ARC-Hunyuan-Video-7B also correctly identifies the relationship between the characters in the video and describes the complex interaction details between the characters in more detail.

## References

- 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.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.
- Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplant-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.
- 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.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- 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.
- Yuchen Li, Hao Zhang, Yongqi Zhang, Xinyu Ma, Wenwen Ye, Naifei Song, Shuaiqiang Wang, Haoyi Xiong, Dawei Yin, and Lei Chen. M2oerank: Multi-objective mixture-of-experts enhanced ranking for satisfaction-oriented web search. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 4441–4454. IEEE Computer Society, 2025.
- Shu-Hsien Liao, Retno Widowati, and Yu-Chieh Hsieh. Investigating online social media users' behaviors for social commerce recommendations. *Technology in Society*, 66:101655, 2021.
- Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.
- Yuqi Qin, Pengfei Wang, and Chenliang Li. The world is binary: Contrastive learning for denoising next basket recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 859–868, 2021.
- Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.
- Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.
- Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*, 2025.
- Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.
- Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*, 2025.
- Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3650–3660, 2021.

---

Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4425–4445, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Dezhi Ye, Bowen Tian, Jiabin Fan, Jie Liu, Tianhua Zhou, Xiang Chen, Mingming Li, and Jin Ma. Improving query correction using pre-train language model in search engines. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2999–3008, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Xue Zhao, Dayiheng Liu, Junwei Ding, Liang Yao, Mahone Yan, Huibo Wang, and Wenqing Yao. Self-supervised product title rewrite for product listing ads. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 79–85, 2022.

Lixin Zou, Shengqiang Zhang, Hengyi Cai, Dehong Ma, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. Pre-trained language model based ranking in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4014–4022, 2021.