VideoVista-CulturalLingo: 360° Horizons-Bridging Cultures, Languages, and Domains in Video Comprehension

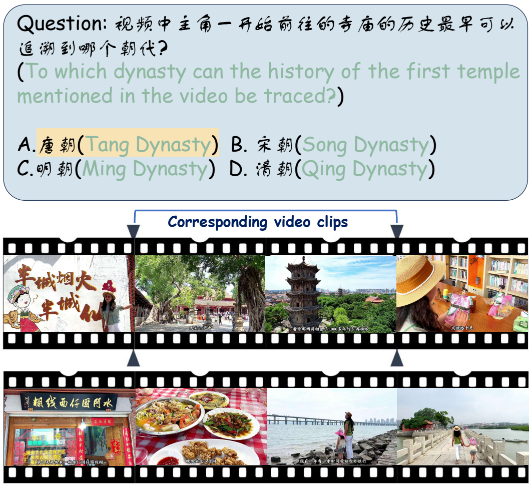
Xinyu Chen[1], Yunxin Li[1], Haoyuan Shi[1], Baotian Hu[1*],
Wenhan Luo[2], Yaowei Wang[1], Min Zhang[1],
[1]Harbin Institute of Technology, Shenzhen, China,
[2]Division of AMC and Department of ECE, HKUST,
{chenxinyuhitsz, liyunxin987}@163.com
{hubaotian, zhangmin2021}@hit.edu.cn

Abstract

Assessing the video comprehension capabilities of multimodal AI systems can effectively measure their understanding and reasoning abilities. Most video evaluation benchmarks are limited to a single language, typically English, and predominantly feature videos rooted in Western cultural contexts. In this paper, we present **VideoVista-CulturalLingo**, the first video evaluation benchmark designed to bridge cultural, linguistic, and domain divide in video comprehension. Our work differs from existing benchmarks in the following ways: 1) **Cultural diversity**, incorporating cultures from China, North America, and Europe; 2) **Multilinguistics**, with questions presented in Chinese and English—two of the most widely spoken languages; and 3) **Broad domain**, featuring videos sourced from hundreds of human-created domains. VideoVista-CulturalLingo contains 1,389 videos and 3,134 QA pairs, and we have evaluated 24 recent open-source or proprietary video large models. From the experiment results, we observe that: 1) Existing models perform worse on Chinese-centric questions than Western-centric ones, particularly those related to Chinese history; 2) Current open-source models still exhibit limitations in temporal understanding, especially in the Event Localization task, achieving a maximum score of only 45.2%; 3) Mainstream models demonstrate strong performance in general scientific questions, while open-source models demonstrate weak performance in mathematics. [1]

1 Introduction

Large Multimodal Models (LMMs) built upon Large Language Models (LLMs) have demonstrated unprecedented capabilities across various domains, including text, image, video, and audio over several years. Particularly in the past year, there has been a surge in the development of LMMs capable of processing video inputs. The dramatic expansion in the length of video frame sequences—from just a few frames to several hundred—demonstrates significant progress in video understanding capabilities. Meanwhile, video evaluation benchmarks have also emerged, evolving from early-stage basic video question answering tasks (Yu et al., 2019; Xu et al., 2017) to general video evaluation benchmarks (Fu et al., 2024; Zhou et al., 2024; Wang et al., 2024b). However, existing video evaluation benchmarks predominantly select videos from sources such as YouTube, Shutterstock, or established video datasets like Ego4D (Grauman et al., 2022) and Movie101 (Yue et al., 2023). These datasets are primarily Western-centric, with a limited representation of Chinese-centric videos as shown in Figure 1. In addition, current video evaluation benchmarks tend to focus on specific events within the videos, neglecting the cultural context and connotations embedded in the content while overlooking the scientific principles and in-



Question: 视频中主角一开始前往的寺庙的历史最早可以追溯到哪个朝代?
(To which dynasty can the history of the first temple mentioned in the video be traced?)

A.唐朝(Tang Dynasty) B. 宋朝(Song Dynasty)
C.明朝(Ming Dynasty) D. 清朝(Qing Dynasty)

Figure 1: **An example of Traditional Chinese Culture from VideoVista-CulturalLingo.** The ground truth is highlighted in yellow.

[*] Corresponding author.

[1] Evaluation Codes and Data are available at https://videovista-culturallingo.github.io/

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27102–27128
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Category	Size
Task Classes	4
Subtask Classes	14
Video Sources	1,389
Video Clips	2,052
Max Duration	1,877.7
Average Duration	267.5
YouTube Video Domains	30
RedNote Video Domains	104
BiliBili Video Domains	12
Chinese Question Number	1,446
English Question Number	1,668
Chinese Culture QA Number	231
American Culture QA Number	200
European Culture QA Number	200
Average Question Length	18
Average Option Length	13
Average Choice Number	4
Total Samples	3,134
Total Questions	3,134

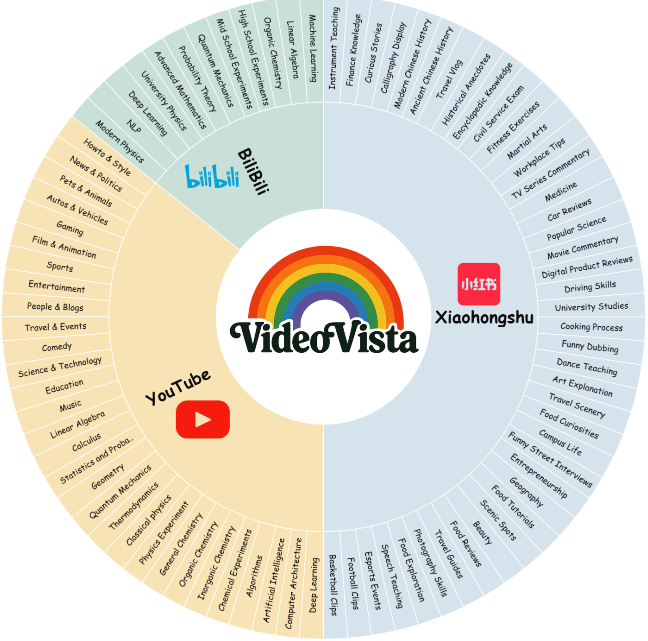


Figure 2: **(Left)** Comprehensive statistics from different perspectives. The durations reported are based on the statistics from the 2,052 video clips. The question and answer length is count in tokens; **(Right)** Videos in VideoVista-CulturalLingo is sourced hundreds of domains from **3** popular video websites across the world. In the video sourced from Xiaohongshu (RedNote), we only present 42 of the all domains.

formation that the videos are intended to convey.

To advance the development of LMMs, we introduce VideoVista-CulturalLingo, the first video evaluation benchmark designed to bridge cultures, languages, and domains in video comprehension. In Figure 2, we present detailed statistics on the questions and videos in VideoVista-CulturalLingo. It comprises 3,134 questions organized into 14 tasks, spanning 2,052 video clips of varying lengths and reflecting both Western and Chinese cultures. English-language videos are sourced from YouTube, while Chinese videos are collected from Xiaohongshu and BiliBili. These videos cover hundreds of distinct domains, ranging from everyday life topics—such as news reports, travel recommendations, sports events, and vlogs—to scientific topics, including calculus, deep learning, organic chemistry, and quantum mechanics.

To efficiently annotate such a large-scale video dataset, we employ a hybrid annotation framework that combines the strengths of both (M)LLMs and human efforts. This framework leverages the powerful capabilities of existing large models, such as Qwen2-VL (Wang et al., 2024a) and DeepSeek-R1 (DeepSeek-AI et al., 2025), to generate an initial pool of question-options-answer (QA) pairs. Human annotators then select the high-quality questions from generated QA pairs and further refine them to enhance clarity and quality.

We have evaluated 24 state-of-the-art (SOTA) LMMs, including proprietary LMMs such as GPT-4o, Gemini-2.0-Flash, as well as open-source video LMMs like Qwen2.5-VL (Team, 2025) and VideoLLaMA3 (Zhang et al., 2025), and image LMMs such as Molmo (Deitke et al., 2024) and DeepSeek2-VL (Wu et al., 2024). Experimental results show that Gemini-2.0-Flash demonstrates the strongest performance among all models, achieving an accuracy score of **76.3%**. Among open-source video LMMs, Qwen2.5-VL-72B achieves the highest score of 61.3%, with a large performance gap compared to Gemini-2.0-Flash in video location tasks. Interestingly, Qwen2.5-VL performs best on cultural understanding, yet still achieves only 65.8% in Chinese cultural understanding. In summary, the main contributions are as follows:

- We present the first video evaluation benchmark that covers diverse domains, languages, and cultures in video comprehension.
- We introduce an autonomic video annotation framework, harnessing the strengths of (M)LLMs (including Qwen2-VL and DeepSeek-R1) and visual recognition tools (including SAM2) to improve the efficiency of video annotation.
- We conduct extensive experiments and in-depth

analysis with VideoVista-CulturalLingo, revealing the limitations of existing LMMs in videos with different cultural or linguistic contexts.

2 Related Work

Development of Video LMMs. Unified modalities encodings have become the mainstream approach adopted by LMMs (Li et al., 2025c) over the past year. LongVA (Zhang et al., 2024a) utilizes a unified encoding method, Uni-Res, which allows models trained solely on image datasets to demonstrate strong potential in video evaluation tasks. Qwen2-VL (Wang et al., 2024a) and Qwen2.5VL (Team, 2025) introduce the M-RoPE positional encoding, incorporating temporal, height, and width components, enabling unified positional modeling across text, image, and video modalities. LLaVA-Video (Zhang et al., 2024b) draws inspiration from the SlowFast approach, encoding video frames at varying granularities into visual sequences of different lengths, effectively addressing the issue of excessively long sequences during video encoding. Current LMMs (Chen et al., 2024c; Yao et al., 2024; Li et al., 2024a, 2025b, 2024c) are capable of unified encoding for image and video modalities, leveraging rich image modality data to enhance visual capabilities and demonstrate strong performance in video evaluation tasks.

Progress of Video Benchmark. Video evaluation benchmarks have also made significant progress. Previously, evaluation datasets (Yu et al., 2019; Xu et al., 2017) typically involved posing broad questions and having the model generate a one or a few-word answer, which was then assessed for accuracy and scored by LLMs (Maaz et al., 2024). The videos used in these datasets were often limited to just a few dozen seconds or minutes in length. Recent video benchmarks (Li et al., 2024b) have seen considerable improvements, both in the variety of evaluation tasks and the duration of the videos. Video-MME (Fu et al., 2024) has extended the evaluation video length to an hour, while also introducing twelve distinct evaluation tasks, including Temporal Reasoning and Information Synopsis. MLVU (Zhou et al., 2024) includes videos of varying lengths, ranging from 3 minutes to 2 hours, covering nine different evaluation tasks, such as Needle Question-Answering. The process of video benchmarks (Fang et al., 2024; Wang et al., 2024b; Liu et al., 2024a) have undoubtedly provided a significant boost to the development of LMMs.

3 VideoVista-CulturalLingo

3.1 Video Collecting and Preprocessing

The videos in our study can be divided into two categories: non-scientific and scientific videos. Non-scientific English videos are randomly crawled from YouTube, while their Chinese counterparts are collected from Xiaohongshu to ensure diversity within the dataset. The domains of these videos come from the original categories on the video platforms. For scientific videos, we first identified four major disciplines: mathematics, physics, chemistry, and computer science. Within each discipline, we further defined four representative sub-disciplines, such as linear algebra in mathematics and quantum mechanics in physics. Domains of these videos are derived from search keywords. These sub-disciplines guide the collection of English scientific videos via the YouTube Data API. For Chinese scientific videos, human annotators manually collected videos from BiliBili.

All videos undergo audio extraction via FFmpeg, followed by transcription using Whisper-Large-v3 with sentence-level timestamp alignment. An audio quality assessment pipeline is implemented using Qwen2.5-32B (Yang et al., 2024), evaluating three dimensions: logical coherence, continuity, and information density. Videos are subsequently classified as either audio-rich (high-quality speech) or audio-noisy (including silent videos). For audio-rich videos, the Qwen2.5-72B model segments transcriptions into contextually coherent paragraphs, which are synchronized with visual content through Whisper's sentence-level alignment to generate short video clips. Audio-noisy videos are processed using the semantics-aware video splitting algorithm from Panda-70M (Chen et al., 2024b), which utilizes visual features to partition videos into semantically consistent segments. This process is illustrated in Figure 3 (a).

To address the challenges of Chinese homophone ambiguity in transcriptions, we develop a context-aware refinement module using Qwen2.5-72B. This module performs three key operations: (1) disambiguation of homophones through semantic analysis, (2) correction of domain-specific terminology, and (3) fluency enhancement, while strictly preserving original semantic content.

3.2 Automatic QA Annotation

The annotation framework comprises four distinct tasks: Event, Culture, Object, and Science. Our

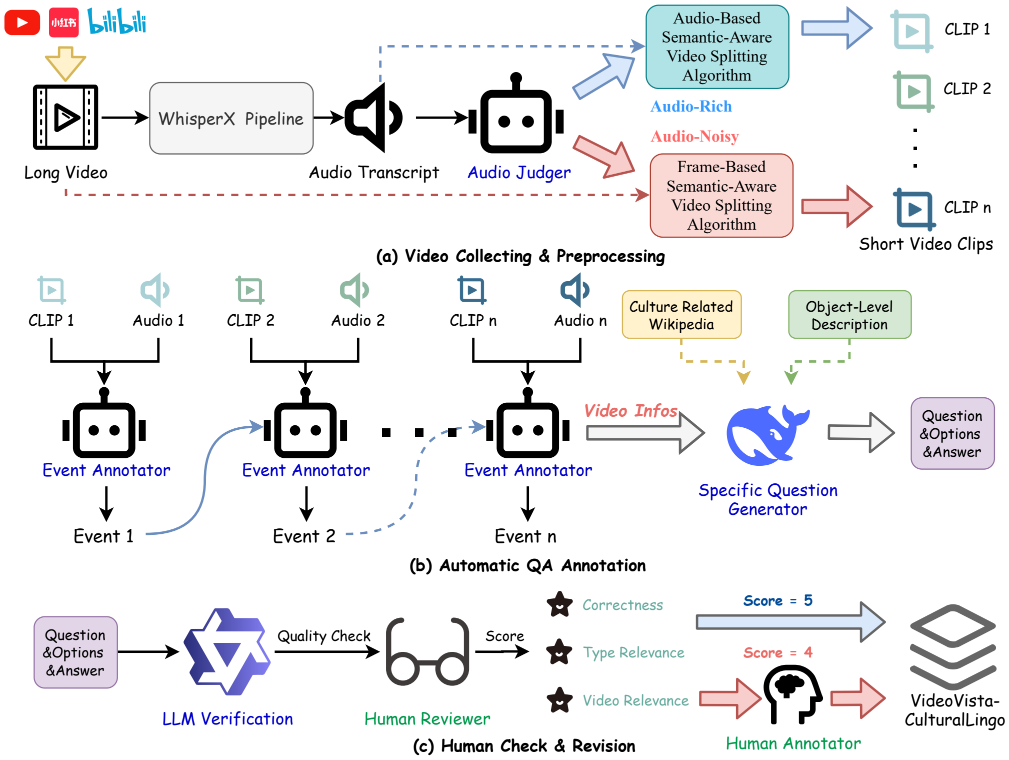


Figure 3: **The three-stage annotation process of VideoVista-CulturalLingo.**

Benchmarks	**#Videos**	**#Clips**	**Len.(s)**	**#QA Pairs**	**Anno.**	**M.L.**	**M.C**	**M.D**	**Open.**
MSRVTT-QA (Xu et al., 2017)	2,990	2,990	15.2	72,821	A	✗	✗	✗	✓
MSVD-QA (Xu et al., 2017)	504	504	9.8	13,157	A	✗	✗	✗	✓
TGIF-QA (Li et al., 2016)	9,575	9,575	3.0	8,506	A&M	✗	✗	✗	✓
ActivityNet-QA (Yu et al., 2019)	800	800	111.4	8,000	M	✗	✗	✗	✗
TVQA (Lei et al., 2018)	2,179	15,253	11.2	15,253	M	✗	✗	✗	✗
NExT-QA (Xiao et al., 2021)	1,000	1,000	39.5	8,564	A	✗	✗	✗	✓
MVBench (Li et al., 2023)	3,641	3,641	16.0	4,000	A	✗	✗	✗	✓
EgoSchema (Mangalam et al., 2024)	5,063	5,063	180.0	5,063	A&M	✗	✗	✗	✗
TempCompass (Liu et al., 2024a)	410	500	11.4	7,540	A&M	✗	✗	✗	✓
Video-MME (Fu et al., 2024)	900	900	1024.0	2,700	M	✗	✗	✓	✓
VideoVista (Li et al., 2024b)	894	3,402	131.0	24,906	A	✗	✗	✓	✓
MLVU (Zhou et al., 2024)	1,323	1,323	720	2,593	A&M	✗	✗	✓	✓
LVBench (Wang et al., 2024b)	500	500	4,101.0	1,549	M	✗	✗	✓	✓
MMBench-Video (Fang et al., 2024)	600	600	165.4	1,998	M	✗	✗	✓	✓
VideoVista-CulturalLingo	1,389	2,052	267.5	3,134	A&M	✓	✓	✓	✓

Table 1: The comparison of various benchmarks involves several key aspects: total number of videos (**#Videos**), number of clips (**#Clips**), average video duration (**Len.**), number of QA pairs (**#QA Pairs**), annotation method (**Anno.**, where M/A indicates manual/automatic annotation), whether the videos span multiple language (**M.L.**),whether the videos span multiple culture background (**M.C.**) ,whether the videos span multiple duration levels (**M.D.**), and if the videos are sourced from diverse open domains (**Open.**)

pipeline employs Qwen2-VL-72B as the primary annotator, Qwen2.5-72B for text-only annotation tasks, and paraphrase-multilingual-MiniLM-L12-v2 for embedding generation. For non-scientific tasks, DeepSeek-V3 (DeepSeek-AI et al., 2024) is employed as the question generator, while DeepSeek-R1 (DeepSeek-AI et al., 2025) is used for generating scientific questions. During the annotation process, while generating questions, four options and the correct answer are also created. The process of automatic QA annotation is illustrated in Figure 3 (b). *The details and prompt for annotation is provided in Appendix D.*

Event. We input the segmented video clips and refined audio transcriptions into the event annotator to label the events occurring in each video segment. For the i-th segment, the model receives historical event annotations from the previous $i-1$ segments to maintain temporal consistency. Each annotated segment follows the structure $(event, audio, start, end)$, where $start$ and end denote the timestamps marking the beginning and conclusion of the current video segment within the full video. The aggregated event sequence is then fed into the question generator, which generates questions of the corresponding task, along with four options for each question and correct answer. Specifically, for event prediction questions, the model is instructed to select the segment that is most logically related to the preceding context as the predicted content. During this process, each task is associated with a specific prompt.

Object. We feed videos into the object classifier to filter those videos that meet three criteria: real-world content, richness in objects, and motion in objects. The filtered videos are then processed by the object extractor to identify three to five primary objects followed by frame-wise presence detection via InternVL2-8B at 1fps sampling. The detected objects are processed through a pipeline combining Grounding-DINO (Liu et al., 2023a) for bounding box prediction and SAM2 (Ravi et al., 2024) for image segmentation. The resulting information is then fed into the object description annotator to generate object-level descriptions that capture both the temporal and spatial aspects of each object. Finally, the object-level descriptions, along with the aggregated event sequence, are input into the question generator to generate the questions.

Culture. We input videos and audio transcriptions into the cultural classifier to evaluate their relationship to Chinese, American, and European cultures individually. Culturally relevant videos are then processed by the cultural concept extractor to identify the two most prominent cultural concepts. These cultural concepts are subsequently encoded into embeddings, which are used to retrieve the entries from pre-encoded Wikipedia data. Using these entries, along with a local backup of Wikipedia, we can retrieve Wikipedia articles corresponding to the identified cultural concepts. By combining this external knowledge with the aggregated event sequence, we input the data into the specific question generator to generate the questions.

Science. The video is input into the science classifier to evaluate its quality based on scientific thematic relevance and knowledge density. After filtering, the aggregated event sequence of the video is fed into the question generator, DeepSeek-R1, to generate questions. In our initial experiments, we observed two recurring issues: generated questions either relied excessively on domain knowledge—detached from the video itself and thus answerable without viewing—or exhibited distractor choices that were either too divergent or too similar, producing items that were trivial or ambiguous. To resolve these shortcomings, we impose deterministic, rule-based constraints that (i) require every question to depend on video context for its solution and (ii) ensure a balanced, pedagogically meaningful set of answer options. Specifically, each question presents four choices: Correct Option, Video Comprehension Error, Domain Knowledge Error, and Dual Error. This structured design rigorously evaluates a model's ability to integrate visual comprehension with scientific reasoning.

3.3 Human Check and Revision

All candidate questions are first filtered linguistically using Qwen2.5-7B with the CircularEval strategy (Liu et al., 2023b) to remove any video-agnostic items. We then establish a Gradio-based annotation platform that includes three assessment dimensions: correctness, type relevance, and video relevance. The correctness score ranges from 0 to 1, assessing whether the model-generated answer is correct; the type relevance score ranges from 0 to 2, evaluating the degree of relevance between the question and task type; and the video relevance score ranges from 0 to 2, determining the degree of relevance between the question and the video content, ensuring that questions are not unrelated to the video frames. Questions achieving maximum scores (score=5) across all dimensions are selected. For borderline cases (score=4), we utilize differentiated handling: first, for the question with wrong answer (correctness=0), we manually correct the answers; second, for the question with suboptimal type or video relevance, we manually refine the questions, options, and answers based on the original questions. We have illustrated this process in the Figure 3 (c). Specifically for cultural questions, two annotators—one of whom is a native speaker of the relevant culture—independently assess each question to ensure cross-validation. Overall, this hybrid automatic/manual pipeline eliminates approximately 60 % of low-quality questions.

3.4 Statistic and Analysis

As shown in Figure 2, VideoVista-CulturalLingo consists of 2,052 video clips or full videos derived from 1,389 original videos, with an average duration of 267.5 seconds. Additionally, VideoVista-CulturalLingo contains 1,446 questions in Chinese

Model	LLM	Frames	Overall	Event	Object	Culture	Science
Open-source Video LMMs							
ShareGPT4Video (Chen et al., 2024a)	Vicuna-7B-v1.5	16f	25.6	23.2	18.9	31.4	34.1
VideoChat2-Mistral (KunChang et al., 2023)	Mistral-7B-Instruct-v0.2	16f	29.6	27.5	25.9	34.7	33.1
Video-LLaVA (Lin et al., 2023a)	Vicuna-7B-v1.5	8f	38.2	42.2	34.4	34.5	41.1
VideoLLaMA2 (Cheng et al., 2024)	Mistral-7B-Instruct-v0.2	32f	31.4	33.6	23.3	34.9	36.6
LLaVA-OneVision (Li et al., 2024a)	Qwen2-7B-Instruct	32f	41.8	43.9	33.8	38.8	53.5
MiniCPM-V 2.6 (Yao et al., 2024)	Qwen2-7B-Instruct	1fps(64)	42.9	44.1	24.1	49.4	62.9
mPLUG-Owl3 (Ye et al., 2024)	Qwen2-7B-Instruct	1fps(128)	49.9	54.4	41.9	45.0	60.1
Oryx-1.5 (Liu et al., 2024b)	Qwen2.5-7B-Instruct	128f	41.4	43.8	32.2	37.6	55.8
LLaVA-Video (Zhang et al., 2024b)	Qwen2-7B-Instruct	1fps(64)	51.0	57.9	39.1	48.8	60.3
Qwen2-VL (Wang et al., 2024a)	Qwen2-7B-Instruct	1fps(300)	49.7	50.1	33.8	54.8	68.0
InternVL2.5 (Chen et al., 2024c)	Internlm2.5-7b-Chat	64f	52.0	56.5	35.5	56.1	65.7
MiniCPM-o 2.6 (Yao et al., 2024)	Qwen2.5-7B-Instruct	1fps(64)	49.0	52.9	28.5	55.9	67.1
TPO (Li et al., 2025a)	Qwen2-7B-Instruct	1fps(96)	50.6	57.2	37.8	49.6	60.4
InternVideo2.5 (Wang et al., 2025)	Internlm2.5-7b-Chat	1fps(512)	52.0	52.5	38.1	58.2	65.9
VideoLLaMA3 (Zhang et al., 2025)	Qwen2.5-7B-Instruct	1fps(180)	60.7	58.0	**66.4**	53.1	64.4
Qwen2.5-VL-7B (Team, 2025)	Qwen2.5-7B-Instruct	1fps(300)	54.3	56.7	38.9	55.2	73.3
Qwen2.5-VL-72B (Team, 2025)	Qwen2.5-72B-Instruct	1fps(300)	**61.3**	**61.0**	40.5	**71.2**	**83.3**
Open-source Image LMMs							
VILA1.5-13B (Lin et al., 2023b)	Vicuna-13B-v1.5	1f	33.3	33.3	29.2	33.9	39.2
VILA1.5-13B (Lin et al., 2023b)	Vicuna-13B-v1.5	8f	36.9	38.2	31.3	38.2	41.9
Molmo 7B-D (Deitke et al., 2024)	Qwen2-7B-Instruct	1f	38.3	44.5	25.3	39.8	46.5
Molmo 7B-D (Deitke et al., 2024)	Qwen2-7B-Instruct	8f	40.3	44.3	30.1	41.8	48.0
DeepSeek2-VL (Wu et al., 2024)	DeepSeekMoE-27B	1f	40.9	44.3	**32.2**	39.3	50.5
DeepSeek2-VL (Wu et al., 2024)	DeepSeekMoE-27B	8f	**42.6**	**47.0**	27.2	**44.4**	**57.5**
Proprietary LMMs							
GPT-4o-2024-11-20	GPT-4o	1fps(128)	56.7	53.4	38.2	**68.0**	78.3
Gemini-1.5-Flash	Gemini-1.5-Flash	1fps	69.4	70.0	65.8	59.0	84.7
Gemini-2.0-Flash-Lite	Gemini-2.0-Flash-Lite	1fps	70.7	63.1	71.6	63.1	82.1
Gemini-2.0-Flash	Gemini-2.0-Flash	1fps	**76.3**	**74.0**	**77.1**	**68.0**	**87.4**

Table 2: **Evaluation results on VideoVista-CulturalLingo benchmark.** The large language model used by LMMs (**LLM**), frames sample strategy (**Frames**), overall evaluation scores (**Overall**), evaluation scores in Event Task(**Event**), evaluation scores in Object Task (**Object**), evaluation scores in Culture Task (**Culture**), evaluation scores in Science Task (**Science**). -[Nf] indicates this LMM task N frames uniformly sampled from a video as input. -[Nfps(M)] indicates this LMM uses N frames per second uniformly sampled from a video as input, with a max frames number M. We have highlighted the highest results in each tasks using **bold**. Meanwhile, the highest results within the 7B/8B open-source Video LMMs are highlighted with an underline.

and 1,668 questions in English, with a comparable number of questions in both languages. In Table 1, we compare the key characteristics of our benchmark with others. Notably, VideoVista-CulturalLingo includes the largest collection of raw videos, totaling 1,389, among benchmarks that have videos multiple duration levels. These 1,389 original videos encompass a diverse range of languages and cultural backgrounds, a feature that sets our benchmark apart from previous ones. Additionally, we employed a annotation approach that combines (M)LLM preliminary annotation with human verification and question revision.

4 Experiment

4.1 Baselines

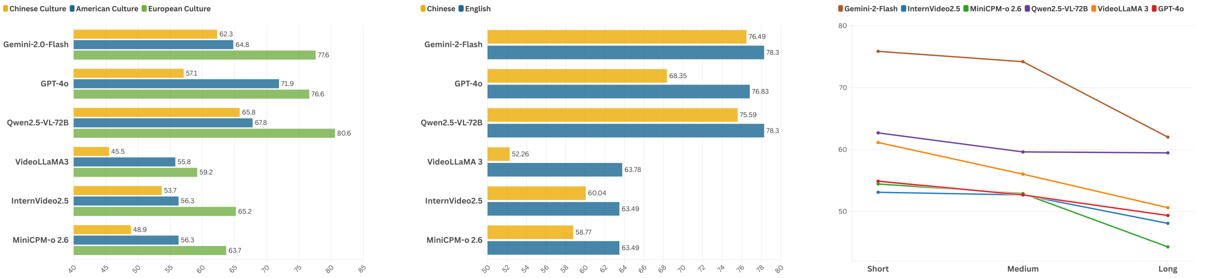
We conducted evaluations on 17 open-source video LMMs, 3 image LMMs, and 4 proprietary LMMs, including the recently released Gemini-2.0-Flash, Qwen2.5-VL (Team, 2025), VideoLLaMA3 (Zhang et al., 2025), DeepSeek2-VL (Wu et al., 2024), among others. *The detailed experiment settings are shown in Appendix B.*

4.2 Main Results

As shown in Table 2, Qwen2.5-VL-72B exhibits the best performance among all open-source video LMMs, achieving an overall score of 61.3%. Additionally, VideoLLaMA3 demonstrates the best performance among all 7B/8B models, with an overall score of 60.7%. This is primarily due to VideoLLaMA3's exceptional capabilities in fine-grained object tasks, making it the only open-source LMM that can compete with proprietary LMMs in this task. In the event task, VideoLLaMA3 also outperforms all other 7B models. Among the open-source image LMMs, DeepSeek2-VL achieved the highest score of 42.6% under 8-frame uniform sam-

Model	Event				Object			Culture			Science			
	ED	EP	ES	EL	OTL	OTS	OSL	CC	AC	EC	SS	COM	AP	SP
MiniCPM-o 2.6	**83.6**	55.0	53.1	35.2	20.1	52.4	35.7	48.9	56.3	63.7	72.1	61.3	69.5	52.7
InternVideo2.5	80.5	52.7	60.3	33.0	37.1	61.2	31.8	53.7	56.3	65.2	72.1	61.3	64.0	54.8
VideoLLaMA3	77.9	57.4	61.7	45.2	**72.1**	64.1	**56.6**	45.5	55.8	59.2	70.2	54.7	64.0	55.9
Qwen2.5-VL-72B	79.2	**60.5**	**78.9**	42.1	31.5	**67.0**	49.7	**65.8**	**67.8**	**80.6**	**86.4**	**85.3**	**79.3**	**79.6**
GPT-4o	86.3	47.3	70.3	28.6	29.4	61.2	46.5	57.1	71.9	76.6	81.6	77.3	80.5	65.6
Gemini-2.0-Flash	92.9	51.9	73.7	70.7	87.2	74.8	59.1	62.3	64.8	77.6	88.2	87.8	81.7	90.7

Table 3: **Detailed Evaluation results on VideoVista-CulturalLingo benchmark.** We only showcase 6 mainstream LMMs. Abbreviations used in the table: Event Description (**ED**), Event Prediction (**EP**), Event Sequence (**ES**), Event Localization (**EL**), Object Temporal Localization (**OTL**), Object Temporal Sequence (**OTS**), Object Spatial Localization (**OSL**), Chinese Culture (**CC**), American Culture (**AC**), European Culture (**EC**), Summarization & Synthesis (**SS**), Comparison & Contrast (**COM**), Application & Procedure (**AP**), Scientific Principle (**SP**). *The full evaluation results are provided in the Appendix C.5, and an introduction to tasks is presented in Appendix E.*



(a) Culture-based Evaluation Results.

(b) Language-based Evaluation Results.

(c) Duration-based Evaluation Results.

Figure 4: **The LMMs performance divided by Culture, Language and Duration.** The Duration in (c): <2 minutes (**Short**), 2-10 minutes (**Medium**), >10 minutes (**Long**).

pling, demonstrating its superior generalization capacity on sequential image data. However, this still shows a gap compared to the leading open-source video LMMs, indicating that questions in VideoVista-CulturalLingo generally require longer video durations to answer. Among proprietary LMMs, Gemini-2.0-Flash clearly outperforms all others, surpassing the strongest open-source video LMM, Qwen2.5-VL-72B, by 15.0%. The largest performance gap between these two models is observed in fine-grained object understanding tasks.

4.3 Detailed Analysis

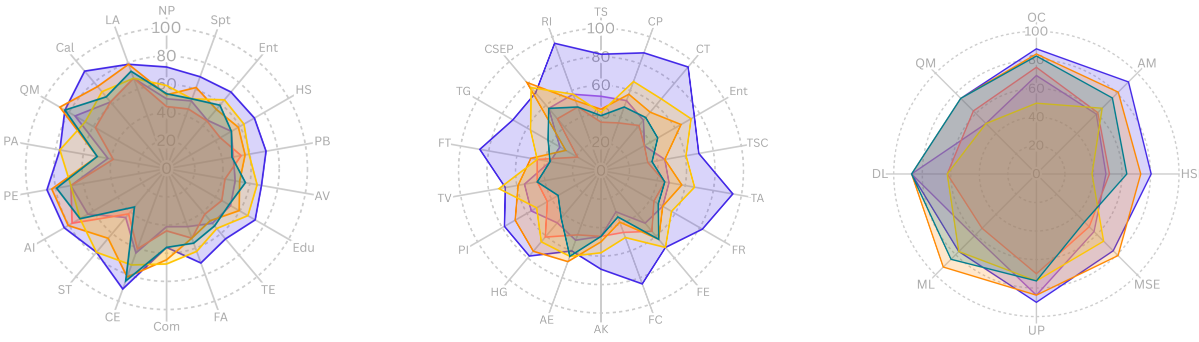
We present the detailed evaluation results of 6 mainstream models across 14 sub-tasks in Table 3.

Event. The Event task consists of four sub-tasks: Event Description, Event Prediction, Event Sequence, and Event Localization, all of which require the model to have a coarse-grained understanding of video content. Current open-source video LMMs exhibit performance comparable to that of proprietary LMMs on the first three sub-tasks, but there remains a gap in the Event Localization task when compared to Gemini-2.0-Flash, with a performance difference of up to 25.5%.

Object. The Object task consists of three sub-tasks: Object Temporal Localization, Object Temporal Sequence, and Object Spatial Localization, which assess the LMMs' ability to perceive the spatial-temporal aspects of fine-grained objects in videos. Video-LLaMA3 and Gemini-2.0-Flash demonstrate strong temporal localization capabilities in the Object Temporal Localization task, achieving scores more than 30% higher than those of other LMMs. Additionally, both LMMs exhibit commendable spatial understanding in the Object Spatial Localization task.

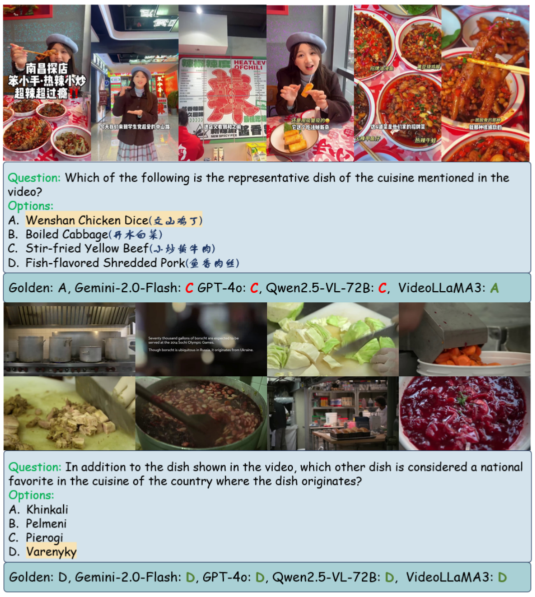
Culture. The Culture task consists of three sub-tasks: Chinese Culture, American Culture, and European Culture, primarily evaluating the model's understanding and generalization abilities across different regional cultures. As shown in Figure 4a, compared to the more prevalent Western cultures in the training data, current LMMs exhibit relatively weaker recognition of Chinese Culture.

Science. The Science task consists of four sub-tasks: Summarization & Synthesis, Comparison & Contrast, Application & Procedure, and Scientific Principle. The first three sub-tasks involve



(a) Domains in YouTube (b) Domains in Xiaohongshu (c) Domains in BiliBili

Figure 5: **The LMMs performance divided by domains from 3 video sources:** Gemini-2.0-Flash, GPT-4o, Qwen2.5-VL-72B, VideoLLaMA3, InternVideo2.5, MiniCPM-o 2.6. In Figures 5a and Figures 5b, we present only the 18 domains with the highest number of videos. In Figure 5c, we exclude domains containing fewer than 10 videos. *The domains in these figures are represented by abbreviations, as described in Appendix A.2.*


南昌探店
笨小手·热辣小炒
超辣超过瘾!!
Seventy thousand gallons of borscht are expected to be served at the 2014 Sochi Olympic Games.
Though borscht is ubiquitous in Russia, it originates from Ukraine.

Question: Which of the following is the representative dish of the cuisine mentioned in the video?
Options:
A. Wenshan Chicken Dice(文山鸡丁)
B. Boiled Cabbage(开水白菜)
C. Stir-fried Yellow Beef(小炒黄牛肉)
D. Fish-flavored Shredded Pork(鱼香肉丝)

Golden: A, Gemini-2.0-Flash: C GPT-4o: C, Qwen2.5-VL-72B: C, VideoLLaMA3: A

Question: In addition to the dish shown in the video, which other dish is considered a national favorite in the cuisine of the country where the dish originates?
Options:
A. Khinkali
B. Pelmeni
C. Pierogi
D. Varenyky

Golden: D, Gemini-2.0-Flash: D, GPT-4o: D, Qwen2.5-VL-72B: D, VideoLLaMA3: D

Figure 6: **Two cases from VideoVista-CulturalLingo.** The ground truth is highlighted in yellow. The correct answers of LMMs are highlighted in green, and the incorrect answers are highlighted in red.

course-oriented educational videos, while the last one focuses on experimental videos. This task primarily evaluates the model's ability to summarize, comprehend, and apply scientific knowledge from videos. The difficulty level covers general knowledge areas rather than in-depth specialized topics. The questions are relatively simple and can be answered with one or two-hop reasoning, so most models perform well in these tasks. We observe that existing open-source LMMs perform comparably to proprietary LMMs across most disciplines. However, there remains a noticeable gap in performance within math. *The detailed comparison is presented in the Appendix C.1.*

4.4 Ablation Study

Language. In Figure 4b, we present the performance differences of 6 mainstream LMMs on Chinese and English. The results in the figure are based on 7 subtasks from the culture and science tasks, as these subtasks contain more domain-specific terms, providing a more accurate assessment of an LMM's capabilities in each respective language. The experiments reveal a noticeable performance gap between the majority of mainstream LMMs when evaluated on Chinese versus English.

Duration. In Figure 4c, we compare the performance of 6 mainstream LMMs across 4 subtasks of event task from videos of varying lengths. The experimental results indicate that as the video duration increases, the performance of model tends to decrease, including Gemini-2.0-Flash.

Domain. In Figure 5, we illustrate the performance of LMMs across different video domains on various video websites. It can be observed that Gemini-2.0-Flash demonstrates strong performance across all domains of videos.

4.5 Case Study

Figure 6 presents two cultural examples alongside evaluation results: the top case illustrates a Chinese cuisine scenario, and the bottom case illustrates a European cuisine scenario. In the Chinese example, the majority of LMMs erroneously choose "C. Stir-fried Yellow Beef," a hallmark Hunan dish. This mistake likely arises from conflation between Jiangxi and Hunan cuisines—both characterized by liberal use of chili peppers—and from the greater domestic and international visibility of Hunan cooking. Such errors reveal that Video-LMMs tend to default to dominant cultural representations, overlooking more localized culinary nuances. By contrast, in the European example all LMMs correctly select option D, indicating robust performance on Western culinary content. Together, these cases exemplify a systematic bias: Video-LMMs achieve higher accuracy on Western cultural contexts but underperform on non-Western ones, such as videos rooted in Chinese culture.

5 Conclusion

In this paper, we introduce the benchmark VideoVista-CulturalLingo, the first video evaluation benchmark that spans multiple languages, cultures, and domains. VideoVista-CulturalLingo includes comprehensive evaluation metrics, ranging from coarse-grained event understanding to fine-grained object recognition, and from exploring the cultural context of videos to uncovering their scientific implications, enabling a comprehensive assessment of current LMMs' capabilities on video tasks. Through our extensive experiments, we highlight weaknesses in the spatial-temporal localization abilities of existing open-source video LMMs and their limitations in recognizing Chinese culture. We hope that VideoVista-CulturalLingo will inspire the development and advancement of video LMMs.

6 Acknowledge

We thank editor and reviewers for their efforts to help improve the quality of our paper. This work was jointly supported by grants: Natural Science Foundation of China (No. 62422603), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515110078), Shenzhen Science and Technology Program (No. ZDSYS20230626091203008). This work was also supported in part by HKUST-MetaX Joint Laboratory for Advanced AI Computing (Grant No. METAX24EG01-D)

Limitations

The proposed benchmark has several limitations: 1) The scientific questions in the benchmark lack domain-specific depth, which prevents them from effectively showcasing the model's performance in specialized scientific fields. In future versions, we plan to incorporate more human expert annotators to enhance the professionalism and complexity of the scientific questions. 2) Due to limitations in the linguistic proficiency and backgrounds of the annotators, the benchmark questions are restricted to two major languages, Chinese and English. This excludes other widely spoken languages such as Spanish, Portuguese, German, and Japanese.

References

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*.

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

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing

Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.

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

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012.

Li KunChang, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*.

Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. 2025a. Temporal preference optimization for long-form video understanding. *Preprint*, arXiv:2501.13919.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024b. Videovista: A versatile benchmark for video understanding and reasoning. *Preprint*, arXiv:2406.11303.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. 2024c. Lmeye: An interactive perception network for large language models. *IEEE Transactions on Multimedia*, 26:10952–10964.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, Zheng Zhang, Baotian Hu, and Min Zhang. 2025c. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023b. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024a. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv: 2403.00476*.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024b. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *Preprint*, arXiv:2408.00714.

Qwen Team. 2025. Qwen2.5-vl.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. Lvbench: An extreme long video understanding benchmark. *Preprint*, arXiv:2406.08035.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

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

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

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A new movie understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

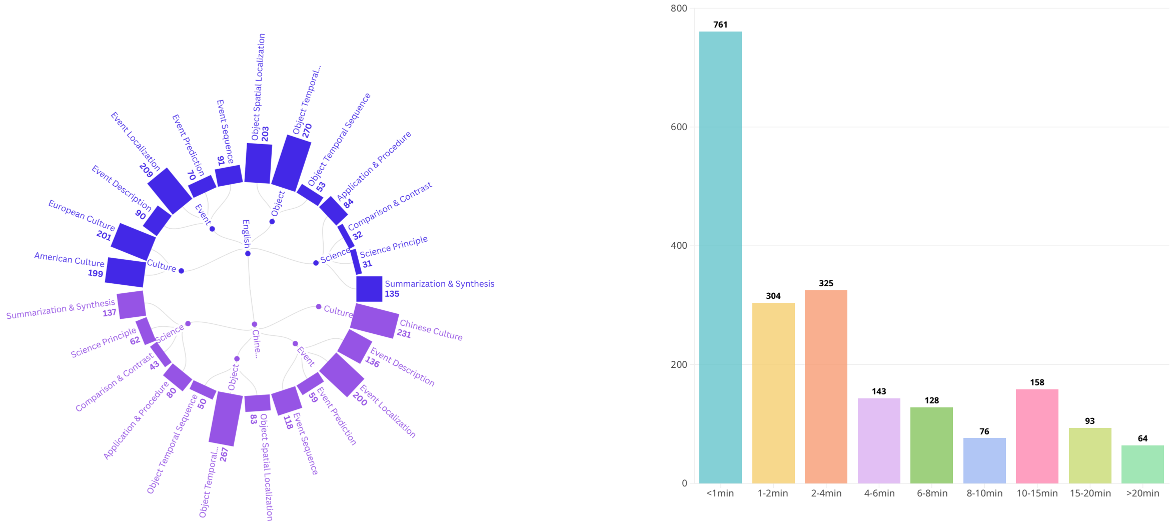
Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

A Additional Dataset Statistics

A.1 Further Statistics

In Figure 7a, we present the statistics for all task categories in VideoVista-CulturalLingo. In VideoVista-CulturalLingo, the number of English questions is slightly higher than that of Chinese questions, with an additional 222 English questions. The task type with the fewest questions in the dataset is "Comparison & Contrast", with a total of only 75 questions, while the task type with the most questions is "Object Temporal Localization," with a total of 537 questions. Figure 7b (b) shows the temporal distribution of video clips. Due to the fine-grained object recognition task, the selected videos are often short segments of longer videos, resulting in a larger proportion of videos that are



(a) The statistics of 14 subtasks divided by languages.

(b) The statistics of duration of videos in VideoVista-CulturalLingo.

Figure 7: (a) shows the quantity statistics for the 14 task categories under both Chinese and English languages. (b) presents the duration statistics of all video clips in VideoVista-CulturalLingo, measured in minutes.

under one minute in length in the dataset. However, VideoVista-CulturalLingo still contains 315 videos longer than 10 minutes, with these long videos primarily concentrated in the Event and Science task categories.

A.2 Abbreviations of Domains

We provided the abbreviations of domains in Figure 5 in Table 4.

B Detailed Experiment Setting

B.1 Open-source Video LMMs

We evaluated the newly released Qwen2.5-VL (Team, 2025), VideoLLaMA3 (Zhang et al., 2025), InternVideo2.5 (Wang et al., 2025), and TPO (Li et al., 2025a) from 2025. Additionally, we evaluated several popular video-capable LMMs introduced in the past two years, including InternVL2.5 (Chen et al., 2024c), LLaVA-Video (Zhang et al., 2024b), mPLUG-Owl3 (Ye et al., 2024), and others.

In evaluating open-source video LMMs, we use the default hyperparameters specified in their respective open-source implementations for inference. The temperature is generally set to 0 or 0.2, num_beamsis set to 1, do_sampleis set to False, and top_pis set to 1.0. The frame sampling methods for different video models are provided in the Table 2. Specifically, for the Qwen2.5-VL and Qwen2-VL models, we set the maximum resolution per frame to 224x224 to avoid excessively long sequence lengths.

B.2 Open-source Image LMMs

We also evaluated three open-source image LMMs on our benchmarks, including VILA 1.5 (Lin et al., 2023b), DeepSeek2-VL (Wu et al., 2024), and Molmo (Deitke et al., 2024). For open-source image LMMs, we employed two video input methods: uniform sampling of 1 frame and uniform sampling of 8 frames.

In evaluating these open-source image LMMs, we also adopted the hyperparameter settings provided in the implementations for inference. Regardless of whether single-frame or eight-frame input is used for evaluation, all images are presented at their original resolution without compression. Specifically, due to an error in the official code of the Molmo model when inputting eight images simultaneously, we concatenated the eight images horizontally into a single image and noted this in the prompt. An example of this image is Figure 8.

B.3 Proprietary LMMs

For proprietary LMMs, we evaluated the newly released Gemini 2.0-Flash and Gemini 2.0-Flash-Lite in February, which are currently the workhorse models of the Google Gemini series. Additionally, we conducted evaluations on other prominent proprietary LMMs, including GPT-4o and Gemini 1.5-Flash.

In evaluating proprietary LMMs, we optimize API resource usage and accelerate the evaluation

Figure 8: **An example of eight images combined in a horizontal layout.**

Full Name	Abbreviation
YouTube Domains	
News & Politics	**NP**
Sports	**Spt**
Entertainment	**Ent**
Howto & Style	**HS**
People & Blogs	**PB**
Autos & Vehicles	**AV**
Education	**Edu**
Travel & Events	**TE**
Film & Animation	**FA**
Comedy	**Com**
Chemical Experiments	**CE**
Science & Technology	**ST**
Artificial Intelligence	**AI**
Physics Experiment	**PE**
Pets & Animals	**PA**
Quantum Mechanics	**QM**
Calculus	**Cal**
Linear Algebra	**LA**
Xiaohongshu Domains	
Travel Scenery	**TS**
Cooking Process	**CP**
Cooking Tutorial	**CT**
Entrepreneurship	**Ent**
TV Series Commentary	**TSC**
Tourist Attractions	**TA**
Food Review	**FR**
Food Exploration	**FE**
Food Curiosities	**FC**
Astronomy Knowledge	**AK**
Art Explanation	**AE**
Historical Gossip	**HG**
Product Information	**PI**
Travel Vlog	**TV**
Fashion Trends	**FT**
Travel Guide	**TG**
Civil Service Exam Preparation	**CSEP**
Relationship Issues	**RI**
BiliBili Domains	
Organic Chemistry	**OC**
Advanced Mathematics	**AM**
High School Experiments	**HSE**
Mid School Experiments	**MSE**
University Physics	**UP**
Machine Learning	**ML**
Deep Learning	**DL**
Quantum Mechanics	**QM**

Table 4: **Abbreviations of domains from different video websites in Figure 5.** The Chinese domains have been translated into English using GPT-4o.

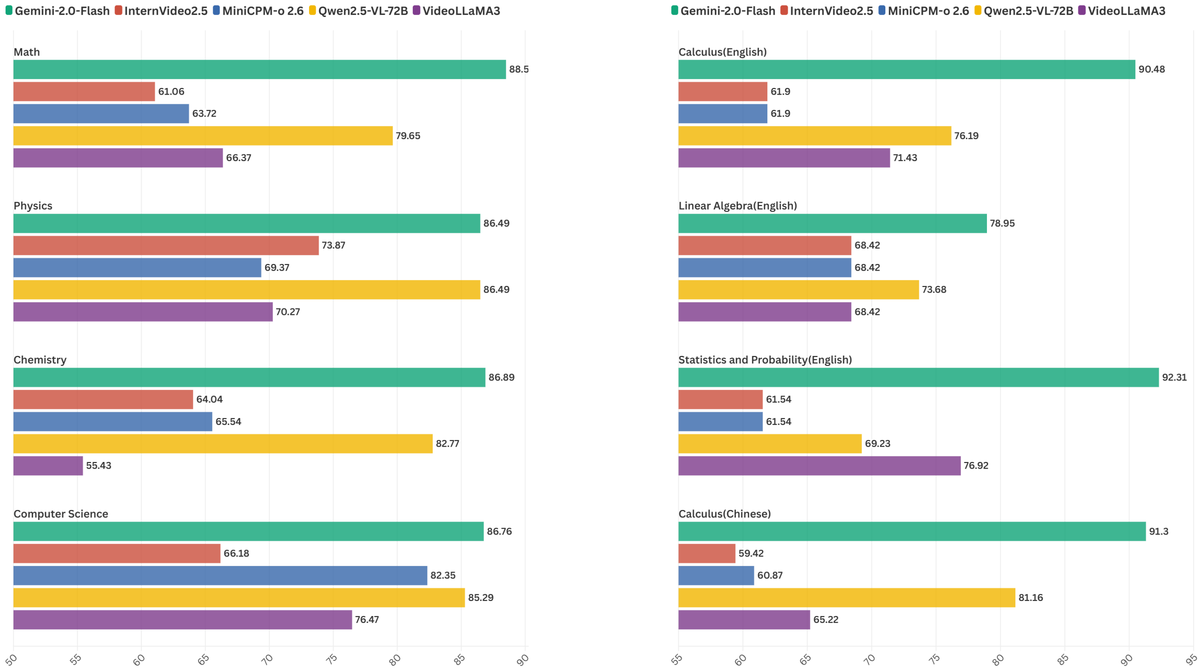
process by input multiple questions for each video. Thanks to the powerful instruction-following capability of Proprietary LMMs, they are able to return a dictionary in the format of {"question id": "prediction"} accurately. Although this may introduce some evaluation bias, Proprietary LMMs still demonstrated exceptional performance on our benchmark. Additionally, when evaluating the GPT-4 model, we compressed all video frames to a resolution of 512x512 for input.

C Further Experiments

C.1 Model Performance in Science

For the third finding discussed in the abstract, we present detailed experimental results in Figure 9. We present a performance comparison between the four best-performing open-source video LMMs and the strongest proprietary model, Gemini-2.0-Flash. As shown in Figure 9a, the primary performance gap between open-source Video LMMs and proprietary LMMs in scientific tasks is observed in the Mathematics disciplines. Specifically, for Physics, Chemistry, and Computer Science questions, the top-performing open-source Video LMM, Qwen2.5-VL-72B, exhibits a performance gap of less than 5% compared to Gemini-2.0-Flash. However, for Math questions, the gap between the two models increases to nearly 10%.

In Figure 9b, we further compare the performance differences of various models across specific math sub-disciplines. It is evident that, regardless of whether the questions are in Chinese or English, existing open-source video LMMs still exhibit a performance gap when compared to the proprietary LMM Gemini-2.0-Flash. The largest gaps are observed in the Calculus (English) and Statistics and Probability (English) categories, where the leading open-source video LMMs show a performance difference exceeding 10% compared to Gemini-2.0-Flash.



(a) Science-based Evaluation Results.

(b) Math-based Evaluation Results.

Figure 9: **The Evaluation results in 4 disciplines and 4 math sub-disciplines.** The experimental results in the figure represent the average values of the four scientific sub-tasks. In (a),we have list the four disciplines covered by the scientific videos in VideoVista-CulturalLingo: Math, Physics, Chemistry, and Computer Science ; In (b), we have listed four math sub-disciplines with a larger number of questions: Calculus (English), Linear Algebra (English), Statistics and Probability (English), and Calculus (Chinese)/Advanced Mathematics.

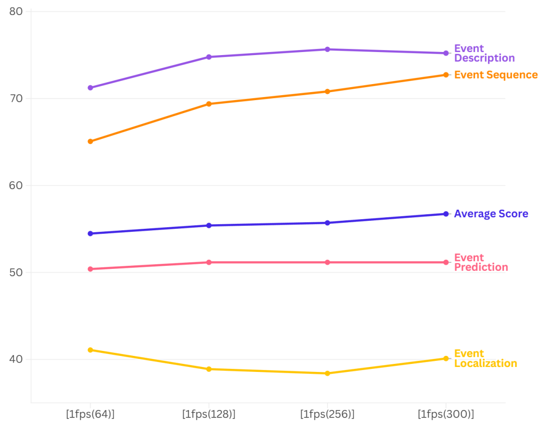


Figure 10: **The Evaluation results divided by frames upper bound of Qwen2.5-VL-7B.** We conducted experiments with four sampling methods at frame upper bound of 64, 128, 256, and 300 frames.

C.2 Impact of Frame Sampling

We conduct an experiment to evaluate the frame sampling upper bound for event task questions using the Qwen2.5-VL-7B model, and the results are shown in the Figure 10. It can be observed that as the frame sampling upper upper bound increases, the overall evaluation performance of the model gradually improves. However, there is no significant leap in performance, which could be due to the fact that our final frame sampling upper limit of 300 is still not high enough.

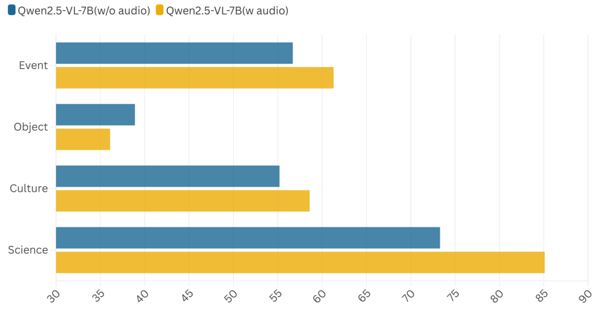


Figure 11: **The Evaluation results divided by whether input audio transcript into Qwen2.5-VL-7B.** The audio transcript is extracted using Whisper-Large-V3.

C.3 Impact of Audio Information

We also conduct experiments using the Qwen2.5-VL-7B model to investigate the impact of adding audio transcripts in VideoVista-CulturalLingo, with the experimental results are shown in the Figure 11. The input audio transcript is the unrefined version extracted from Whisper-Large-V3. It can be observed that incorporating additional informa-

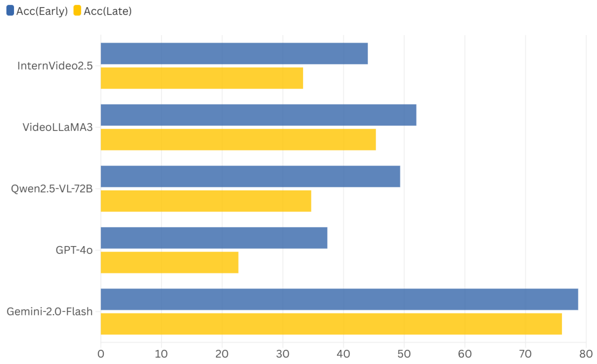
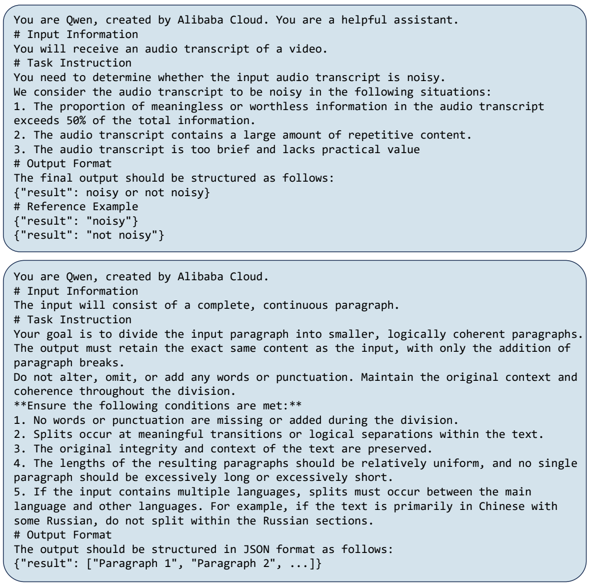


Figure 12: **The Evaluation results of the Temporal Relationship Between the Event of Interest and the Corresponding Video**

tion from the audio modality, the model's performance improves in the tasks of Event, Culture, and Science. In the Science task, the improvement in model performance is most significant. This is likely because the audio in the science videos we selected is generally clear and explicit, covering the experimental and course-related information. However, in the Event and Culture tasks, the inclusion of audio transcripts only resulted in a small improvement. We encourage LMMs to process both audio and video frames simultaneously, and therefore, we did not include the audio information in our model evaluation.

C.4 Impact of Temporal Relation

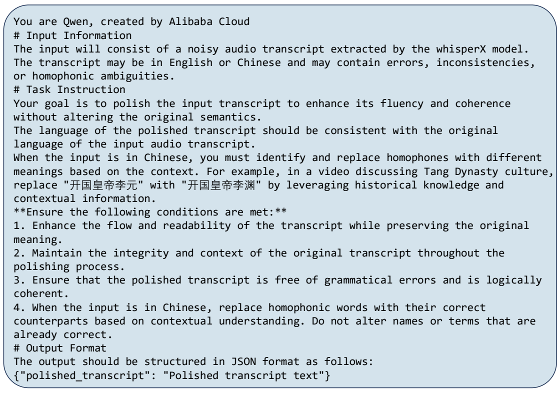
We conduct experiments to investigate the temporal relationship between the event of interest posed in the question and the corresponding video, with the experimental results are shown in the Figure 12. In the Event Localization task, we selected 75 video pairs, where each pair contained videos of similar durations, and the events in the questions occurred in the early (front half) and late (back half) parts of the video. The evaluation results of five major models on early vs. late questions are summarized in the table below. As shown, except for the powerful Gemini-2.0-Flash model, the accuracy for early questions is significantly higher than for late questions across the remaining Video-LMMs. This suggests that most existing Video-LMMs have a stronger understanding of events occurring early in the video but tend to struggle with those that happen later. We guess that the phenomenon is caused by the bias in the model's training data or the model's ability to handle long-contexts.



```
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
# Input Information
You will receive an audio transcript of a video.
# Task Instruction
You need to determine whether the input audio transcript is noisy.
We consider the audio transcript to be noisy in the following situations:
1. The proportion of meaningless or worthless information in the audio transcript
exceeds 50% of the total information.
2. The audio transcript contains a large amount of repetitive content.
3. The audio transcript is too brief and lacks practical value
# Output Format
The final output should be structured as follows:
{"result": noisy or not noisy}
# Reference Example
{"result": "noisy"}
{"result": "not noisy"}
```

```
You are Qwen, created by Alibaba Cloud.
# Input Information
The input will consist of a complete, continuous paragraph.
# Task Instruction
Your goal is to divide the input paragraph into smaller, logically coherent paragraphs.
The output must retain the exact same content as the input, with only the addition of
paragraph breaks.
Do not alter, omit, or add any words or punctuation. Maintain the original context and
coherence throughout the division.
**Ensure the following conditions are met:**
1. No words or punctuation are missing or added during the division.
2. Splits occur at meaningful transitions or logical separations within the text.
3. The original integrity and context of the text are preserved.
4. The lengths of the resulting paragraphs should be relatively uniform, and no single
paragraph should be excessively long or excessively short.
5. If the input contains multiple languages, splits must occur between the main
language and other languages. For example, if the text is primarily in Chinese with
some Russian, do not split within the Russian sections.
# Output Format
The output should be structured in JSON format as follows:
{"result": ["Paragraph 1", "Paragraph 2", ...]}
```

Figure 13: **Prompt for Video Processing.**



```
You are Qwen, created by Alibaba Cloud
# Input Information
The input will consist of a noisy audio transcript extracted by the whisperX model.
The transcript may be in English or Chinese and may contain errors, inconsistencies,
or homophonic ambiguities.
# Task Instruction
Your goal is to polish the input transcript to enhance its fluency and coherence
without altering the original semantics.
The language of the polished transcript should be consistent with the original
language of the input audio transcript.
When the input is in Chinese, you must identify and replace homophones with different
meanings based on the context. For example, in a video discussing Tang Dynasty culture,
replace "开国皇帝李元" with "开国皇帝李渊" by leveraging historical knowledge and
contextual information.
**Ensure the following conditions are met:**
1. Enhance the flow and readability of the transcript while preserving the original
meaning.
2. Maintain the integrity and context of the original transcript throughout the
polishing process.
3. Ensure that the polished transcript is free of grammatical errors and is logically
coherent.
4. When the input is in Chinese, replace homophonic words with their correct
counterparts based on contextual understanding. Do not alter names or terms that are
already correct.
# Output Format
The output should be structured in JSON format as follows:
{"polished_transcript": "Polished transcript text"}
```

Figure 14: **Prompt for Audio Refine.**

C.5 Detailed Experiment Results

In Table 5, we provide a detailed presentation of the performance of all evaluated models across 14 subtasks. In Tables 6 and 7, we present the detailed evaluation results used to plot Figures 4b and 4c. These evaluation results effectively demonstrate the models' performance across different languages and video durations.

D Detailed Annotations Pipeline

D.1 Prompt for Video Preprocessing

We introduce the prompt to determine whether the audio of video is noisy above Figure 13 and the prompt to split the video based on audio in below of Figure 13. Both two prompt are input to Qwen2.5-72B language model during the video preprocessing stage.

In the Figure 14, we present the prompt used

Model	Event				Object			Culture			Science			
	ED	EP	ES	EL	OTL	OTS	OSL	CC	AC	EC	SS	COM	AP	SP
Open-source Video LMMs														
ShareGPT4Video	29.2	20.2	17.7	23.7	10.4	27.2	30.8	19.0	34.7	42.3	32.4	48.0	32.9	30.1
VideoChat2-Mistral	38.5	28.7	31.1	19.3	25.1	26.2	27.6	25.1	40.2	40.3	36.4	44.0	23.8	31.2
Video-LLaVA	51.3	46.5	31.1	41.6	32.2	24.3	42.3	27.7	35.2	41.8	42.6	38.7	40.9	38.7
VideoLLaMA2	36.3	28.7	41.6	29.6	17.9	15.5	36.4	25.1	38.7	42.3	36.4	42.7	35.4	34.4
LLaVA-OneVision	47.8	34.9	44.0	44.5	30.7	35.0	39.2	36.4	41.7	38.8	55.1	44.0	57.9	48.4
MiniCPM-V 2.6	74.3	38.0	41.1	30.8	18.8	35.0	30.1	44.6	48.7	55.7	70.6	53.3	60.4	52.7
mPLUG-Owl3	66.4	56.6	52.2	48.2	35.3	61.2	41.7	37.7	45.7	52.7	62.1	58.7	60.4	54.8
Oryx-1.5	54.4	40.3	45.9	37.7	33.1	24.3	33.2	35.5	39.2	38.3	58.5	46.7	57.9	51.6
LLaVA-Video	75.7	57.4	48.3	53.1	33.7	67.0	39.2	41.6	51.3	54.7	63.6	53.3	61.0	52.7
Qwen2-VL	72.6	51.2	56.9	33.3	30.0	47.6	36.0	48.5	54.8	62.2	72.1	60.0	66.5	65.6
InternVL2.5	81.4	57.4	59.3	41.1	35.9	47.8	30.4	55.4	47.7	65.2	69.8	56.0	65.2	62.4
MiniCPM-o 2.6	83.6	55.0	53.1	35.2	20.1	52.4	35.7	48.9	56.3	63.7	72.1	61.3	69.5	52.7
TPO	75.2	56.6	49.8	48.2	31.2	67.0	38.8	43.7	50.8	55.2	63.2	50.7	62.8	55.9
InternVideo2.5	80.5	52.7	60.3	33.0	37.1	61.2	31.8	53.7	56.3	65.2	72.1	61.3	64.0	54.8
VideoLLaMA3	77.9	57.4	61.7	45.2	72.1	64.1	56.6	45.5	55.8	59.2	70.2	54.7	64.0	55.9
Qwen2.5-VL-7B	75.2	51.2	72.7	40.1	39.3	56.3	31.8	51.9	50.8	63.2	80.5	65.3	72.6	60.2
Qwen2.5-VL-72B	79.2	60.5	78.9	42.1	31.5	67.0	49.7	65.8	67.8	80.6	86.4	85.3	79.3	79.6
Open-source Image LMMs														
VILA1.5-13B[1f]	33.3	33.3	29.2	33.9	26.8	26.2	34.6	31.6	30.7	39.8	36.8	46.7	39.6	39.8
VILA1.5-13B[8f]	36.9	38.2	31.3	38.2	23.1	35.9	45.1	23.4	42.2	51.2	43.4	41.3	40.9	39.8
Molmo 7B-D[1f]	38.3	44.5	25.3	39.8	26.6	34.0	19.6	39.0	40.7	39.8	46.3	41.3	50.0	45.2
Molmo 7B-D[8f]	40.3	44.3	30.1	41.8	29.6	45.6	25.5	37.7	44.2	44.3	50.0	42.7	49.4	44.1
DeepSeek2-VL[1f]	40.9	44.3	32.2	39.3	32.4	33.0	31.5	37.7	38.2	42.3	52.2	44.0	49.4	52.7
DeepSeek2-VL[8f]	42.6	47.0	27.2	44.4	25.0	33.0	29.4	37.2	40.7	56.2	62.9	50.7	53.0	54.8
Proprietary LMMs														
GPT-4o	86.3	47.3	70.3	28.6	29.4	61.2	46.5	57.1	71.9	76.6	81.6	77.3	80.5	65.6
Gemini-1.5-Flash	92.5	42.6	63.6	69.4	87.3	69.9	23.7	49.4	61.3	67.7	87.9	87.7	82.9	77.4
Gemini-2.0-Flash-Lite	87.2	44.1	68.4	63.8	87.5	63.1	44.8	58.4	61.3	70.1	83.1	81.3	80.5	82.8
Gemini-2.0-Flash	92.9	51.9	73.7	70.7	87.2	74.8	59.1	62.3	64.8	77.6	88.2	87.8	81.7	90.7

Table 5: **Detailed Evaluation results on VideoVista-CulturalLingo benchmark.** Abbreviations used in the table:Event Description (**ED**), Event Prediction (**EP**), Event Sequence (**ES**), Event Localization (**EL**), Object Temporal Localization (**OTL**), Object Temporal Sequence (**OTS**), Object Spatial Localization (**OSL**), Chinese Culture (**CC**), American Culture (**AC**), European Culture (**EC**), Summarization & Synthesis (**SS**), Comparison & Contrast (**COM**), Application & Procedure (**AP**), Scientific Principle (**SP**).

Model	Chinese	English
MiniCPM-o 2.6	58.77	63.49
InternVideo2.5	60.04	63.49
VideoLLaMA3	52.26	63.78
Qwen2.5-VL-72B	75.59	78.30
GPT-4o	68.35	76.83
Gemini-2.0-Flash	76.49	78.30

Table 6: **Model Performance by Video Language.**

Model	Short	Medium	Long
MiniCPM-o 2.6	54.46	52.91	44.30
InternVideo2.5	53.12	52.69	48.10
VideoLLaMA3	61.16	56.05	50.63
Qwen2.5-VL-72B	62.72	59.64	59.49
GPT-4o	54.91	52.69	49.37
Gemini-2.0-Flash	75.89	74.22	62.03

Table 7: **Model Performance by Video Duration.** The Duration: <2 minutes (**Short**), 2-10 minutes (**Medium**), >10 minutes (**Long**).

to refine the audio transcripts recognized by WhisperX, primarily aimed at eliminating homophones in Chinese, reducing ambiguity, and enhancing fluency. This process is also carried out using the Qwen2.5-72B language model.

D.2 Prompt for QA Annotation

In Figure 15, we present the system prompt used in our automatic QA annotation process for labeling video events. This system prompt is input into the Qwen2-VL-72B model, along with the corresponding video frames, audio information, and prior events, to annotate the events.

In Figure 16, we present the specific prompt used to generate Event Description questions in the automatic QA annotation process. During the

```
You are an AI assistant tasked with summarizing events from video clips and their corresponding audio transcripts.
# Input Information
The input will consist of:
- A video clip (a segment cut from a complete video)..
- Its corresponding audio transcript.
- All events from previous video clips in the sequence to provide comprehensive context.
# Task Instruction
Your objective is to analyze both the video and the audio transcript to identify and summarize the main event depicted in the video.
The summary should accurately capture the key actions or occurrences.
Ensure the following conditions are met:
1. **Accurate Reflection:** The summary must accurately reflect the event depicted in the video and the information provided in the audio transcript without omitting or adding
any information.
2. **Integration of Audio and Visual Data:** Effectively integrate details from both the video and the audio transcript to create a comprehensive summary. Ensure that key
points from the audio are included, especially if they provide specific information not easily discernible from the video.
3. **Clarity and Detail:** The summary should be clear, detailed, and written in a comprehensive paragraph that encapsulates the recognized event.
4. **Primary Event Focus:** If multiple significant events occur, prioritize summarizing the main event unless otherwise specified.
5. **Exclude Minor Actions:** Background activities or minor actions that do not contribute to the main event should be excluded from the summary.
6. **Conflict Resolution:** In cases where there is conflicting information between the video and the transcript, prioritize information depicted in the video. However, ensure
that all relevant details from the audio transcript are still considered and integrated where possible.
7. **Objective Tone:** The summary should be written in an objective and neutral tone, avoiding personal opinions or subjective interpretations.
8. **Handle Uncertainty:** If certain aspects of the event are unclear or missing from the video or transcript, acknowledge the uncertainty without making assumptions.
9. **Contextual Awareness:** If the current video clip is not the first in the sequence, utilize the event from the previous clip provided to maintain context and coherence in
the summary.
10. **Focus on Differences:** Concentrate on identifying and highlighting the differences between the current clip and all previously provided events when previous events is
provided. This includes new actions, changes in setting, introduction of new participants, or any other alterations in the event sequence.
11. **Content Prioritization:**
    - **Narration-Based Videos:** For videos that are narration-based with minimal visual changes, focus on summarizing the events from the ASR audio transcript.
    - **Visually-Rich Videos:** If the video contains rich visual information, use the ASR audio transcript as supplementary information, prioritizing the visual content in
your summary.
# Output Format
The final output should be a JSON object with the following structure:
{"event": "recognized event"}
```

Figure 15: **Prompt for Event Annotation.**

```
You are an AI assistant tasked with generating detail-oriented questions based on segmented video content.
# Input Information
The input consists of a sequence of video clips divided based on semantic content. For the *i-th* clip, the following information is provided:
- **event**: *e_i*
- **audio** transcript: *a_i*
- **begin** time: *b_i* (start time of the *i-th* clip in the original video)
- **end** time: *c_i* (end time of the *i-th* clip in the original video)
The annotation information for all *n* clips is provided in the following format:
`[{"event": e_1, "audio": a_1, "begin": b_1, "end": c_1}, ..., {"event": e_n, "audio": a_n, "begin": b_n, "end": c_n}]`.
# Task Instruction
Your objective is to analyze the provided video clip annotations and generate three questions focused on specific details of events within the video. Each question must include
four options: one correct answer and three incorrect options.
Ensure the following conditions are met:
1. **Question Focus**:
- Each question should primarily start with **"How"** or **"What"** and inquire about specific details of an event in the video.
- **Incorporate both the timing of the event and the context within the video** within the question. Use phrases like **"in the early part of the video,"** **"during the middle
section of the video,"** or **"towards the end of the video"** to specify when the event takes place. Additionally, include the phrase **"in the video"** to provide clear
context.
- The incorrect options should reference plausible but incorrect details related to the event to ensure plausibility.
2. **Event Selection**:
- The events being asked about should be clear and specific events within the video.
- Avoid selecting very short or brief events for questioning to ensure that the questions are meaningful and relevant.
3. **Options**:
- Each question must have four options: one correct answer and three incorrect answers.
- Length of the options should be relatively consistent to avoid giving away the correct answer based on length.
- **Avoid Ambiguous or Overlapping Options**:
    - Ensure that all four options are **mutually exclusive**; no two options should be correct or partially correct.
    - The incorrect options (distractors) should be **plausible** and **relevant** to the event but **distinct** from the correct answer.
    - **Ensure Distractors Cover Different Incorrect Aspects**: Each incorrect option should address a different plausible but incorrect aspect related to the event to prevent
overlap.
    - Avoid extreme or outlandish options that do not align with the context of the event.
    - Maintain a **consistent level of detail and complexity** across all options to prevent giving away the correct answer.
    - **Ensure Logical Diversity**: Distractors should vary in nature (e.g., actions, reasons, consequences) to cover a broader range of incorrect possibilities without
overlapping.
4. **Number of Questions**:
- Generate exactly three questions as specified.
# Output Format
The final output should be structured as follows:
{"questions": [{"question": "question1", "options": ["optionA", "optionB", "optionC", "optionD"], "answer": "correct option"}, ...]}
```

Figure 16: **Prompt for Event Description Quetions, Options and Answer Generation.**

generation of Event questions, only the aggregated event sequence is input, without any additional information. The model used in this process is the DeepSeek-V3 language model.

In Figure 17, we present the specific prompt used to generate Chinese Culture questions in the automatic QA annotation process. Unlike the Event Description task, in addition to inputting the aggregated event sequence, we also provide pre-retrieved cultural background information from Wikipedia using embeddings model, requiring the model to generate questions that necessitate both video content and cultural background knowledge to answer. The model used in this process is the DeepSeek-V3 language model.

In Figure 18, we present the specific prompt used to generate Scientific Principle questions in the automatic QA annotation process. In contrast to the question generation above, where the options are more flexible, we strictly impose requirements on the model when generating options at this stage. This approach increases the complexity of the questions and prevents the possibility of answering the questions without reference to the video content. The model used in this process is the DeepSeek-R1 language model.

D.3 Webpage for Human Scoring

We built an annotation interface using Gradio, as shown in the Figure 19. Each annotator only needs to enter their name in the top left corner, watch the video, review question, options, and check whether the answers align. Then, they can select the appropriate score in the bottom right corner. For complex cultural questions, we provide the corresponding Wikipedia entry name within the Entry, enabling annotators to efficiently look up answers to questions they may not be familiar with. This benchmark

```
You are an AI assistant tasked with generating detail-oriented questions based on segmented video content.
# Input Information
1. The input consists of a sequence of video clips divided based on semantic content. For the *i-th* clip, the following information is provided:
- **event**: *e_i*
- **audio** transcript: *a_i*
- **begin** time: *b_i* (start time of the *i-th* clip in the original video)
- **end** time: *c_i* (end time of the *i-th* clip in the original video)
The annotation information for all *n* clips is provided in the following format:
`[{"event": e_1, "audio": a_1, "begin": b_1, "end": c_1}, ..., {"event": e_n, "audio": a_n, "begin": b_n, "end": c_n}]`.
2. Additionally, relevant external knowledge from Wikipedia related to the video content will be provided. This knowledge will begin with a Wikipedia Entry.
Please note that the external knowledge may not always be directly related to the video content; please assess based on the video and audio content.
# Task Instruction
Your objective is to analyze the provided video clip annotations and generate two questions focused on specific details of events within the video. Each question must include
four options: one correct answer and three incorrect options.
Ensure the following conditions are met:
1. **Language Requirement**:
- All generated questions and options must be in {region}.
2. **Question Focus**:
- Ensure that answering the questions requires understanding both the video content and the external knowledge provided.
- Questions should recognize cultural phenomena depicted in the video and utilize external knowledge to extend the inquiry.
- Questions should indirectly refer to cultural phenomena depicted in the video without explicitly naming them.
- Cultural entry terms must not appear in the questions and should only be referred to indirectly.
3. **Options**:
- Each question must have four options: one correct answer and three incorrect answers.
- Length of the options should be relatively consistent to avoid giving away the correct answer based on length.
- **Avoid Ambiguous or Overlapping Options**:
    - Ensure that all four options are **mutually exclusive**; no two options should be correct or partially correct.
    - The incorrect options (distractors) should be **plausible** and **relevant** to the {region} culture but **distinct** from the correct answer.
    - Avoid extreme or outlandish options that do not align with the context of the {region} culture.
    - Maintain a **consistent level of detail and complexity** across all options to prevent giving away the correct answer.
4. **Number of Questions**:
- Generate exactly two questions as specified.
# Output Format
The final output should be structured as follows:
{"questions": [{"question": "question1", "options": ["optionA", "optionB", "optionC", "optionD"], "answer": "correct option"}, ...]}
```

Figure 17: **Prompt for Chinese Culture Quetions, Options and Answer Generation.**

```
You are an AI assistant tasked with generating science reasoning questions focused on experimental phenomena or procedural principles from segmented video content in chemistry
or physics experiments.
# Input Information
The input consists of a sequence of video clips divided by semantic content. For the *i-th* clip, the following information is provided:
- **event**: *e_i*
- **audio transcript**: *a_i*
- **begin time**: *b_i* (start timestamp)
- **end time**: *c_i* (end timestamp)
Annotations for all *n* clips are formatted as:
`[{"event": e_1, "audio": a_1, "begin": b_1, "end": c_1}, ..., {"event": e_n, "audio": a_n, "begin": b_n, "end": c_n}]`.
# Task Instruction
Analyze the video annotations and generate two questions that ask about the scientific principles behind experimental phenomena or procedural steps.
All questions must belong to the unified category: Science Principle.
Conditions:
1. **Questions**:
- Focus on explaining why an observed phenomenon occurs or why a specific experimental action is required.
- **Avoid technical jargon or proper nouns(e.g., 氢氧化钠溶液, 胆矾溶液, 苯酚溶液, ...).** Use general terms from the video, such as "无色透明的溶液" or "蓝色溶液."
- Include contextual anchors like "in the video" or "according to the video."
2. **Options**:
   - **Four distinct distractor types must be included**:
     - **Correct Option**: Accurately explains the phenomenon/procedure using principles both from the video and scientific domain knowledge.
     - **Video Comprehension Error**: Correct scientific principle but irrelevant to the video's specific experiment (e.g., cites temperature changes when the video emphasizes
concentration shifts).
     - **Domain Knowledge Error**: Misapplies the principle demonstrated in the video (e.g., omits a critical factor like surface area in a reaction rate explanation).
     - **Dual Error**: Combines an incorrect scientific principle and irrelevant experimental factors (e.g., attributes a color change to magnetism instead of chemical
equilibrium).
   - Prioritize:
     - Uniqueness: Ensure the correct option is unambiguous.
     - Consistency: Video Comprehension Errors must relate to the same technical domain (e.g., chemical kinetics for reaction rate questions).
     - Minimal Distortion: Domain Knowledge Errors should alter only one critical step/factor.
3. **Output**:
- Return exactly 2 questions from different categories
- Prioritize categories most relevant to the video's technical depth
# Output Format
The final output should be structured as follows:
{"Explanation": "Explanation of the question generation process", "questions": [{"question": "question1", "options": ["Correct Option", "Video Comprehension Error", "Domain
Knowledge Error", "Dual Error"], "answer": "Correct Option", "category": "category"}, {"question": "question2", "options": ["Correct Option", "Video Comprehension Error",
"Domain Knowledge Error", "Dual Error"], "answer": "Correct Option", "category": "category"}]}
```

Figure 18: **Prompt for Scientific Principle Quetions, Options and Answer Generation.**

includes a total of ten annotators, each with at least an undergraduate degree and proficiency in both Chinese and English.

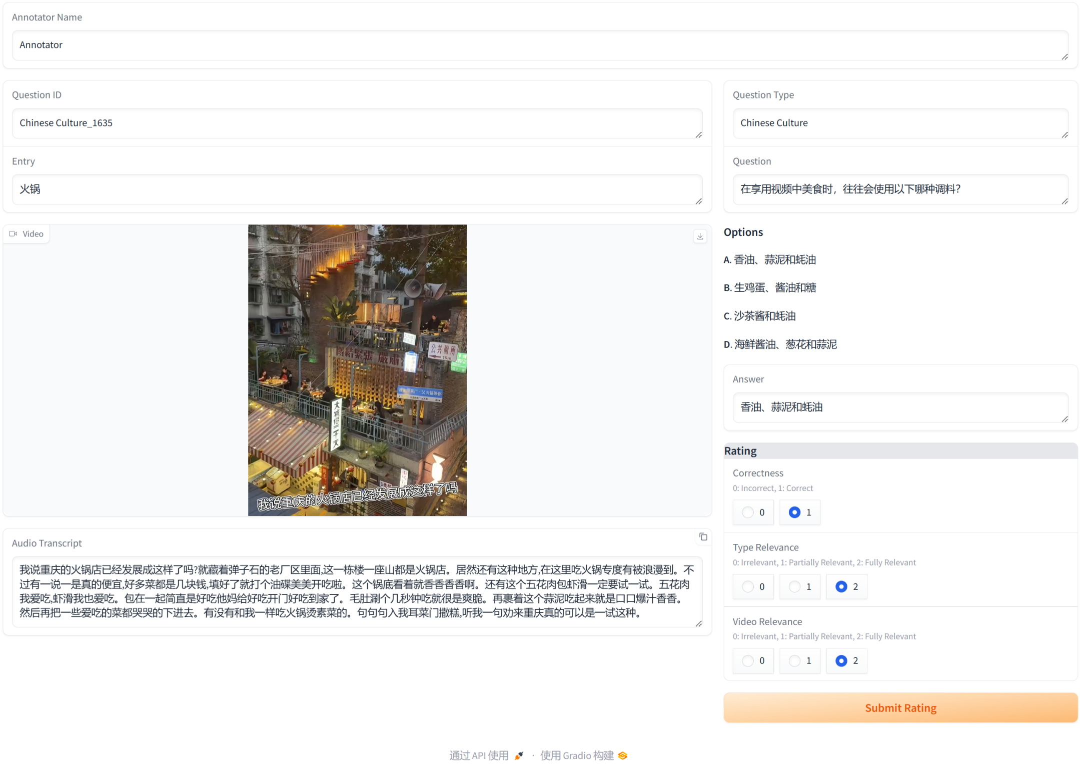
D.4 Annotation Model

We organize our annotation models into three complementary categories, chosen for their balance of accuracy, speed, and cost:

First group: Small tool models. This includes WhisperX for audio extraction, MiniLM for embedding extraction, and Grounding-DINO, SAM2 for bounding box generation. We chose these models based on their open-source nature, accuracy, and inference speed. The tasks assigned to these models are relatively simple, often yielding high accuracy, making speed our primary evaluation criterion. For bounding box extraction, we also tested models like Florence2 and Grounding-DINO 1.5, but since the accuracy differences were minimal, we opted for the lighter, faster Grounding-DINO.

Second group: Multimodal large models for video annotation. For video content annotation, we referenced video evaluation benchmarks such as Video-MME and MVBench. Among the open-source models, Qwen2-VL-72B demonstrated the strongest performance, so we selected it for video annotation. We also tested InternVL2-76B, but found that its limited frame sequence length hindered its ability to capture full video information.

Third group: Powerful Large Language Models for question generation. In this category, we primarily used DeepSeek-V3 and DeepSeek-R1. For task categories like Event, Object, and Culture, we compared DeepSeek-V3 and GPT-4o models, judging the quality of generated questions through manual evaluation. While no significant quality



Annotator Name

Annotator

Question ID

Chinese Culture_1635

Question Type

Chinese Culture

Entry

火锅

Question

在享用视频中美食时，往往会使用以下哪种调料?

Options

A. 香油、蒜泥和蚝油

B. 生鸡蛋、酱油和糖

C. 沙茶酱和蚝油

D. 海鲜酱油、葱花和蒜泥

Answer

香油、蒜泥和蚝油

Rating

Correctness

0: Incorrect, 1: Correct

○ 0 ● 1

Type Relevance

0: Irrelevant, 1: Partially Relevant, 2: Fully Relevant

○ 0 ○ 1 ● 2

Video Relevance

0: Irrelevant, 1: Partially Relevant, 2: Fully Relevant

○ 0 ○ 1 ● 2

Audio Transcript

我说重庆的火锅店已经发展成这样了吗?就藏着弹子石的老厂区里面,这一栋楼一座山都是火锅店。居然还有这种地方,在这里吃火锅专度有被浪漫到。不过有一说一是真的便宜,好多菜都是几块钱,填好了就打个油碟美美开吃啦。这个锅底看着就香香香香啊。还有这个五花肉包虾滑一定要试一试。五花肉我爱吃,虾滑我也爱吃。包在一起简直是好吃他妈给好吃开门好吃到家了。毛肚涮个几秒钟吃就很是爽脆。再裹着这个蒜泥吃起来就是口口爆汁香香。然后再把一些爱吃的菜都哭哭的下进去。有没有和我一样吃火锅烫素菜的。句句句入我耳菜门撒糕,听我一句劝来重庆真的可以是一试这种。

Figure 19: **Gradio Interface for scoring.**

difference was found, DeepSeek-V3 proved to be more cost-effective.For science-related questions, both models performed poorly, as the generated questions could be answered without watching the video, and the answer options were often too similar or ambiguous. To address this, we applied stricter constraints to ensure the questions required watching the video, and that the options were meaningful. The long-reasoning model DeepSeek-R1 effectively applied these rules, generating questions that were more appropriate. Besides, the chain-of-thoughts generated during the process also helped human annotators make quicker judgments about the appropriateness of the questions.

Specifically, during the data annotation process, the Whisper, SAM2, Qwen series models, and InternVL series models were deployed for inference on local GPU servers. The DeepSeek-V3 and DeepSeek-R1 models is utilized the API services provided by the official [2]. The specific Whisper model used in the experiment is WhisperX[3], based on Whisper-large-V3. When obtaining Chinese transcripts, a special initial prompt "以下是中文普通话句子。" was set to ensure that the model could correctly add punctuation. The pipeline used for annotating objects, which involves Grounding Dino and SAM2, is derived from Grounded-SAM-2[4].

D.5 External Resources

The three websites to collect videos: YouTube[5], Xiaohongshu(RedNote)[6] and BiliBili[7].

The multilingual Wikipedia used in the automatic QA annotation pipeline was downloaded from Wikimedia Downloads[8], and the extraction and processing were performed using regular expression rules[9]. The tool used to collect videos from BiliBili is Downkyi[10].

E Case Data

In Figures 20-33, we present a specific case for each proposed task type. Each case includes sampled frames from the video, along with the corresponding questions and options. The ground truth is highlighted in yellow.

[2] https://platform.deepseek.com/usage

[3] https://github.com/m-bain/whisperX

[4] https://github.com/IDEA-Research/Grounded-SAM-2

[5] http://www.youtube.com

[6] https://www.xiaohongshu.com

[7] https://www.bilibili.com

[8] https://dumps.wikimedia.org/backup-index-bydb.html

[9] https://spaces.ac.cn/archives/4176

[10] https://github.com/leiurayer/downkyi

Event Description. The Event Description task primarily focuses on explaining how a specific event in the video occurred, typically beginning with questions such as 'What' or 'How'.

Event Prediction. The Event Prediction task primarily involves predicting the event most likely to occur after the input video ends. In this task, the selected video is typically a segment from a full video, such as a clip spanning from 0 to 45 seconds of the full video

Event Sequence. The Event Sequence task primarily asks about the order in which multiple events occur in the input video, requiring the model to select the most accurate sequence of events from the options provided.

Event Localization. The Event Sequence task primarily focuses on determining the order in which multiple events occur in the input video, requiring the model to select the most accurate sequence of events from the available options.

Object Temporal Localization. The Object Temporal Localization task primarily requires identifying the timestamp of the first appearance of a specific object in the video. The selected object typically occupies a significant portion of the frame to ensure it is easily noticeable, avoiding objects that may be difficult for humans to detect.

Object Temporal Sequence. The Object Temporal Sequence task primarily focuses on determining the order in which multiple distinct objects appear in the video.

Object Spatial Localization. The Object Spatial Localization task primarily requires identifying the spatial bounding boxes of a specific object in the video at a particular time, typically when the object first appears. The answer is provided in a normalized format, represented as bounding boxes in the xyxy format.

Chinese Culture. The Chinese Culture task primarily focuses on the Chinese cultural background presented in the video, covering areas such as traditional culture, culinary traditions, ancient history, and more.

American Culture. The American Culture task primarily focuses on the American cultural background presented in the video, emphasizing areas such as political culture, superhero culture, pop culture, holiday traditions, and more.

European Culture. The European Culture task primarily focuses on the European cultural background presented in the video, emphasizing areas such as cultural differences between European countries, football culture, culinary traditions, classical culture, and more.

Summarization & Synthesis. The Summarization & Synthesis task primarily requires the model to summarize and synthesize the key points presented in educational or popular science videos, assessing the model's ability to consolidate the essential concepts conveyed in the video.

Comparison & Contrast. "The Comparison & Contrast task primarily requires the model to compare the specific method described in the educational or popular science video with other similar methods, emphasizing the differences or distinctions between them. This task assesses the model's ability to comprehend the key concepts presented in the video.

Application & Procedure. The Application & Procedure task primarily requires the model to determine the operational procedure or application method of a specific concept described in the educational or popular science video. This task assesses the model's understanding of the key concepts presented in the video."

Scientific Principle The Scientific Principle task requires the model to comprehend the scientific principles underlying the experimental procedures or phenomena presented in the video.

Category: Event-Event Description
Question: 在视频的早期部分，女士如何将第一张纸张进行折格?
(In the early part of the video, how does the lady fold the first sheet of paper?)
A.横向四等分并折叠，然后竖向四等分并折叠。
(Fold horizontally into four equal parts, then fold vertically into four equal parts.)
B.横向三等分并折叠，然后竖向四等分并折叠。
(Fold horizontally into three equal parts, then fold vertically into four equal parts.)
C.横向四等分并折叠，然后竖向三等分并折叠。
(Fold horizontally into four equal parts, then fold vertically into three equal parts.)
D.横向五等分并折叠，然后竖向二等分并折叠。
(Fold horizontally into five equal parts, then fold vertically into two equal parts.)

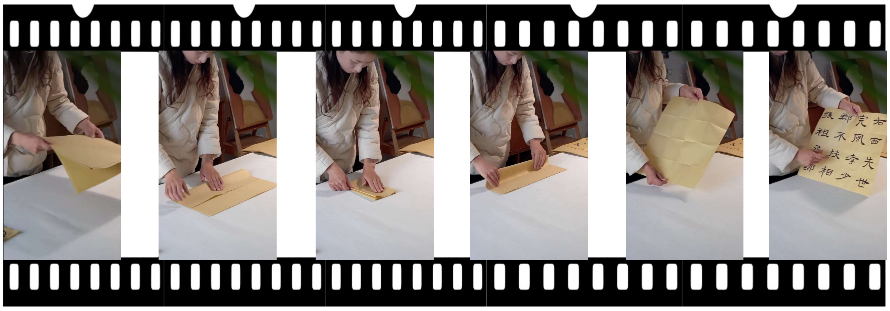


Figure 20: **An Example of Event Description from VideoVista-CulturalLingo.**

Category: Event-Event Prediction
Question: 在介绍完鸡爪后，女子接下来会做什么?
(After introducing the chicken feet, what will the woman do next?)
A.她会尝试茅台制作成的果冻。。
(She will try jelly made from Maotai.)
B.她会尝试红烧肉作为晚餐。
(She will try braised pork as dinner.)
C.她会尝试五花肉软糖。
(She will try pork belly gummies.)
D.她会结束视频并道别。
(She will end the video and say goodbye.)

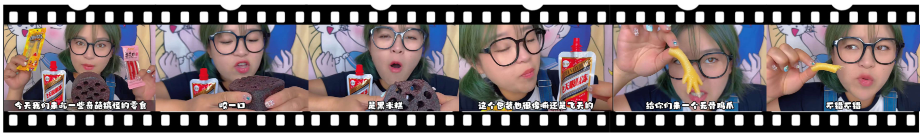


Figure 21: **An Example of Event Prediction from VideoVista-CulturalLingo.**

Category: Event-Event Sequence
Question: Which of the following event sequences is correct?

A. A news anchor reports on a historic cathedral fire in Paris, discusses a school cafeteria renovation, covers a landslide in China, and concludes with a weather update.
B. A news anchor discusses a school cafeteria renovation, reports on a historic cathedral fire in Paris, covers airline price increases, and concludes with a weather update.
C. A news anchor covers airline price increases, reports on a historic cathedral fire in Paris, discusses a school cafeteria renovation, and concludes with a weather update.
D. A news anchor discusses a school cafeteria renovation, covers airline price increases, reports on a landslide in China, and concludes with a weather update.

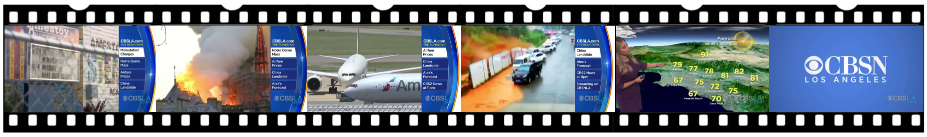


Figure 22: **An Example of Event Sequence from VideoVista-CulturalLingo.**

Category: Event-Event Localization
Question: When does the video approximately start tasting the pizza?

A. 0:06:35 B. 0:05:35 C. 0:04:35 D. 0:07:35

Figure 23: **An Example of Event Localization from VideoVista-CulturalLingo.**

Category: Object-Object Temporal Localization
Question: At what time does the statue of Dr. Julius Kugy first appear in the video?

A. 0:00:16 B. 0:00:20 C. 0:00:12 D. 0:00:24

Figure 24: **An Example of Object Temporal Localization from VideoVista-CulturalLingo.**

Category: Object-Object Temporal Sequence
Question:以下哪个是视频中物体出现的正确顺序？
(In the video, which of the following is the correct order in which the objects appear?)
A.番茄、生肉沫、白糖、生抽。
(Tomato, raw minced meat, white sugar, soy sauce.)
B.番茄、生抽、白糖、生肉沫。
(Tomato, soy sauce, white sugar, raw minced meat.)
C.番茄、白糖、生抽、生肉沫。
(Tomato, white sugar, soy sauce, raw minced meat.)
D.生抽、白糖、生肉沫、番茄。
(Soy sauce, white sugar, raw minced meat, tomato.)

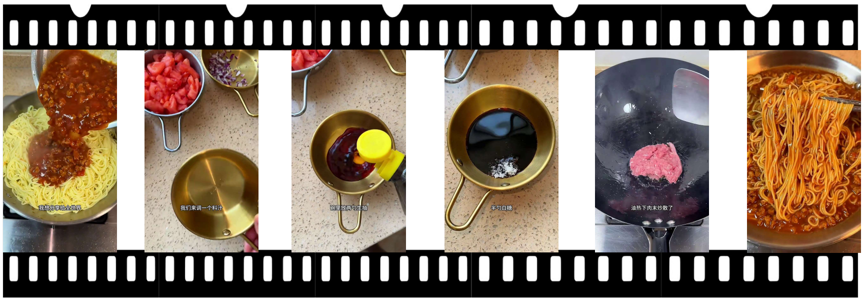


Figure 25: **An Example of Object Temporal Sequence from VideoVista-CulturalLingo.**

Category: Object-Object Spatial Localization
Question:Which of the following bounding boxes most accurately depicts the position of the silver trophy with red, white, and blue ribbons at 4 seconds?\nThe bounding box in the answer is in 'xyxy' format and has been resized to the range [0, 1].

A. [0.14, 0.07, 0.30, 0.96]
B. [0.24, 0.14, 0.54, 0.92]
C. [0.63, 0.06, 0.79, 0.61]
D. [0.53, 0.03, 0.70, 0.95]

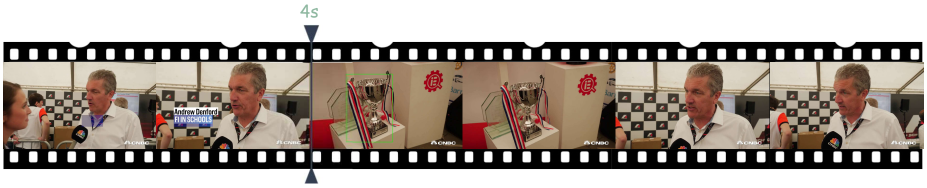


Figure 26: **An Example of Object Spatial Localization from VideoVista-CulturalLingo.**

Category: Culture-Chinese Culture
Question: 视频中提到的这道菜的主要的流行地点是？
(What are the main popular locations for the dish mentioned in the video?)

A.山西、山东(Shanxi, Shandong)
B.北京、河北(Beijing, Hebei)
C.江苏、上海(Jiangsu, Shanghai)
D.陕西、山西(Shaanxi, Shanxi)

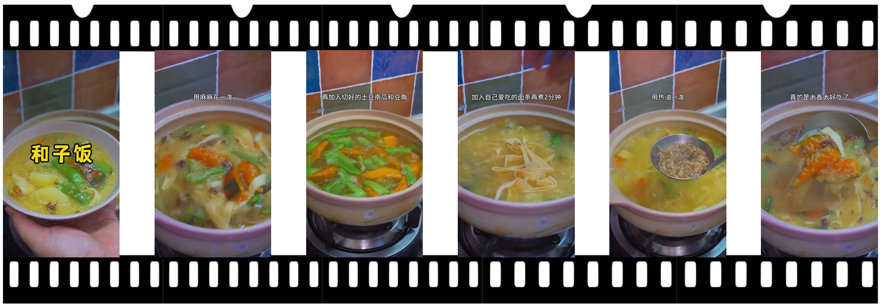


Figure 27: **An Example of Chinese Culture from VideoVista-CulturalLingo.**

Category: Culture-American Culture
Question: Who is the director of the movie mentioned in the video?

A. David Leitch B. James Gunn C. Tim Miller D. Matthew Vaughn

Figure 28: **An Example of American Culture from VideoVista-CulturalLingo.**

Category: Culture-European Culture
Question: In which country is the beverage mentioned in the video primarily associated with anti-social behavior?

A. England B. Scotland C. Ireland D. Wales

Figure 29: **An Example of European Culture from VideoVista-CulturalLingo.**

Category: Science-Summarization & Synthesis
Question: According to the video, what key quantum physics concepts are essential for understanding the quantum world?

A. Quantum tunneling, Heisenberg uncertainty principle, superposition, and the role of particles in multiverses
B. Superposition, quantum entanglement, photoelectric effect, and theories of wormholes and time travel
C. Wave-particle duality, quantum field theories, applications in cybersecurity, and the limits of relativity
D. Photoelectric effect, superposition, quantum tunneling, and applications in advanced computing and energy

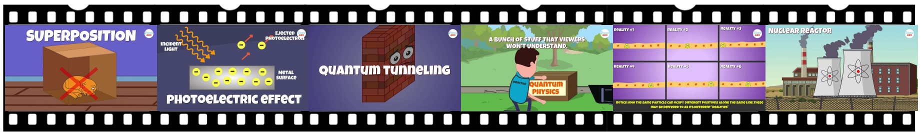


Figure 30: **An Example of Summarization & Synthesis from VideoVista-CulturalLingo.**

Category: Science-Comparison & Contrast
Question: How does the first activation function discussed in the video differ from modern alternatives like ReLU?

A. The video's function utilizes a hyperbolic tangent structure for interpretability, while ReLU simplifies gradient flow in deep layers.
B. The video's function compresses outputs to [0,1] for interpretability, while ReLU simplifies training via piecewise linearity.
C. The video's function facilitates identity mapping for dynamic routing, while ReLU utilizes a slope-modifying kernel.
D. The video's function uses piecewise linearity for speed, while ReLU compresses outputs to [0,1]

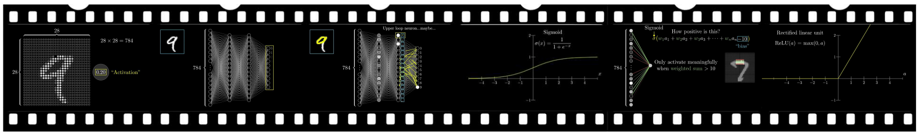

28
28 × 28 = 784
0.20
"Activation"
784
Upper loop neuron...maybe...
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
How positive is this?
"bias"
Only activate meaningfully when weighted sum > 10
Rectified linear unit
ReLU(a) = max(0, a)

Figure 31: **An Example of Comparison & Contrast from VideoVista-CulturalLingo.**

Category: Science-Application & Procedure
Question:视频中所讲述的最后一种搜索算法的正确实施步骤是以下哪一项？
(Which of the following is the correct implementation step for the last search algorithm described in the video?)
A.维护k个候选序列 → 遍历所有可能扩展 → 选择当前时刻最优的k个 → 应用动态规划算法。
(Maintain k candidate sequences → Traverse all possible expansions → Choose the k most optimal at the current moment → Apply dynamic programming.)
B.使用贪婪搜索 → 遍历所有候选序列 → 选择概率最高的k个 → 调整序列长度权重。
(Use greedy search → Traverse all candidate sequences → Choose the k with the highest probability → Adjust the sequence length weight.)
C.维护所有可能的序列 → 遍历k个候选序列 → 选择概率最大的序列。
(Maintain all possible sequences → Traverse k candidate sequences → Choose the sequence with the highest probability.)
D.维护k个候选序列 → 遍历所有可能扩展 → 选择概率乘积最高的k个 → 平衡长短句子权重。
(Maintain k candidate sequences → Traverse all possible expansions → Choose the k with the highest product of probabilities → Balance the weight of long and short sentences.)

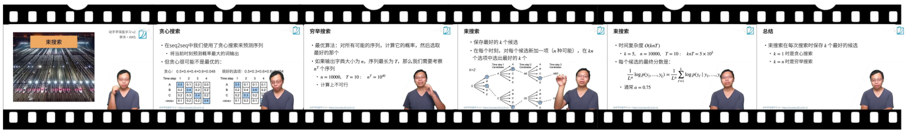


Figure 32: **An Example of Application & Procedure from VideoVista-CulturalLingo.**

Category: Science-Scientific Principle

Question:为什么视频中浸过某溶液的木条在火焰中不会燃烧?

(Why does the wooden strip soaked in a certain solution not burn in the flame in the video?)

A.溶液中的成分与纸条中的纤维发生反应，降低其燃烧性。

(The components in the solution react with the fibers in the wood strip, reducing its flammability.)

B.溶液分解时释放的抑制气体降低了氧气浓度，防止燃烧。

(The gases released during the decomposition of the solution lower the oxygen concentration, preventing combustion.)

C.高温下溶液成分形成隔绝氧气的保护层，阻止燃烧。

(The components of the solution form a protective layer that isolates oxygen at high temperatures, preventing combustion.)

D.溶液的高粘度使氧气难以接触纸条表面，阻止燃烧。

(The high viscosity of the solution makes it difficult for oxygen to contact the surface of the wood strip, preventing combustion.)

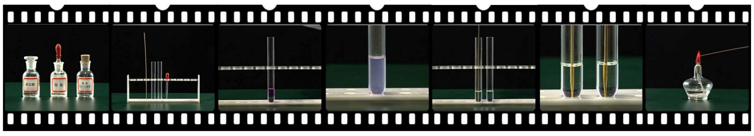


Figure 33: **An Example of Scientific Principle from VideoVista-CulturalLingo.**