This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

HUMANVBENCH: Probing Human-Centric Video Understanding in MLLMs with Automatically Synthesized Benchmarks

Ting Zhou[1], Daoyuan Chen[2], Qirui Jiao[1], Bolin Ding[2], Yaliang Li[2], Ying Shen[1,3,4*]

[1]Sun Yat-Sen University, [2]Alibaba Group, [3] Peng Cheng Laboratory,
[4] Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology
{zhout88,jiaoqr3}@mail2.sysu.edu.cn, sheny76@mail.sysu.edu.cn
{daoyuanchen.cdy,bolin.ding,yaliang.li}@alibaba-inc.com

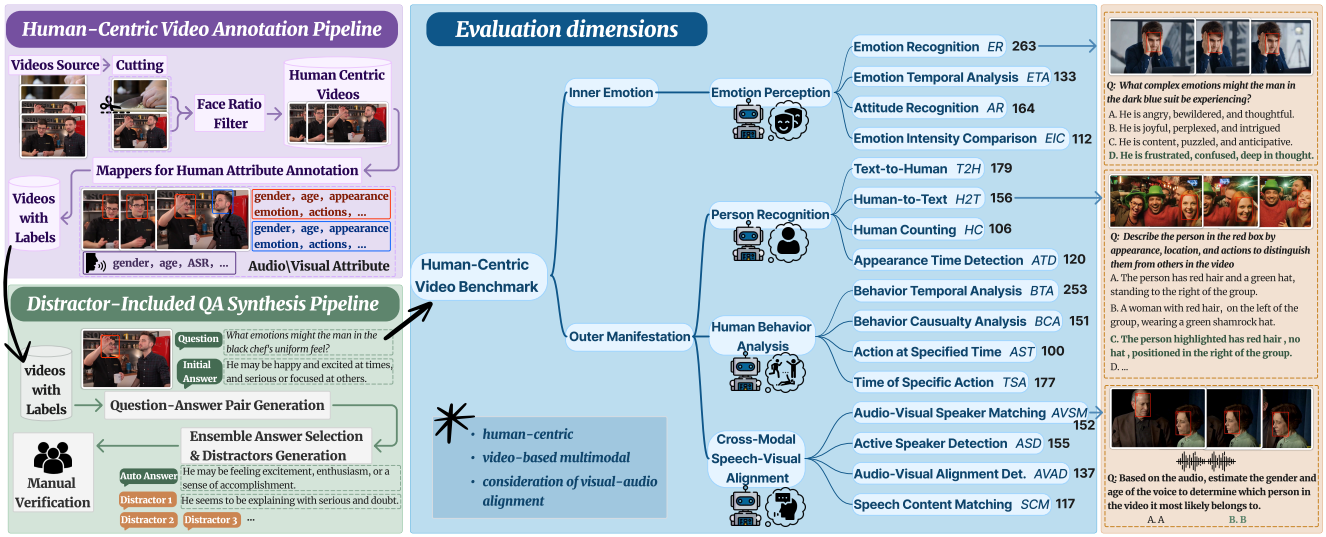


Figure 1. HUMANVBENCH is a benchmark for human-centric video understanding that features 16 fine-grained tasks, each denoted by its acronym and the number of included QA instances (middle blue box). It is synthesized with minimal human intervention by using a Video Annotation Pipeline (upper left, purple box) and a Distractor-Included QA Synthesis Pipeline (lower left, green box).

Abstract

Evaluating the nuanced human-centric video understanding capabilities of Multimodal Large Language Models (MLLMs) remains a great challenge, as existing benchmarks often overlook the intricacies of emotion, behavior, and cross-modal alignment. We introduce HUMANVBENCH*, a comprehensive video benchmark designed to rigorously probe these capabilities across 16 fine-grained tasks. A cornerstone of our work is a novel and scalable benchmark construction methodology, featuring two automated pipelines that synthesize high-quality video annotations and challenging multiple-choice questions with minimal human labor. By leveraging state-of-the-art models for annotation and systematically converting model-induced errors into plausible distractors, our framework provides a generalizable "machine" for creating nuanced evaluation suites. Our extensive evaluation of 30 leading MLLMs on* HUMANVBENCH *reveals critical deficiencies, particularly in perceiving subtle emotions and aligning speech with visual cues, with even top proprietary models falling short of human performance. We open-source* HUMANVBENCH *and our synthesis pipelines to catalyze the development of more socially intelligent and capable video MLLMs.*

1. Introduction

In recent years, Multimodal Large Language Models have emerged as pivotal advancements, extending traditional language models to process diverse modalities such as text,

*Corresponding author.

images, and videos [22–24, 26, 27, 40]. Video-oriented MLLMs, in particular, have drawn growing interest for their potential to interpret video content in ways closely aligned with human perception [46]. However, the extent to which these models truly achieve human-like understanding, especially in complex human-centric scenarios, remains an open question.

Human-centric scenes in videos are rich in emotional, behavioral, and interactive content, requiring robust foundational human perceptual skills from video understanding models. Existing MLLM benchmarks mainly target general content comprehension, but often overlooking the fine-grained and multi-dimensional evaluation of essential human perceptual abilities, like nuanced understanding of emotions, behaviors, and person recognition. Synchronizing speech with visual cues, a key aspect of human perception, is also frequently neglected. Unlike humans who easily detect audio-visual mismatches, models struggle with speaker identification and lip-sync alignment. These foundational skills underpin advanced human-related reasoning tasks, including narrative reasoning, intention inference, and social intelligence. Therefore, a foundational human-centric video understanding benchmark is necessary, which requires not only new evaluation dataset but also scalable and systematic frameworks for its creation.

To address this, we introduce a novel and scalable framework for automated video benchmark creation. This framework is powered by two core pipelines: 1) the Human-Centric Video Annotation Pipeline, which leverages over twenty state-of-the-art (SOTA) data processing operators to produce dense, multimodal, automated annotations; and 2) the Distractor-Included QA Synthesis Pipeline, which generates high-quality multiple-choice questions with semantically plausible distractors. As a powerful instantiation of our framework, we present HumanVBench, a pioneering benchmark featuring 16 fine-grained QA tasks specifically tailored for human-centric analysis in video MLLMs. These tasks are organized around two core dimensions: inner emotion and outer manifestation, as depicted in Figure 1.

A key innovation of our framework lies in its model-driven design. Unlike conventional benchmarks that rely on laborious human annotation, our approach automates the multi-modal, fine-grained annotation process by integrating diverse task-specific algorithms and models. The Distractor-Included QA Synthesis Pipeline further ensures the challenging nature of the questions by leveraging multi-model ensembles and their common failure modes to generate semantically deceptive distractors. The strategy transforms the human role from manual creation to efficient quality review, significantly boosting both construction efficiency and scalability. Moreover, our method is easily adaptable to "in-the-wild" video data, enabling benchmarks creation beyond controlled or domain-specific settings.

Through HUMANVBENCH, we comprehensively evaluate 30 state-of-the-art video MLLMs, covering both open-source and commercial models. Our results reveal notable gaps between current model capabilities and human-level video understanding, particularly in nuanced emotion perception. While the proprietary models demonstrate closer human-like accuracy, open-source models often misclassify emotions due to temporal noise, underscoring the need for further architectural improvements and refined datasets.

In summary, our contributions are as follows:

• We introduce HUMANVBENCH, a novel video benchmark for MLLMs that emphasizes fine-grained human comprehension in videos, focusing on emotion perception, person recognition, behavioral analysis, and speech-visual alignment.

• We propose a novel and scalable framework for automated video benchmark construction, featuring two advanced pipelines for multi-modal annotation and distractor-aware QA synthesis. Our framework is generalizable and significantly reduces manual annotation effort.

• Our comprehensive evaluation of 30 SOTA video MLLMs offers key insights, facilitating in-depth discussions regarding their performance, strengths, and areas for enhancement.

• We release our benchmark, including data, evaluation, and synthesis codes at *https://github.com/datajuicer/data-juicer/tree/HumanVBench* , to foster further evolution of future human-centric video analysis systems.

2. Related Works

Video Benchmarks for MLLMs. Recent years have witnessed the rapid development of benchmarks for evaluating MLLMs, particularly in the video domain. Representative efforts [17, 33, 60, 65] such as Video-MME primarily target general video understanding abilities like temporal reasoning and broad scene comprehension. Although these include some human-centric content, their evaluation is often dispersed and lacks the structured, fine-grained focus necessary for deep diagnostics of human-centric perception.

To better assess human understanding, benchmarks often focus on observable activities like action recognition, pose estimation, or motion analysis [21, 45, 52], but typically neglect affective states, person recognition, or multimodal synchronization. In video emotion understanding, emotion-centric datasets [4, 39, 42] like VEATIC mostly rely on discrete emotion classification with fixed emotion categories, lacking multi-dimensional tasks such as emotion dynamics or cross-person intensity comparison. Broader MLLM benchmarks with emotion tasks often lack clear granularity and rigorous evaluation protocols. At a higher cognitive level, benchmarks [29, 30, 56] like Social-IQ target narrative reasoning, intention inference, and social intelligence, intertwining perception with higher-order reasoning

and thus not isolating foundational perceptual competence.

In contrast, HumanVBench systematically targets the foundational perceptual layer of human-centric video understanding. It offers a comprehensive suite of 16 fine-grained tasks spanning person recognition, behavior and emotion perception, and audio-visual alignment, covering around 2.5k carefully synthesized multi-choice questions. This structured and multimodal evaluation serves as a prerequisite for higher-level cognitive reasoning. Quantitative comparisons with existing benchmarks are in Appendix B.

Benchmark Construction and QA Synthesis. Methods for constructing QA pairs in existing video benchmarks fall into three categories: Some use fully manual annotation, with human-written questions and answers [17, 51, 60]. Others repurpose existing datasets, as in ActivityNet-QA [54] built from ActivityNet [11], or HumanBench [45], which aggregates multiple public datasets for tasks like pose estimation and attribute recognition. Some benchmarks also incorporate automated QA generation, either by using commercial LLMs to produce QA pairs directly [34], or by synthesizing reasoning and temporal questions from existing annotations like captions, temporal labels, or scene graphs [33, 44, 55]. However, these approaches typically rely on extensive manual or historical annotations and struggle to scale to in-the-wild video data from scratch. Moreover, the automated generation of challenging distractors remains underexplored. Existing methods—often based on general semantic similarity [16, 28, 41, 43] or a single fine-tuned LLM for distractor generation [5, 38], lack mechanisms to systematically capture models' actual inference weaknesses in the complex video domain.

To address these limitations, our method introduces a scalable, model-driven framework that constructs benchmarks directly from raw videos. This framework leverages a multi-operator approach for dense, fine-grained annotation and employs an MLLM ensemble to synthesize challenging and discriminative evaluations.

3. The Proposed HUMANVBENCH

3.1. Task Design and Definition

Human observers naturally focus on individuals in videos, attending to their appearance, emotions, and behaviors. This intrinsic focus underpins our design of 16 fine-grained, human-centric tasks, aimed at evaluating MLLMs' ability to mimic human-like perception in video analysis (Figure 1). Each task definitions and examples are provided in Appendix J. The tasks can be grouped into two categories based on content observability: *Inner Emotion* and *Outer Manifestation*.

Inner emotions are often less observable in real-world videos, requiring abstract skills to interpret subtle facial expressions, body language, and verbal cues. The tasks in this category, which we term **Emotion Perception**, evaluate MLLMs' ability to detect and interpret these emotional nuances. Specifically, this category includes: *Emotion Recognition* (ER) identifies the most fitting emotional description for an individual. *Emotion Temporal Analysis* (ETA) tracks emotional changes of a person over time. *Attitude Recognition* (AT) assesses the inferred attitude (positive, negative, or neutral) of individuals toward specific events or entities. *Emotion Intensity Comparison* (EIC) differentiates and quantifies the emotional intensity of various individuals.

In contrast to inner emotions, outer manifestations deal with more tangible aspects such as identifying individual(s), causality reasoning, and synchronization of video elements like speech or singing. Guided by these aspects of human observation, we formulated three task categories:

Person Recognition evaluates the model's ability to identify individuals in complex scenes, similar to "person-finding": *Text-to-Human* (T2H) identifies a person based on a textual description. *Human-to-Text* (H2T) assesses the accuracy of a textual explanation attributed to a target person. *Human Counting* (HC) detects, tracks, and counts distinct individuals. *Appearance Time Detection* (ATD) identifies the specific time a specified individual appears.

Behavior Analysis scrutinizes a model's ability to understand and analyze individual behaviors, with four tasks: *Behavior Temporal Analysis* (BTA) tracks changes in a specified individual's behaviors over time. *Behavior Causality Analysis* (BCA) examines causal reasoning of behaviors in the video context. *Action at Specified Time* (AST) identifies the precise actions occurring at a given time. *Time of Specific Action* (TSA) detects the exact moments when specific actions occur.

Speech-Visual Alignment is a critical cross-modal category designed to test capabilities like speaker identification and lip-syncing. This category includes: *Audio-Visual Speaker Matching* (AVSM) correlates audio features to identify individuals and their visual appearance (gender, age). *Active Speaker Detection* (ASD) identifies the individual currently speaking by integrating visual and audio inputs. *Audio-Visual Alignment Detection* (AVAD) detects synchronization points and coherence between lip movements and audio. *Speech Content Matching* (SCM) analyses spoken content against text, evaluating transcription or lip-reading capabilities.

Together, these tasks offer a comprehensive assessment of MLLMs' emotion, identity, behavior, and speech-visual alignment understanding, driving towards human-like video comprehension.

3.2. Human-Centric Video Annotation Pipeline

Creating task-specific questions for the aforementioned tasks hinges on extensive human-centric annotations within

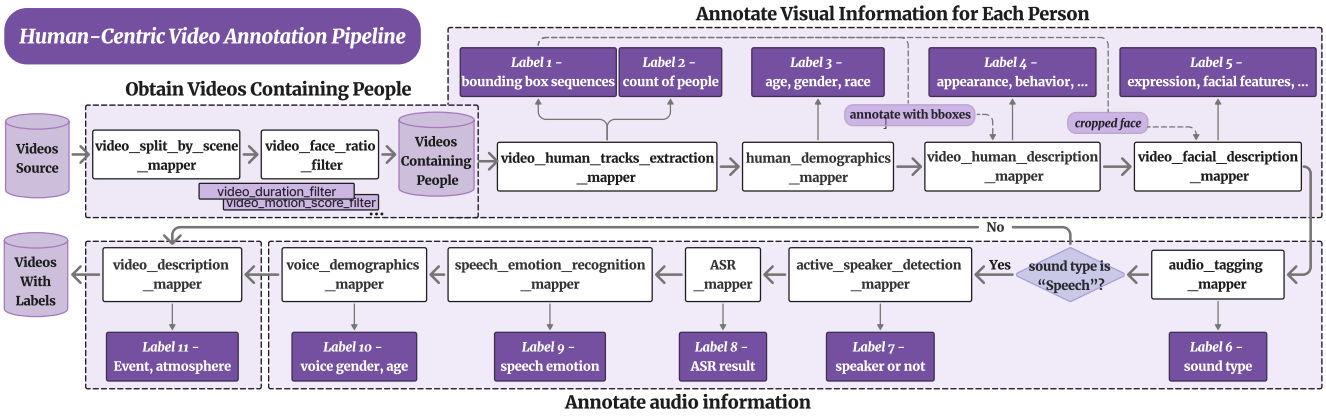


Figure 2. The Human-Centric Video Annotation Pipeline involves obtaining videos featuring people and annotating both visual and auditory information as well as overall event atmospheres.

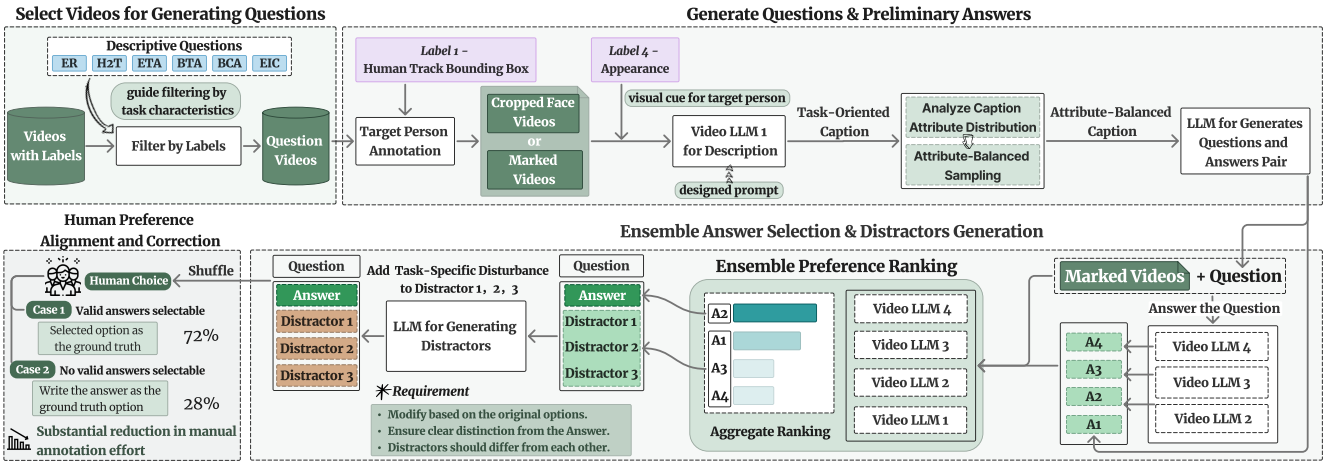


Figure 3. The Distractor-Included QA Synthesis Pipeline facilitates four steps: selecting "question videos", generating Questions, ensemble-based options generation with model-derived errors as distractors, and manually verifying multiple-choice questions.

videos. Our video annotation pipeline, illustrated in Figure 2, emphasizes multi-modal, granular annotation pathways, supported by the Data-Juicer framework [6, 8]. While benefiting from existing operators, we've also developed innovative operators to enhance the open-source community. Detailed annotations and examples are in the Appendix K.

Collecting Videos Containing People. We primarily sourced copyright-free videos from Pexels, which offers diverse in-the-wild footage under varied scenes and capture conditions. Movies from the MF2 benchmark [57], licensed under Public Domain 1.0, were also included to further increase content diversity. Importantly, we used only the raw visual data without any pre-existing captions or metadata. All annotations were generated internally by our pipelines, thereby preventing risks of data leakage. We begin by splitting each video into scenes with `video_split_by_scene_mapper` to facilitate subsequent accurate human tracking. The resulting clips are then filtered by attributes such as duration and optical flow, after which `video_face_ratio_filter` is applied to ensure that faces are visible in most frames. This guarantees that the selected videos are primarily human-centric and ready for further annotation.

Human-Centric Video Annotation Pipeline. We design a set of operators to extract and enrich human-centered video information. First, `video_human_tracks_extraction_mapper` links detected faces and bodies across consecutive frames using overlap thresholds, producing reliable person tracks and an approximate count of individuals in a shot. These tracks support subsequent tasks such as person highlighting and appearance/action descriptions.

From the tracks, `human_demographics_mapper` extracts face crops and applies attribute models to infer demographic labels (e.g., age, gender, race). Flexible bounding box crops (full-body or face-only) further allow generating individual-focused clips. Based on this, `video_human_description_mapper` cap-

tures appearance and posture using an MLLM, while `video_facial_description_mapper` focuses on facial expressions and changes. This ensures descriptions are unaffected by background or other people, while preserving facial detail.

For audio, `audio_tagging_mapper` first classifies sound types. If speech is present, additional operators are invoked: `active_speaker_detection_mapper` fuses audio-visual cues to localize speakers, `asr_mapper` transcribes speech, `speech_emotion_recognition_mapper` detects emotions, and `voice_demographics_mapper` profiles vocal attributes (e.g., gender, age).

Finally, `video_description_mapper` summarizes overall atmosphere and narrative using MLLMs. These multimodal annotations enable automated task construction. For example, in the *Audio-Visual Speaker Matching* task, videos are selected where exactly one individual's visual demographics (*Label-3*) match the audio demographics (*Label-10*), while other individuals' features mismatch. The matched individual serves as ground truth, with IDs assigned via *Label-1* tracking. Implementation details for each operator and full task construction appear in Appendices K and L.

3.3. Distractor-Included QA Generation Pipeline

For tasks with exact answers (e.g., restricted categories, numbers, or letters), questions, correct answers, and distractors are constructed using tailored templates and annotations from Section 3.2. For open-ended questions, we developed a pipeline to generate questions, answers, and distractors, applied to six tasks in HUMANVBENCH: *Emotion Recognition, Emotion Temporal Analysis, Emotion Intensity Comparison, Human-to-Text, Behavior Temporal Analysis*, and *Behavior Causality Analysis*. The pipeline (Figure 3) consists of four stages:

Selecting Videos. Videos are filtered by task criteria. For example, *Behavior Temporal Analysis* requires longer clips, while *Human-to-Text* requires at least two people.

Question and Preliminary Answer Synthesis. Firstly, target persons are localized using bounding box tracks. For facial emotion-focused tasks, we reconstruct the video with cropped face regions; for others, the target person is highlighted with a red bounding box to create "marked videos". These marked videos are fed into a Video-MLLM with prompt designed for the specific task, so that the model produces targeted descriptions such as emotional states or behavior summaries. Since Video-MLLMs may not always focus on the highlighted individual, we incorporate a visual cue from *Label-4* (Figure 2) into the prompt. The generated descriptions are then analyzed to extract task-relevant attributes, and their distributions are adjusted to ensure balance (e.g., maintaining similar proportions of positive and negative emotions). Based on these captions, an LLM (GPT-4 in our workflow) is used to generate questions and preliminary answers.

Ensemble Answer Selection and Distractors Generation. To mitigate errors from single-model answers, we adopt an ensemble strategy. Multiple MLLMs (Gemini [47], VideoLLaMA3 [58], ShareGPT4Video [10]) produce candidate answers, which are ranked via preference voting. The top choice is taken as the correct answer, while the others are converted into distractors: if semantically distinct (valid answer component), they are retained; otherwise an LLM (GPT-4 in our workflow) introduces task-specific perturbations. Retaining these erroneous answers allows the distractors to reflect typical model mistakes, thereby ensuring plausibility and distractive power rather than being arbitrarily fabricated, which increases the overall difficulty of the questions.

Manual Verification and Correction Given the annotation models' limitations, systematic biases and errors may occur in the generated multiple-choice questions. To address this concern and ensure the benchmark's reliability, we incorporate a human verification stage: all candidate options are shuffled, and annotators are instructed to select the correct one; if none of the options is fully accurate, they directly write the correct answer, which is then incorporated as the ground truth.

The Distractor-Included QA Synthesis Pipeline handles large-scale question and distractor generation, while humans serve as final arbiters of correctness. This substantially reduces manual workload—annotators only needed to rewrite the correct answer in about 25% of cases, while in the remaining 75% they confirmed one of the generated options (see Section 4.4). To further ensure reliability, we conducted an inter-annotator agreement (IAA) study on 240 randomly sampled questions from the final verified questions, where two independent annotators achieved a Cohen's Kappa of 0.8833, demonstrating the reliability of our benchmark. Additionally, we also provide distractor quality evaluation in Appendix E.

3.4. Post-Processing

To address the prevalent issue of answer leakage in multimodal evaluation datasets, we adopt the approach proposed in [9]. Specifically, we test models without visual input and remove frequently correct QAs (around 6%), ensuring random accuracy. This preserves visual relevance while mitigating leakage risks. In total, HUMANVBENCH contains 2475 problem instances; per-task counts are shown in Figure 1, with additional statistics in Appendix C.

Table 1. Summary of performance on HUMANVBENCH. We report average accuracy (%) across four main categories. Full results for all 30 models and 16 tasks are in Appendix D. The best overall results are shown in bold, while the best open-source model results are underlined "-" means the model recognizes its lack of required capabilities for the task and thus refuses to answer.

Model	Mod.	**Emotion**	**Person Recog.**	**Behavior**	**Avg.**	**Speech-Visual Alignment**					**Overall Avg.**
		4 tasks	4 tasks	4 tasks	12 tasks	AVSM	ASD	AVAD	SCM	4 tasks	16 tasks
Random Guess		24.4	25.2	22.9	24.2	42.8	23.6	33.3	25.0	31.2	25.9
Qwen-VL3 (7B)	V	43.2	67.6	54.3	<u>55.0</u>	<u>71.1</u>	<u>65.8</u>	<u>34.3</u>	22.2	48.3	<u>53.4</u>
InternVL3.5 (8B)	V	<u>47.5</u>	57.4	58.8	54.6	63.8	63.2	31.4	23.1	45.4	52.3
VideoLLaMA3 (7B)	V	39.7	<u>68.5</u>	<u>55.8</u>	54.7	64.6	63.1	<u>34.3</u>	17.9	45.0	52.3
VideoLLaMA (7B)	V+A	21.1	23.8	25.7	23.5	40.1	26.6	33.1	26.2	31.5	25.5
VideoLLaMA2.1 (7B)	V+A	36.5	36.4	43.6	38.8	44.0	31.6	32.1	23.7	32.9	37.4
Qwen2.5-Omni (7B)	V+A	35.5	44.5	38.3	39.4	<u>71.1</u>	48.4	27.0	<u>71.8</u>	<u>54.6</u>	43.2
GPT-4o	V	33.6	50.9	62.1	48.9	-	-	-	-	-	-
GPT-5	V	46.8	69.5	67.3	61.2	-	-	-	-	-	-
Gemini-1.5-Pro	V+A	50.9	71.4	60.7	61.0	90.1	76.8	66.4	84.6	79.5	65.6
Gemini-2.5-flash	V+A	**53.0**	75.7	64.9	64.5	**98.7**	**80.6**	57.7	**99.1**	84.0	64.5
Gemini-2.5-Pro	V+A	52.9	**83.5**	**70.7**	**69.0**	96.7	78.7	**72.3**	98.3	**86.5**	**73.4**
Human (Graduate)		84.6	88.5	87.0	86.7	96.0	96.1	87.0	98.3	94.4	88.6

4. Evaluation and Insights

4.1. Experimental Settings

We evaluate 30 of the most popular video MLLMs, including visual-only MLLMs from the Qwen-VL series, InternVL series, and VideoLLaMA3, as well as audio-visual MLLMs like Qwen2.5-Omni [53] and Video-LLaMA series [15, 59]. We also evaluated commercial models from the GPT-4o, GPT-5, and Gemini series. All tasks were cast as multiple-choice questions (N-choose-1, with variable N), with human and random guesses as baselines. For more implementation details and the full results of all 16 tasks on 30 MLLMs , please refer to Appendix I and D.

4.2. Comprehensive Evaluation of MLLMs

Table 1 summarizes the performance evaluation results for our benchmark, revealing several key insights from analyzing vision-only tasks (emotion perception, person recognition, and behavior analysis tasks) and cross-modal tasks (speech-visual alignment tasks).

4.2.1. Performance in Vision-Only Tasks

Open-Source Video-MLLMs: Overall, current models still exhibit a clear gap from human-level performance, especially in Emotion Perception tasks, underscoring their limited ability to capture emotional nuances. Nevertheless, among open-source models, Qwen3-VL leads in video human understanding, and achieves a modest 55.0% mean accuracy across 12 vision-only tasks. Although it still lags behind human performance, it has surpassed GPT-4o and approaches other commercial counterparts, highlighting the immense potential of open-source models.

Proprietary MLLMs: Overall, with the exception of GPT-4o, proprietary models substantially outperform open-source ones across tasks, though their capability in emotion perception remains relatively weak, highlighting the need for better nuanced emotion understanding; notably, Gemini-2.5-Pro stands far ahead, achieving near-human proficiency in person recognition, whereas GPT-4o performs unexpectedly poorly; its subpar performance in the Emotion Understanding category and in specific Person Recognition tasks (T2H and ATD) results in its average performance being lower than that of several leading open-source models.

4.2.2. Performance in Speech-Visual Alignment Tasks

Lip-Reading Ability of Visual-Only MLLMs For AVSM and ASD tasks, many videos include dubbing or multi-speaker conversations with single-speaker audio, making lip movements alone insufficient for accurate responses. Therefore, when answering these two tasks, visual-only models essentially degrade to speech action recognition but can still get some questions correct, with QwenVL3 outperforming other models. However, for the AVAD and SCM task, almost all models perform at a random level, indicating that current models lack precise lip-reading (lip translation) ability. Future models could focus on enhancing lip-reading capabilities.

Audio-Video MLLMs: Although most open-source MLLMs can process both audio and visual inputs, they—except for Qwen2.5-Omni—perform near-randomly on Speech-Visual Alignment tasks, particularly AVSM and ASD, where they even lag behind visual-only models. This deficiency may stem from two key factors: a limited ability to visually interpret intricate lip movements (as further detailed in Appendix Section F) and a general lack of a suitable temporal processing mechanism to facilitate fine-grained speech-to-lip alignment, further exacerbated by scarce datasets providing audio-visual lexical mappings. Consequently, while dedicated ASR models can effectively handle SCM, most open-source video-MLLMs struggle acutely with correlating speech to lip movements. In contrast, the Gemini series shows exceptional cross-modal alignment, performing robustly even on AVAD, a task no other models can handle.



Proper Option Content: Concern or Impatience	
MLLMs	**Wrong Option Content**
Gemini-1.5-pro, GPT-4o, LLaVA-Video, Qwen-VL2, videollama3、Qwen-VL2.5, ...	Surprise and concern (×)
InternVL2, ShareGPT4Video	Surprise (×)

Proper Option Content: Pride and Excitement	
MLLMs	**Wrong Option Content**
InternVL2.5, LLaVA-Video, Qwen-VL2.5, ShareGPT4Video, videollama3, ViLAMP, ...	Surprise, confusion, and possibly frustration (×)

Figure 4. Two examples of 8-frame speaker videos sampled evenly for the emotion recognition task, with responses from various MLLMs.

(a) *Emotion Recognition* Accuracy for the full dataset and the subset where target individuals are in the speaking state.

Models	ER Accuracy	
	Full	Speaking Subset
CogVLM2	38.4	34.7(↓3.7)
VideoLLaMA3	34.6	30.3(↓4.3)
LLaVAOneVision	36.9	34.7(↓2.2)
InternVL3.5	43.7	40.9(↓2.8)
InternVideo	35.7	31.0(↓4.7)
Qwen-VL2.5	43.0	41.0(↓2.0)
ShareGPT4Video	35.0	31.6(↓3.4)
ChatBridge	28.9	24.1(↓4.8)
Qwen2.5-Omni	42.2	40.3(↓1.9)
Average	36.9	33.5(↓3.3)

(b) Timestamp integration effect on Video-MLLMs in *Appearance Time Detection* , *Action at Specific Time* and *Time of Specific Action*.

Models	Time-Specific Tasks		
	ATD	AST	TSA
Chat-UniVi	22.1 (↑2)	40 (↑26)	14.1 (↑7)
CogVLM2	40.8 (↑1)	35 (↓1)	10.7 (↓3)
VILA	38.3 (↑18)	33 (–)	48 (↓3)
Video-LLaVA	36.7 (↑5)	25 (↓2)	45.8 (↑5)
LLaVAOneVision	10 (–)	41 (↑4)	47.5 (–)
InternVL2	31.7 (↑11)	45 (↑7)	52.5 (↑19)
Video-LLaMA	24.2 (↑4)	25 (↑7)	9 (–)
Video-LLaMA2.1	24.2 (↑7)	25 (↓2)	29.4 (↓4)
ChatBridge	17.5 (↑9)	40 (↓1)	14.1 (↓3)
Average	26.8 (↑6)	34.3 (↑4)	30.1 (↑2)

(c) Effectiveness of the Distractor-Included QA Synthesis Pipeline in generating six types of descriptive multiple-choice questions. The Efficiency Gain is calculated as 1 / (1 - No Edit Rate).

Question Type	No Edit	Efficiency Gain
Behavior Temporal Analysis	68%	3.1
Emotion Intensity Comparison	63%	2.7
Emotion Recognition	81%	5.3
Behavior Causality Analysis	74%	3.8
Emotion Temporal Analysis	83%	5.9
Human-to-Text	65%	2.9
Average	72.3%	3.6

4.3. Dissecting Fine-grained Insights

4.3.1. Discussion on Speaker Emotion Recognition

A detailed analysis reveals that models frequently misattribute speaker emotions as "surprise" or "shock". This issue primarily stems from frame sampling processes, where frames depicting "mouth-opening" motions are misclassified. Figure 4 illustrates this phenomenon using the commonly adopted eight-frame fixed sampling approach [36, 63], which introduces temporal noise and leads to erroneous emotional judgments. The speaker-centric evaluations presented in Table 2a further confirm a consistent decline in accuracy across all models, underscoring the inherent challenges of this task for video MLLMs.

4.3.2. Timestamps Impact on Time-Specific Tasks

Our findings show that many open-source video-MLLMs struggle with time-specific tasks due to their inability to correlate video events with a timeline (Appendix Table D). While top performers like VideoLLaMA3 and LLaVA-Video natively incorporate textual timestamps, we explored this capability by adding timestamps to models without native support. As shown in Table 2b, this intervention generally improves temporal inference accuracy, though its effectiveness varies due to models' limited exposure to timestamp data. Despite these gains, a significant gap remains compared to MLLMs that natively incorporate time-event alignment, underscoring the necessity of integrating temporal reasoning mechanisms from the outset of model design.

4.4. Quantifying the Reduction of Manual Labor in Question Generation

We report the distribution of manual work required for the 6 descriptive multiple-choice questions generated by the pipeline in Table 2c. On average, 72% of the generated questions contain the correct answer, requiring no revision, with humans only needing to select the correct answer. The remaining 28% lack a correct answer and require a simple manual input, typically one sentence, which is much faster than creating questions from scratch. Notably, the overall proportion of questions containing the correct answer exceeds the best model's accuracy in the final evaluation (e.g., the "no-edit" rate for ER is 81%, while the best model accuracy is 49.4%), as options are drawn from multiple models, increasing correct answer coverage and reflecting the benefit of ensemble-based design. Besides, incorrect model responses are repurposed as plausible distractors, enhancing question difficulty. For other rule-based non-descriptive tasks, over 70% of the questions can be directly adopted. Overall, while human oversight is still needed, our partial automation significantly boosts efficiency, allowing a person using the pipeline to complete 3 times more questions within the same timeframe as manual creation from scratch, substantially reducing manual workload.

4.5. Generalizability of the Synthesis Pipelines Beyond Human-Centric Tasks

A key contribution of our work is the general-purpose nature of our automated pipelines, which extends far beyond

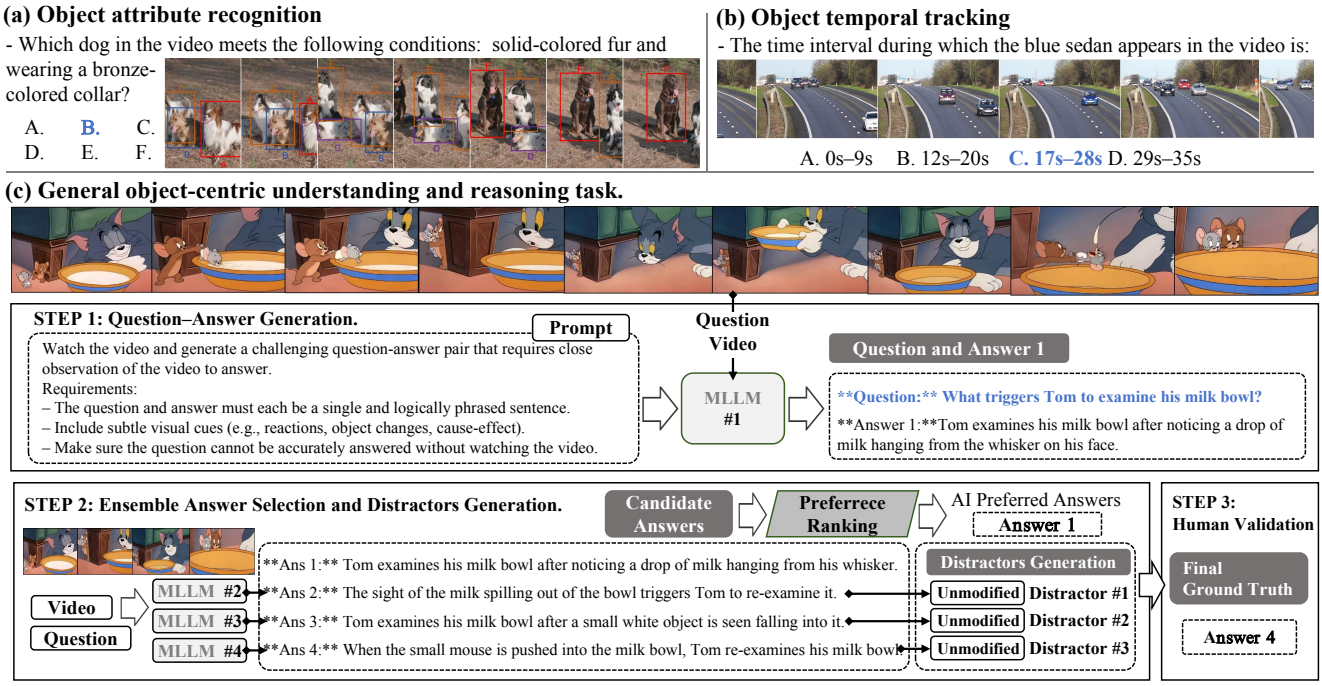


Figure 5. Demonstration of our pipeline's generalizability. (a) Object attribute recognition for pets. (b) Object temporal tracking for cars. (c) General causal reasoning QA generated for an object-centric event, with the MLLMs used in this case being Gemini-1.5-pro, VideoLLaMA3, LLaVA-Video, and gpt-4o. This highlights the adaptability of our framework.

the human-centric domain. Its modular architecture allows for easy adaptation to new domains by simply swapping initial detectors. We demonstrate this by generating diagnostic questions for non-human entities, such as fine-grained attribute recognition for pets and temporal tracking for vehicles, with examples shown in Figure 5.

For object attribute recognition (e.g., a specific dog; Figure 5 (a)), we adapt our pipeline by: *(i)* replacing the human detector with a dog detector (YOLOv11) to extract instance tracks (*Label-1*); *(ii)* using an MLLM to annotate attributes (e.g., fur pattern, accessories) from cropped tracks; *(iii)* selecting a uniquely described dog as the ground truth; and *(iv)* rendering bounding boxes from *Label-1* as candidates. Similarly, for object temporal tracking (e.g., a car's appearance interval, Figure 5 (b)), we: *(i)* replace the human detector with a vehicle detector for tracking; *(ii)* annotate attributes (e.g., color, type) for each track; *(iii)* select a unique car as ground truth; and *(iv)* use track frames to determine the appearance interval as ground truth. These procedures parallel our *Text-to-Human* and *Appearance Time Detection* tasks, proving the framework's broad applicability to object-centric attribute recognition and temporal tracking.

Furthermore, our Distractor-Included QA Generation Pipeline is inherently transferable, enabling the synthesis of diverse descriptive questions and answer choices beyond human-centric scenarios, as exemplified by the general causal reasoning question in Figure 5 (c). This transferability underscores the broader value of our methodology as a scalable solution for accelerating the creation of high-quality MLLM evaluations across diverse domains.

5. Conclusion

We present HUMANVBENCH to address the pressing need for improved assessment of human-centric video understanding in MLLMs. By incorporating extensive evaluation dimensions through 16 fine-grained tasks, HUMANVBENCH provides a systemic view into both successes and critical shortcomings of video MLLMs, particularly in emotion perception and speech-visual alignment. Experimental findings across over twenty leading video MLLMs illustrate that while proprietary models approach human accuracy in some tasks, substantial advancements are required, particularly in cross-modal domains where alignment between speech and visual elements proves challenging.

Beyond the benchmark itself, the core contribution of this work lies in its extensibility and efficient methodology. Our framework transforms model-generated errors into high-quality distractors and shifts the human role from creating tasks from scratch to efficiently validating and correcting them, thereby multiplying the benchmark's production efficiency. By open-sourcing the benchmark and underlying methodologies, we hope HUMANVBENCH can foster collaborative efforts aimed at advancing the frontiers of human-like video understanding capabilities in MLLMs.

Acknowledgments

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123003), the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund (Artificial Intelligence Field) Key Support Projects (U25B2072), and The Major Key Project of PCL (Grant No. PCL2025A17).

References

[1] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms, 2024. 8

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 8

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 2

[5] Grace Byun and Jinho D Choi. D-gen: Automatic distractor generation and evaluation for reliable assessment of generative model. *arXiv preprint arXiv:2504.13439*, 2025. 3

[6] Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, et al. Data-juicer: A one-stop data processing system for large language models. In *Companion of the 2024 International Conference on Management of Data*, pages 120–134, 2024. 4

[7] Daoyuan Chen, Yilun Huang, Xuchen Pan, Nana Jiang, Haibin Wang, Ce Ge, Yushuo Chen, Wenhao Zhang, Zhijian Ma, Yilei Zhang, Jun Huang, Wei Lin, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-juicer 2.0: Cloud-scale adaptive data processing for foundation models, 2024. 7

[8] Daoyuan Chen, Haibin Wang, Yilun Huang, Ce Ge, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-juicer sandbox: A comprehensive suite for multimodal data-model co-development. *arXiv preprint arXiv:2407.11784*, 2024. 4

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *CoRR*, 2024. 5

[10] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 5, 7

[11] Shizhe Chen, Yuqing Song, Yida Zhao, Qin Jin, Zhaoyang Zeng, Bei Liu, Jianlong Fu, and Alexander Hauptmann. Activitynet 2019 task 3: Exploring contexts for dense captioning events in videos. *arXiv preprint arXiv:1907.05092*, 2019. 3

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1

[13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 1

[14] Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation. In *Forty-second International Conference on Machine Learning*, 2025. 1

[15] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6, 1, 7

[16] Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Ornelas, and Andrew Lan. Exploring automated distractor generation for math multiple-choice questions via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082, 2024. 3

[17] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2, 3

[18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1

[19] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities

with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 1

[20] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 1

[21] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8450–8460, 2025. 2

[22] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. From training-free to adaptive: Empirical insights into mllms' understanding of detection information. *arXiv preprint arXiv:2401.17981*, 2024. 2

[23] Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9296–9307, 2025.

[24] Qirui Jiao, Daoyuan Chen, Yilun Huang, Xika Lin, Ying Shen, and Yaliang Li. Detailmaster: Can your text-to-image model handle long prompts? *arXiv preprint arXiv:2505.16915*, 2025. 2

[25] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding, 2024. 1

[26] Jiayi Kuang, Yinghui Li, Chen Wang, Haohao Luo, Ying Shen, and Wenhao Jiang. Express what you see: Can multimodal llms decode visual ciphers with intuitive semiosis comprehension? In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12743–12774. Association for Computational Linguistics, 2025. 2

[27] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Comput. Surv.*, 57(8):190:1–190:36, 2025. 2

[28] Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy, Shashank Goyal, et al. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225:120022, 2023. 3

[29] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379, 2018. 2

[30] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8211–8225, 2020. 2

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1

[32] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1

[33] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. 2, 3

[34] Keliang Li, Zaifei Yang, Jiahe Zhao, Hongze Shen, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Herm: Benchmarking and enhancing multimodal llms for human-centric understanding. *arXiv preprint arXiv:2410.06777*, 2024. 3

[35] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22932–22941, 2023. 7

[36] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7

[37] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1

[38] Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. Better distractions: Transformer-based distractor generation and multiple choice question filtering. *arXiv preprint arXiv:2010.09598*, 2020. 3

[39] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019. 2

[40] Zhen Qin, Daoyuan Chen, Wenhao Zhang, Liuyi Yao, Yilun Huang, Bolin Ding, Yaliang Li, and Shuiguang Deng. The synergy between data and multi-modal large language models: A survey from co-development perspective. *arXiv preprint arXiv:2407.08583*, 2024. 2

[41] Fanyi Qu, Hao Sun, and Yunfang Wu. Unsupervised distractor generation via large language model distilling and counterfactual contrastive decoding. *arXiv preprint arXiv:2406.01306*, 2024. 3

[42] Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. Veatic: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4467–4477, 2024. 2

[43] Katherine Stasaski and Marti A Hearst. Multiple choice question generation utilizing an ontology. In *Proceedings*

of the 12th workshop on innovative use of NLP for building educational applications, pages 303–312, 2017. 3

[44] Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*, 2025. 3

[45] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, and Wanli Ouyang. Humanbench: Towards general human-centric perception with projector assisted pretraining, 2023. 2, 3

[46] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[49] 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. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. 1

[50] Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. Mio: A foundation model on multimodal tokens, 2025. 1

[51] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 3

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[53] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. 6, 1

[54] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 3

[55] Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: a large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23191–23200, 2023. 3

[56] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 2

[57] Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, Stella Frank, Alessandro Suglia, Chrysoula Zerva, Desmond Elliott, Mariella Dimiccoli, Mohit Bansal, Oswald Lanz, Raffaella Bernardi, Raquel Fernández, Sandro Pezzelle, Vlad Niculae, and André F. T. Martins. Movie facts and fibs (mf^2): A benchmark for long movie understanding, 2025. 4

[58] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 5

[59] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023. 6, 1

[60] Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. Lvbench: A benchmark for long-form video understanding with versatile multi-modal question answering. *International Journal of Computer Vision*, pages 1–22, 2025. 2, 3

[61] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 7

[62] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. 1

[63] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models, 2024. 7

[64] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 1

[65] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025. 2