VIDEONORMS: Benchmarking Cultural Awareness of Video Language Models

Anonymous ACL submission

Abstract

As Video Large Language Models (VideoLLMs) are deployed globally, they require understanding of and grounding in the relevant cultural background. To properly assess these models' cultural awareness, adequate benchmarks are needed. We introduce VIDEONORMS, a benchmark of over 1000 (video clip, norm) pairs from US and Chinese cultures annotated with socio-cultural norms grounded in speech act theory, norm adherence and violations labels, and verbal and non-verbal evidence. To build VIDEONORMS, we use a human-AI collaboration framework, where a teacher model using theoretically-grounded prompting provides candidate annotations and a set of trained human experts validate and correct the annotations. We benchmark a variety of open-weight VideoLLMs on the new dataset which highlight several common trends: 1) models performs worse on norm violation than adherence; 2) models perform worse w.r.t Chinese culture compared to the US culture; 3) models have more difficulty in providing non-verbal evidence compared to verbal for the norm adhere/violation label and struggle to identify the exact norm corresponding to a speech-act; and 4) unlike humans, models perform worse in formal, non-humorous contexts. Our findings emphasize the need for culturally-grounded video language model training — a gap our benchmark and framework begin to address. [1]

1 Introduction

AI systems based on large language models (LLMs), including video large language models (VideoLLMs), are deployed globally, requiring adaptation to differing cultural contexts. However, cultural competence of VideoLLMs has not been given the same attention as general object/action recognition (Xu et al., 2016), temporal reasoning (Li and et al., 2024; Fu et al., 2025; Xiao et al., 2021; Jang et al., 2017), or narrative understanding (Tapaswi et al., 2016). In the textual modality, frameworks for evaluating cultural competence are largely based on understanding social norms – rules of thumb for human behavior, such as "It is rude to run a blender at 5am" (Forbes et al., 2020; Emelin et al., 2021; Hendrycks et al., 2021; Ziems and et al., 2023) in differing cultural contexts (Fung et al., 2023; Huang and Yang, 2023; Li et al., 2023; CH-Wang et al., 2023). In images, benchmarks testing the models' cultural knowledge (Winata et al., 2025) and culturally-aligned image generation have been proposed (Nayak et al., 2025). In videos, recent work has explored norms grounded in the physical world (Rezaei et al., 2025).

Expanding on these approaches, we propose to evaluate the cultural competence of VideoLLMs through their understanding of cultural norms: societal rules or standards that "delineate an accepted and appropriate behavior within a culture" (American Psychological Association, 2025), when the input modality is a video. For example, does the video language model understand that when two people introduce themselves and shake hands, they adhere to the greeting norms in US culture? This is particularly challenging in the multimodal context, where understanding whether a cultural norm is present and adhered to requires correct interpretation of implicit meaning expressed through both nonverbal cues (e.g., gaze, gestures) (Pang et al., 2024; Hessels et al., 2025), and verbal features (e.g., sarcasm) (Rakov and Rosenberg, 2013; Tepperman et al., 2006). Additional challenge stems from the cross-cultural aspect, due to the documented differences in norms and values across cultures (Inglehart, 2018; Hofstede, 1984; Gelfand et al., 2011). To systematically evaluate cultural norm understanding, we introduce the VIDEONORMS benchmark and propose 3 tasks: (1) binary classification, where given a video clip the model has to

[1] We will release the code and data.

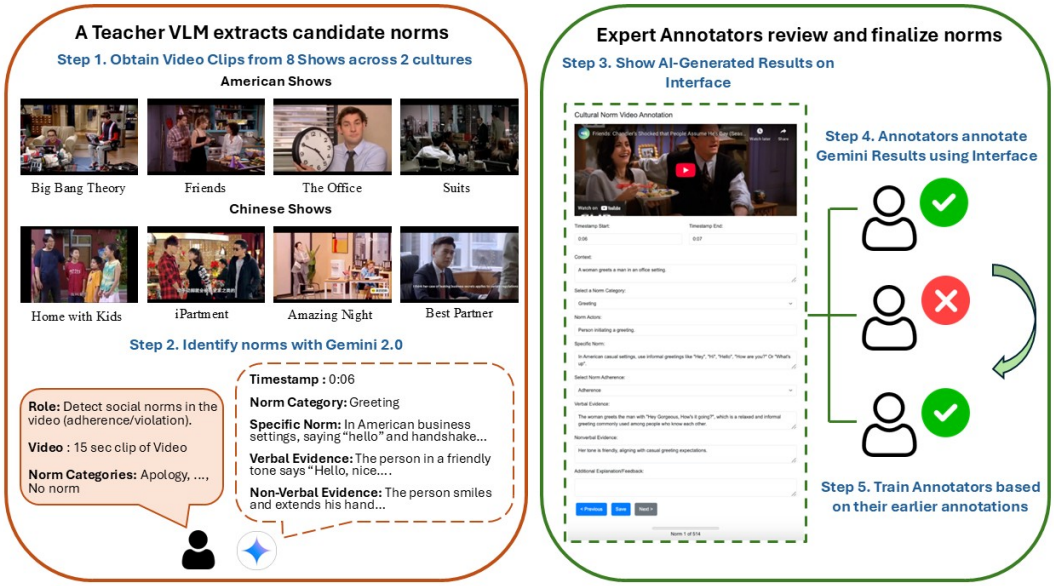


Figure 1: VIDEONORMS Dataset Construction: left panel shows teacher VideoLLM generations using speech act theory prompting; right panel shows the expert annotator editing process.

predict whether a particular cultural norm was adhered to or violated; (2) an explanation task where the model also has to provide verbal and non-verbal evidence to support its adherence or violation label; and (3) generation of an applicable cultural norm (if any) for the video clip. Our contributions are the following:

- **Methodology.** To efficiently construct a dataset of videos annotated for cultural norms, we utilize a two-stage human-AI collaboration framework: (1) given a 15 second video clip, a teacher VideoLLM, using a prompt grounded in speech act theory (Austin, 1962; Searle, 1969), extracts candidate norm category, the associated socio-cultural norm if exists, an adherence/violation label, and the relevant verbal and/or non-verbal evidence (see the left panel on Figure 1); (2) expert annotators with a relevant cultural background review and edit errors in these candidate annotations (right panel on Figure 1).
- **Analysis of disagreements** between the teacher model and human annotators, as well as among the human raters.
- **VIDEONORMS Benchmark**, the first benchmark to evaluate cross-cultural (US and Chinese) norm understanding in VideoLLMs containing video clips from 4 pairs of comparable US-Chinese TV shows in workplace/informal social settings and drama/comedy genres. Each clip is paired with human-edited judgments on norm category, adherence/violation label, and the associated (non)verbal evidence totaling over 1K (video clip, norm) pairs.
- **Empirical analysis** of modern open-weight VideoLLMs' cultural competence, revealing that models underperform in the following settings: 1) detecting norm violations compared to norm adherence; 2) when the social context is formal rather than informal; 3) when the setting is w.r.t Chinese culture rather than US culture; 4) when providing evidence for the adherence/violation label and identifying the actual norms relevant to a speech act in a video.

2 Related Work

Social Norms, Morality, and Cultural Knowledge in NLP ATOMIC and Social IQa modeled everyday commonsense and social interactions (Sap et al., 2019a,b). Social Chemistry 101 formalized computational investigation of social norms with language models (Forbes et al., 2020). Moral Stories added structured narratives to probe norm adherence vs. violation and consequences (Emelin et al., 2021). ETHICS/Delphi framed moral judgments at scale but also prompted critical reflection on dataset design, value pluralism, and the pitfalls of single "gold" labels (Hendrycks et al., 2021;

Jiang et al., 2025; Talat et al., 2022). NormBank compiled 155k situational norms grounded in roles and settings, moving beyond decontextualized rules (Ziems and et al., 2023). Complementary cross-cultural resources, GeoMLAMA, XCOPA, Candle/CCSK, and CulturalBench, probe geographic and cultural variation, showing that model performance and knowledge can vary widely across regions and languages (Yin et al., 2022; Ponti and et al., 2020; Nguyen et al., 2022; Chiu et al., 2024). In multimodal contexts, most work has focused on cultural knowledge (Winata et al., 2025; Shafique et al., 2025) and culturally-aligned generation (Nayak et al., 2025; Havaldar et al., 2025) recently also exploring norm understanding (Fung and Ji, 2025; Rezaei et al., 2025). Our main contribution is a dataset for cross-cultural social norm understanding requiring fine-grained comprehension of implicit meaning in speech and non-verbal cues across two cultures.

Video Understanding and Video-Language Models Modern VideoLLMs connect powerful vision encoders with LLMs with instruction tuning (Alayrac et al., 2022; Liu et al., 2023). Extensions to video unify image/video tokenization and improve temporal reasoning (Lin et al., 2024). Open families such as Qwen2-VL and InternVL further expand input resolution, multi-granular perception, and tool use (Wang and et al., 2024; Zhu et al., 2025). However, benchmarks predominantly target temporal or narrative understanding: MVBench, Video-MME, NExT-QA, TGIF-QA, MSRVTT-QA, MSVD-QA, and MovieQA (Li and et al., 2024; Fu et al., 2025; Xiao et al., 2021; Jang et al., 2017; Xu et al., 2016; Tapaswi et al., 2016). Closer to social reasoning, MovieGraphs annotates human-centric situations (relations, emotions, motivations) in movies (Vicol et al., 2018). To our knowledge, none of these explicitly evaluate cultural norm adherence/violation.

3 VIDEONORMS Benchmark

To build VIDEONORMS, we designed a multi-stage human-AI collaboration framework that integrates video sampling, AI-assisted norm extraction, and human refinement. Figure 1 illustrates the overall process. The collaboration framework has 3 steps: (1) select clips from eight US and Chinese television shows spanning both formal and informal contexts (Section 3.1); (2) use a teacher VideoLLM (Gemini 2.0) with *speech act theory prompting* to extract candidate norm category, adherence/violation, and verbal/non-verbal evidence (Section 3.2); (3) refine the model-generated data by obtaining edits from 3 trained annotators with a relevant cultural background on each instance (Section 3.3).

3.1 Video Selection

To construct a comparable cross-cultural benchmark for social norm recognition, we selected eight popular television shows, four from the United States and four analogous shows from China, representing both formal and informal social settings. For the US dataset, the two workplace-context shows are *Suits* (drama genre), which centers around professional interactions in a corporate law firm, and *The Office* (comedy genre), a workplace mockumentary that, while comedic, provides a rich depiction of office dynamics and hierarchical relationships. For informal settings, we chose *Friends* and *The Big Bang Theory*, both of which portray casual, peer-to-peer interactions among close friends in everyday social contexts. For the Chinese dataset, we similarly selected four well-known Chinese TV shows, evenly split between formal and informal scenarios. The two formal-context shows are *The Best Partner* (drama genre), which focuses on professional conduct and workplace interactions within a law firm, and *Amazing Night* (comedy genre), a comedy set in a business environment that serves a functionally similar role to *The Office* in the US dataset. For informal scenarios, we selected *iPartment*, a sitcom about young adults living together, and *Home With Kids*, a family sitcom that highlights casual interactions among siblings, parents, and children. For each TV show, we sample 15 second clips as the input for the video language model (see Appendix A.1).

3.2 Teacher Model Annotation with Speech Act Theory Prompting

To generate candidate annotations of video clips we use `Gemini-2.0-Pro` model (DeepMind, 2024), as it was the only large video model offering integrated video and audio long-context understanding at the time of the study.

Speech Act Theory Prompting We ground our prompt for cultural norm annotation in speech act theory (Searle, 1969; Austin, 1962). Classic speech act theory: distinguishes between different types of illocutionary acts, identifying five broad classes: *assertives*, *directives*, *commissives*, *expressives*,

and *declarations*. We select the following potential norm categories based on these speech act categories: *Thanks*, *Apology*, *Admiration*, *Greeting*, and *Farewell* (mapping to *expressives*); *Requesting Information* and *Rejecting a Request* (corresponding to *directives*); *Granting a Request* and *Agreement* (function as *commissives*), and *Disagreement* is a form of *assertive* that negates another speaker's claim. These categories also align with prior work on norm understanding (Li et al., 2023) based on the Linguistic Data Consortium taxonomy (Linguistic Data Consortium, 2022). We also include a *Custom Category* which provides flexibility to capture culturally specific or emergent speech acts.

After selecting the most relevant category, the model is asked to generate an applicable cultural norm within that category. Our pormpt provides examples of norms in both formal and casual contexts for each category (for example, one can apologize with "Ooops!" in an informal context, and "I would like to express my apology" in a formal context). Besides identifying the norm and the norm category, the model is asked to generate the context in which the situation occurs, the subjects of the norm, whether the norm is adhered to or violated, and provide an explanation with verbal and non-verbal evidence from the clip for the adherence/violation label (see Table 1). The prompt was also adapted for Chinese culture, including the translation to Mandarin. See full prompts in Appendix, Table 5.

Candidate data statistics Applying these prompts, Gemini produced 514 unique (video clip, norm) instance candidate annotations across the US dataset and 501 across the CN dataset (see overall statistics in Appendix, Table 8).

3.3 Human Verification

Each candidate annotation generated above was further verified by 3 annotators. Annotators validate or modify each field in Table 1 by checking whether the selected timestamps isolate the correct behavior, confirming the norm category, and ensuring the verbal and nonverbal evidence supports the adherence or violation label. The interface includes an Additional Explanation/Feedback field where annotators briefly document the rationale for any modifications or confirm agreement with the teacher model's output. See a screenshot of the interface in Appendix B.3. Figure 2 compares the teacher model's initial outputs with annotator refinements for both US and Chinese shows.

English Field	Mandarin Field	Explanation
Timestamp	时间戳	Start and end time of the event.
Context	情境描述	Brief description of the setting and the social hierarchy between participants (e.g., colleagues, friends, siblings).
Norm Category	规范类别	Selected from the predefined list or dynamically generated under the Custom Category.
Norm Subject	行为主体	The individual to whom the norm is applied (without using character names).
Specific Norm	具体规范	A precise articulation of the expected behavior for the given context.
Norm Adherence	规范遵循情况	Indicates whether the behavior represents adherence or violation.
Explanation	解释说明	A breakdown of verbal and nonverbal cues (e.g., dialogue, tone, facial expression, and body language) that justify the assessment. Includes subfields: 语言证据 (verbal evidence) and 非语言证据 (nonverbal evidence).

Table 1: English and Mandarin output fields, with explanations.

We ensured that each annotator's relevant mono-cultural identity by confirming their country of residence, primary language, mono-cultural self-identification, earliest language in life, and other factors (see Appendix B.1). Detailed instructions were provided (see Appendix B.3) and each annotator was first trained on initial 20 instances, where their answered were reviewed to ensure high-quality completion and instruction following. All annotators were fairly compensated above the local minimum wage guidelines.

3.4 Analysis of disagreements

We report Fleiss's κ to measure how often annotators agree across violation/adherence binary labels beyond chance (Fleiss et al., 1971). We also report the combined annotator edit percentage, representing the fraction of instances edited by the annotators. As shown in Table 2, agreement remained relatively high across US shows, with *Suits* achieving the highest agreement ($\kappa = 0.76$), followed by Big Bang Theory ($\kappa = 0.71$). The combined annotator

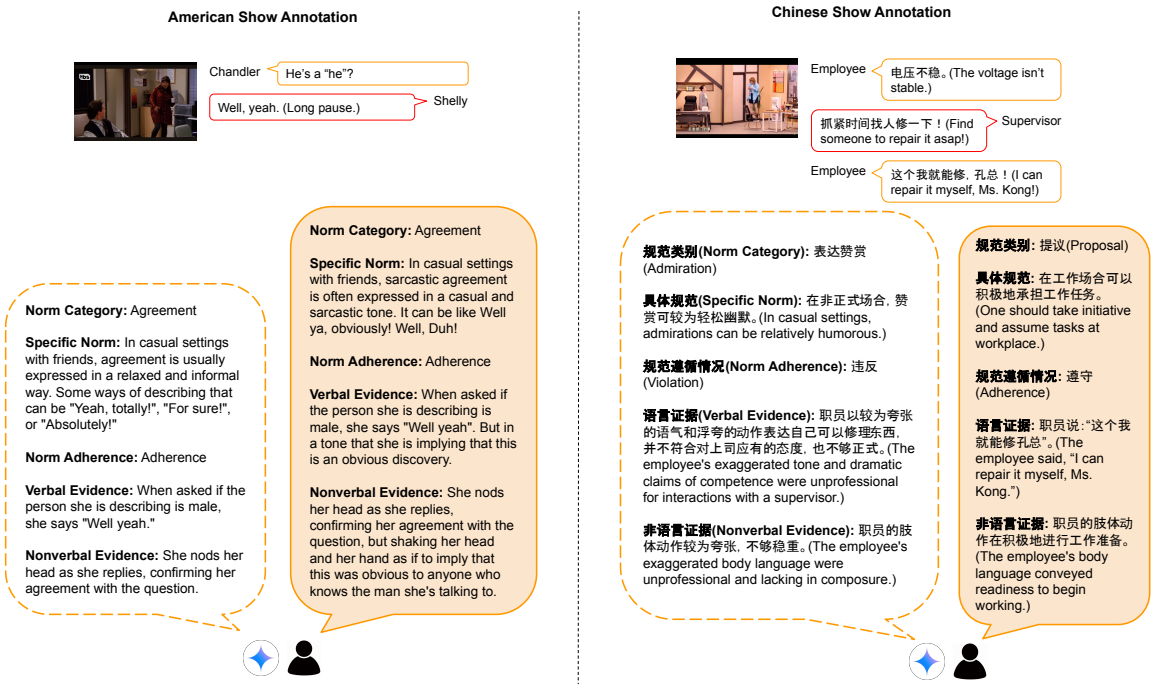


Figure 2: Examples of Gemini-generated normative behavior annotations and corresponding human refinements for US (left) and Chinese (right) shows, as recorded through the annotation interface.

change percentages[2] for US shows ranged from 18.28% to 26.04%, reflecting relatively accurate labeling by the teacher model. In contrast, combined annotator change percentages for Chinese shows were notably higher, ranging from 42.35% to 53.39%, indicating lower performance of the model in Chinese context. This aligns with findings in previous research that LLMs perform more effectively in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) cultural contexts (Mihalcea et al., 2025; Liu et al., 2025). Consequently, annotators in the Chinese dataset needed to modify a larger fraction of candidate annotations, which naturally contributes to lower inter-annotator agreement metrics ($0.27 - 0.45$).

Dataset	Genre	Setting	Show	Fleiss's κ	Change %
US	Comedy	Informal	Friends	0.61	23.36%
	Comedy	Informal	Big Bang Theory	0.71	18.28%
	Comedy	Workplace	The Office	0.59	26.04%
	Drama	Workplace	Suits	0.76	20.62%
Chinese	Comedy	Informal	iPartment	0.33	50.00%
	Comedy	Informal	Home with Kids	0.27	46.89%
	Comedy	Workplace	Amazing Night	0.42	53.39%
	Drama	Workplace	Best Partner	0.45	42.35%

Table 2: Fleiss's κ Scores and Annotator Change Percentages by Genre and Setting.

[2]Total number of fields across all norms and all annotators that have been changed divided by the total number of fields. See Appendix A.4 for breakdown by field.

When comparing across genre (comedy/drama) and setting (workplace/informal), both US and Chinese subsets display higher agreement for workplace drama shows (Suits $\kappa = 0.76$ and Best Partner $\kappa = 0.45$), while showing lower agreement on comedy shows. Unlike for US shows, the agreement for Chinese shows in informal settings yielded lower agreement than workplace settings ($\kappa = 0.27 - 0.33$ vs. $\kappa = 0.42 - 0.45$), aligning with findings conform with psychology studies Sino-US workplace flexibility (Lai et al., 2022; Wei and Royle, 2024).

4 Benchmarking Open-Weight Video Models for Cultural Awareness

We evaluate cultural awareness with three progressively challenging tasks, each grounded in a localized 15-second video segment. Transcripts for each clip are obtained with `whisper-large-v3` (Radford et al., 2022) and provided as part of the prompt, ensuring grounding for verbal cues. Each instance specifies the clip's start timestamp to bind predictions to the correct sub-span rather than the entire video. Tasks are evaluated in US and Chinese (CN) cultural contexts, and each is situated within an explicit *norm category* (e.g., GREETING, REQUESTING INFORMATION). For all F1 scores, we

report 95% confidence intervals (CIs) computed with `confidenceinterval` package (Gildenblat, 2023) using the Takahashi et al. (2022) method. For average verbal/nonverbal and Task 3 scores, we report 95% CIs based on the standard error of the mean.

4.1 Video Cultural Norm Understanding Tasks and Metrics

Task 1: Predicting Adherence or Violation Given a video segment, transcript, norm category, and a specific norm, the model must predict whether the observed behavior *adheres to* or *violates* that norm (binary classification). We report, separately for US/Chinese norms, per-class F1 for Adherence (pos) and Violation (neg).

Task 2: Predicting Adherence/Violation and Extracting Evidence Building on Task 1, the model must (i) predict Adherence/Violation and (ii) generate two evidence strings: (a) *verbal evidence* referencing the spoken content (e.g., quoted phrases), and (b) *nonverbal evidence* referencing visual social cues (e.g., gaze, gesture, posture, interpersonal distance, facial expression). We report the label F1 and for evidence quality, we adopt an LLM-as-judge protocol: a GPT-5 grader compares verbal/nonverbal rationales (for correctly predicted labels) to the reference evidences[3] in our dataset using a 5-point rubric (from unsupported to perfectly grounded, see Appendix C.3). Two of the authors verified the performance of LLM-as-a-Judge on 20 instances, out of which only 2 did not fully adhere to the rubric. The grader provides a rationale plus numeric verdict, which we average over all cases for reporting in the *Verbal* and *Nonverbal* columns. This design emphasizes whether models cite *diagnostic, modality-appropriate* cues rather than merely producing the correct label.

Task 3: Predicting a Cultural Norm Given a video segment, transcript, and norm category, the model must generate a *specific norm* that captures the exhibited behavior (e.g., "offer a brief handshake and a friendly greeting in business introductions"). Generated norms are evaluated by the GPT-5 grader using a 5-point rubric judging how well the generated norm matches the reference norm in our dataset. The mean scores are reported in the *Score* column.

[3]When multiple evidences are present due to annotator edits, we take the maximum similarity score among them.

Statistic	**US**	**CN**
Number of Video Clips	266	249
Number of Norm Categories	137	247
Task 1 & 2: Adherence Classification		
Total Cases	724	1113
Adherence	60.9%	80.0%
Violation	39.1%	20.0%
Task 3: Specific Norm Generation		
Total Cases	744	1119
Top Norm Categories		
Expressing criticism / 提出批评	27.2%	3.9%
Requesting information / 请求信息	20.2%	12.9%
Admiration / 表达赞赏	7.4%	5.6%
Greeting / 问候	6.0%	5.0%
Agreement / 表示同意	2.7%	4.5%
Disagreement / 表达异议	2.7%	10.9%
Granting request / 拒绝请求	2.6%	2.2%
Rejecting request / 商务谈判	2.6%	1.3%
Apology / 表示感谢	2.2%	1.7%
Thanks / 邀请	1.9%	1.3%

Table 3: Dataset statistics across US and Chinese norms for Tasks 1–3. Values represent percentages of total samples per language; Task 3 lists the ten most frequent norm categories.

4.2 Dataset Statistics and Metadata

Table 3 presents the statistics for VIDEONORMS across both languages and tasks. Our annotation process involves each norm generated by the teacher model being judged and edited by three annotators. For Task 1, we aggregate by taking unique combinations of norm category and specific norm among the three annotations, with adherence labels determined by majority vote among annotators sharing identical category-norm pairs. Cases with two annotations and without clear majority agreement are excluded, though these represent a negligible portion of the dataset. For Task 2, we additionally compare the generated evidence to all candidate evidences. Task 3 includes all unique category-norm combinations across the three annotations.

4.3 Models

We benchmark recent commonly used open-weight VideoLLMs:

- LLaVA-family models (`LLaVA-Next-Video` (Zhang et al., 2024), `LLaVA-OneVision` (Li et al., 2024)) extend image-text pretraining to video by sampling frames and using linear scaling with Rotary Position Embeddings (Su

Models	Task 1			Task 2				Task 3
	F1 (pos)	F1 (neg)	Avg. F1	F1 (pos)	F1 (neg)	Macro F1	Verbal / Nonverbal	Avg. Score
US Norms								
Llava-Next-Video	$75.4_{72.3-78.5}$	$18.6_{12.4-24.7}$	$47.0_{40.5-53.4}$	$58.4_{54.0-62.8}$	$62.4_{58.5-66.6}$	$60.4_{56.9-64.2}$	$2.17_{\pm0.11}/2.04_{\pm0.11}$	$2.24_{\pm0.08}$
Llava-OneVision	$\mathbf{78.8}_{75.7-82.0}$	$45.4_{39.3-51.5}$	$62.1_{55.9-68.3}$	$\mathbf{76.0}_{73.1-78.7}$	$51.1_{45.7-56.2}$	$63.6_{60.1-67.0}$	$2.59_{\pm0.11}/2.14_{\pm0.10}$	$2.07_{\pm0.07}$
Intern3-VL	$56.0_{51.5-60.5}$	$60.1_{56.9-65.1}$	$58.5_{52.2-64.8}$	$50.0_{45.6-54.3}$	$61.8_{58.0-65.6}$	$55.9_{52.4-59.4}$	$2.72_{\pm0.12}/\mathbf{2.32}_{\pm0.12}$	$2.29_{\pm0.07}$
Intern3.5-VL	$75.5_{71.8-79.1}$	$\mathbf{65.3}_{60.8-69.9}$	$\mathbf{70.4}_{64.3-76.4}$	$65.7_{61.7-69.4}$	$\mathbf{64.9}_{60.9-68.9}$	$\mathbf{65.3}_{62.0-68.5}$	$2.64_{\pm0.11}/2.22_{\pm0.10}$	$2.37_{\pm0.08}$
Qwen2-VL	$73.3_{69.6-77.0}$	$62.5_{57.9-67.2}$	$67.9_{61.8-74.0}$	$67.2_{63.2-70.7}$	$60.0_{55.6-64.2}$	$63.6_{60.2-66.9}$	$\mathbf{2.75}_{\pm0.11}/2.09_{\pm0.10}$	$2.33_{\pm0.07}$
Qwen2.5-VL	$49.3_{44.5-54.2}$	$62.8_{59.0-66.6}$	$56.1_{49.7-62.4}$	$34.8_{30.1-39.7}$	$61.4_{57.6-64.9}$	$48.1_{44.8-51.6}$	$2.47_{\pm0.12}/2.07_{\pm0.12}$	$\mathbf{2.39}_{\pm0.07}$
VideoChatR1	$68.5_{64.6-72.5}$	$65.0_{60.7-69.3}$	$66.8_{60.6-72.9}$	$61.0_{56.9-64.7}$	$64.0_{59.9-67.7}$	$62.5_{59.2-65.6}$	$2.67_{\pm0.11}/2.21_{\pm0.11}$	$2.35_{\pm0.07}$
Chinese (CN) Norms								
Llava-Next-Video	$\mathbf{89.2}_{87.1-91.2}$	$7.0_{0.21-12.0}$	$48.1_{42.9-53.3}$	$\mathbf{89.1}_{87.4-90.6}$	$25.4_{18.5-32.5}$	$57.2_{53.4-61.2}$	$1.78_{\pm0.07}/1.56_{\pm0.06}$	$1.75_{\pm0.05}$
Llava-OneVision	$78.5_{75.9-81.1}$	$47.4_{42.7-52.2}$	$\mathbf{63.0}_{58.0-67.9}$	$66.3_{63.4-69.1}$	$42.8_{37.9-46.5}$	$54.5_{51.4-57.2}$	$2.65_{\pm0.10}/2.15_{\pm0.09}$	$2.40_{\pm0.06}$
Intern3-VL	$40.8_{37.1-44.6}$	$37.7_{33.8-41.6}$	$39.3_{33.9-44.6}$	$33.0_{29.7-36.5}$	$37.4_{33.9-40.7}$	$35.2_{32.5-37.8}$	$\mathbf{2.90}_{\pm0.12}/\mathbf{2.38}_{\pm0.11}$	$\mathbf{2.70}_{\pm0.06}$
Intern3.5-VL	$72.4_{69.6-75.1}$	$45.6_{41.1-50.1}$	$59.0_{53.9-64.0}$	$73.3_{70.7-75.8}$	$47.6_{43.2-51.8}$	$60.4_{57.6-63.3}$	$2.74_{\pm0.09}/2.37_{\pm0.08}$	$2.65_{\pm0.06}$
Qwen2-VL	$75.4_{72.8-78.1}$	$\mathbf{48.6}_{44.1-53.1}$	$62.0_{57.0-67.0}$	$67.9_{65.0-70.7}$	$45.6_{41.4-49.5}$	$56.8_{53.9-59.6}$	$2.73_{\pm0.09}/2.25_{\pm0.09}$	$2.67_{\pm0.06}$
Qwen2.5-VL	$51.7_{48.3-55.1}$	$42.1_{38.2-46.0}$	$46.9_{41.7-52.1}$	$47.5_{43.8-51.1}$	$41.3_{37.6-45.0}$	$44.4_{41.4-47.4}$	$2.87_{\pm0.11}/2.26_{\pm0.09}$	$2.63_{\pm0.06}$
VideoChatR1	$75.9_{73.2-78.5}$	$48.0_{43.4-52.5}$	$61.9_{56.9-66.9}$	$74.7_{72.3-77.1}$	$\mathbf{48.4}_{43.7-52.6}$	$\mathbf{61.6}_{58.5-64.5}$	$2.86_{\pm0.09}/2.28_{\pm0.09}$	$2.52_{\pm0.06}$

Table 4: Task evaluation results for **US** and **Chinese** cultural norms. Highest values per column are bolded separately for each culture. 95% CIs in the underscore.

et al., 2023) to achieve long-context understanding.

- `InternVL-3` (Zhu et al., 2025) and `InternVL-3.5` (Wang et al., 2025) use variable visual position encoding (Ge et al., 2024) for longer multimodal contexts as well as a native multimodal pre-training approach.
- `Qwen2-VL` (Wang and et al., 2024) and `Qwen2.5-VL` (Qwen et al., 2025) employ dedicated Multimodal Rotary Position Embeddings (M-RoPE) to represent temporal and spatial information, enabling the model to comprehend dynamic video content.
- `VideoChatR1` (Li et al., 2025) uses Reinforcement Fine-Tuning (RFT) with GRPO (Shao et al., 2024) for video MLLMs to achieve state-of-the-art performance on spatio-temporal perception tasks.

Each model is evaluated under identical protocols per task and language: for segment-level tasks, inputs include localized timestamps and aligned transcripts, while evidence and norm-induction tasks require structured outputs (labels with verbal/nonverbal rationales or specific-norm statements). We do not apply task-specific fine-tuning; hyperparameters and inference settings are detailed in Appendix C.2 for reproducibility.

4.4 Results and Findings

Table 4 reports evaluation results for US and Chinese norms across the three tasks. We highlight the following key observations:

Detecting norm violation is more challenging than adherence For the CN subset, all models show significantly lower performance on the norm violation detection task compared to adherence detection. For the US subset, only 2 models showed a slightly higher performance for adherence detection compared to violation detection.

Models perform worse on CN culture compared to US All models performed worse or within 1 F1 score improvement on the CN subset compared to the US one on the adherence/violation identification task (Task 1), indicating the over-alignment with WEIRD cultures (Mihalcea et al., 2025) similarly to the teacher model. Notably, the gap is not improved with models specifically advertising multilingual performance, such as Qwen (5.9 average F1 gap for `Qwen2-VL` and 9.2 gap for `Qwen2.5-VL`) and InternVL (19.2 and 11.4 gap for Intern3-VL and Intern3.5-VL, respectively).

Non-verbal evidence is harder to extract than verbal evidence, indicating a lack of understanding of non-verbal cues in videos. Non-verbal evidences received lower scores on average across all models and cultures. Across both verbal and non-verbal scores, the average score for all models falls below 3 (on the scale of 1-5) for both Task 2 (evidence extraction) and Task 3 (norm identification), indicating unsatisfactory responses in the majority of cases.

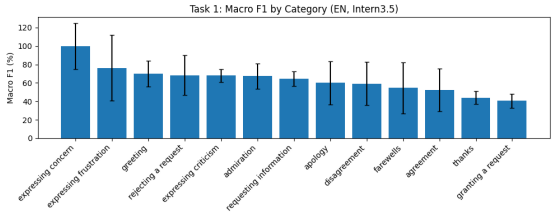
Models perform worse in formal cultural contexts Across US and CN cultures and across models, we see that for drama shows set in workplace

contexts the violation detection is worse than for more informal contexts, such as shows set in the comedy genre. Full comparison tables are in Appendix, Table 12 (US) and Table 13 (CN). Interestingly, *while humans achieve the highest inter-annotator agreement on such shows, models show the lowest performance.*

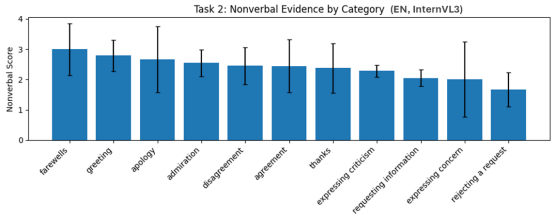
Models largely perform close to each other with confidence intervals overlapping significantly. For US norms, Intern3.5-VL on average has slightly better performance at classification tasks. For CN norms, Llava-OneVision is better at classification, VideoChatR1 is able to preserve classification performance while generating explanations, and Intern3-VL performs best at generation tasks: providing verbal and non-verbal evidence (Task 2) as well as identifying a specific norm (Task 3).

Error analysis For the best model for each task in the US subset, we explore how performance differs by norm category. The distribution of F1 scores by category for the best model for Task 1 is shown in Figure 3a, where it can be seen that THANKS and GRANTING A REQUEST categories received markedly lower scores compared to categories like EXPRESSING CONCERN, GREETING, REJECTING A REQUEST.

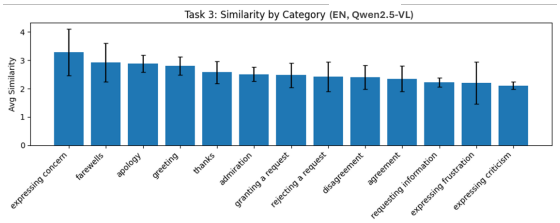
For non-verbal evidence scores (Figure 3b), US cultural norms that include hand gestures like FAREWELLS, GREETING achieve higher scores than those requiring complex emotional understanding such as REJECTING A REQUEST, EXPRESSING CONCERN, REQUESTING INFORMATION. The distribution per norm category for Task 3 performance (applicable category prediction) is displayed in Figure 3c, for which EXPRESSING CRITICISM, REQUESTING INFORMATION receive much lower scores compared GREETING, APOLOGY, FAREWELLS, EXPRESSING CONCERN. Overall, our dataset can help diagnose VideoLLMs' gaps in cultural norm understanding.



(a) Task 1 score distribution per category.



(b) Task 2 score distribution per category.



(c) Task 3 score distribution per category.

Figure 3: F1 score distributions with 95% CIs by norm category for the US, for Tasks 1, 2 (non-verbal), 3.

5 Conclusions

To evaluate cultural competence of VideoLLMs, we introduce a benchmark VIDEONORMS, containing video clips from popular US and Chinese TV shows annotated for cultural norms, adherence and violation labels and verbal and non-verbal evidence. The dataset was constructed via speech act theory prompting with a strong teacher VideoLLM and subsequent edits by 3 trained annotators for every instance. Analysis of annotator disagreements showed that the teacher model was less accurate for Chinese culture, and inter-annotator agreement was highest for shows set in more formal contexts (drama genre set in workplace compared to comedies). We benchmark a variety of open-source VideoLLMs and find the following trends: 1) detecting norm violations is more challenging than adherence; 2) models perform worse on Chinese culture compared to US culture; 3) models have more difficulty in providing non-verbal evidence compared to verbal for the norm adhere/violation label and struggle to identify the exact norm corresponding to a speech-act; and 4) models perform worse in formal cultural contexts, unlike humans. Overall, we hope our research serves as an important step towards understanding the cultural capabilities of video models.

6 Limitations

Cultural norm understanding is a nuanced topic and all research on it would have certain limitations. One of them is the ecological fallacy (Brewer and Venaik, 2014), or incorrect inference of individual-level traits based on aggregated national-level cul-

ture data. We note that we do not make any inferences about individuals but rather attempt to capture a subset of each countries' cultural norms. To mitigate the issues of teacher model bias or inaccuracy (Bender et al., 2021), we theoretically motivate our prompt and ensure extensive human validation by hiring 3 annotators from the respective cultures to verify each of the generated instances. We note that due to inherent trade-offs between the number of annotators and the number of annotated instances, our annotator sample may not be representative of a particular culture due to the intra-cultural variation and demographic factor impacts (Plepi et al., 2022; Wan et al., 2023), however, we mitigated this as much as possible by ensuring every annotator passes the screening questions (see Section B.1) and by including disagreements in our evaluation settings where practical (see Section 4.2). Moreover, since this is not a longitudinal study, we do not capture norm variation across time. Finally, while the shows that were picked for the study are broadly popular in respective cultures, they may not capture all possible cultural norms. Therefore, we note that our dataset is not meant to be a final collection of cultural norms for any country, but rather a proxy or a subset of such norms aiding contemporary evaluation efforts. As with any benchmark, however, a potential risk is over-optimization to the metric, where high performance does not necessarily mean that the model is culturally competent. We hope that our dataset paves the way for larger-scale, representative, longitudinal annotations of cultural norms using our framework.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

American Psychological Association. 2025. Cultural norm. https://dictionary.apa.org/cultural-norm. APA Dictionary of Psychology. Retrieved 2025-10-03.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Paul Brewer and Sunil Venaik. 2014. The ecological fallacy in national culture research. *Organization Studies*, 35(7):1063–1086.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 2025-10-03.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

J.L. Fleiss and 1 others. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Yi R. Fung and Heng Ji. 2025. Normlens: Massively multicultural MLLM reasoning with fine-grained social awareness. In *First Workshop on Social Simulation with LLMs*.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. 2024. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *Preprint*, arXiv:2412.09616.

Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, and 1 others. 2011. Differences between tight and loose cultures: A 33-nation study. *science*, 332(6033):1100–1104.

Jacob Gildenblat. 2023. A python library for confidence intervals. https://github.com/jacobgil/confidenceinterval.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32213–32230, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations (ICLR)*. ETHICS benchmark.

Roy S. Hessels, Toshiki Iwabuchi, and Diederick C. Niehorster. 2025. Gaze behavior in face-to-face interaction: A cross-cultural investigation between japan and the netherlands. *Cognition*, 263:106174.

G. Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Ronald F. Inglehart. 2018. *Cultural Evolution*. Cambridge University Press.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T. Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, Jon Borchardt, Taylor Sorensen, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2025. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, 7:145–160.

Lei Lai, Elyssa Besen, Natalia Sarkisian, and Qingwen Xu. 2022. A sino-u.s. comparison on workplace flexibility: evidence from multinational firms. *The International Journal of Human Resource Management*, 33(3):561–593.

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

Kun Li and et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *Preprint*, arXiv:2504.06958.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Linguistic Data Consortium. 2022. Ccu ta1 mandarin/chinese development annotation. LDC Catalog No. LDC2022E18.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28657–28670.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stańczak, and Aishwarya

Agrawal. 2025. Culturalframes: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. *Preprint*, arXiv:2506.08835.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2022. Extracting cultural commonsense knowledge at scale. *arXiv preprint arXiv:2210.07763*.

Hio Tong Pang, Xiaolin Zhou, and Mingyuan Chu. 2024. Cross-cultural differences in using nonverbal behaviors to identify indirect replies. *Journal of Nonverbal Behavior*, 48:323–344.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. *arXiv preprint arXiv:2210.14531*.

Edoardo Maria Ponti and et al. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Rachel Rakov and Andrew Rosenberg. 2013. "sure, i did the right thing": a system for sarcasm detection in speech. In *Interspeech 2013*, pages 842–846.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. 2025. EgoNormia: Benchmarking physical-social norm understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19256–19283, Vienna, Austria. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. *Preprint*, arXiv:2506.07032.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. 2022. Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores. *Applied Intelligence*, 52(5):4961–4972.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 769–779, Seattle, Washington, USA. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. yeah right: sarcasm recognition for spoken dialogue systems. In *Interspeech 2006*, pages paper 1821–Wed2BuP.13.

Paul Vicol, Makarand Tapaswi, Lluis Castrejón, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruyuan Wan, Alireza Mohammadshahi, Simran Arora, Dhairya Shah, Daniel Khashabi, Greg Durrett, Eduard Hovy, Amin Mirzaei, Byron C. Wallace, and Eunsol Choi. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Peng Wang and et al. 2024. Qwen2-vl: Enhancing vision-language model's visual perception and understanding. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Wei Wei and Tony Royle. 2024. Confucian values and zero-hour contracts: Sensemaking in workplace regimes at mcdonald's in china and the uk. *Economic and Industrial Democracy*, page 0143831X241265065.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

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)*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Caleb Ziems and et al. 2023. Normbank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Candidate Data Details

A.1 Video Clipping and Segmentation

We collected 5-7 clips of 2-3 minutes each from YouTube for every TV show in our dataset. YouTube clips were selected instead of full episodes because they are easier to download and process into smaller segments. After collection, a Python script was used to divide each 2-3 minute clip into 15-second sub-clips. This length was chosen because a 15-second segment typically contains one distinct social norm, allowing the model to focus on a single interaction or event. Shorter segments also help Gemini generate more precise and detailed norm outputs, which are described in the following section.

A.2 Prompting details

Table 5 shows the full speech act theory-based prompt for candiate cultural norm annotation.

Table 5: Speech Act Theory Cultural Norm Annotation Prompt

US Prompt	Chinese Prompt
Role: You are a culturally aware system with knowledge of norms in US culture. Your task is to analyze video content and detect instances where specific social norms occur, adhering to cultural expectations in the US. You will provide a timestamp for each instance, determine whether the character adheres to or violates the norm, and explain your assessment based on both verbal and nonverbal cues. Generate the norm based on the verbal/nonverbal cues. Instructions: If norm_category is chosen to be Custom category, replace the norm_category with the generated norm category. Do not output Custom category. Norm Categories: 1. Apology: Formal Context: business conversation between colleagues Example Specific Norm: In American business settings, use formal apology like "I am truly sorry for...“ or "I take full responsibility for..." and maintain eye contact to convey attentiveness and interest . Casual Context: conversation between college students Example Specific Norm: For minor disturbances or inconveniences, a common phrase used to express apologies in American culture can be "I'm sorry" "Oops", "My bad" or "Sorry about that." 2. Greeting: Formal Context: business meeting between executives Example Specific Norm: In American business settings, use formal greetings like "Hello", "Good morning", "Good afternoon", and "Good evening". Handshake is also common. Casual Context: greetings between friends Example Specific Norm: In American casual settings, use informal greetings like "Hey", "Hi", "Hello", "How are you?" Or "What's up". 3. Thanks: Formal Context: business meeting between a junior and senior employee Example Specific Norm: In American formal culture, expressing gratitude can be through phrases like "I sincerely appreciate your time/effort/help" or "I am truly grateful for your consideration" while maintaining eye contact to convey attentiveness and interest. Casual Context: conversation between friends Example Specific Norm: For minor disturbances or inconveniences, common phrases used to express grateful in American culture can be "thanks a lot," "thanks a bunch," "you're the best," "you rock," or "I appreciate it,"	角色说明: 你是一个具备中国文化规范意识的系统。你的任务是分析视频内容，检测是否出现了特定的社会规范，并判断是否符合中国文化的预期。你需要为每个实例提供时间戳，判断角色是否遵守规范或违反规范，并根据语言和非语言线索解释你的判断。你还需根据这些线索生成对应的社会规范。 指引说明: 如果 规范类别 被设为“自定义类别”，请用你生成的规范类别替换 规范类别，**不能输出“自定义类别”**。 社会规范类别: 1. 道歉: 正式场合: 同事之间的商务对话 示例规范: 在中国商务场合，正式的道歉用语如“非常抱歉给您带来了困扰，我会立即改正”或“对于此次给您带来的不便，我们深感歉意”，有时还会伴随轻微鞠躬。 非正式场合: 大学生之间的对话 示例规范: 对于轻微打扰或造成不便，中国人常说“对不起”“不好意思”“原谅我啦 ”或“我错了”。 2. 问候: 正式场合: 高管之间的商务会议 示例规范: 在正式场合中，人们常说“您好”、“您贵姓？”、“早上好”、“久仰大名”等，搭配握手、点头或轻微鞠躬；年长者或较高职位者之间还可能双手握手以示尊重。 非正式场合: 朋友之间的寒暄 示例规范: 常见用语有“你好”、“嗨”、“早啊”、“最近怎么样？”或“吃了吗？”。 3. 表示感谢: 正式场合: 初级与高级员工之间的会议 示例规范: 正式表达感谢时常用“谢谢您”、“承蒙关照”等，并伴随点头或轻微鞠躬。在递交名片、文件或礼物时，需双手递交。 非正式场合: 朋友之间的交流 示例规范: 常用“谢谢”、“麻烦你了”、“多谢”、“谢啦”或“辛苦啦”。

US Prompt	Chinese Prompt
4. Admiration: Formal Context: colleagues offer feedback Example Specific Norm: In American culture, a formal compliment typically focuses on specific achievements, skills, or contributions, using phrases like "I was impressed by your presentation" or "Thank you for your hard work on this project." [0.5em] Casual Context: conversation between friends Example Specific Norm: In American culture, casual compliments are often more personal and emotional, such as "You always know how to make me smile" or "I really admire your creativity." [0.5em] 5. Requesting Information: Formal Context: meeting between employee and supervisor Example Specific Norm: Asking coworker's personal finances. Questions about a person's salary, wealth, or how much things cost are considered an invasion of privacy and very rude. Casual Context: conversation between friends Example Specific Norm: In American culture, casual requests are typically made directly and with a casual tone, often using "could you" or "would you mind" followed by the request. 6. Granting a Request: Formal Context: conversation with a colleague Example Specific Norm: Agree to the request with a positive response such as, "Of course, I'd be happy to help," accompanied by a nod or smile to reinforce willingness. Casual Context: conversation between siblings Example Specific Norm: In American casual settings, granting requests are typically positive tone, often using "of course" or "no problem". 7. Disagreement: Formal Context: business discussion between employees Example Specific Norm: Use respectful language such as "I see your point, but I'd like to offer another perspective," while maintaining a calm tone and open body language to show that the disagreement is friendly and constructive. Casual Context: conversation between siblings Example Specific Norm: In American culture, informal disagreements are often handled with directness and a focus on resolving the issue, rather than avoiding confrontation or resorting to indirect language. 8. Agreement: Formal Context: business meeting Example Specific Norm: Use respectful language such as "I second that motion" or "I concur with that statement" for endorsement and approval. Casual Context: conversation between siblings Example Specific Norm: In casual settings with friends, agreement is usually expressed in a relaxed and informal way. Some ways of describing that can be "Yeah, totally!", "For sure!", or "Absolutely!" 9. Farewells: Formal Context: business meeting Example Specific Norm: Use respectful language such as "Goodbye", "Until next time", "Farewell" or "Take care". Casual Context: conversation between siblings Example Specific Norm: In casual settings with friends, farewell is usually expressed in a relaxed and informal way. Some ways of describing that can be "See ya!", "Bye!", or "Later!"	4. 表达赞赏: 正式场合: 同事之间的反馈交流 示例规范: 赞美多围绕专业素养、能力或成果，如“从您身上学到了很多”、“您的专业素养令人钦佩”。会议中向上级或嘉宾致意时应起立表达敬意。 非正式场合: 朋友之间的对话 示例规范: 赞赏可较为轻松幽默，如“你太厉害了！”、“哇，真牛！”等。 5. 请求信息: 正式场合: 员工与主管之间的会议 示例规范: 表达时较为委婉，例如“方便的话.....”、“我想了解一下.....”或“打扰一下.....”，语气应礼貌客气。 非正式场合: 朋友之间的对话 示例规范: 常用“你知道.....吗？”、“我可以问你个事吗？”等温和语气。 6. 同意请求: 正式场合: 与同事的交流 示例规范: 常用“可以的，我会协助您完成”或“好的，我来处理”，并辅以微笑或点头。 非正式场合: 兄妹或朋友之间 示例规范: 常用“行啊”、“没问题”、“好说好说”等语句。 7. 表达异议: 正式场合: 商务讨论 示例规范: 用“我理解您的意思，不过我有些不同的考虑”或“是否可以从另一个角度考虑？”表达异议，语气应平和，肢体语言开放，突出理性沟通。 非正式场合: 兄妹或朋友间 示例规范: 通常直接表达不同意见，如“我不这么觉得”、“你这想法不太行”。 8. 表示同意: 正式场合: 商务会议 示例规范: 用“我同意”、“好的”、“没问题”等语言表示认同。 非正式场合: 日常对话 示例规范: 常用“对！”、“嗯”、“可以啊”之类词语表示认可。 9. 道别: 正式场合: 商务场合结束时 示例规范: 常用“再见”、“保持联系”、“改日再聊”、“祝您一路顺风”等。 非正式场合: 朋友之间 示例规范: “拜拜”、“走啦”、“回头见”等。

US Prompt	Chinese Prompt
10. Rejecting a Request: Formal Context: conversation with a colleague Example Specific Norm: A formal way to reject a request is to be polite, clear, and professional. Express appreciation first, then provide a clear but polite rejection, or offer a brief explanation or suggest an alternative. Casual Context: conversation between siblings Example Specific Norm: In American casual settings, rejecting requests can be more relaxed but still polite and considerate. Some rejection might use humor if appropriate. 11. <Custom category>: When the above categories do not apply, generate a new norm category. Replace the C̆ustom categoryẅith the newly generated norm category. Do not output C̆ustom categoryäs the norm_category. Example: norm_category: Expressing criticism Casual Context: business meeting Example Specific Norm: Offer any criticism in a way that emphasizes a person’s strengths and highlights ways they could easily improve. 12. No norm: When no social norm can be applied. Output Format (JSON): “‘json ["timestamp": "start": "MM:SS", "end": "MM:SS" , “context": “Brief description of the setting and hierarchy between the participants", "norm_category": "Category from the above list of norms", “norm_subject": “Specifically to whom the norm is applied to in this context (no character name)", “specific_norm": “Specific norm in the norm category that is applicable to this context like the Example Specific Norm", "norm_adherence": "adherence/violation", "explanation": "verbal_evidence": "Description of verbal cues supporting adherence/violation.", "nonverbal_evidence": "Description of nonverbal cues supporting adherence/violation." , ...] """	10. 拒绝请求: 正式场合: 工作中回绝他人请求 示例规范: 避免直接说“不”，可用“目前这个时间点不太合适”、“我需要先请示领导”等委婉方式拒绝，同时表达未来愿意协助的态度。 非正式场合: 与熟人交流 示例规范: “这个好像不太方便”、“我现在有点忙”、“再看看吧”。 11. 自定义类别: 当以上类别不适用时，**可根据情境生成一个新的社会规范类别，并将其填入规范类别 字段**。**不能输出“自定义类别”**。 示例: 规范类别: 提出批评 非正式场合: 商务会议 示例规范: 提出批评或反馈时应使用礼貌、含蓄、不伤面子的表达方式，体现对“和谐”、“人情”与“面子”的重视。 12. 无规范: 当场景中不存在任何适用的社会规范时使用此类别。 输出格式: “‘json ["时间戳": "开始": "MM:SS", "结束": "MM:SS" , "情境描述": "对场景和参与者间地位关系的简要描述", "规范类别": "上方规范类别之一", "行为主体": "此规范执行于谁（不使用具体角色名）", "具体规范": "在该情境中应遵循的具体社会规范", "规范遵循情况": "遵守 或 违反", "解释说明": "语言证据": "支持该行为是否符合规范的语言线索", "非语言证据": "支持该行为是否符合规范的非语言线索" , ...] """

A.3 Candidate Data Statistics

Table 8 shows detailed distribution of candidate data annotations.

Total number of norm instances detected		
	US	**China**
(Video Clip, Norm) Pairs	514	501
Distribution Across Norm Categories		
Requesting Information	37.18%	17.06%
Admiration	16.25%	12.97%
Greeting	13.72%	16.38%
No Norm	7.94%	15.36%
Agreement	4.33%	6.14%
Thanks	4.33%	6.83%
Farewells	3.97%	1.37%
Apology	3.97%	3.75%
Rejecting a Request	3.61%	5.12%
Granting a Request	2.53%	0.34%
Disagreement	2.17%	14.68%
Custom Categories	46.11%	41.52%
Expressing Criticism	71.73%	23.08%
Expressing Concern (US) / Invitation (CN)	2.95%	3.37%
Expressing Frustration (US) / Suggestion (CN)	1.27%	1.92%
Gift Giving	0.84%	1.44%
Reassurance (US) / Business Language (CN)	0.84%	1.44%
Adherence vs. Violation		
Adherence	58.95%	61.08%
Violation	35.99%	28.54%
N/A	5.06%	10.38%

Table 8: Summary of teacher model annotations (with Top 5 custom categories).

A.4 Annotator Change Percentages by Field

Table 9 shows percentages of changes per field by U.S. and Chinese annotators.

Field	US (%)	China (%)
timestampStart	2.40	4.86
timestampEnd	3.05	11.78
context	4.99	56.82
normCategory	14.59	50.83
normActors	29.44	53.83
specificNorm	16.99	57.88
normAdherence	13.62	26.28
verbalEvidence	23.35	64.54
nonverbalEvidence	19.71	53.89

Table 9: Overall Percentage of Changes per Field by U.S. and Chinese Annotators

B Annotation Details

B.1 Screening Questions and Answers

Below are the screening questions used for both American and Chinese annotators:

- Were you raised monolingual?
- What's your earliest language in life, and what's your primary language?
- Do you identify yourself as monocultural or multicultural?
- What's your highest education level completed?
- What is your country of birth? How many years have you lived in your current country of residence?

For the U.S. annotators, all three were raised monolingual. English is the first and primary language of all three annotators. All three identify as monocultural and have completed a Bachelor's degree. All were born in the United States and have lived there for their entire lives.

For the Chinese annotators, all three were raised monolingual. Mandarin is the first and primary language of all three annotators. All three identify as monocultural and have completed a Bachelor's degree. All were born in China and have lived there for their entire lives.

B.2 Annotator Survey

After the annotators received their offers, we asked them to complete a short follow-up survey. This survey included questions about their demographic background and prior exposure to the shows used in the study. Specifically, annotators were asked to provide their age, indicate which of the four shows they had previously watched, and rate their familiarity with each show.

The U.S. annotators were 27, 28, and 29 years old, while the Chinese annotators were between 25, 28, and 29 years old.

For the U.S. annotators, two had watched *Friends* and *Big Bang Theory*, while all three had watched *The Office* and *Suits*. Of the two annotators who had watched *Friends*, one gave a familiarity rating of $5/5$ and the other gave a $2/5$. For *Big Bang Theory*, both annotators gave a familiarity rating of $5/5$. For *The Office*, two annotators gave a familiarity rating of $5/5$, and the third gave a $4/5$. For *Suits*, two annotators gave a familiarity rating of $5/5$, and the third gave a $3/5$.

For the Chinese annotators, all three had

watched *iPartment* and *Home with Kids*, while only one had watched *Amazing Night* and none had watched *Best Partner*. For *iPartment*, two annotators gave a familiarity rating of $5/5$, and the third gave a $4/5$. For *Home with Kids*, the first annotator gave a familiarity rating of $5/5$, the second gave a $4/5$, and the third gave a $3/5$. The annotator who had watched *Amazing Night* gave a familiarity rating of $3/5$.

B.3 Annotation Instructions and User Interface

Figure 4 shows the detailed instructions provided to the annotators. Figure 5 shows the annotation interface.

C Experimental details

C.1 Open-Weight Model Inference prompts

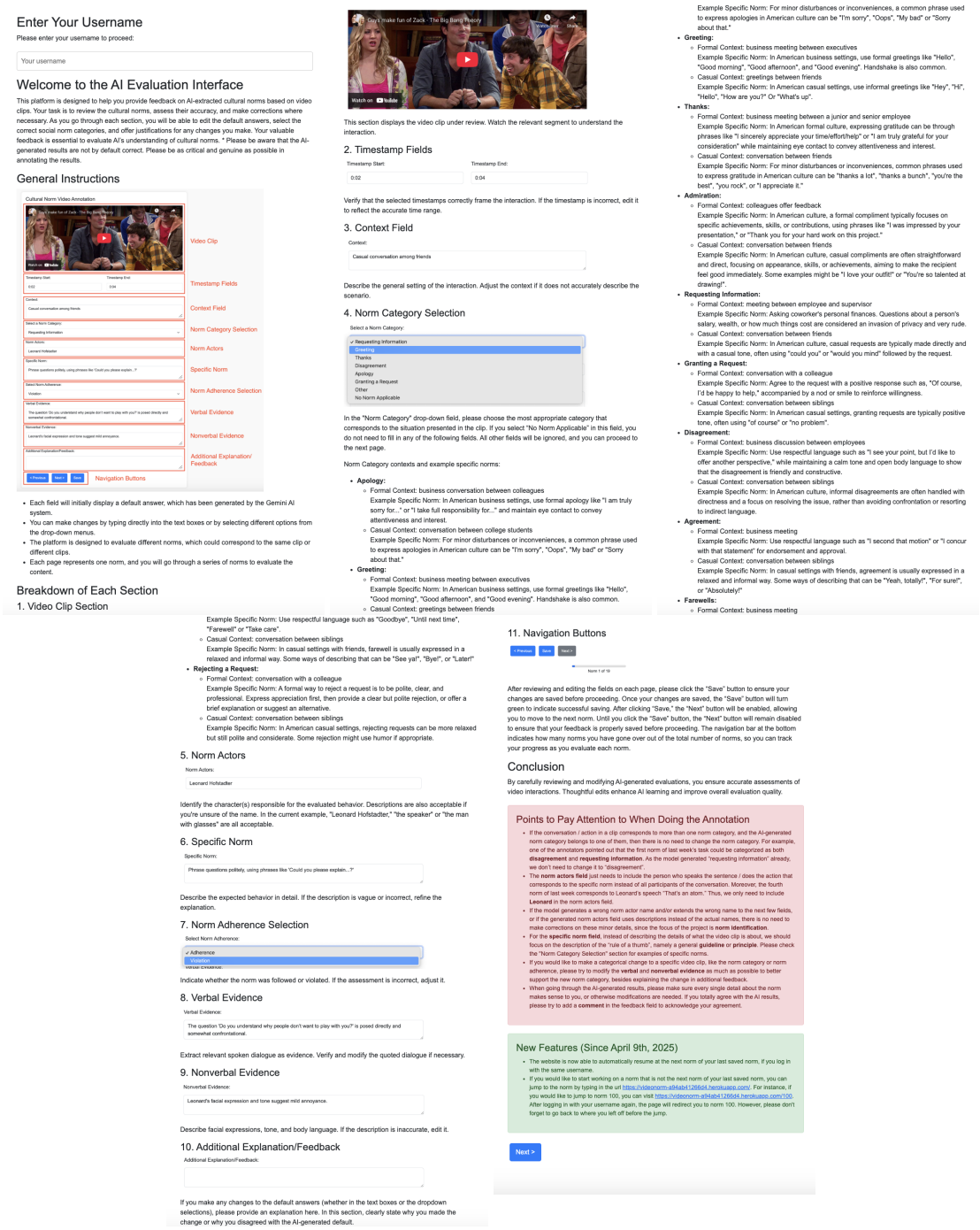


Figure 4: Detailed instructions provided on the first page of the user interface. The page is cut into 5 screenshots, from left to right, in this figure.

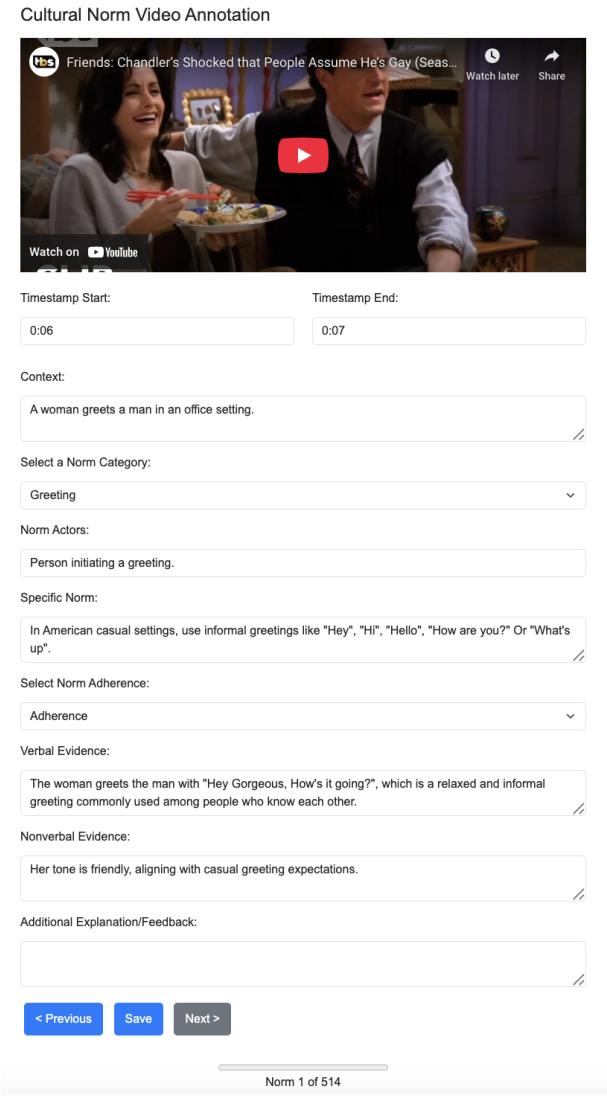


Figure 5: User Interface shown to annotators. It comprises of a 15 second clip with an option to edit the predicted norm from Teacher model.

Table 10: Prompt templates for cultural norm analysis tasks in Chinese and English contexts.

Culture	Task	Prompt Template
CN	1	您是一个具有中国社会规范和文化意识的系统，请用中国文化背景来分析视频内容。 注意：下面提供了片段开始时间，请基于该时间点的视频片段进行分析，而不是整个视频。 任务：根据视频、视频文本、给定文化规范类别和具体的文化规范，判断该规范是否被遵守或违反。 给定片段开始时间：{ts} 给定视频文本：{transcript} 给定规范类别：{norm_category} 给定具体规范：{specific_norm} 请根据视频内容写出 "遵守" 或 "违反"。 输出格式： 规范类别：{norm_category} 遵守/违反：[遵守/违反]
CN	2	您是一个具有中国社会规范和文化意识的系统，请用中国文化背景来分析视频内容。 注意：下面提供了片段开始时间，请基于该时间点的视频片段进行分析，而不是整个视频。 任务：根据视频、视频文本、给定文化规范类别和具体的文化规范，判断该规范是否被遵守或违反，并根据视频内容提供简要的语言和非语言证据来支持你的判断。 给定片段开始时间：{ts} 给定视频文本：{transcript} 给定规范类别：{norm_category} 给定具体规范：{specific_norm} 输出格式： 规范类别：{norm_category} 遵守/违反：[遵守/违反] 语言证据：[来自视频的简短语言证据，例如：说的话、语气等] 非语言证据：[来自视频的简短非语言证据，例如：肢体语言、表情等]
CN	3	您是一个具有中国社会规范和文化意识的系统，请用中国文化背景来分析视频内容。 注意：下面提供了片段开始时间，请基于该时间点的视频片段进行分析，而不是整个视频。 任务：根据视频、视频文本、给定文化规范类别，请提供一个与该类别相关的具体规范。 给定片段开始时间：{ts} 给定视频文本：{transcript} 给定规范类别：{norm_category} 输出格式： 规范类别：{norm_category} 具体规范：[来自视频的简短具体规范]
US	1	You are a culturally aware system with knowledge of US social norms. Analyze video content using US cultural context. Note: The timestamp start is included so that your analysis is tied to the correct segment of the video, not the entire clip. Task: Given the video, its transcript, cultural norm category and a specific cultural norm, determine if the norm is adhered to or violated. Given Timestamp Start: {ts} Given Transcript: {transcript} Given Norm Category: {norm_category} Given Specific Norm: {specific_norm} Output format: Timestamp Start: {ts} Norm Category: {norm_category} Adherence/Violation: [Adherence/Violation]

Continued on next page

Table 10 (continued): Prompt templates for cultural norm analysis tasks

Language	Task	Prompt Template
US	2	You are a culturally aware system with knowledge of US social norms. Analyze video content using US cultural context. Task: Given the video, its transcript, cultural norm category and a specific cultural norm, determine if the norm is adhered to or violated, and provide verbal and nonverbal evidence for your decision based on the video. Given Timestamp Start: {ts} Given Transcript: {transcript} Given Norm Category: {norm_category} Given Specific Norm: {specific_norm} Output format: Timestamp Start: {ts} Norm Category: {norm_category} Adherence/Violation: [Adherence/Violation] Verbal Evidence: [Verbal evidence from the video, e.g., spoken phrases, tone] NonVerbal Evidence: [NonVerbal evidence from the video, e.g., body language, expressions]
US	3	You are a culturally aware system with knowledge of US social norms. Analyze video content using US cultural context. Task: Given the video, its transcript and cultural norm category, provide a specific norm related to the below category exhibited in the video. Given Timestamp Start: {ts} Given Transcript: {transcript} Given Norm Category: {norm_category} Output format: Timestamp Start: {ts} Norm Category: {norm_category} Specific Norm: [Brief specific norm from the video]

C.2 Hyperparameters

We evaluate models in the 7B-8B parameter range to ensure fair comparison across architectures: `LLaVA-Next-Video-7b-hf`, `Llava-Onevision-qwen2-7b`, `InternVL3-8B`, `InternVL3.5-8B`, `Qwen2-VL-7B-Instruct`, `Qwen2.5-VL-7B-Instruct`, and `VideoChat-R1-7B`. All models are loaded in bfloat16 precision using Flash Attention 2 for efficient inference, with each model deployed on a single GPU. For video processing, we uniformly sample 32 frames from each clip as input. We employ an adaptive temperature strategy: generation begins at temperature 0.0 for deterministic outputs, and if the response fails to conform to the required task structure (e.g., missing labels or evidence fields), we incrementally increase temperature by 0.2 per retry attempt. This approach balances output consistency with generation flexibility when models struggle with structured formatting requirements.

C.3 LLM-as-a-Judge Prompts

Table 11 shows the prompts used for the LLM-as-a-Judge GPT-5 evaluator.

Task	Prompt
Norm Similarity (Task 3)	For norm similarity: Your task is to evaluate cultural awareness of an AI model. Grade how well the generated norm matches the reference norm. Use the following rubric: 1 - Neither the setting is captured and the norm content is completely different. 2 - Captures some of the norm content but the setting is very different. 3 - Captures some of the norm content and the setting. 4 - Captures the setting and the essential content, but minor differences exist, like differences in wording. 5 - Captures the setting and perfectly captures the content of the norm without introducing hallucinations. First output your reasoning, then your final verdict after the words "Final Verdict:" Reference norm: {ref_norm} Generated norm: {gen_norm}
Verbal Evidence Similarity (Task 2)	For verbal: Your task is to evaluate cultural awareness of an AI model. Grade how well the generated verbal evidence matches the Gold verbal evidence. Use the following rubric: 1 - Neither the content of characters' speech is captured, nor the reasoning from the gold evidence. 2 - Does not capture the content, but contains a similar line of reasoning as the gold evidence. 3 - Essentially captures the content, but minor differences exist, and no reasoning is provided. 4 - Essentially captures the content, but minor differences exist, and contains the same line of reasoning as the gold evidence. 5 - Perfectly captures the content of the speech without introducing hallucinations and contains the same line of reasoning as the gold evidence. First output your reasoning, then your final verdict after the words "Final Verdict:". Gold verbal evidence: {gold} Generated verbal evidence: {gen}
Non-verbal Evidence Similarity (Task 2)	For non-verbal: Your task is to evaluate cultural awareness of an AI model. Grade how well the generated non-verbal evidence matches the Gold non-verbal evidence. Use the following rubric: 1 - Neither the characters' actions or emotions are captured, nor the reasoning from the gold evidence. 2 - Does not capture actions or emotions, but contains a similar line of reasoning as the gold evidence. 3 - Essentially captures actions or emotions, but minor differences exist, and no reasoning is provided. 4 - Essentially captures actions or emotions, but minor differences exist, and contains the same line of reasoning as the gold evidence. 5 - Perfectly captures actions or emotions without introducing hallucinations and contains the same line of reasoning as the gold evidence. First output your reasoning, then your final verdict after the words "Final Verdict:". Gold verbal evidence: {gold} Generated verbal evidence: {gen}

Table 11: LLM-as-a-Judge prompts to compare the similarity of generations with dataset references.

C.4 Evaluation breakdown by shows

Table 12 shows the US results broken down by show, and Table 13 shows the CN results.

		Task 1			Task 2					Task 3	
Model	**Show**	**F1 (pos)**	**F1 (neg)**	**Samples**	**F1 (pos)**	**F1 (neg)**	**Verbal**	**Nonverbal**	**Samples**	**Score**	**Samples**
Llava-Next-Video	friends	74.9	23.8	204	56.7	66.4	2.275	1.985	214	2.124	210
	bbt	73.8	11.1	158	61.3	65.2	2.303	1.936	172	2.123	162
	office	70.1	20.0	184	57.9	67.3	2.157	2.043	182	2.223	189
	suits	82.2	16.4	174	57.9	46.5	1.882	2.215	176	2.500	178
Intern3-VL	friends	62.0	63.5	204	58.8	65.5	2.771	2.389	213	2.262	210
	bbt	55.9	63.6	158	51.1	66.0	2.680	2.184	172	2.198	162
	office	67.1	71.1	182	59.0	68.0	2.640	2.439	178	2.446	185
	suits	37.0	45.2	174	28.2	47.3	2.797	2.217	176	2.270	178
Intern3.5-VL	friends	80.2	66.2	204	69.5	63.6	2.773	2.177	214	2.257	210
	bbt	73.9	65.2	161	65.1	67.4	2.517	2.103	175	2.248	165
	office	77.2	75.1	181	68.3	72.3	2.714	2.230	179	2.529	188
	suits	69.5	51.9	174	59.0	55.6	2.485	2.396	176	2.466	178
Qwen2-VL	friends	75.6	59.7	204	66.7	59.8	2.925	2.134	214	2.495	210
	bbt	75.1	58.8	158	68.0	55.6	2.620	2.157	172	2.173	162
	office	72.6	73.5	182	63.4	69.4	2.775	2.042	180	2.325	187
	suits	69.2	54.3	174	70.0	51.1	2.618	2.009	175	2.299	177
Qwen2.5-VL	friends	62.7	69.1	204	47.9	66.7	2.608	2.040	214	2.294	211
	bbt	51.5	63.3	158	38.7	65.5	2.271	2.010	172	2.259	162
	office	55.9	71.5	182	37.5	65.5	2.480	2.190	180	2.473	187
	suits	23.6	46.1	174	12.0	46.6	2.508	2.051	176	2.522	178
Llava-Onevision	friends	79.1	41.1	204	77.9	49.2	2.774	2.055	214	2.052	211
	bbt	75.5	31.3	158	69.3	42.0	2.757	2.068	172	1.988	162
	office	80.3	66.7	182	76.4	66.2	2.623	2.085	180	2.129	187
	suits	80.1	28.9	174	79.2	41.3	2.211	2.358	176	2.124	178
VideoChatR1	friends	76.0	71.7	204	65.7	68.8	2.704	2.190	214	2.493	211
	bbt	69.0	64.9	158	61.0	64.4	2.713	2.176	172	2.185	162
	office	70.0	70.7	182	64.3	68.8	2.608	2.133	180	2.258	187
	suits	57.8	50.9	174	52.6	52.0	2.637	2.363	174	2.416	178

Table 12: US results per show

		Task 1			Task 2					Task 3	
Model	**Show**	**F1 (pos)**	**F1 (neg)**	**Samples**	**F1 (pos)**	**F1 (neg)**	**Verbal**	**Nonverbal**	**Samples**	**Score**	**Samples**
Intern3-VL	ipartment	33.3	46.1	275	24.7	45.2	2.730	2.119	279	2.611	275
	legal	36.6	18.9	253	30.8	19.2	3.030	2.758	259	2.573	253
	hwk	59.1	19.0	313	50.4	21.6	3.361	2.603	311	2.938	306
	amazing_night	17.4	53.6	271	11.6	53.2	2.453	2.142	273	2.629	272
Intern3.5-VL	ipartment	70.1	53.7	275	71.6	54.9	2.606	2.171	278	2.457	276
	legal	70.5	26.2	254	69.9	26.7	2.760	2.571	258	2.545	255
	hwk	84.6	30.1	313	84.1	32.2	2.891	2.467	310	2.766	312
	amazing_night	54.1	55.8	271	60.1	60.1	2.642	2.256	273	2.796	269
Qwen2-VL	ipartment	70.2	52.1	275	66.9	51.3	2.630	2.232	279	2.451	275
	legal	78.4	26.1	254	65.7	24.7	2.540	2.118	259	2.838	253
	hwk	87.1	32.0	313	83.2	31.1	3.022	2.449	311	2.666	305
	amazing_night	53.2	60.5	271	38.6	57.6	2.552	2.052	273	2.717	272
Qwen2.5-VL	ipartment	50.4	50.7	275	45.4	49.1	2.531	2.242	279	2.511	274
	legal	37.6	19.0	254	34.4	19.7	3.056	2.264	259	2.756	254
	hwk	70.8	26.0	313	67.0	26.0	3.254	2.518	311	2.630	311
	amazing_night	31.4	58.6	271	27.0	57.8	2.587	1.953	273	2.622	270
Llava-Next-Video	ipartment	85.2	16.9	274	85.8	35.6	1.796	1.444	249	1.757	276
	legal	95.1	0.0	247	94.8	8.7	1.741	1.426	212	1.690	252
	hwk	96.3	0.0	306	94.7	17.1	1.946	1.790	291	1.748	313
	amazing_night	77.3	2.0	266	78.3	23.1	1.529	1.474	236	1.796	269
Llava-Onevision	ipartment	73.8	52.3	275	60.9	46.6	2.653	2.170	279	2.330	276
	legal	87.1	24.3	254	70.9	21.4	2.521	1.987	259	2.323	254
	hwk	82.8	31.7	313	76.5	28.8	2.904	2.386	308	2.377	310
	amazing_night	63.9	58.4	269	46.7	57.0	2.433	1.971	269	2.590	268
VideoChatR1	ipartment	71.2	54.5	275	71.3	54.6	2.511	2.109	279	2.402	276
	legal	75.7	25.6	254	75.6	26.4	2.713	2.221	259	2.408	255
	hwk	87.6	36.9	313	85.3	37.3	3.331	2.504	311	2.751	309
	amazing_night	59.7	57.6	271	58.0	58.6	2.673	2.209	271	2.479	240

Table 13: Chinese results with breakdown by shows