CultureVidBench: Benchmarking Cultural Understanding in Text-to-Video Generation

Xianjing Han[1], Yuhan Su[2], Yang Deng[3], Dong Ma[4], Wee Peng Tay[1], Bin Zhu[3] †
[1] Nanyang Technological University [2] Centrale Supélec
[3] Singapore Management University [4] University of Cambridge
Correspondence: binzhu@smu.edu.sg
https://hanxjing.github.io/CultureVidBench/

Abstract

Text-to-video (T2V) generation models have advanced rapidly, yet their ability to represent diverse cultural contexts remains underexplored. Existing benchmarks mainly focus on perceptual quality, physical plausibility, and text-video alignment, but do not directly assess whether generated videos capture culturally specific objects, actions, rituals, visible text, or audio cues. We introduce CultureVidBench, a comprehensive benchmark for evaluating cultural understanding in T2V generation. CultureVidBench contains 1,000 curated prompts covering 12 countries, 6 continents, 8 cultural regions, and 14 cultural aspects organized into three categories: material culture, social practice & performance, and ritual & ceremony. Designed specifically for video generation, CultureVidBench emphasizes dynamic and multimodal cultural representation, including social interactions, ritual procedure, and culturally appropriate visible text and audio. We evaluate seven representative T2V models through human user studies and MLLM-based automatic assessment across cultural faithfulness, multimodal cultural rendering, semantic adherence, and perceptual quality. Results show that although current models achieve strong semantic adherence and visual quality, they often fail to faithfully capture fine-grained cultural details, particularly for underrepresented regions, rituals, and multimodal cultural cues.

1 Introduction

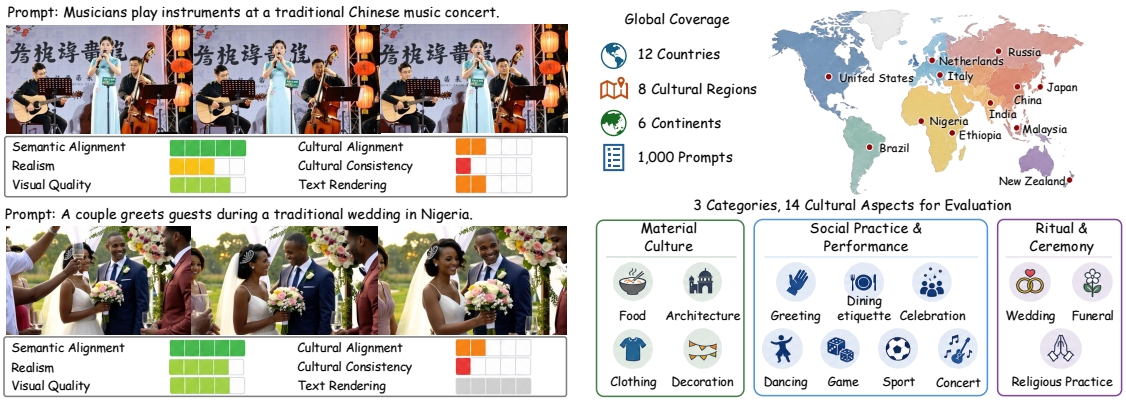
Recent advances in text-to-video (T2V) generation have enabled models to produce increasingly realistic, temporally coherent, and semantically aligned videos from natural language prompts (Wan et al., 2025; Google DeepMind, 2025; Alibaba Cloud, 2026). These models are rapidly adopted to support a growing range of applications, including filmmaking, social media content creation, advertising, and education (Zheng et al., 2025; Chen et al., 2026b). As generated videos become widely consumed across different societies, an important question emerges: *can T2V models faithfully understand and represent cultural contexts?*

Culture is deeply embedded in real-world videos. It shapes not only visible objects such as clothing, food, architecture, and decorations, but also human actions, social interactions, ritual procedures, performances, music, and symbolic practices (Geertz, 2017). Therefore, cultural understanding in T2V generation cannot be limited to general prompt following or visual realism. As shown in Figure 1(a), a generated video may appear visually plausible and semantically related to the prompt while still misrepresenting the intended culture. For example, a model may depict a traditional Chinese music performance using culturally inconsistent instruments or text, or generate a Nigerian wedding scene with generic attire and ceremonial elements. Such videos may score highly for visual quality and coarse semantic alignment, but they remain culturally inaccurate, misleading, or inappropriate for viewers in the target culture. These failures suggest that cultural understanding is a central requirement for trustworthy T2V generation, rather than a secondary aesthetic preference.

Despite its importance, cultural understanding remains largely underexplored in existing T2V evaluation. Current benchmarks primarily assess perceptual quality (Huang et al., 2024), motion consistency (Liao et al., 2024), physical plausibility (Gu et al., 2025), and text-video semantic alignment (Sun et al., 2025). While these dimensions are essential, they do not directly evaluate whether a generated video correctly captures culturally specific objects, behaviors, rituals, visible text, or audio cues. Recent cultural benchmarks for text-to-image generation have begun to evaluate cultural competence in static images (Kannen et al., 2024; Shi et al., 2025; Malakouti et al., 2026), but they cannot fully capture the video-specific nature

†Corresponding author and project lead.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 49814–49832
October 24-29, 2026 ©2026 Association for Computational Linguistics



(a) Sampled videos failed in cultural understanding.

(b) Evaluation coverage and cultural aspects of CultureVidBench.

Figure 1: Cultural failures in T2V generation and overview of CultureVidBench. (a) Examples of semantically aligned and visually plausible generated videos with cultural inaccuracies. (b) CultureVidBench cultural coverage.

of culture. Unlike images, videos require models to represent culturally grounded actions and interactions unfolding over time, often involving multiple modalities such as written language, music, chanting, and ambient sound.

To address this gap, we introduce CultureVidBench, a comprehensive benchmark for evaluating cultural understanding in text-to-video generation. As shown in Figure 1 (b), CultureVidBench covers 12 countries spanning 6 continents and 8 world cultural regions, and organizes cultural content into three major categories: material culture, social practice & performance, and ritual & ceremony. These categories further span 14 cultural aspects, such as food, architecture, greetings, games, weddings, and religious rituals. In total, CultureVidBench contains 1,000 curated prompts constructed from culturally grounded elements and activities in CulturalAtlas (Mosaica, 2024) and Wikipedia (Wikipedia, 2026). Particularly, unlike benchmarks that focus primarily on static cultural artifacts, CultureVidBench emphasizes dynamic and multimodal cultural representation, including culturally specific actions, social interactions, procedural rituals, visible text, and audio cues.

We further design a comprehensive evaluation framework for cultural video generation across four complementary dimensions: cultural faithfulness, multimodal cultural rendering, semantic adherence, and perceptual quality. Cultural faithfulness assesses whether the target cultural elements, practices, and contexts are faithfully represented, while also checking whether the video avoids introducing culturally inconsistent artifacts or symbols from unrelated contexts. Multimodal cultural rendering evaluates whether visible text and audio are appropriate for the target cultural scenario. Semantic adherence and perceptual quality are included to distinguish cultural failures from general prompt-following or low-quality generation failures. We evaluate seven representative T2V models using both human studies and MLLM-based automatic evaluation. To improve interpretability, we introduce a target-explicit evaluation protocol that explicitly identifies expected cultural elements from prompts before evaluating video evidence.

Experiments reveal several important findings: 1) While current T2V models often achieve strong semantic adherence and visual quality, they still struggle to faithfully capture culturally specific details. 2) This limitation is more pronounced in underrepresented cultural regions, where models perform substantially worse than in high-resource cultures. 3) Material culture is generally easier to generate than social practices and rituals, suggesting that current models are better at rendering cultural objects than culturally grounded actions and procedures. 4) Multimodal cultural rendering remains particularly challenging, especially for culturally appropriate visible text and audio.

Our contributions can be summarized as follows:

- We introduce CultureVidBench, a benchmark for evaluating cultural understanding in text-to-video generation across diverse countries, cultural regions and cultural categories.
- We propose a comprehensive cultural evaluation framework covering cultural faithfulness, multimodal cultural rendering, semantic adherence, and perceptual quality.
- We benchmark seven SOTA T2V models and reveal systematic cultural disparities, showing that current models remain biased toward better-represented cultures and struggle with culturally specific practices, rituals, and multimodal cues.

2 Related Work

Text-to-Video Generation Benchmarks. Recent advances in T2V generation have highlighted the need for systematic and reliable evaluation (Sun et al., 2025; Huang et al., 2025). Early benchmarks, such as VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024), assess T2V models from multiple general aspects, including visual quality, temporal consistency, motion quality, and text-video alignment. Beyond general quality assessment, recent benchmarks have examined more specific model capabilities. For example, DEVIL (Liao et al., 2024) focuses on dynamic properties in generated videos, and OSCBench (Han et al., 2026) evaluates object state change reasoning. Other benchmarks, including VideoPhy (Bansal et al., 2024), PhyGenBench (Meng et al., 2024), and PhyWorldBench (Gu et al., 2025), evaluate whether generated videos follow physical commonsense and plausible world dynamics. RoboTrustBench (Li et al., 2026) further evaluates video world models for robotic manipulation under normal, constraint-sensitive, counterfactual, and adversarial instructions. Despite these advances, existing T2V benchmarks still underexplore cultural understanding, which is essential for generating videos that are faithful and appropriate across diverse social contexts. To address this gap, we introduce CultureVidBench, a benchmark specifically designed for cultural evaluation in T2V generation.

Cultural Benchmarks. As generative models and vision-language models (VLMs) are increasingly deployed worldwide, evaluating their ability to represent diverse cultures has become an important research direction. Existing cultural benchmarks have studied this problem across several domains (Bhatt and Diaz, 2024; Chen et al., 2026a). In text-to-image generation, benchmarks such as CUBE-1K (Kannen et al., 2024), CULTIVate (Malakouti et al., 2026), ViSAGe (Jha et al., 2024), CultureBench (Shi et al., 2025), CulturalFrames (Nayak et al., 2025), and CuRe (Rege et al., 2025) evaluate whether generated images faithfully reflect cultural concepts, artifacts, and social contexts. In vision-language understanding, CulturalVQA (Nayak et al., 2024), CulturalGround (Nyandwi et al., 2025), CURVE (Singh et al., 2026), and VideoNorms (Varimalla et al., 2025) examine the cultural knowledge and grounding ability of VLMs. However, T2V generation poses distinct challenges for cultural evaluation. Unlike static images, videos require models to represent culture through temporal actions, social interactions, rituals, and multimodal cues such as visible text and audio. Therefore, protocols designed for T2I generation or VLM understanding cannot fully capture video-specific cultural challenges, especially those involving human actions, ritual procedures, and audio cultural expression. CultureVidBench addresses this gap by providing a dedicated benchmark for culturally grounded T2V generation.

3 Benchmark Construction

The goal of CultureVidBench is to provide a comprehensive benchmark for evaluating cultural understanding in T2V generation. This requires broad coverage of cultural regions, comprehensive representation of cultural aspects, and prompts designed for dynamic video content. In this section, we construct CultureVidBench through country selection, cultural taxonomy, and prompt generation.

3.1 Data Source and Cultural Taxonomy

Country Selection. To ensure broad cultural coverage, we select countries based on the cultural regions defined by the World Values Survey (World Values Survey Association, 2023). Specifically, CultureVidBench covers 8 world cultural regions, including Confucian, West and South Asian, African-Islamic, Orthodox Europe, Catholic Europe, Protestant Europe, English-speaking, and Latin America. Based on this regional coverage, we include 12 countries across 6 continents, as shown in Figure 1(b). This design enables systematic evaluation of T2V models across culturally distinct societies while maintaining manageable benchmark scale.

Cultural Taxonomy. Culture is a multifaceted concept that includes material expressions, social practices, collective behaviors, and ritual traditions shared within a society (Geertz, 2017). Such dimensions are also commonly used in cultural knowledge sources such as CulturalAtlas (Mosaica, 2024) and Wikipedia (Wikipedia, 2026), where cultures are described through artifacts, customs, social practices, festivals, ceremonies, and religious traditions. Based on this representation, we organize these sources into three major categories: material culture, social practice & performance, and ritual & ceremony. These categories further cover 14

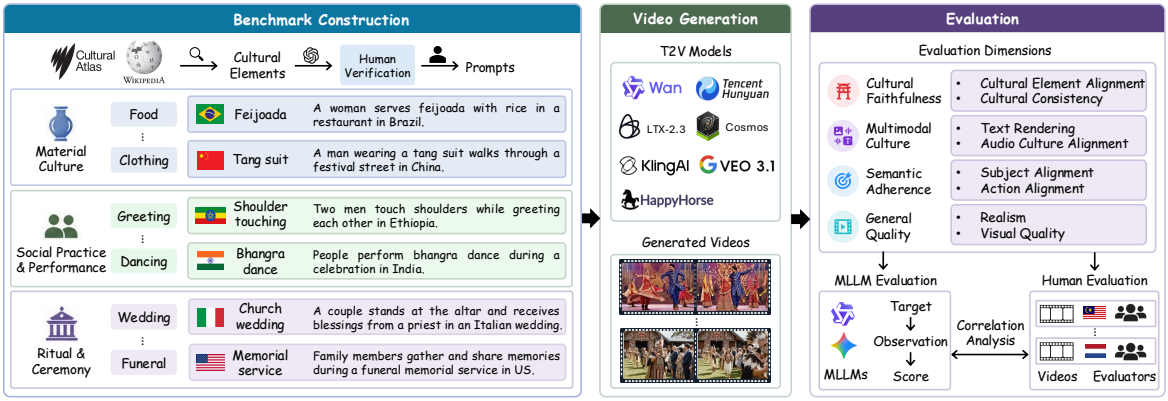


Figure 2: Framework of the CultureVidBench construction and evaluation pipeline. We collect cultural elements for each country and cultural aspect, and construct text prompts for video generation. The generated videos are evaluated by both humans and MLLMs across multiple dimensions, and their correlations are analyzed to assess the reliability of automatic evaluation.

cultural aspects, as shown in Figure 1 (b).

Cultural Element Collection. For each country and cultural aspect, we systematically collect representative cultural elements from CulturalAtlas and Wikipedia. Examples of cultural elements are shown in Figure 2. For material culture and social practice & performance, we mainly extract representative cultural elements. In contrast, rituals and ceremonies often involve structured social procedures that cannot be fully captured by a single element. Therefore, for the ritual & ceremony category, we additionally extract representative sub-activities from cultural descriptions to better capture procedural cultural knowledge. For example, for the China-wedding-tea ceremony, we extract sequential activities such as kneeling before parents, serving tea, and receiving red envelopes.

3.2 Prompt Generation

Since CultureVidBench focuses on T2V generation, all prompts are designed to emphasize actions and interactions in dynamic scenarios. Therefore, for the collected cultural elements in each category, we construct prompts using category-specific templates. For material culture, we explicitly add human actions to embed cultural elements into video scenarios, using the template <subject> <specific action> <material cultural element> <country/cultural context>. For social practice & performance, we focus on culturally grounded interactions, activities, and performances, and use the template <subject> <social practice & performance> <country/cultural context>. For ritual & ceremony, we generate prompts from representative sub-activities to better reflect culturally grounded behaviors. We use the template <subject> <sub-activity> <sub-activity> <country/cultural context>.

We provide GPT-5.4 (OpenAI, 2026) with the collected cultural elements and the category-specific templates to generate three candidate prompts for each cultural element. All candidate prompts are then manually verified and revised by two human annotators guided by cultural element references to ensure linguistic naturalness, cultural consistency, diversity, and video generation feasibility. For material culture and social practice & performance categories, we manually select the most natural and culturally representative prompt. Since ritual and ceremony activities often contain richer procedural content, we retain two prompts with different activity focuses for each ritual element to improve diversity and coverage. In addition, to evaluate multimodal cultural rendering in T2V models, some prompts are designed with explicit visible-text or audio cues, such as festival banners and stage performance.

3.3 Benchmark Statistics

CultureVidBench contains 828 cultural elements across 12 countries and 14 cultural aspects. Based on these elements, we generate and manually curate 1,000 prompts, including 272 for material culture, 386 for social practice & performance, and 342 for ritual & ceremony. Figure 3 shows the prompt distribution across countries and cultural categories. Although the number of prompts varies slightly by country, the overall distribution remains relatively balanced. This design provides broad coverage across world cultural regions while preserving diversity in cultural categories and activities.

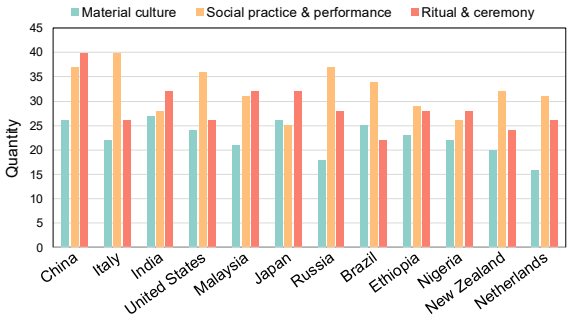


Figure 3: Prompt distribution across different countries and cultural categories.

4 Evaluation

We design an evaluation protocol that separately assesses cultural faithfulness, multimodal cultural rendering, semantic adherence, and general quality. We further combine scalable MLLM-based automatic evaluation with human user studies, where human judgments are used to validate the reliability of automatic assessment.

4.1 Evaluation Dimensions

Cultural Faithfulness. This dimension evaluates whether the generated video faithfully represents the target cultural context. It includes two aspects. (1) *Cultural Element Alignment* measures whether the cultural elements specified in the prompt are correctly presented. It evaluates not only visible cultural elements such as clothing and food, but also culturally specific interactions and procedures. For example, an Ethiopian shoulder-touching greeting should be represented with the appropriate greeting gesture rather than a generic interaction. A Japanese Shinto wedding should include key ritual elements such as the shrine setting and the correctly performed sake-sharing ceremony, rather than showing a modern wedding. (2) *Cultural consistency* measures whether the video avoids culturally inconsistent elements from other countries or cultural backgrounds. This is important because a video may include the correct cultural element while still hallucinating foreign symbols, resulting in a mixed or misleading cultural representation.

Multimodal Cultural Rendering. Culture may also be expressed through visible text and audio, beyond visual objects and actions. We therefore evaluate whether culture-related multimodal content is rendered appropriately from two aspects. (1) *Text Rendering* evaluates whether visible text, when present, is linguistically and culturally appropriate. (2) *Audio Culture Alignment* evaluates whether culturally relevant audio, such as music, chanting, or performance sounds, matches the intended cultural background and scenario. Audio is evaluated only for aspects where cultural sound is expected, including celebration, dancing, concerts, weddings, funerals, and religious practices.

Semantic Adherence. Although our benchmark focuses on cultural understanding, semantic alignment is necessary for determining whether the generated video follows the prompt. We evaluate two aspects. (1) *Subject Alignment* measures whether the people or participants described in the prompt are correctly presented. (2) *Action Alignment* measures whether the actions or activities described in the prompt are accurately depicted. This dimension helps distinguish cultural failures from general prompt-following failures.

Perceptual Quality. We also evaluate the perceptual quality of the generated videos. This dimension includes two aspects. (1) *Realism* measures whether the video appears natural and realistic, without obvious artifacts, distorted humans or objects, or unnatural motion. (2) *Visual Quality* measures the overall visual presentation, including color, clarity, and aesthetic quality. These criteria help examine whether cultural failures persist even when videos are visually plausible.

4.2 MLLM Evaluation

Text-video similarity models, such as CLIP (Radford et al., 2021) or ViCLIP (Wang et al., 2024), mainly capture coarse semantic consistency between prompts and generated videos. However, cultural evaluation requires more fine-grained reasoning over region-specific elements, contextual appropriateness, and possible foreign cultural hallucinations. Cultural cues may also appear across visual content, visible text, and audio, making multimodal reasoning necessary.

Recent MLLMs (Google, 2026; OpenAI, 2026) provide a promising alternative because they can analyze visual content, interpret text in frames, and reason about audio. However, directly asking MLLMs to assign scores may lead to entangled judgments across dimensions. For example, a low-quality video may receive lower cultural scores even when the target cultural element is present. To reduce such interference, we design a target-explicit evaluation procedure that separates target identification, observation extraction, and score assignment when applicable.

For dimensions with explicit evaluation targets, including cultural element alignment, subject align-

Evaluator	Models	Cultural Faithfulness		Multimodal Cultural Rendering		Semantic Adherence		Perceptual Quality	
		Alignment	Consistency	Text Rendering	Audio Alignment	Subject	Action	Realism	Visual Quality
Human	*Open-source models*								
	LTX-2.3-22B (HaCohen et al., 2026)	0.359	0.385	0.177	0.408	0.489	0.408	0.552	0.638
	Cosmos-Predict-2.5-14B (Ali et al., 2025)	0.403	0.464	0.178	-	0.545	0.466	0.409	0.491
	HunyuanVideo-1.5 (Team, 2025)	0.475	0.509	0.102	-	0.584	0.511	0.550	0.607
	Wan-2.2-14B (Wan et al., 2025)	0.485	0.499	0.170	-	0.579	0.489	0.560	0.644
	Proprietary models								
	Kling-3.0 (KlingAI, 2026)	0.701	0.703	0.264	0.589	0.768	0.709	0.675	0.769
	Veo-3.1-Fast (Google DeepMind, 2025)	**0.766**	0.754	0.430	**0.703**	**0.808**	0.730	0.628	0.781
	HappyHorse-1.0 (Alibaba Cloud, 2026)	0.742	**0.757**	**0.434**	0.674	0.802	**0.753**	**0.698**	**0.808**
MLLM	*Open-source models*								
	LTX-2.3-22B (HaCohen et al., 2026)	0.454	0.581	0.214	0.566	0.730	0.502	0.513	0.697
	Cosmos-Predict-2.5-14B (Ali et al., 2025)	0.518	0.653	0.207	-	0.872	0.523	0.410	0.507
	HunyuanVideo-1.5 (Team, 2025)	0.624	0.729	0.263	-	0.897	0.663	0.558	0.698
	Wan-2.2-14B (Wan et al., 2025)	0.603	0.727	0.297	-	0.909	0.650	0.671	0.861
	Proprietary models								
	Kling-3.0 (KlingAI, 2026)	0.883	0.909	0.402	0.809	0.992	**0.885**	0.776	0.935
	Veo-3.1-Fast (Google DeepMind, 2025)	**0.915**	0.931	**0.456**	0.783	**0.994**	0.874	0.745	0.926
	HappyHorse-1.0 (Alibaba Cloud, 2026)	0.907	**0.943**	0.439	**0.896**	0.992	0.883	**0.792**	**0.948**

Table 1: Human evaluation and Gemini-3.1-Pro-based MLLM evaluation results. All scores are normalized to 0-1 for comparison. "-" indicates that the model does not support audio generation.

ment, and action alignment, the MLLM follows three steps. (1) *Target Identification*: the model identifies the expected evaluation target from the prompt, target country, metadata, and current dimension. For example, the cultural element for cultural element alignment and the expected participants for subject alignment. (2) *Observation Extraction*: the model examines the generated video and extracts observations relevant to the identified target and current dimension. (3) *Score Decision*: the model assigns a discrete score with a short justification based on the identified target and extracted observations. We apply this structured protocol to all adopted MLLMs. This makes the evaluation more independent and interpretable across dimensions. Detailed prompts for MLLM-based evaluation are provided in the Appendix C.

4.3 Human Evaluation

Human user studies provide direct judgments and serve as a reference for validating the reliability of MLLM-based evaluation. Since exhaustive human evaluation over all generated videos is time-consuming, we follow existing T2V benchmarks (Feng et al., 2025) and evaluate a representative subset of CultureVidBench. To preserve benchmark diversity, we randomly sample one prompt under each cultural aspect from each country, resulting in 168 prompts. For each selected prompt, we evaluate videos generated by all seven adopted T2V models.

For each country, we recruit three native and culturally familiar evaluators from that country. The evaluation set for each country contains 98 generated videos from seven models. Each video is rated according to the evaluation dimensions described above using a 1-5 Likert scale. The final human score for each video is obtained by averaging the ratings from the three evaluators. Details of the human evaluation are provided in the Appendix C.

5 Experiments

5.1 Experimental Setup

We evaluate seven representative T2V models on CultureVidBench, including four open-source models, LTX-2.3-22B (HaCohen et al., 2026), Cosmos-Predict-2.5-14B (Ali et al., 2025), HunyuanVideo-1.5 (Team, 2025), and Wan-2.2-14B (Wan et al., 2025), and three proprietary models, Kling-3.0 (KlingAI, 2026), Veo-3.1-Fast (Google DeepMind, 2025), and HappyHorse-1.0 (Alibaba Cloud, 2026). Cosmos-Predict-2.5-14B is originally designed for world simulation, and we use its text-to-world generation capability for evaluation. Detailed generation settings are provided in Appendix B.

For automatic evaluation, we use both similarity-based and MLLM-based evaluators. ViCLIP (Wang et al., 2024) is used to measure coarse text-video semantic similarity, while Qwen3-VL-32B (Bai et al., 2025), GPT-5.4 (OpenAI, 2026), and Gemini-3.1-Pro (Google, 2026) evaluate videos across our proposed dimensions. Qwen3-VL-32B and GPT-5.4 take 16 uniformly sampled frames as input, whereas Gemini-3.1-Pro supports direct video understanding and evaluates both visual content and audio-related cultural cues. All evaluation scores are normalized from a five-point scale to the range of 0-1 for model comparison.

Models	Cultural Faithfulness				Multimodal Cultural Rendering				Semantic Adherence				Perceptual Quality			
	Alignment		Consistency		Text Rendering		Audio Alignment		Subject		Action		Realism		Visual Quality	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
ViCLIP (Wang et al., 2024)	0.137	0.158	-	-	-	-	-	-	0.135	0.159	0.224	0.325	-	-	-	-
Qwen3-VL-32B (Bai et al., 2025)	0.427	0.518	0.389	0.463	0.101	0.120	-	-	0.375	0.446	0.396	0.470	0.188	0.223	0.154	0.178
GPT-5.4 w/o Target	0.519	0.614	0.441	0.513	0.437	0.472	-	-	0.311	0.363	0.395	0.467	0.191	0.221	0.115	0.134
GPT-5.4 (OpenAI, 2026)	**0.546**	**0.644**	0.469	0.545	0.441	0.487	-	-	0.335	0.384	0.411	0.487	0.202	0.235	0.110	0.125
Gemini-3.1-Pro w/o Target	0.522	0.616	0.516	0.607	0.513	0.684	**0.373**	**0.441**	0.388	0.437	0.406	0.487	0.352	0.426	0.353	0.421
Gemini-3.1-Pro (Google, 2026)	0.529	0.629	**0.521**	**0.614**	**0.527**	**0.702**	0.364	0.431	**0.409**	**0.468**	**0.413**	**0.492**	**0.362**	**0.441**	**0.365**	**0.440**
Inter-human	0.591	0.689	0.576	0.672	0.606	0.694	0.509	0.583	0.448	0.529	0.477	0.566	0.449	0.535	0.437	0.513

Table 2: Correlation between human and MLLM-based automatic evaluations in terms of country-averaged Kendall's τ and Spearman's ρ. The last row reports the mean inter-human correlation for reference.

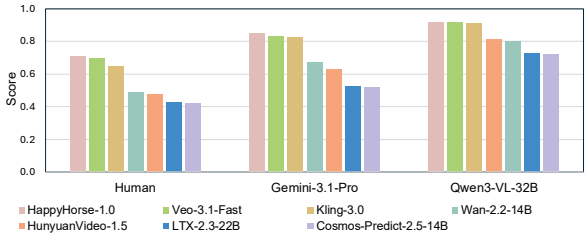


Figure 4: Overall performance rankings of T2V models from human evaluator and MLLM-based evaluators.

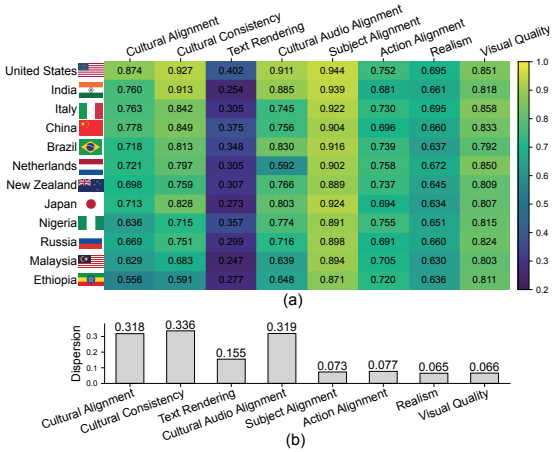


Figure 5: (a) Country-level results averaged over all T2V models evaluated by Gemini-3.1-Pro. (b) Dispersion denotes the gap between the highest and lowest country-level scores.

5.2 Performance Comparison

Table 1 reports the human and MLLM-based evaluation results of seven T2V models across all dimensions. Overall, both evaluators show a clear gap between proprietary and open-source models, with proprietary models consistently achieving higher scores across most evaluation dimensions. Moreover, current T2V models still show limited cultural understanding. Compared with semantic adherence and perceptual quality, the scores on cultural faithfulness and multimodal cultural rendering are generally lower, indicating that models can generate visually plausible and semantically related videos while missing culturally specific details. Multimodal cultural rendering is particularly challenging, as even the strongest proprietary model, HappyHorse-1.0, still shows clear limitations in producing culturally appropriate visible text.

To examine whether MLLM-based evaluators are consistent with human judgments, Table 2 reports the correlations between automatic and human evaluation using country-averaged Kendall's τ and Spearman's ρ, with inter-human agreement included as a reference. Compared with ViCLIP, MLLM-based evaluators show higher correlations with human evaluation, suggesting the need for fine-grained multimodal reasoning in cultural evaluation. Moreover, our target-explicit structured evaluation further improves both Gemini-3.1-Pro and GPT-5.4 on dimensions with clearly defined evaluation targets, such as cultural element alignment, subject alignment, and action alignment.

Among automatic evaluators, Gemini-3.1-Pro with full-video input achieves the highest overall correlation with human judgments, showing the benefit of direct video-level input for multimodal understanding and cultural reasoning. However, its correlations remain relatively limited on more subjective dimensions, such as realism and visual quality, suggesting that these aspects are still challenging for MLLM-based evaluators. Notably, inter-human agreement is also limited on these subjective dimensions, partly because the fine-grained rating scale requires annotators to distinguish subtle differences between adjacent scores.

We further visualize the model rankings produced by human evaluation, Gemini-3.1-Pro, and Qwen3-VL-32B across all evaluation dimensions in Figure 4. Gemini-3.1-Pro produces rankings consistent with human judgments, while Qwen3-VL-32B can also distinguish proprietary from open-source models. These results demonstrate that MLLM-based evaluation can provide a reliable and scalable approximation of human assessment.

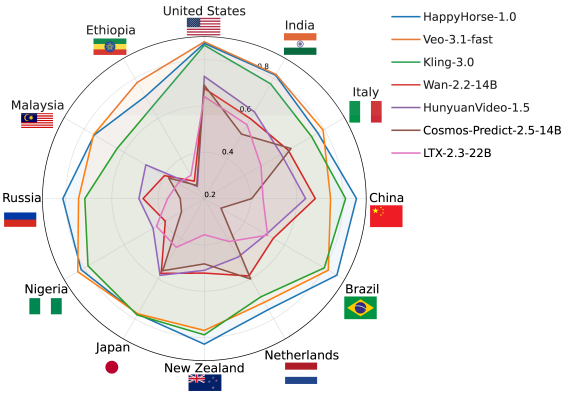


Figure 6: Country-level cultural scores of different T2V models evaluated by Gemini-3.1-Pro. Scores are averaged over cultural faithfulness and multimodal cultural rendering.

Figure 7: Cultural errors in videos generated by T2V models, such as incorrect visible text and inaccurate cultural elements.

5.3 Cross-country Cultural Performance

We then analyze the performance of T2V models across countries. Figure 5 (a) shows the country-level results averaged over all evaluated models across different evaluation dimensions. To better quantify cross-country variation, we also report the dispersion of each evaluation dimension in Figure 5 (b). The results show that current T2V models perform relatively consistently across countries on semantic adherence and perceptual quality. In contrast, culture-related dimensions exhibit much larger cross-country disparities. We can observe that T2V models tend to achieve stronger performance on high-resource cultural contexts, such as the United States, while performing worse on relatively underrepresented cultural contexts, such as Ethiopia and Malaysia.

We further examine the country-level cultural scores of different T2V models in Figure 6. The results reveal clear cross-country variation across models. Open-source models show larger performance fluctuations and generally perform worse on relatively underrepresented cultural contexts, including Ethiopia, Malaysia, and Nigeria. Proprietary models achieve higher and more stable cultural scores overall, but they still show visible drops on these countries.

To provide a more intuitive analysis, Figure 7 shows representative cultural errors made by different T2V models. In the Japan-Shogi example, HappyHorse-1.0 captures the shogi scenario, but still produces inaccurate visible text on the pieces, while HunyuanVideo-1.5 generates a chess-like scene instead of the correct Shogi element. In the Ethiopia-Crossing oneself example, all models struggle to faithfully represent the culturally specific religious gesture, which requires pinching the fingers together when making the sign of the cross (Wikipedia, 2026). Overall, these findings show that current T2V models can often capture the coarse semantic content of the prompt, such as “two players” and “move pieces”, but still have limitations in culturally grounded generation, especially for underrepresented cultural contexts.

5.4 Cultural Category Analysis

We further analyze model performance across different cultural categories and aspects in Figure 8. Overall, semantic adherence is consistently higher than cultural performance across all cultural aspects, showing the limitations of current T2V models in culturally grounded generation. Among the three cultural categories, material culture generally achieves stronger cultural performance, especially for aspects such as architecture, clothing, and food, which are strongly associated with distinctive visual appearances. In contrast, social practice & performance and ritual & ceremony aspects are more challenging, as they require models to capture culturally grounded actions, interactions, and procedures rather than only visual objects. This suggests that cultural video generation remains particularly difficult when culture is expressed through human activities and ritual processes.

6 Conclusion

In this paper, we presented CultureVidBench, a comprehensive benchmark for evaluating cultural understanding in T2V generation. CultureVidBench covers diverse countries, cultural regions, and cultural aspects, and evaluates generated videos across cultural faithfulness, multimodal cul-

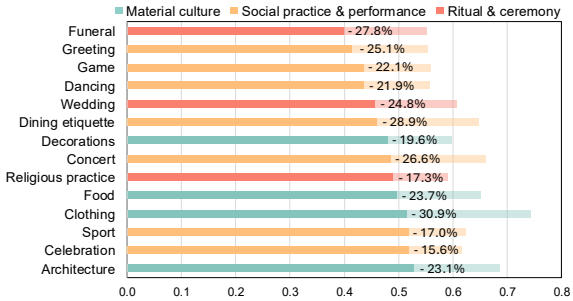


Figure 8: Cultural scores (dark-colored bars) and semantic adherence scores (light-colored bars) across cultural aspects evaluated by human. Cultural scores are averaged over cultural faithfulness and multimodal cultural rendering. Percentages indicate the relative decrease from semantic adherence to cultural score.

tural rendering, semantic adherence, and perceptual quality through both human evaluation and MLLM-based assessment. We evaluate both proprietary and open-source T2V models and validate the reliability of MLLM-based evaluation. Our results show that current T2V models achieve strong semantic adherence and visual quality, but still struggle with cultural faithfulness and multimodal cultural rendering, especially for underrepresented regions, rituals, and multimodal cultural cues such as visible text and audio. These findings highlight the importance of systematic cultural evaluation, and CultureVidBench can facilitate future research toward more culturally aware and trustworthy T2V generation systems.

Limitations

CultureVidBench currently covers 12 countries selected to span multiple continents and cultural regions, but it does not exhaustively cover all countries or cultural contexts. This design allows us to evaluate cultural understanding across diverse and representative scenarios, but may still miss many local, regional, and minority cultural practices. In future work, CultureVidBench can be further expanded to include a larger set of countries, more fine-grained cultural contexts, and broader community or expert contributions.

Ethical Considerations

CultureVidBench is designed for evaluating cultural understanding in T2V generation. During benchmark construction, we conducted safety checks to avoid prompts involving harmful, offensive, violent, or discriminatory content. CultureVidBench will be released for research and evaluation purposes only. The selected prompts are intended to evaluate whether models can faithfully represent culturally grounded objects, actions, rituals, visible text, and audio cues, rather than to rank cultures themselves.

For human evaluation, we use a representative subset of CultureVidBench to reduce the annotation burden while maintaining coverage across countries and cultural aspects. Human evaluators are recruited through Prolific and are compensated for their participation. The evaluation focuses on generated videos rather than personal information, and no private or sensitive participant data is collected beyond what is required by the study platform. Participants were informed that their responses would be used for research purposes to evaluate generated videos.

Acknowledgments

This research/project is supported by the Ministry of Education (MOE), Singapore, under its Academic Research Fund (AcRF) Tier 2 (Proposal ID: T2EP20125-0048). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore. This research/project is also supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016).

References

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, and 1 others. 2025. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*.

Alibaba Cloud. 2026. Happyhorse-1.0-t2v.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. 2025. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31137–31155.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Xiaolin Chen, Xuemeng Song, Wenhao Shi, Xianjing Han, Mong-Li Lee, and Wynne Hsu. 2026a. Arteculture: Benchmarking culture-conditioned visual emotion understanding in multimodal large language models. *arXiv preprint arXiv:2608.03358*.

Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2026b. T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6474–6485.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. 2025. Tc-bench: Benchmarking temporal compositionality in conditional video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4638–4662.

Clifford Geertz. 2017. *The Interpretation of Cultures*. Basic Books.

Google. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks.

Google DeepMind. 2025. Veo 3.1.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and 1 others. 2025. Phyworldbench: A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, and 1 others. 2026. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*.

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. 2026. Oscbench: Benchmarking object state change in text-to-video generation. In *ACL 2026*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, and 1 others. 2025. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji B Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Advances in Neural Information Processing Systems*, 37:13716–13747.

KlingAI. 2026. Klingai.

Huiqiong Li, Jiayu Wang, Zhiting Mei, Anirudha Majumdar, Jingjing Chen, and Bin Zhu. 2026. Robotrustbench: Benchmarking the trustworthiness of video world models for robotic manipulation. *arXiv preprint arXiv:2606.01600*.

Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, and 1 others. 2024. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. 2026. Culture in action: Evaluating text-to-image models through social activities. In *The Fourteenth International Conference on Learning Representations*.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. 2024. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*.

Mosaica. 2024. The Cultural Atlas. https://culturalatlas.sbs.com.au/.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal llms with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24198–24242.

OpenAI. 2026. Gpt-5.4.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. 2025. Cure: Cultural gaps in the long tail of text-to-image systems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15680–15691.

Chuancheng Shi, Shangze Li, Shiming Guo, Simiao Xie, Wenhua Wu, Jingtong Dou, Chao Wu, Canran Xiao, Cong Wang, Zifeng Cheng, and 1 others. 2025. Where culture fades: Revealing the cultural gap in text-to-image generation. *arXiv preprint arXiv:2511.17282*.

Darshan Singh, Arsha Nagrani, Kawshik Manikantan, Harman Singh, Dinesh Tewari, Tobias Weyand, Cordelia Schmid, Anelia Angelova, and Shachi Dave. 2026. Curve: A benchmark for cultural and multilingual long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32860–32871.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416.

Tencent Hunyuan Foundation Model Team. 2025. Hunyuanvideo 1.5 technical report. *Preprint*, arXiv:2511.18870.

Nikhil Reddy Varimalla, Yunfei Xu, Arkadiy Saakyan, Meng Fan Wang, and Smaranda Muresan. 2025. Videonorms: benchmarking cultural awareness of video language models. *arXiv preprint arXiv:2510.08543*.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, and 1 others. 2024. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pages 42055–42079.

Wikipedia. 2026. Wikipedia.

World Values Survey Association. 2023. WVS Cultural Map: 2023 Version Released. https://www.worldvaluessurvey.org/WVSNewsShow.jsp?ID=467.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, and 1 others. 2025. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*.

A CultureVidBench Details

CultureVidBench contains 1,000 prompts across 14 cultural aspects. To reduce cultural stereotyping, we avoid prompts based on physical appearance, racialized descriptions, or essentialized portrayals of people. Instead, we prioritize concrete cultural objects, practices, performances, and ritual activities grounded in cultural sources. The prompts are concise yet descriptive, with an average length of 12.8 words, providing sufficient cultural context for video generation while avoiding unnecessary linguistic complexity.

To further position CultureVidBench among existing cultural benchmarks, Table 4 compares it with representative benchmarks for cultural evaluation. Existing benchmarks mainly focus on text-to-image generation, covering cultural artifacts, social activities, multilingual visual concepts, or cultural stereotypes. While these benchmarks provide valuable resources for evaluating static cultural representation, they do not address the dynamic and multimodal nature of video generation. In contrast, CultureVidBench is designed specifically for T2V generation. Unlike prior benchmarks that mainly focus on static visual content or limited cultural dimensions, CultureVidBench emphasizes culturally grounded actions, social interactions, ritual procedures, visible text, and audio cues in video generation.

Table 5 further compares the prompt design of CultureVidBench with CUBE-1K (Kannen et al., 2024) and CulturalFrames (Nayak et al., 2025). While prior benchmarks mainly use static descriptions or simple human actions, CultureVidBench includes culturally grounded interactions, multi-step ritual activities, and explicit visible-text cues. Such prompt designs enable a more comprehensive evaluation of whether T2V models can generate culturally appropriate human actions and social interactions, follow the temporal and procedural structure of cultural activities, and accurately render multimodal cultural elements in both visual and audio modalities.

B T2V Generation Setting

For each T2V model, we generate videos for all prompts in CultureVidBench, resulting in 7,000 videos in total. We follow the official implementations and default generation settings of each model whenever possible. Table 3 reports the detailed settings, including resolution, number of frames, audio availability, and video duration. For LTX-2.3-22B, we use its recommended two-stage generation pipeline. For Cosmos-Predict-2.5-14B, which is designed for world simulation, we use its text-to-world generation function.

Models	Resolution	Frames	Audio	Duration (s)
LTX-2.3-22B	1920×1088	121	✓	5
Cosmos-Predict-2.5-14B	1280×704	93		6
HunyuanVideo-1.5	1280×720	121		5
Wan-2.2-14B	1280×720	81		5
Kling-3.0	1280×720	121	✓	5
Veo-3.1-Fast	1280×720	144	✓	6
Happyhorse-1.0	1280×720	121	✓	5

Table 3: Generation settings of the evaluated T2V models, including resolution, total number of frames, audio availability, and video duration.

C Evaluation

C.1 Human Evaluator Selection

We construct the human evaluation questionnaires using Qualtrics* and recruit evaluators through Prolific†. Participants are asked to evaluate videos corresponding to their own cultural background. To improve cultural reliability and reduce language barriers, each questionnaire is translated into the official language of the corresponding country. We further screen participants based on nationality, current residence, and primary language to increase the likelihood that they are familiar with the target cultural context. To ensure annotation quality, we include attention-check questions and remove invalid submissions that fail these checks. Participants are compensated at an hourly rate of £6.

Figure 10 shows the task instructions and detailed evaluation criteria provided to participants during the human evaluation. Figure 11 shows the rating interface, where participants view the cultural element, prompt, and generated video, and then assign scores across the evaluation dimensions.

C.2 MLLM-based Evaluation Details

We conduct MLLM-based evaluation to provide scalable automatic assessment for generated videos. We adopt both open-source and proprietary MLLMs, including Qwen3-VL-32B, GPT-5.4, and Gemini-3.1-Pro. Among them, Gemini-3.1-Pro supports full-video input and audio understanding, which allows it to directly evaluate both visual and

*https://www.qualtrics.com/
†https://www.prolific.com/

Benchmark	Model	# Countries	# Cultural Aspects	# Prompts	Cultural Faithfulness	Cultural Stereotype	Perceptual Quality	Semantic Alignment	Multimodal Rendering
ViSAGe (Jha et al., 2024)	T2I	135	1	405		✓			
CUBE-1K (Kannen et al., 2024)	T2I	8	3	1,000	✓		✓		
CultureBench (Shi et al., 2025)	T2I	15	-	7,932	✓		✓	✓	
CulturalFrames (Nayak et al., 2025)	T2I	10	5	983	✓	✓	✓	✓	
CultDiff (Bayramli et al., 2025)	T2I	10	3	1,500	✓		✓	✓	
CuRe (Rege et al., 2025)	T2I	64	32	300	✓		✓	✓	
CULTIVate (Malakouti et al., 2026)	T2I	16	9	576	✓			✓	
CultureVidBench (ours)	T2V	12	14	1,000	✓		✓	✓	✓

Table 4: Comparison of CultureVidBench with existing cultural benchmarks.

	CUBE-1K	CulturalFrames	CultureVidBench (Ours)
Evaluated Model	T2I	T2I	T2V
Prompt with Human Actions	×	A bride and groom exchange vows at a Hindu wedding in India.	Two players move pieces on a shogi board at a table in Japan.
Prompt with Multi-Step Activities	×	×	A bride and groom sit near a sacred fire and offer ritual items into the flames during a wedding in India.
Prompt with Text Cues	×	×	Families decorate homes for Christmas with a festive banner in the United States.
Prompt with Audio Cues	×	×	Performers sing and dance to Afrobeats music on stage in Nigeria.

Table 5: Prompt comparison of CultureVidBench with existing T2I cultural benchmarks.

audio-related cultural dimensions. The detailed prompts used for Gemini-3.1-Pro-based evaluation are provided in Table 6.

To make the evaluation more interpretable and reduce interference across different dimensions, we use a target-explicit structured evaluation protocol. This structure encourages the MLLM to evaluate dimensions independently rather than relying on a general impression of the video. Figure 9 presents an example of target-explicit evaluation using Gemini-3.1-Pro. In this example, the cultural element should be pelmeni, a traditional Russian meat dumpling, as shown in the reference image. The MLLM correctly observes that the generated dumplings are large and crescent-shaped with crimped edges, which do not match the small, round, ear-like shape of pelmeni, and therefore assigns a low score for cultural element alignment. At the same time, it gives high scores for subject alignment and action alignment, since the woman is clearly visible and is placing dumplings on a plate. The MLLM scores are close to the averaged human scores from three evaluators, showing that the structured protocol helps MLLMs provide dimension-specific judgments that are consistent with human assessment.

C.3 Correlation Analysis

We measure the consistency between MLLM-based evaluation and human evaluation using Kendall's τ and Spearman's ρ. These rank-based metrics indicate whether automatic evaluators produce model rankings that agree with human judgments. Since human evaluation is conducted separately for each country, directly computing a single global correlation may be affected by cross-country differences in evaluator scoring styles. Therefore, we compute correlations independently within each country and then average the country-level correlations:

$$\tau = \frac{1}{|\mathcal{C}|}\sum\nolimits_{c\in\mathcal{C}} \tau_c, \quad \rho = \frac{1}{|\mathcal{C}|}\sum\nolimits_{c\in\mathcal{C}} \rho_c, \qquad (1)$$

where $\mathcal{C}$ denotes the set of evaluated countries. τ_c and ρ_c denote Kendall's τ and Spearman's ρ computed within country c, respectively.

D Qualitative Examples

To provide a more intuitive analysis, we present qualitative examples across the three cultural categories in CultureVidBench: material culture, social practice & performance, and ritual & ceremony. For each category, we select representative prompts from different countries and provide video frames generated by seven T2V models.

Material Culture. For this category, we select two decoration-related cases: US-Christmas decoration and Malaysia-Hari Raya decoration, representing a high-resource cultural context and a relatively underrepresented cultural context, respectively. As shown in Figure 12, most models can capture the general visual appearance of Christmas decorations, such as Christmas trees, lights, and

You are an expert evaluator for cultural capability in text-to-video generation. You are given a 5-second AI-generated video.

Your task is to evaluate how well the video matches the prompt and the target cultural context.

Metadata:
- Target country/culture: Netherlands
- Cultural element: Cheese
- Video prompt: A vendor cuts large cheese wheels into slices at an open-air market with a sign board in the Netherlands.

Rules:
- Evaluate only what is visible in the video.
- Use the full 1-5 scale.
- For text rendering, use "NA" if no visible text appears.
- For audio cultural alignment, use "NA" only if the video has no audible audio, or if the audio is purely non-cultural ambient sound.
- Evaluate each question independently.
- The requirement fields should describe the evaluation target, not the model's visual observation.
- The observation fields should describe only visible evidence from the frames.

Evaluation criteria:
{Criteria}

Output Format:
Return strictly valid JSON only. Do not include markdown, explanations outside JSON, or extra text.
{
"1a": { "requirement": "Expected cultural element.", "observation": "", "justification": "", "score": 1, },
"1b": { "observation": "", "justification": "", "score": 1, },
"2a": { "observation": "", "justification": "", "score": 1, },
"2b": { "audio observation": "", "justification": "", "score": 1, },
"3a": { "requirement": "Expected subject or participants.", "observation": "", "justification": "", "score": 1, },
"3b": { "requirement": "Expected action.", "observation": "", "justification": "", "score": 1, },
"4a": { "observation": "", "justification": "", "score": 1, },
"4b": { "observation": "", "justification": "", "score": 1, }
}

Table 6: Prompt template for target-explicit MLLM-based evaluation. Criteria are the same as those used in the human-evaluation interface.

festive banners. However, despite good semantic adherence and visual quality, the generated banners often contain misspelled or unreadable text, showing that visible text rendering remains a major limitation for current T2V models.

The Hari Raya case is more challenging. Hari Raya is one of the most important and widely celebrated festivals in Malaysia, yet only a few models roughly capture its decorative style, such as HappyHorse-1.0. In contrast, several models generate scenes resembling Christmas decorations, such as HunyuanVideo-1.5. This indicates a bias toward high-resource cultural patterns and weaker understanding of underrepresented cultural contexts. Text rendering also remains unsuccessful in all the videos in this case, further highlighting the limitation of current models in generating culturally appropriate visible text.

Social Practice & Performance. For this category, we select two cases: US-Backyard games with frisbees and Ethiopia-Eskista dance. The frisbee case involves relatively fast and large body motions, making it challenging for video generation. As shown in Figure 13, most T2V models can capture the general concept of a casual backyard scene with frisbees. However, videos generated by HunyuanVideo-1.5, LTX-2.3-22B, and Cosmos-Predict-2.5-14B contain noticeable visual artifacts or less realistic motion. These results reflect the performance gap between proprietary and open-source models in generating dynamic social activities.

In contrast, the Ethiopia-Eskista dance case receives consistently low human scores across models. Most models generate generic Ethiopian dance scenes but fail to capture the distinctive Eskista movement, especially the rapid shoulder shaking. This indicates that current T2V models may recognize the broad cultural context, such as Ethiopian performers and dance settings, but still fail to faithfully represent the specific cultural practice. The low human scores further confirm that this type of social performance remains challenging, as its cultural meaning depends on fine-grained body motion and performance style.

Ritual & Ceremony. For this category, we select two wedding-related cases: Italy-Church wedding and China-Tea ceremony. The Italy-Church wedding case requires the correct depiction of key ritual participants and interactions, including the bride and groom standing at the church altar and receiving blessings from a priest. As shown in Figure 14, in the Italy-Church wedding case, several models, including HappyHorse-1.0, Veo-3.1-Fast, Kling-3.0, HunyuanVideo-1.5, and Cosmos-Predict-2.5-14B, reasonably capture the cultural scene, such as the bride and groom standing in a

	Cultural Element Alignment	Cultural Consistency	Subject Alignment	Action Alignment
Target	The video should show pelmeni, traditional Russian meat dumplings.	–	The video should feature a woman.	The woman should be placing pelmeni dumplings on a plate.
Observation	The dumplings appear large and crescent-shaped with crimped edges, rather than the small, round, ear-like shape of pelmeni.	The dumplings do not appear to belong to the target local culture, while other elements, such as the dining table and snowy winter setting, are consistent with the scene.	A woman is clearly visible sitting at the table.	The woman is placing the dumplings with a spoon.
MLLM Score	2-Poor	3-Fair	5-Excellent	5-Excellent
Human Score	2.3	3.0	5.0	5.0

Figure 9: Example of target-explicit evaluation using Gemini-3.1-Pro. The evaluator identifies the target requirement, observes the generated video, and assigns a score for each dimension. Human scores are averaged from three evaluators. We also provide a Wikipedia reference image for the target cultural element.

church and receiving blessings from a priest. However, Wan-2.2-14B generates a more generic indoor wedding ceremony with a weaker sense of the church altar, while LTX-2.3-22B depicts a stage-like scene with rows of people, where the altar is not clearly presented. These results suggest that most models can recognize the overall ritual context, but some still lack fine-grained cultural details in representing religious wedding settings.

The China-Tea ceremony case requires the correct depiction of ritual interactions, such as kneeling before parents and serving tea to them, which involves understanding culturally specific procedures and participant relationships. HappyHorse-1.0 generates a culturally and semantically aligned video with a clear tea-serving action, an appropriate wedding setting, and correct visible text. Veo-3.1-Fast and Kling-3.0 also roughly capture the tea-serving procedure, but their visible text rendering remains inaccurate. However, Wan-2.2-14B, HunyuanVideo-1.5, LTX-2.3-22B, and Cosmos-Predict-2.5-14B show weaker performance, with incorrect procedural steps and culturally inconsistent details, such as Western-style wedding attire. This shows that ritual generation requires not only visual realism, but also an understanding of culturally specific actions, social roles, and procedural order. Such cultural expressions are also susceptible to interference from other cultural patterns, making it difficult for models to maintain cultural consistency throughout the generated video.

E Failure Analysis

Overall, current T2V models remain limited in culturally grounded generation, particularly for underrepresented cultural contexts, rituals, and multimodal cultural rendering.

Underrepresented cultural contexts. Models often omit fine-grained local details and default to more dominant cultural patterns, potentially reflecting imbalanced cultural coverage in their training data. More geographically and linguistically balanced data, together with culture-aware data sampling, may improve representation of underrepresented cultures.

Rituals. Material culture is often associated with distinctive visual patterns that can be learned from repeated examples. In contrast, rituals require models to coordinate culturally specific actions, temporal order, and relationships among participants. These failures may therefore arise from both insufficient cultural knowledge and limited action planning. Procedural annotations, temporally ordered supervision, and explicit action planning may help address these limitations.

Multimodal cultural rendering. Text rendering requires exact symbolic structures rather than approximate visual resemblance, while existing models receive limited glyph-level supervision. Videos must additionally preserve character identity and location across frames, often resulting in distorted or flickering text. Script-aware rendering, OCR-based supervision, and temporal consistency constraints may help.

Task Instructions

In this study, you will evaluate videos generated by text-to-video models based on their **cultural faithfulness**, **multimodal cultural rendering**, **semantic alignment**, and **overall generation quality**.

For each video, please follow these steps:

1. Read the **cultural element** and **text prompt** carefully.
2. Turn on the **sound** and watch the **entire video** from beginning to end.
3. Score each question using the **1–5 scale**. If the video does not contain relevant text or culture-related audio, select **NA** for the corresponding text/audio question.
4. Evaluate each question independently.

> **Example:** When evaluating whether the action is correctly shown, focus only on the action, even if the specific cultural element is imperfect.

Evaluation Criteria

1. Cultural Faithfulness

1a. Cultural Element Alignment

Does the video correctly present the cultural element specified in the prompt?

- **1 Very poor:** The cultural element is completely absent or replaced by something entirely unrelated.
- **2 Poor:** The cultural element appears, but it is clearly incorrect or belongs to the wrong cultural category.
- **3 Fair:** The cultural element is partially recognizable, but it contains visual or semantic errors.
- **4 Good:** The cultural element is mostly correct, with only minor errors or missing details.
- **5 Excellent:** The cultural element is accurately, clearly, and convincingly presented.

1b. Cultural Purity

Does the video avoid showing elements from other countries or cultures that clearly do not belong to the target culture?

- **1 Very poor:** The video contains multiple elements from other countries or cultures, which clearly interfere with or replace the target cultural expression.
- **2 Poor:** The video contains elements that do not belong to the target culture and noticeably affect cultural interpretation.
- **3 Fair:** The video contains a small number of elements from other cultures or questionable elements, but the target cultural content is still recognizable.
- **4 Good:** The video mostly maintains cultural consistency, with only minor or barely noticeable elements from other cultures.
- **5 Excellent:** The video contains no noticeable elements from other countries or cultures, and the target cultural expression is pure and consistent.

2. Multimodal Cultural Rendering

2a. Text Rendering Correctness

If text appears in the video, is it correctly rendered and culturally appropriate? Select **NA** if no text appears in the video.

- **1 Very poor:** The basic composition of the text is incorrect, meaningless, or completely inappropriate for the target culture.
- **2 Poor:** The text is partially visible, but it is mostly incorrect, distorted, or culturally inappropriate.
- **3 Fair:** The text is partially recognizable, but it contains clear errors in language, form, or cultural usage.
- **4 Good:** The text is mostly correct and culturally appropriate, with only minor visual or language issues.
- **5 Excellent:** The text is clear, accurate, readable, and culturally appropriate.

2b. Audio Cultural Alignment

If the video contains culture-related audio or music, does it match the cultural context of the prompt? Select **NA** if no cultural audio in the video.

- **1 Very poor:** The culture-related audio is completely unrelated to the scene or conflicts with the cultural context.
- **2 Poor:** The culture-related audio is inappropriate or noticeably interferes with the target cultural scene.
- **3 Fair:** The culture-related audio is partially appropriate, but there are mismatches with the cultural scene.
- **4 Good:** The culture-related audio mostly fits the cultural context, with only minor issues.
- **5 Excellent:** The culture-related audio strongly matches and enhances the cultural context of the video.

3. Semantic Alignment

3a. Subject / Participant Alignment

Does the video correctly show the subject or participants described in the prompt?

- **1 Very poor:** The required subject or participant is absent or entirely unrelated.
- **2 Poor:** The subject or participant appears, but does not match the category described in the prompt.
- **3 Fair:** The subject or participant is generally correct, but has clear attribute or role errors.
- **4 Good:** The subject or participant is correct, with only minor attribute issues.
- **5 Excellent:** The subject or participant fully matches the prompt in category, role, and attributes.

3b. Action Alignment

Does the action in the video match the action described in the prompt?

- **1 Very poor:** The action is absent or completely different from the action described in the prompt.
- **2 Poor:** The intended action is barely recognizable, but it is clearly physically or logically incorrect.
- **3 Fair:** The correct action is shown, but there are physical, temporal, or logical issues.
- **4 Good:** The action is mostly correct, with only minor unnatural motion or execution issues.
- **5 Excellent:** The action is clearly, naturally, and accurately performed.

4. Realism and Visual Quality

4a. Realism

Does the video look like a real-world video?

- **1 Very poor:** The video looks highly artificial, distorted, or obviously fake.
- **2 Poor:** The video contains many visual artifacts, unrealistic motion, or unnatural textures.
- **3 Fair:** Some parts look realistic, but noticeable artifacts reduce the overall realism.
- **4 Good:** The video appears mostly realistic, with only minor imperfections.
- **5 Excellent:** The video looks convincingly realistic, with natural motion, lighting, and textures.

4b. Visual Quality

Is the video clear, stable, and visually appealing at an acceptable level?

- **1 Very poor:** The video is very blurry, distorted, unstable, or visually unpleasant.
- **2 Poor:** The video has frequent quality issues, such as flickering, deformation, or poor composition.
- **3 Fair:** The video is generally acceptable, but contains visual defects.
- **4 Good:** The video is clear and visually pleasing, with only minor artifacts.
- **5 Excellent:** The video is high-quality, clear, stable, and visually appealing.

Figure 10: Task instructions and evaluation criteria in the human evaluation interface.

Cultural Element and Video

Cultural Element: United States - Christmas lights

Prompt: Families hang colorful Christmas lights around houses at night in United States.

1. Cultural Faithfulness

	Very Poor	Poor	Fair	Good	Excellent
Cultural Element Alignment: Does the video accurately present the cultural element specified in the prompt? (E.g., are the hot dog, wedding ceremony, or funeral ritual shown accurately?)	○	○	○	○	○
Cultural Purity: Does the video avoid showing elements from other countries or cultures that do not belong to the United States? (E.g., does a United States food stall or dining scene avoid Asian tableware, and does a United States funeral or religious ceremony avoid elements from other cultures?)	○	○	○	○	○

2. Multimodal Cultural Rendering

	Very Poor	Poor	Fair	Good	Excellent	NA
Text Rendering Correctness: If text appears in the video, is it correctly rendered and culturally appropriate? Select **NA** if no text appears in the video. (E.g., are shop signs, posters, menus, or labels written in correct English?)	○	○	○	○	○	○
Audio Cultural Alignment: If cultural audio or music is present, does it match the cultural context of the prompt? Select **NA** if no cultural audio is present. (E.g., does the background music or sound in a concert, performance, wedding, funeral, or religious ceremony scene match the corresponding cultural setting?)	○	○	○	○	○	○

3. Semantic Alignment

	Very Poor	Poor	Fair	Good	Excellent
Subject / Participant Alignment: Does the video correctly show the subject or participants described in the prompt (E.g., are the men, women, tourists, performers, etc. in the prompt correctly presented?)	○	○	○	○	○
Action Alignment: Does the action in the video match the action described in the prompt? (E.g., are actions like adding a mustard, walking, or dancing correct?)	○	○	○	○	○

4. Realism and Visual Quality

	Very Poor	Poor	Fair	Good	Excellent
Realism: Does the video look like a realistic real-world video? (E.g., are there no unnatural generation artifacts, such as objects suddenly appearing or disappearing, unrealistic motion jumps, or distorted body parts?)	○	○	○	○	○
Perceptual Quality: Is the video visually clear, stable, and aesthetically acceptable?	○	○	○	○	○

Figure 11: Human evaluation interface used for rating generated videos.

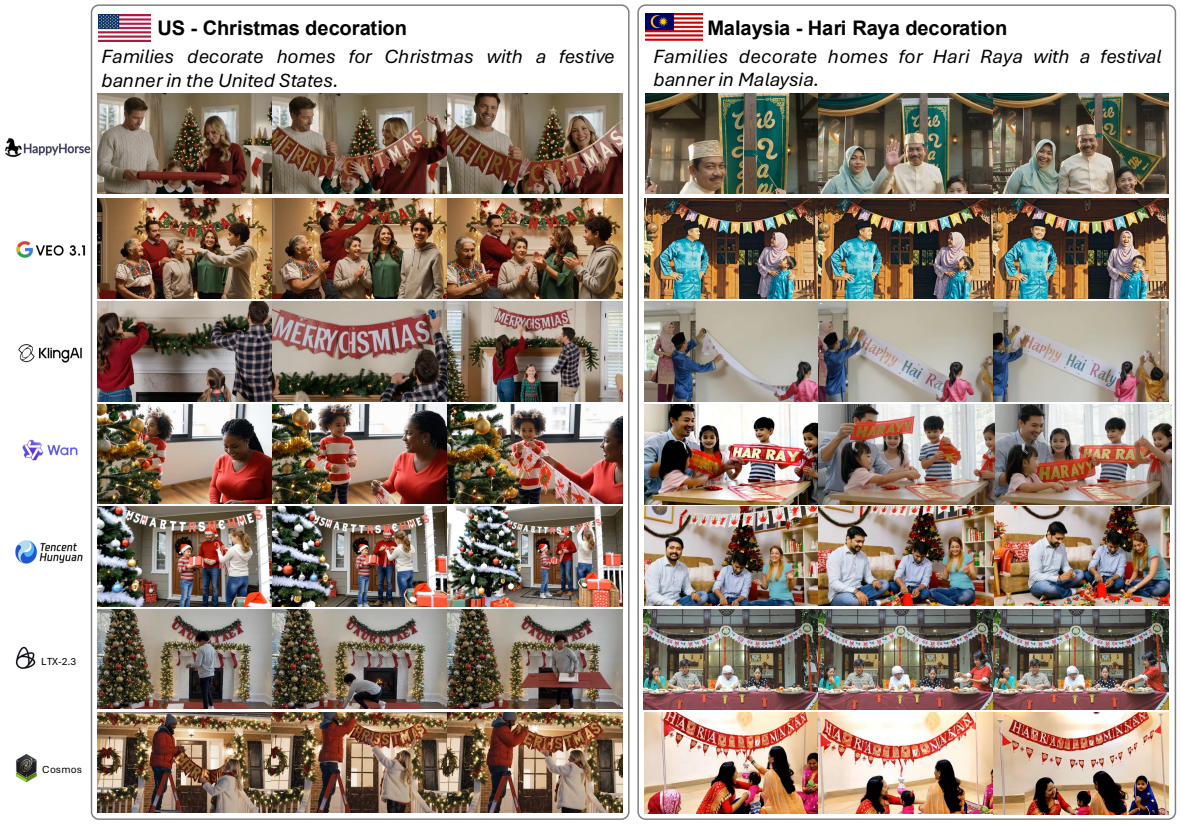


Figure 12: Sampled videos of different models with prompts from material culture category.

Figure 13: Sampled videos of different models with prompts from social practice & performance category.

Figure 14: Sampled videos of different models with prompts from ritual & ceremony category.