
Impossible Videos

Zechen Bai ^{*1} Hai Ci ^{*1} Mike Zheng Shou ¹

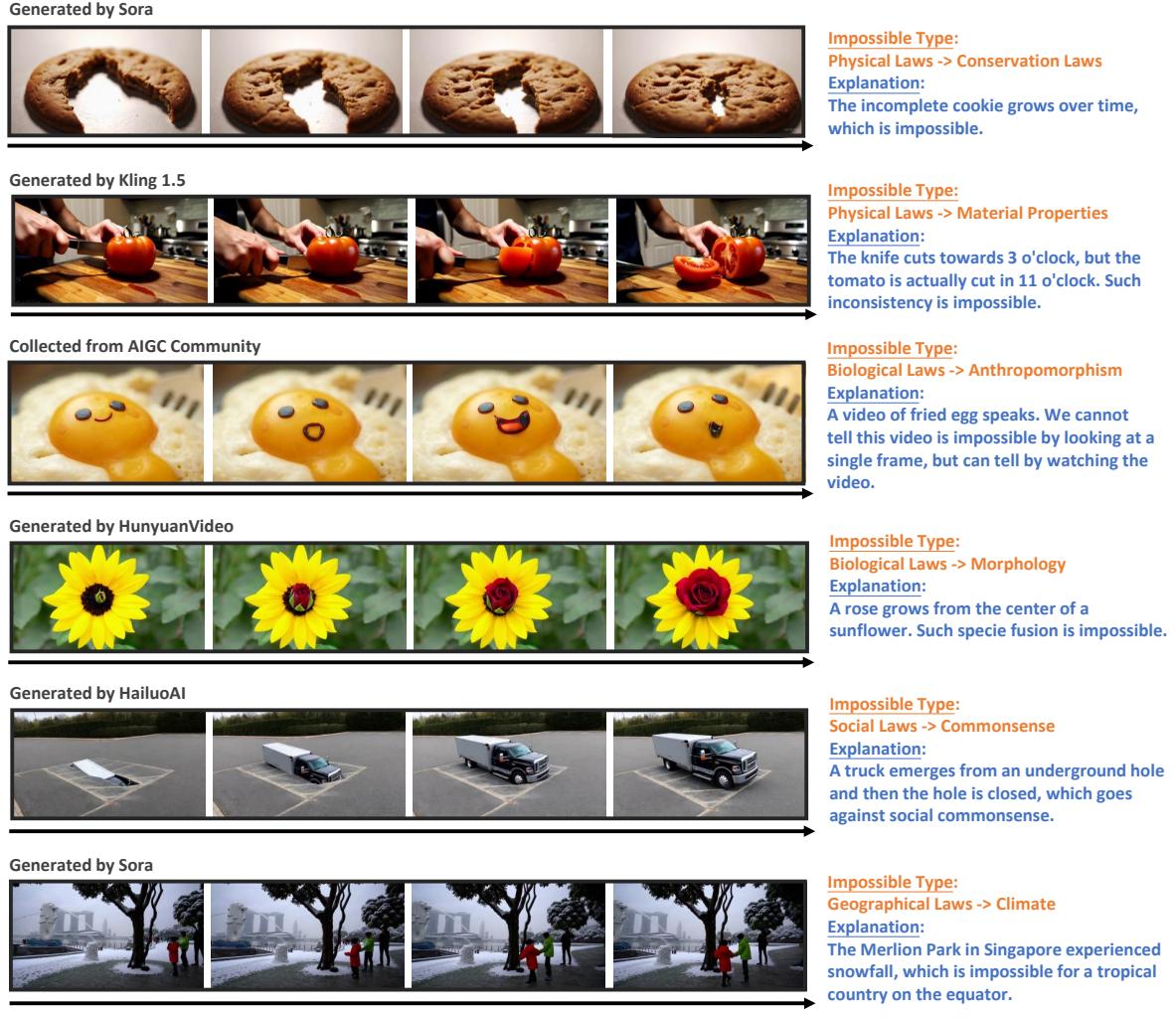


Figure 1: Impossible Video Examples with Impossible Type and Explanation.

Abstract

Synthetic videos nowadays is widely used to complement data scarcity and diversity of real-world videos. Current synthetic datasets primarily replicate real-world scenarios, leaving impos-

^{*}Equal contribution ¹Show Lab, National University of Singapore, Singapore. Correspondence to: Mike Zheng Shou <mike.zheng.shou@gmail.com>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

sible, counterfactual and anti-reality video concepts underexplored. This work aims to answer two questions: 1) Can today's video generation models effectively follow prompts to create impossible video content? 2) Are today's video understanding models good enough for understanding impossible videos? To this end, we introduce IPV-BENCH, a novel benchmark designed to evaluate and foster progress in video understanding and generation. IPV-BENCH is underpinned by a comprehensive taxonomy, en-

compassing 4 domains, 14 categories. It features diverse scenes that defy physical, biological, geographical, or social laws. Based on the taxonomy, a prompt suite is constructed to evaluate video generation models, challenging their prompt following and creativity capabilities. In addition, a video benchmark is curated to assess Video-LLMs on their ability of understanding impossible videos, which particularly requires reasoning on temporal dynamics and world knowledge. Comprehensive evaluations reveal limitations and insights for future directions of video models, paving the way for next-generation video models. Project page: <https://showlab.github.io/Impossible-Videos/>.

1. Introduction

Video data has been a long-standing focus in the research community, offering the potential to capture richer and more structured information compared to text (Yang et al., 2024a; Liu et al., 2024c). Over time, this domain has expanded from early tasks like action recognition (Zhu et al., 2020) to more advanced applications such as video captioning (Bai & An, 2018), question answering, video generation (Ning et al., 2023), and video-based world modeling (Liu et al., 2024c), showcasing its versatility and growing importance in advancing AI capabilities.

Although video data seems to be abundant on the Internet, it suffers from issues of low quality and diversity, etc. Recent efforts have been trying to alleviate these issues by generating videos either from neural generation models or simulation engines (Agarwal et al., 2025). However, the primary goal is still to replicate real-world scenes in a controlled way, which severely limits the broader imagination space and applications. In this work, we introduce the concept of *impossible videos*, which particularly captures counterfactual and anti-reality scenes that are *impossible* in real world. We argue that impossible videos can form an effective testbed to assess video models. As an *out-of-real-world-distribution* data, it requires the model to not simply *memorize* real-world data and *retrieve* similar information based on the input, but to genuinely *learn* from real-world data and *reason* upon the input.

Currently, mainstream video models can be categorized into two major categories based on tasks: understanding and generation (Xie et al., 2024; Wu et al., 2024). Thus, this work aims to probe the boundary of the two types of models by answering the following questions.

First, can today's video generation models effectively follow text prompts to create impossible video content?

Recent research focus of video generation has been evolving from fundamental video quality (Ning et al., 2023) (e.g.,

aesthetic quality, motion smoothness) to advanced semantic quality (e.g., physical laws, subject consistency) (Bansal et al., 2024). State-of-the-art models have been positioned as a world simulator (OpenAI, 2025). Ideally, the model should be able to generate either physically coherent or anti-reality videos with detailed control of the text prompt, enabling wider range of applications, such as filming, advertising, etc. Prompting the model to generate impossible videos challenges the model to "break the rule" yet faithfully following the prompt.

Second, are today's video understanding models good enough for understanding impossible videos? Advanced video understanding models, especially video large language models (Video-LLMs), are mostly trained on large-scale video datasets, demonstrating remarkable performance on popular benchmarks with real-world videos. Impossible videos pose specific challenges on reasoning temporal dynamics and world knowledge.

To achieve these goals, we propose a benchmark, IPV-BENCH, focusing on **ImPossible Videos**. To the best of our knowledge, this is the first work focusing on this topic. We start with building a comprehensive taxonomy covering diverse aspects of impossible videos, including scenes violating physical laws, biological laws, geographical laws, and social laws. Based on the taxonomy, we construct IPV-TXT, a prompt suite that consists of 260 text prompts; IPV-VID, a video set with 902 high-quality videos. Both the text prompts and videos are organized following the structure of the taxonomy, describing or displaying impossible scenes with particular consideration on temporal dynamics. Examples of impossible videos are shown in Fig. 1.

Based on this benchmark, we conduct comprehensive evaluations for mainstream video understanding models and generation models, suggesting that most video models fall short on impossible videos. Extensive analysis further reveals several insights of certain limitation and future direction. We will make the data public to inspire future research. In summary, our contributions includes:

- To the best of our knowledge, this is the first work to identify and investigate impossible videos, which explores a critical yet absent space for advanced video understanding and generation research. To this end, we construct a benchmark IPV-BENCH.
- IPV-BENCH is underpinned by a comprehensive taxonomy. Based on the taxonomy, a prompt suite and a set of high-quality videos is collected and carefully annotated, supporting downstream evaluations.
- Based on IPV-BENCH, we conduct extensive evaluation and analysis on mainstream video understanding and generation models, unveiling current limitations and revealing future directions.

Table 1: Comparison of IPV-BENCH and Existing Benchmarks.

Benchmark	Tasks			Video Data			Text Data		
	AIGC Detection	Video Understanding	Video Generation	Real World Videos	Generated Videos	Impossible Videos	Text Prompts	Text Descriptions	Impossible Text
GenVideo (Chen et al., 2024a)	✓	✗	✗	✓	✓	✗	✓	✗	✗
GenVidBench (Ni et al., 2025)	✓	✗	✗	✓	✓	✗	✓	✗	✗
LOKI (Ye et al., 2024)	✓	✓	✗	✓	✓	✗	✗	✓	✗
VBench (Huang et al., 2024)	✗	✗	✓	✗	✗	✗	✓	✗	✗
VideoPhy (Bansal et al., 2024)	✗	✗	✓	✗	✗	✗	✓	✗	✗
PhyGenBench (Meng et al., 2024a)	✗	✗	✓	✗	✗	✗	✓	✗	✗
SEED-Bench (Li et al., 2023)	✗	✓	✗	✓	✗	✗	✗	✓	✗
Video-Bench (Ning et al., 2023)	✗	✓	✗	✓	✗	✗	✗	✓	✗
MV-Bench (Li et al., 2024)	✗	✓	✗	✓	✗	✗	✗	✓	✗
TempCompass (Liu et al., 2024d)	✗	✓	✗	✓	✗	✗	✗	✓	✗
IPV-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

2. Related Work

This work evaluates video models across two key domains: video understanding and video generation. Tab. 1 presents a comprehensive comparison of IPV-BENCH and existing benchmarks, highlighting their relationships and key differences. We then outline the primary objectives of existing models and benchmarks in each domain.

Video Understanding remains a fundamental challenge in computer vision, encompassing tasks such as action recognition (Zhu et al., 2020), object localization (Fan et al., 2023; Bai et al., 2025), tracking (Zhao et al., 2023b), temporal grounding (Lin et al., 2023b), captioning (Wang et al., 2020; Bai et al., 2021), and, more recently, AI-Generated video detection (Ye et al., 2024). Video Large Language Models (Video-LLMs) (Tang et al., 2023), powered by Large Language Models (LLMs) (Zhao et al., 2023a), leverage language as a universal interface to facilitate a wide range of video-related tasks. Most open-source Video-LLMs extend from multimodal large language models (MLLMs) originally designed for images, such as LLaVA (Liu et al., 2024b). Some closed-source models, such as GPT-4o (Hurst et al., 2024), although initially designed for images, can also function as video models by processing multiple frames as input. Popular benchmarks such as VideoMME (Fu et al., 2024), MV-Bench (Li et al., 2024), assess models on a range of tasks (e.g., multiple-choice and open-ended questions) and scenarios (e.g., daily life, sports, and films). These datasets primarily consist of publicly available videos sourced from the internet. To our knowledge, no existing benchmark explicitly includes a dedicated set of impossible or counterfactual videos for evaluation, which are crucial for assessing model generalization, robustness and reasoning abilities.

Video Generation has garnered significant attention in both academia and industry, with text-to-video generation serving as a foundational task (Wu et al., 2023). Notable open-source models include Stable Video Diffusion (Blattmann et al., 2023), CogVidX (Yang et al., 2024b), Open-Sora (Zheng et al., 2024a), Show-1 (Zhang et al., 2024), and HunyuanVideo (Kong et al., 2024), among oth-

ers. Proprietary models include Kling (KLING, 2025), Sora (OpenAI, 2025), and Hailuo (Hailuo, 2025), among others. One of the primary challenges in video generation is achieving high visual quality, including factors such as resolution, realism, and temporal consistency (Huang et al., 2024). To address this, benchmarks such as VBench (Ning et al., 2023) comprehensively evaluate various aspects of visual quality. However, with advancements in video generation models, research focus has shifted towards ensuring semantic coherence, particularly in maintaining adherence to physical laws. Recent benchmarks such as PhyGenBench (Meng et al., 2024a) and VideoPhy (Bansal et al., 2024) have emerged to evaluate models’ ability to generate physically plausible videos. Beyond physics-constrained generation, an often-overlooked aspect is physics-defying content, or more broadly, the generation of impossible scenes, which plays a significant role in creative domains such as film and advertising. Creativity is tangentially related to this challenge; however, it is only briefly considered in some comprehensive benchmarks (Zeng et al., 2024) and has yet to be systematically studied.

3. IPV-BENCH

We first develop a taxonomy that systematically categorizes various types of impossible scenes. This taxonomy serves as the foundation for two critical components of the benchmark: 1) IPV-TXT, a suite of high-quality text prompts describing impossible scenes that cannot occur in the real world. 2) IPV-VID, a curated collection of high-quality videos depicting impossible scenes, each with corresponding annotations. An overview of the taxonomy and the roles of its components is presented in Fig. 2.

3.1. IPV Taxonomy and Prompt Suite

Overview. As illustrated in Fig. 2, the taxonomy is structured around four major categories: “Physical Laws”, “Biological Laws”, “Geographical Laws”, and “Social Laws”. Each category is divided into multiple subcategories, which are further subdivided, forming a detailed hierarchical struc-

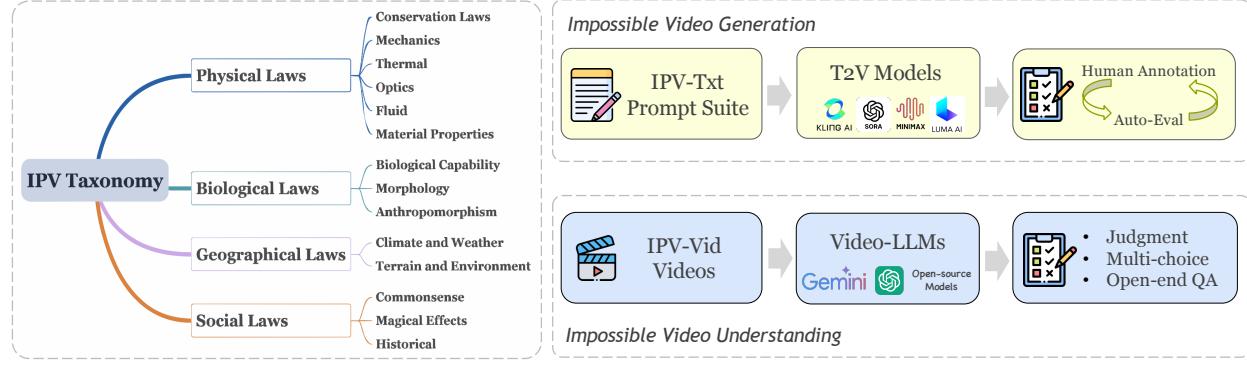


Figure 2: **Overview of the IPV-BENCH Benchmark.** IPV-BENCH is structured with a comprehensive taxonomy, enabling the creation of a diverse prompt suite (IPV-TXT) and a high-quality video dataset (IPV-VID). These components facilitate the evaluation of popular video generation and understanding models.

ture. Building upon this hierarchy, our IPV-TXT benchmark incorporates 260 high-quality text prompts that describe various counterfactual and other unusual scenarios.

1. “Physical Laws” covers 6 common laws: Mechanics, Thermal, Optics, Fluid, Material Properties, and Conservation Laws. This categorization considers most physical phenomena. We instantiate text prompts to explicitly describe a scene defying a specific physical law. For example, “*A person pours milk into a glass cup half filled with milk, but the amount of milk in the glass cup does not change at all*” describes a video that violates the law of conservation of mass.
2. “Biological Laws” categorizes 3 sub-categories: 1) Biological Capability covers scenes that exceeds human or animals capabilities. For example, “*A person flapped his arms like wings and successfully flew into the sky.*”; 2) Morphology consider impossible body composition. E.g., “*A horse walks on the grassland, gradually growing from four legs to eight legs*”; and 3) Anthropomorphism includes non-living objects exhibit anthropomorphic behavior, e.g., “*A fried egg with a face on it is opening its mouth and speaking something*”.
3. “Geographical Laws” considers the impossible phenomena in natural environment, including Climate and Weather anomalies, and Terrain and Environmental anomalies. For example, “*A mountain flattens into a perfect plateau, leaving all its trees and snow intact on the new flat surface*”.
4. “Social Laws” includes 3 types of counterfactual phenomena violating social laws: 1) Commonsense defines unusual scenes that violates our daily routine or customs. For example, “*The programmer in front of the computer suddenly started eating the keyboard and quickly ate half of it*”. 2) Magical Effects emphasize creative content and effects that cannot be easily interpreted by certain scientific law. For example, “e.g., A hand turns on a flashlight and shines it on a glass cup,

and the cup immediately breaks into pieces.” 3) Historical anomalies highlight a interesting type of scene that displays impossible combinations of human-object or human-human across different historical periods. For example, “*Albert Einstein and Donald Trump are shaking hands at the White House*”.

Construction. We propose a comprehensive framework for developing both the taxonomy and the IPV-TXT benchmark. During the development process, the taxonomy structure and text prompts evolve together, informing and refining each other.

1) Initialization. We establish the initial taxonomy structure by reviewing relevant literature (Bansal et al., 2024; Meng et al., 2024a;b; Zheng et al., 2024b) and leveraging large language models (LLMs), such as GPT-4o, for additional insights. Beyond the widely studied domain of physical laws, we expand our taxonomy to include broader categories, incorporating everyday scenarios and intriguing effects that can be applied in downstream applications.

2) Iterative Refinement. We refine the taxonomy and prompts through an iterative process. At each iteration, we collect a set of text prompts to expand the existing taxonomy structure. If certain text prompts do not fit within the structure, this suggests the need for adjustments to the taxonomy, ensuring that both the taxonomy and prompts evolve together. New prompts drive the taxonomy toward greater comprehensiveness, while the updated taxonomy encourages human to explore more intriguing examples.

LLM Refinement. We carefully design prompts to guide various LLMs in generating a broader range of example scenarios. We encourage the models to generate original and creative examples rather than merely substituting nouns in existing examples. We employ three LLMs—namely GPT-4o, Claude-3.5, and Gemini-1.5—to maximize diversity.

Crowdsourcing Refinement. To further improve the taxonomy, we aim to collect more “outlier” samples to challenge

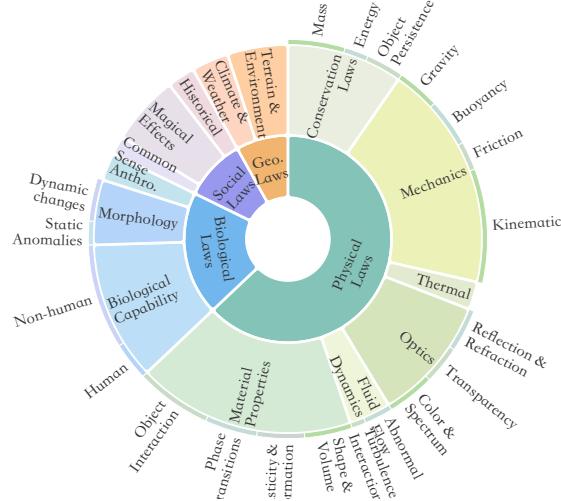


Figure 3: Distribution of the Prompt Suite Across the Taxonomy.

and refine the current structure until it accommodates the majority of samples. To this end, we employ crowdsourcing to collect data samples from a diverse set of contributors. We employ a questionnaire (see Appendix Fig. 10) to collect data from 15 participants from diverse academic and professional backgrounds, including computer science, economics, arts, and education. The participants have acknowledged that the collected data will be used for research purposes. On average, participants required 31 minutes to complete the questionnaire. This time commitment demonstrates the complexity and value of the collected text prompts. Finally, we have collected 150 impossible text prompts, which contribute to refining the taxonomy. It is essential to explicitly specify “LLMs are not allowed” in the questionnaire to guarantee the authenticity of collected data samples.

3) Quality Control. We conduct a thorough review of the prompts across multiple dimensions. *Clarity.* The text prompts should be expressed in a clear and comprehensible manner. Complicated or confusing prompts will be manually revised or removed. *Accessibility.* The described scene should be accessible in daily life. We exclude some obscure or esoteric samples. *Relevance.* The text prompt should be relevant to its belong category (and sub-category) in the taxonomy. We also merge some similar or duplicated samples at this step. *Visualizability.* The described scene should be able to be displayed by a video. Some non-visual scenes, such as sound-related descriptions, will be excluded.

4) Linguistic Enhancement. We further refine the linguistic clarity and expressiveness of the text prompts to enhance their suitability for video generation models. Specifically, we reference popular prompt rewriting strategies (Kong et al., 2024; Yang et al., 2024b) and instruct GPT-4o to enhance the text prompts. Compared to the original descriptions, the rewritten prompts feature relatively longer sequences, greater detail, and a more structured format.

This methodology yields a robust and comprehensive taxonomy, and a prompt suite, which can be used for assessing T2V models’ capability on generating impossible videos, providing a valuable tool for advancing this domain. The distribution of the prompt suite is shown in Fig. 3.

3.2. IPV-VID

We construct IPV-VID, a novel video benchmark designed to assess the capabilities of popular VideoLLMs in reasoning about temporal dynamics and world knowledge using impossible videos. IPV-VID is structured according to the established taxonomy. Given the unique nature of video data, IPV-VID places particular emphasis on temporal anomalies, such as motion and stage changes, which are challenging to identify from individual frames alone.

3.2.1. VIDEO COLLECTION.

T2V Generation. We prompt 10 state-of-the-art T2V models to generate a comprehensive set of videos, including open-sourced models (Open-Sora (Zheng et al., 2024a), HunyuanVideo (Kong et al., 2024), CogVidX (Yang et al., 2024b), Mochi 1 (GenmoAI, 2025), LTX (GenmoAI, 2025), and Pyramid-Flow (Jin et al., 2024)) and closed-sourced models (Sora (OpenAI, 2025), Kling (KLING, 2025), Luma (LumaLabs, 2025), and Hailuo (Hailuo, 2025)). Using 260 high-quality text prompts from the IPV-TXT suite, we generate a total of 2,600 synthetic videos.

Web Video Collection. To enhance the scale and diversity of the benchmark, we supplement our dataset with videos collected from the Internet. The primary sources include community websites for commercial video generation models, such as Sora and Hailuo, among others. Additionally, we gather videos from Twitter (X) shared by users, explicitly mentioning that they were generated by specific AI models. During the manual collection process, we adhere to an implicit criterion: videos must exhibit phenomena that are impossible in the real world. From these sources, we collect a total of 155 videos.

Real-world Videos. To mitigate evaluation bias and ensure a balanced distribution, we incorporate real-world videos into the benchmark. These videos are used alongside synthetic videos to evaluate the performance of VideoLLMs. We utilize OpenVid (Nan et al., 2024) as the foundational dataset for real-world videos. Using AI-generated videos as queries, we leverage the CLIP (Radford et al., 2021) model to retrieve content-consistent videos from OpenVid. During retrieval, we filter out unsuitable videos based on video length, aspect ratio, and aesthetic score. Additionally, we exclude videos with cartoon-style visuals, conspicuous logos, subtitles, and similar features. Totally, we collect 650 real-world videos to include in the benchmark.

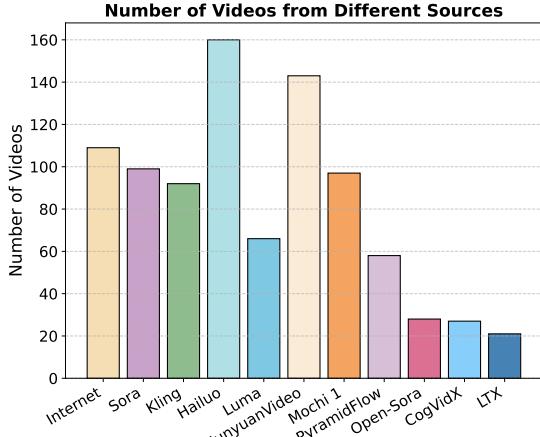


Figure 4: Sources of Impossible Videos.

3.2.2. HUMAN ANNOTATION.

Step 1: Video Filtering. Our objective is to curate a collection of high-quality videos that depict impossible or counterfactual scenes. To ensure data quality, we develop a custom annotation tool and perform human annotation on the collected videos. A detailed description of the annotation tool is provided in the Appendix A.1. The main criteria of IPV-VID contains two aspects: 1) *Visual Quality*. Videos of low visual quality are excluded, including but not limited to spatial blurring, temporal flickering, poor aesthetic quality, insufficient motion, and inconsistent style. 2) *Semantic Plausibility*. In line with the benchmark’s objective, we retain videos with semantically counterfactual content—scenes that are impossible in the real world.

This ensures that the selected videos exhibit high visual quality while maintaining low semantic plausibility, thereby requiring models to reason based on semantic content rather than low-level visual features. Fig. 4 illustrates the distribution of retained impossible videos after filtering.

Step 2: Detailed Annotation. For videos satisfying the aforementioned criteria, we perform detailed annotation with the following labels: 1) *Spatial or Temporal Anomaly*. This field requires annotators to determine whether the impossibility can be identified through spatial semantic information or necessitates temporal reasoning; 2) *Taxonomy Category*. Annotators assign a category label based on the IPV taxonomy; 3) *Explanation*. Annotators provide a brief textual description of the specific impossible phenomenon depicted in the video.

3.2.3. TASK DESIGN.

Judgment Task requires models to classify the input video as either synthetic or real by answering the question, “Is the provided video generated by AI?” To minimize the influence of visual elements and style, ensuring models focus on semantic content, we use synthetic videos without watermarks

and exclude real-world videos with cartoon-style visuals, conspicuous logos, subtitles, or similar features. To ensure a balanced evaluation, the dataset maintains a 1 : 1 ratio of synthetic to real-world videos. This task is framed as a binary classification problem and evaluated using average Accuracy and F1-score. Additionally, we report the ‘yes rate’ in Appendix B.1 to facilitate model diagnosis.

Multi Choice Task (MCQA) task requires models to identify the description that best captures the impossible phenomenon depicted in the video. The question is formulated as follows: ‘Select the best answer to the following multiple-choice question based on the video’. To create effective distractors that challenge the model, we carefully design an instructional prompt and leverage the GPT-4o model for distractor generation.

The instructional prompt is designed with the following considerations: 1) Ensure all options, including both correct answer and distractors, are similar in length, style, detail degree, and complexity; 2) Ensure distractors also present specific impossible phenomenon; 3) Ensure the impossible phenomenon in distractor shall involve visual elements shown in the given video frame, to avoid the model solve the problem with simple visual element grounding. By incorporating detailed annotations and visual content as references, we mitigate hallucination in GPT-4o, ensuring high-quality distractor generation. A qualitative example of distractors is provided in Fig. 6. The instructional prompt is included in Appendix C. The MCQA task is treated as a multi-class classification problem and evaluated using mean Accuracy.

Open-ended QA Task (OpenQA). We introduce an open-ended impossible explanation task, which requires models to *independently* and *correctly* identify the impossible phenomenon depicted in the video without any hints. The task is framed with the following question: “Based on your observation of the video, what content or event makes the video impossible or unusual in the real world?”. Compared to the MCQA task, the open-ended explanation task is more challenging, as models must generate responses without reference to candidate options. This task more accurately assesses whether models can genuinely perceive and articulate detailed anomalies, rather than relying on guesswork.

For evaluation, we employ an LLM as an evaluator to score model responses by comparing them to the annotated text explanations in the benchmark. Empirically, we observe that directly instructing the LLM to assign scores results in instability. To address this, we propose a *justification-then-score* approach. In this approach, the evaluator first provides a justification by identifying key matches or mismatches between the model’s prediction and the ground truth. Based on this justification, the evaluator assigns a semantic alignment score on a scale from 0 to 1, where 1.0 indicates perfect alignment, 0.8-0.9 indicates good alignment, 0.5-0.7 indi-

Table 2: Evaluation Results of IPV-TXT Across Dimensions. This table compares the performance of state-of-the-art video generation models using the IPV-TXT benchmark as text prompts in the T2V setting. A higher score indicates better performance in a given dimension. **Bold** denotes best, underline denotes second.

Model	Physical		Biological		Social		Geographical		Overall		
	Visual Quality	Prompt Following	IPV Score								
<i>Open-source Models</i>											
LTX (HaCohen et al., 2024)	58.3	14.1	35.3	21.6	44.0	16.0	57.1	23.8	52.3	16.5	10.0
Open-Sora (Zheng et al., 2024a)	63.8	22.1	25.5	29.4	20.0	32.0	57.1	47.6	51.5	26.5	15.8
Pyramid-Flow (Jin et al., 2024)	88.3	15.3	92.2	21.6	72.0	28.0	100.0	52.4	88.5	20.8	19.6
CogVidX-1.5 (Yang et al., 2024b)	38.7	29.4	35.3	<u>52.9</u>	40.0	<u>56.0</u>	81.0	<u>61.9</u>	41.5	<u>39.2</u>	16.9
Mochi 1 (GenmoAI, 2025)	68.7	44.2	64.7	56.6	80.0	60.0	71.4	76.2	69.2	50.8	37.3
HunyuanVid (Kong et al., 2024)	95.1	23.9	88.2	35.3	88.0	40.0	90.5	42.9	92.7	29.2	26.2
<i>Proprietary Models</i>											
Luma (LumaLabs, 2025)	88.3	11.7	90.2	19.6	82.6	17.4	85.7	52.4	88.0	17.1	14.3
Sora (OpenAI, 2025)	98.8	15.3	98.0	43.1	100.0	30.4	<u>95.2</u>	<u>61.9</u>	98.4	26.0	25.2
Kling (KLING, 2025)	98.8	21.5	<u>94.1</u>	33.3	<u>95.7</u>	43.5	81.0	42.9	<u>96.1</u>	27.5	26.7
Hailuo (Hailuo, 2025)	<u>96.3</u>	<u>30.1</u>	<u>94.1</u>	45.1	100.0	52.0	90.5	<u>61.9</u>	95.8	37.7	<u>36.2</u>

cates partial alignment, 0.1-0.4 indicates weak alignment, and 0.0 indicates no alignment. The justification step is critical to ensure fair and stable score assignment. By default, we employ GPT-4o as the primary evaluator. To avoid self-evaluation bias, as GPT-4o is also evaluated as a VideoLLM, we additionally utilize Claude-3.5-Sonnet as an evaluator.

Quality Assessment. Since the task of MCQA and OpenQA is constructed or evaluated with the help of GPT-4o, which risks hallucination, we conduct a human assessment. Specifically, for each task, we randomly select a subset of 100 samples and ask human to answer the questions, imitating the video understanding models. After that, the accuracy or score serve as a golden reference in models’ evaluation.

4. Evaluate Impossible Video Generation

Setup. We evaluate mainstream video generation models, including open-source (Open-Sora 1.2 (Zheng et al., 2024a), HunyuanVideo (Kong et al., 2024), CogVidX (Yang et al., 2024b), Mochi 1 (GenmoAI, 2025), LTX (HaCohen et al., 2024), Pyramid-Flow (Jin et al., 2024)) and closed-source models (Sora (OpenAI, 2025), Kling 1.5 (KLING, 2025), Luma (LumaLabs, 2025), and Hailuo (Hailuo, 2025)).

IPV-Score Metric. To evaluate a model, we first generate a set of videos using the IPV-TXT prompt suite. Human annotators are then employed to label two aspects of each video: 1) *Visual quality*, assessing whether the video meets high-quality standards, and 2) *Impossible prompt following*, determining whether the video accurately depicts the impossible event described in the text prompt. Annotators provide binary labels for each dimension. Using these labels, we introduce the IPV-Score, a novel metric to evaluate a model’s ability to generate high-quality videos that faithfully depict impossible events. The IPV-Score is calculated as the percentage of videos that are both high-quality and faithful to the prompt within the entire generated set.

Human- and Auto-eval. Tab. 2 presents the performance of the evaluated models. Additionally, we propose an au-

tomated evaluation method to assess a model’s capability in generating impossible videos. A detailed analysis is provided in the Appendix B.2.

4.1. Results and Analysis

Can today’s video generation models effectively follow prompts to create impossible video content? As shown in Tab. 2, the top-performing model, Mochi 1, generates high-quality impossible videos in only 37.3% of cases. Other models perform even worse, suggesting that current video generation models remain far from achieving satisfactory performance in generating high-quality impossible videos.

Unbalanced capability between visual quality and impossible prompt following. For instance, Luma demonstrates remarkable visual quality, surpassing most open-source models, but its prompt-following score is significantly lower. In contrast, some open-source models, such as Mochi 1, exhibit superior prompt-following capabilities, even outperforming many proprietary models. An ideal model should excel in both dimensions, achieving a balance that is quantified by our IPV-Score metric.

What limits the the model on impossible video generation? Beyond basic prompt comprehension and following, we identify two unique challenges posed by impossible text prompts, as illustrated in Fig. 5. *First, impossible prompts may trigger low visual quality.* While the model attempts to follow the prompt, the unusual nature of impossible prompts often results in visual artifacts or generation failures. This is likely because impossible prompts represent out-of-distribution data for the model. *Second, an overemphasis on adhering to physical laws may constrain the model’s creative freedom.* In many failure cases, videos accurately capture the semantic elements of the prompt but fail to depict the critical impossible phenomenon. Instead, they depict normal scenes that conform to real-world rules. Future video generation models may consider the above factors to achieve a better model. Additional examples are included in the Appendix E.

Impossible Videos

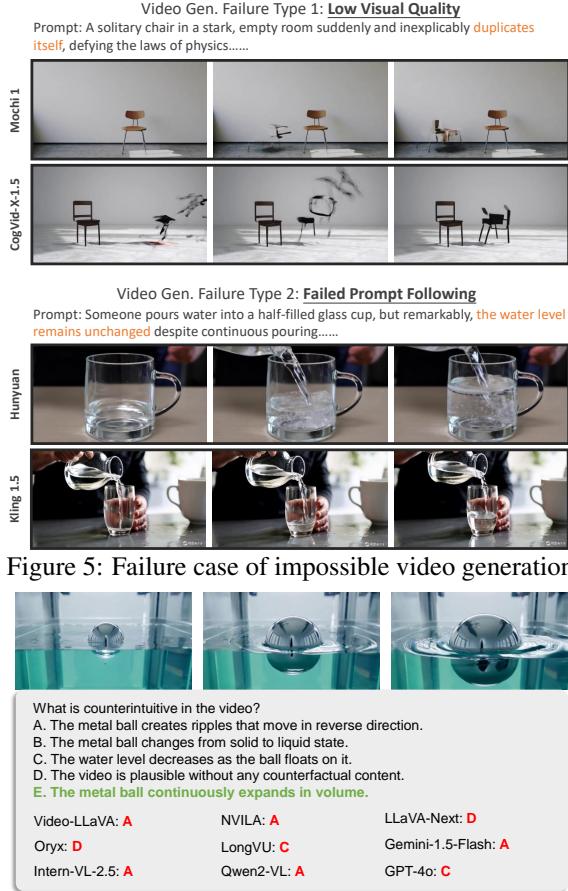


Figure 5: Failure case of impossible video generation.

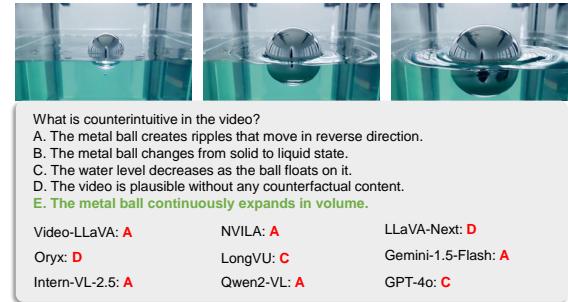


Figure 6: Example of the MCQA task. We highlight the correct option in **red** and the incorrect option in **green**.

5. Evaluate Impossible Video Understanding

We evaluate a diverse set of popular Video-LLMs, including both open-source and proprietary models. The three tasks are organized hierarchically, with increasing difficulty, to comprehensively assess model capabilities. Table 3 reports the evaluation results for impossible video understanding.

5.1. Results

Judgment Task. Most models achieve comparable Accuracy and F1-scores on this task. Qwen2-VL achieves the highest accuracy at 76.2%, even outperforming Gemini by 3.1 percentage points. Empirically, we observe that some models exhibit bias, predominantly answering ‘Yes’ (or ‘No’) for the majority of videos. To account for this bias, we also report F1-scores, with Video-LLaVA achieving the highest performance. Detailed results and additional analysis are provided in the Appendix B.1. GPT-4o reject answering “Yes” or “No” for this task. To ensure fair comparison, we do not further tune the text prompt for it.

Multi Choice Task. Model performance on this task exhibits significant variation. The top-performing model, LLaVA-Next, achieves an accuracy of 86.4%, surpassing both GPT-4o and Gemini. In contrast, Video-LLaVA



LongVU: The video shows a **cat eating from a bowl of food** that is placed on the floor. **This is a normal and expected behavior** for a cat, so there is no content or event that makes the video impossible or unusual in real-world.

Qwen2-VL: The video shows a cat eating a large amount of cat food in a short amount of time. **This is unusual because cats typically eat smaller amounts of food at a time and do not eat as much as shown in the video.** Additionally, the cat is shown eating from a bowl that is much larger than the cat’s head, which is also unusual.

NVILA: The video depicts a **cat eating a large amount of food** in a short period of time, which is unusual and counterintuitive in real-world scenarios.

LLaVA-Next: The video is impossible in real-world because the cat is not eating the food, but instead, it is **pouring the food into the bowl**.

Intern-VL-2.5: The video depicts a cat defying the laws of physics by lifting a large quantity of kibble with its mouth and then **pouring it into a bowl**. This is impossible in the real world as **cats do not have the physical ability to lift and pour such a large amount of kibble**.

GPT-4o: The video shows a **cat seemingly pouring an endless stream of kibble from its mouth onto a plate**, creating an unusually large pile. This is counterintuitive and impossible in the real world because:

1. **Endless Supply:** **A cat cannot store or produce such a large amount of kibble in its mouth.**
2. **Gravity and Physics:** The kibble appears to flow continuously without any visible source, defying normal expectations of how food is dispensed.
3. **Cat Behavior:** **Cats do not typically dispense food in this manner.**

Human Annotation: **A cat unnaturally generates cat food from its head, which is physically impossible in reality.**

Figure 7: Example of the OpenQA task. We ask state-of-the-art video understanding models to analyze whether the video is impossible or not. We highlight the correct analysis in **red** and the incorrect analysis in **green**.

achieves only 26.8% accuracy, which is close to the random baseline. Most open-source models exhibit substantial room for improvement. Fig. 6 presents an example of the MCQA task, where most models fail to select the correct option.

Open-ended QA Task. Among the three tasks, the OpenQA task is the most conceptually straightforward yet practically challenging (Fig. 7). The last two column of Tab. 3 indicate that most current models struggle to independently identify counterfactual phenomena in videos. Among open-source models, LLaVA-Next achieves superior scores with both GPT-4o and Claude-3.5 evaluators, aligning with its strong performance on the MCQA task. GPT-4o attains the highest performance with both evaluators, underscoring its robust visual understanding and reasoning capabilities. The overall scores suggest that most models struggle to comprehend impossible videos, highlighting a promising direction for future research.

5.2. Analysis

Are today’s video understanding models good enough for understanding impossible videos? The results from the MCQA and OpenQA tasks collectively provide insights into this question. Overall, proprietary models show promising potential, outperforming most open-source models on this task. However, their ability to independently identify impossible videos remains suboptimal, as evidenced by the OpenQA scores. Most open-source models perform poorly on this task, indicating significant room for improvement.

Unbalanced model capabilities across domains and tasks. Tab. 3 reveals that many models exhibit unbalanced capabilities across domains and tasks. Across domains, “Physical”

Table 3: Evaluation Results for Impossible Video Understanding. This table compares the performance of sota VideoLLMs using the *IPV-Vid* benchmark. A higher score indicates better performance in a given dimension. **Bold** denotes best, underline denotes second, Open(C) denotes using Claude as the evaluator.

Model	Judgement		Physical		Biological		Social		Geographical		Overall		
	Acc.	F1	MC	Open	MC	Open	MC	Open	MC	Open	MC	Open	Open(C)
<i>Open-source Models</i>													
Random	50.0	50.0	20.0	-	20.0	-	20.0	-	20.0	-	20.0	-	-
Human	-	-	-	-	-	-	-	-	-	-	94.0	82.7	-
Video-LLaVA (Lin et al., 2023a)	72.7	72.9	23.0	13.3	34.6	29.7	31.8	16.9	24.1	25.9	26.8	14.2	18.7
Oryx (Liu et al., 2024e)	58.6	44.6	52.0	16.7	68.8	33.9	70.3	32.6	83.9	35.2	60.4	18.7	22.7
Intern-VL-2.5 (Chen et al., 2024b)	56.5	69.7	61.9	34.9	64.2	59.6	65.5	49.6	77.0	48.4	62.4	33.0	32.5
NVILA (Liu et al., 2024f)	72.6	64.0	60.2	26.4	63.3	52.3	68.9	36.6	69.0	42.9	62.6	26.8	30.6
LongVU (Shen et al., 2024)	70.3	68.0	69.3	21.8	79.6	38.3	77.0	35.6	77.0	35.6	73.3	21.9	25.4
Qwen2-VL (Wang et al., 2024)	76.2	<u>71.1</u>	69.1	32.8	75.8	56.3	75.0	48.3	75.9	53.0	71.4	31.7	33.7
LLaVA-Next (Liu et al., 2024a)	73.2	70.4	82.8	37.7	92.9	57.2	90.5	51.3	90.8	51.4	86.4	34.4	<u>38.6</u>
<i>Proprietary Models</i>													
Gemini-1.5-Flash (Team et al., 2024)	73.1	64.0	<u>80.5</u>	48.5	90.8	66.4	<u>89.2</u>	<u>59.3</u>	93.1	<u>65.3</u>	<u>84.4</u>	<u>42.5</u>	36.8
GPT-4o (Hurst et al., 2024)	-	-	76.7	58.2	83.8	<u>75.5</u>	84.5	64.9	92.0	71.1	79.7	49.1	<u>45.1</u>

Table 4: Video understanding evaluation results on two categories of videos: 1) videos that can be understood through **spatial scene understanding and world knowledge, and 2) videos that require **temporal** reasoning for comprehension.**

Model	Spatial		Temporal	
	MC	Open	MC	Open
<i>Open-source Models</i>				
Random	20.0	-	20.0	-
Video-LLaVA (Lin et al., 2023a)	34.6	30.3	23.8	13.7
Oryx (Liu et al., 2024e)	83.2	41.0	51.4	17.5
Intern-VL-2.5 (Chen et al., 2024b)	72.4	56.6	60.6	37.2
NVILA (Liu et al., 2024f)	75.2	51.6	57.6	28.0
LongVU (Shen et al., 2024)	85.7	42.6	68.5	22.7
Qwen2-VL (Wang et al., 2024)	79.7	57.6	68.2	34.4
LLaVA-Next (Liu et al., 2024a)	95.5	55.8	82.7	39.8
<i>Proprietary Models</i>				
Gemini-1.5-Flash (Team et al., 2024)	93.0	67.5	81.1	50.0
GPT-4o (Hurst et al., 2024)	90.6	75.2	75.6	59.1

emerges as the most challenging, with most models achieving the lowest scores in this category. In contrast, the remaining 3 domains exhibit relatively balanced performance across models. We hypothesize that “*Physical*” contains more challenging samples that necessitates temporal dynamic reasoning. Across tasks, performance on the MCQA and OpenQA tasks is closely correlated, whereas the Judgment task appears to be distinct. For instance, Video-LLaVA excels in the Judgment task but underperforms significantly in the other tasks. Discrepancies also exist between MCQA and OpenQA performance. LLaVA-Next shows strong performance on MCQA but underperforms on OpenQA. In contrast, GPT-4o achieves the highest performance on OpenQA but lags slightly behind on MCQA.

What makes a good model for impossible video understanding? We identify two critical factors: temporal dynamic reasoning and world knowledge reasoning. For instance, in the first example from Fig. 1, the cookie appears plausible in individual frames. However, only by analyzing and reasoning across frames can one detect the self-growing phenomenon—an impossible event. Conversely, the last example in Fig. 1 (“snowing in Singapore”) requires world

knowledge for identification—specifically, the understanding that Singapore is a tropical country. Due to space limit, additional case studies are provided in the Appendix F.

Challenge and opportunity on temporal reasoning. Of the two key factors, world knowledge is primarily governed by the LLM, whereas temporal reasoning offers greater design flexibility but remains more challenging. Tab. 4 provides a separate evaluation of spatial- and temporal-focused videos. For all models, scores on temporal-focused videos are consistently lower than those on spatial-focused videos. This clearly demonstrates that temporal dynamic reasoning poses significant challenges for most current models. Video expert models with high frame rates (e.g., LongVU) do not exhibit a significant advantage. Interestingly, the top-performing models (e.g., LLaVA-Next and GPT-4o) are all image-based. It is worth noting that GPT-4o is evaluated using only 1 FPS. This observation suggests that more sophisticated temporal modules, rather than simply expanding the context window, may be key to understanding and reasoning about impossible videos.

6. Conclusion

In this work, we introduce the concept of *Impossible Videos* as a novel testbed to challenge and advance video understanding and generation models. Unlike real-world videos, impossible videos depict counterfactual or physically implausible scenarios, requiring models to move beyond memorization and retrieval toward deeper reasoning and generalization. To support research in this direction, we propose IPV-BENCH, a comprehensive benchmark consisting of a structured taxonomy, a diverse prompt suite (IPV-TXT), and a high-quality video dataset (IPV-VID). Extensive evaluations show that current video models struggle with impossible videos, revealing critical limitations in their reasoning abilities under non-realistic conditions. These findings shed light on existing gaps and point to promising directions for future research.

Impact Statement

This work introduces *Impossible Videos*, a benchmark for evaluating video models on counterfactual and anti-reality scenarios. By focusing on violations of physical, biological, geographical, and social laws, it pushes video generation and understanding models beyond imitation toward deeper reasoning and generalization. Our findings reveal current limitations in temporal dynamics and world knowledge reasoning, and offer a new testbed for safer, more robust video AI. However, benchmarks involving implausible or surreal content may be misused to deliberately generate deceptive or misleading media. To mitigate this risk, our benchmark is designed for research purposes, with a focus on transparency, educational value, and responsible model development.

Acknowledgment

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-030).

References

- Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- Bai, S. and An, S. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.
- Bai, Z., Nakashima, Y., and Garcia, N. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5422–5432, 2021.
- Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., and Shou, M. Z. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2025.
- Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.-W., and Grover, A. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.
- Blattmann, A., Dockhorn, T., Kulal, S., Mendelievitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Chen, H., Hong, Y., Huang, Z., Xu, Z., Gu, Z., Li, Y., Lan, J., Zhu, H., Zhang, J., Wang, W., et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024a.
- Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.
- Fan, K., Bai, Z., Xiao, T., Zietlow, D., Horn, M., Zhao, Z., Simon-Gabriel, C.-J., Shou, M. Z., Locatello, F., Schiele, B., et al. Unsupervised open-vocabulary object localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13747–13755, 2023.
- Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mmme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- GenmoAI. Mochi 1: A new sota in open-source video generation models. <https://www.genmo.ai/blog>, 2025. Accessed: 2025-01-28.
- HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.
- Hailuo. Hailuo ai: Transform idea to visual with ai. <https://hailuoai.video/>, 2025. Accessed: 2025-01-28.
- Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., and Lin, Z. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.
- KLING. Kling ai: Next-generation ai creative studio. <https://klingai.com/>, 2025. Accessed: 2025-01-28.
- Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.
- Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.
- Lin, K. Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A. J., Yan, R., and Shou, M. Z. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023b.
- Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.
- Liu, H., Yan, W., Zaharia, M., and Abbeel, P. World model on million-length video and language with ringattention. *arXiv e-prints*, pp. arXiv–2402, 2024c.
- Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., and Hou, L. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024d.
- Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., and Rao, Y. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024e.
- Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024f.
- LumaLabs. Luma dream machine. <https://lumalabs.ai/dream-machine>, 2025. Accessed: 2025-01-28.
- Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., and Luo, P. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024a.
- Meng, F., Shao, W., Luo, L., Wang, Y., Chen, Y., Lu, Q., Yang, Y., Yang, T., Zhang, K., Qiao, Y., et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024b.
- Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., and Tai, Y. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.
- Ni, Z., Yan, Q., Huang, M., Yuan, T., Tang, Y., Hu, H., Chen, X., and Wang, Y. Genvidbench: A challenging benchmark for detecting ai-generated video. *arXiv preprint arXiv:2501.11340*, 2025.
- Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., and Yuan, L. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.
- OpenAI. Sora. <https://openai.com/sora/>, 2025. Accessed: 2025-01-28.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.
- Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.
- Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Wang, L., Bai, Z., Zhang, Y., and Lu, H. Show, recall, and tell: Image captioning with recall mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12176–12183, 2020.

- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.
- Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.
- Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., and Shou, M. Z. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., and Schuurmans, D. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024a.
- Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.
- Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024.
- Zeng, A., Yang, Y., Chen, W., and Liu, W. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024.
- Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., and Shou, M. Z. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pp. 1–15, 2024.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.
- Zhao, Z., Wang, J., Horn, M., Ding, Y., He, T., Bai, Z., Zietlow, D., Simon-Gabriel, C.-J., Shuai, B., Tu, Z., et al. Object-centric multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16601–16611, 2023b.
- Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024a.
- Zheng, Z., Yan, X., Chen, Z., Wang, J., Lim, Q. Z. E., Tenenbaum, J. B., and Gan, C. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024b.
- Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., and Li, M. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020.

A. Additional Details of Benchmark Construction

A.1. Video Annotation Tool

Fig. 8 illustrates a screenshot of our data annotation tool. It has been divided into two zones: Display Zone showing the essential information for annotation, including original text prompt, the taxonomy label of the prompt, and the video. Annotation Zone provides several fields for annotations, including data for both IPV-VID curation and labels for video generation models evaluation. This all-in-one evaluation tool greatly eases the annotation efforts, supporting multiple ways for downstream usage. Specifically, the questions for each field is carefully designed to fit human customs. For example, although we aim to annotate “if this video is impossible”, we find it is more intuitive to answer “if this video is reasonable” in practice.

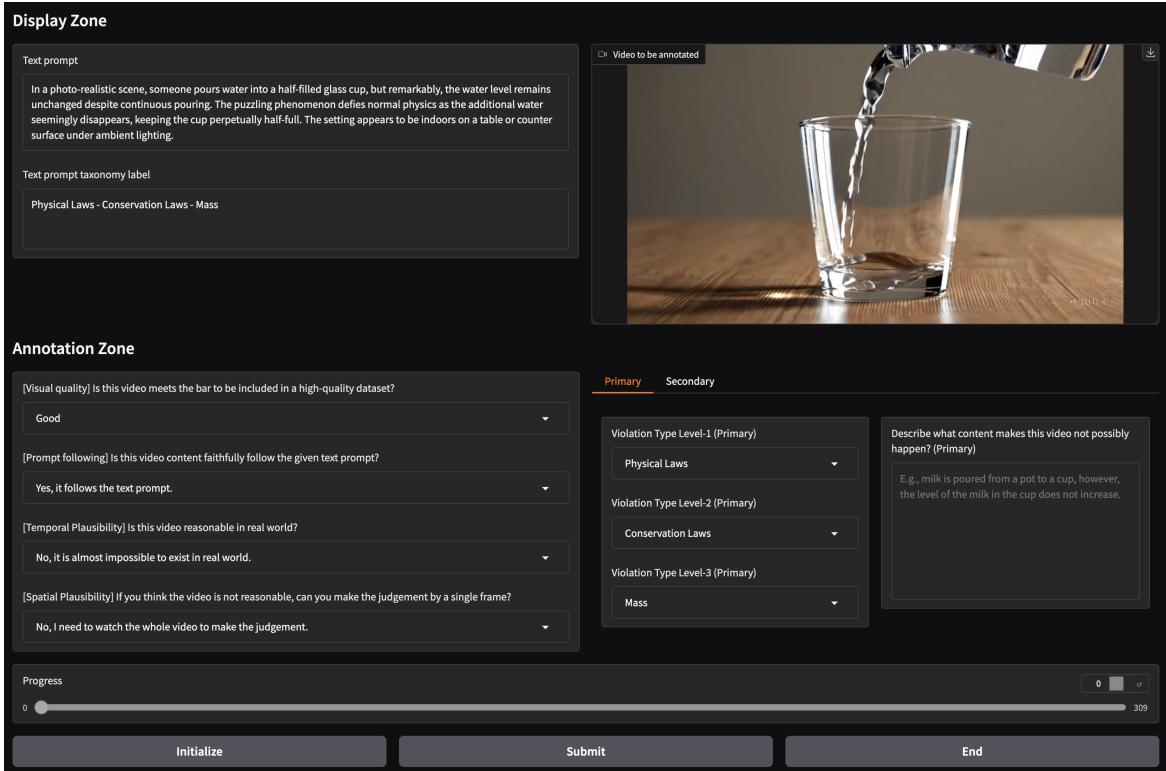


Figure 8: Screenshot of the annotation tool.

B. Additional Results and Analysis

B.1. AI-Generated Video Judgment

Tab. 5 presents more detailed results on the task of AI-Generated video judgment, helping understanding unique properties of each model. Video-LLaVA is a representative balanced model, with similar accuracy on both fake and real videos, and Yes Rate around 50%. In contrast, Intern-VL, NVILA and Gemini are significantly biased. Intern-VL is biased to answer ‘Yes’ with a high Yes Rate 93.5%. NVILA, and Gemini are biased toward the other direction, preferring answer ‘No’ with lower Yes Rate. The trends can also be reflected in the huge accuracy difference between fake and real videos.

B.2. Evaluating Impossible Video Generation – An Automatic Strategy

In this section, we introduce an automatic evaluation strategy as a surrogate to human score presented in the main text. We first generate a collection of videos based on the IPV-Txt prompt suite for each model. Then, we compute the the following metrics to evaluate the visual quality and impossible prompt following capability, respectively.

1) Visual Quality. We use the popular video quality assessment suite, VBench (Huang et al., 2024), to build a compound

Impossible Videos

Table 5: Detailed Evaluation Results for AI-Generated Video Judgment Task.

Model	Overall Acc.	Fake Acc.	Real Acc.	F1-Score	Yes Rate
Video-LLaVA (Ning et al., 2023)	72.7	72.9	72.5	72.9	50.3
Oryx (Liu et al., 2024e)	58.6	33.2	84.3	44.6	24.5
Intern-VL-2.5 (Chen et al., 2024b)	56.5	99.7	12.8	69.7	93.5
NVILA (Ni et al., 2025)	72.6	48.4	97.1	64.0	25.8
LongVU (Shen et al., 2024)	70.3	62.7	78.0	68.0	42.5
Qwen2-VL (Wang et al., 2024)	76.2	58.3	94.3	71.1	32.1
LLaVA-Next (Liu et al., 2024a)	73.2	63.3	83.2	70.4	40.2
Gemini-1.5-Flash (Team et al., 2024)	73.1	47.6	98.8	64.0	24.6

metric that measures overall video quality without considering the prompt. Specifically, we combine the six factors—Subject Consistency, Background Consistency, Motion Smoothness, Aesthetic Quality, Imaging Quality, and Dynamic Degree—from VBench to form our final metric. Similar to VBench, we calculate the weighted average of these six factors as the final evaluation score. We tailor the weights of these factors to better suit our video domain, such as reducing the weight of Aesthetic Quality since our impossible videos mostly follow a realistic style. The weights we use for each factor are: 2.0, 2.0, 0.2, 0.2, 2.0, 1.0.

2) *Impossible Prompt Following*. To assess whether the impossible event described in the text prompt is faithfully represented in the generated video. We utilize GPT-4o to provide a binary judgment for each video and calculate the following ratio as the final score. To achieve accurate judgment, we propose a three-step strategy to break down the task. Specifically, we prompt GPT-4o with the text prompt and frames sampled at 1 FPS. We instruct GPT-4o to: 1) identify and summarize the impossible event in the text prompt, 2) ground the impossible event in the video, and 3) confirm the visibility of all key elements that constitute a violation of common sense, reason, and conclude with a "Yes" or "No." We also provide two additional chain-of-thought examples as demo cases. Please see Fig. 12 for qualitative examples.

We compare the baseline prompting strategy with our prompting strategy on videos generated by Kling, where our approach better aligns with human annotations. See Tab. 6.

Table 6: Comparison between our prompt strategy and baseline prompt strategy for impossible prompt following evaluation.

Human Alignment	
GPT-4o	0.72
GPT-4o + our prompt strategy	0.80
Δ	11.1%

3) *IPV-Score*. We calculate the IPV-score as the product of the visual quality score and the impossible prompt following score, designed to assess the model's ability to generate high-quality, impossible videos. Since both scores are evaluated independently, their product effectively models the joint distribution of these factors. Before performing the multiplication, we further scale the ranges of both to better align them.

B.2.1. HUMAN ALIGNMENT

We calculate the Spearman's correlation coefficient (ρ) between our automatic evaluation score and human annotation score. Fig. 9 shows the corresponding results. We can see that the Spearman's correlation coefficient between the visual quality score, prompt following score, and IPV-score with human annotations remains above 0.8, demonstrating consistency with human annotators.

C. Instructional Prompt of MCQA Task

In Fig. 11, we present the complete instructional prompt used in the MCQA task. It has particularly designed rules for constructing distractors that challenge the reasoning capability of video understanding models.

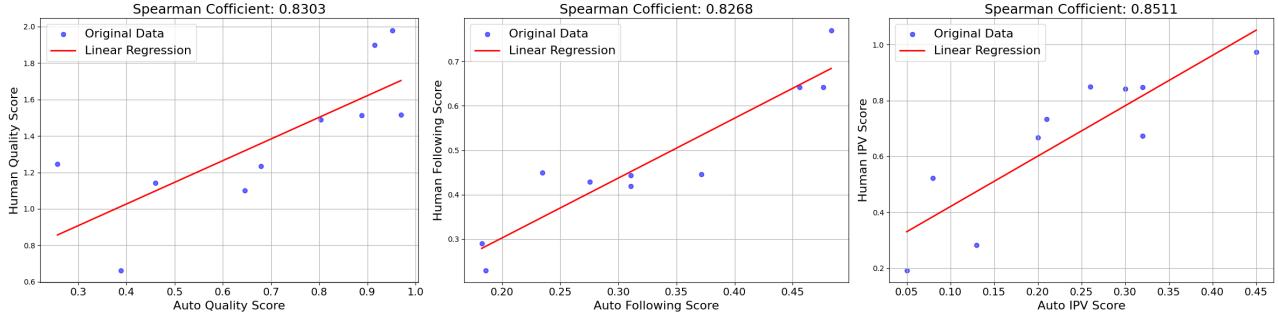


Figure 9: Spearman’s correlation coefficient ρ between automatic evaluation score and human annotation score on 10 different video generation models.

D. Qualitative Examples

Fig. 13 illustrates more video examples from the IPV-BENCH.

E. Case Study of Impossible Video Generation

In this section, we present three common failure examples of impossible video generation in Fig. 14.

Rows 1-2 show cases with high visual quality but that do not faithfully follow the video prompt; the generated videos adhere to the physical laws of the real world.

Rows 3-4 display cases with high visual quality, but they do not faithfully follow the video prompt to generate the designated impossible event, instead introducing other common-sense violations. For example, in the third row, the airplane engine is asymmetrical, and in the fourth row, one corner of the cookie moves independently.

Rows 5-8 show videos that follow the video prompt but suffer from poor visual quality, with issues like blurriness or an animated style.

F. Case Study of Impossible Video Understanding

In this section, we present a series of case studies of impossible video understanding in Fig. 15, Fig. 16, Fig. 17, Fig. 18, Fig. 19, Fig. 20. We observe that either strong open-source model, LLaVA-Next (Liu et al., 2024a) or close-source model, GPT-4o (Hurst et al., 2024) suffer from difficulty on understanding impossible phenomenon, particularly on temporal dynamics.

Scenario Brainstorming Questionnaire

Please come up with 10 dynamic video scenarios that meet the following requirements:

1. The scenario should be impossible (or extremely difficult) to achieve in the real world.
2. The scenario should be a dynamic video scene.
3. The scenario may seem reasonable in a static state but becomes unreasonable when in motion. For example, "a cat growing wings" is already highly unrealistic in a static scene, so such cases should be avoided.
4. Tip: Prioritize using common, everyday objects to create creative and unexpected scenarios. *Example: A face in a cookie starts talking.*
5. Tip: Consider scenarios that violate social norms or expectations. *Example: It snows in Singapore.*
6. Tip: You may also consider scenarios that violate physical laws. *Example: Water flows backward.*
7. Do not use ChatGPT or other large language models — use your own creativity!
8. Avoid overly abstract scenarios — the scene should have a reasonable degree of visual feasibility.

Figure 10: Questionnaire used for collecting impossible text prompts for IPV-TXT.

You are tasked with assisting in the creation of multiple-choice QA data for a video benchmark evaluating Video-Language Models. The goal is to generate a **correct answer** and a set of **plausible distractors** for each video based on the provided single description that emphasizes the counterfactual aspect.

Task Overview

For each video:

1. A description is provided, focusing on the counterfactual or impossible content.
2. An additional tag specifying the "impossible type" label (e.g., physical laws -- mechanics -- kinematics) will also be included. If there is any inconsistency between the description and the type tags, always prioritize the text description. The type tags are only for reference.
3. A single frame from the video will also be provided to aid in curating realistic distractors. The frame is only used to construct distractors and not to infer the correct answer.

You must generate:

- **One correct answer:** A precise, accurate explanation of the counterfactual aspect.
- **Three distractors:** Plausible but incorrect options, designed to test nuanced reasoning.

Guidelines

Correct Answer:

- Must align perfectly with the specific counterfactual event highlighted in the video, based solely on the text description.
- Should be concise and match the length, style, and level of detail of the distractors.
- Reduce unnecessary details or specificity to ensure the correct answer is not conspicuously longer or more descriptive than the distractors.

Distractors:

Ensure distractors have similar description length and, more importantly, detail degree to the correct answer!

1. Diverse Objects and Attributes (Prioritized):

- Focus on various aspects of the scene, including different objects and attributes, such as the motion of objects, colors, sizes, or environmental elements (e.g., "the ball's motion" "the grass's color").
- Leverage the provided video frame to identify objects not directly involved in the counterfactual phenomenon and create distractors around these objects or their attributes.

2. Impossible Phenomenon Option (Prioritized):

- Describes an alternative impossible phenomenon not present in the video. Ensure that the object is detectable in the video frame, but the described phenomenon does not occur. This type of distractor ensures that models rely on both video and text reasoning.
- Prioritize temporal dynamics, state changes, and motions, etc.

General Rules for All Options (including both Correct Answer and Distractors):

- Ensure all options, including both correct answer and distractors, are similar in length, style, detail degree, and complexity.
- Use the provided video frame effectively to curate distractors based on realistic details visible in the scene.
- Ensure distractors challenge reasoning rather than simple visual-text matching. This can be achieved by describing motions or state changes.

Input Format

1. A single description of the video, focusing on its counterintuitive aspect(s).
2. A tag specifying the type of impossibility (e.g., physical laws -- mechanics -- kinematics). The tag is for your reference.
3. A single frame from the video.

Output Format

Produce a dictionary where each key represents a counterfactual event (e.g., `event_1`, `event_2`), and the value is another dictionary with:

- `correct_answer`
- `distractors`

NOTE: You directly output the result without explanation or other words.

Input: {actual input}

Figure 11: Instructional Prompt of MCQA Task.

Impossible Videos

System prompt: You are a helpful video analyzer that can determine if a counter-intuitive or commonsense violation event in a text prompt is faithfully followed in the provided video frames. Your task is to analyze the video frames and check if the event described in the text prompt occurs, providing a detailed explanation for your answer.

User:



Prompt: A melted candle unexpectedly defies nature by regenerating itself upward.

Here is a video generated by AI from the provided text prompt. Please check whether the counter-intuitive or commonsense violation event in the text prompt is faithfully followed by the generated video and explain the reason. You should first identify and brief the counter-intuitive or commonsense violation event from the prompt. Then, carefully check if this event occurs in the video. Finally, summarize the reason and conclude your answer with: [Yes] or [No]. **Do not analyze other factors** such as video quality, etc. **Only focus on the counter-intuitive or commonsense violation event**. If the key elements constituting a violation of common sense cannot be directly observed and confirmed in the video, the violation of common sense is deemed not to have occurred.

Assistant: The prompt describes a scenario where a melted candle regenerates itself upward, defying gravity and appearing to reconstruct as if time is flowing backward.

In the frames provided, the candle appears to grow taller, with the wax extending upward, which aligns with the described counter-intuitive event. The wax seems to be reconstructing itself, suggesting a reverse process of melting.

Therefore, the video does demonstrate the commonsense violation event as described in the prompt.

The answer is: [Yes]

User:



Prompt: A desk Lamp unexpectedly illuminates and grows brighter after someone unplugs its power cord, defying normal electrical behavior.

Here is a video generated by AI from the provided text prompt. Please check whether the counter-intuitive or commonsense violation event in the text prompt is faithfully followed by the generated video and explain the reason. You should first identify and brief the counter-intuitive or commonsense violation event from the prompt. Then, carefully check if this event occurs in the video. Finally, summarize the reason and conclude your answer with: [Yes] or [No]. **Do not analyze other factors** such as video quality, etc. **Only focus on the counter-intuitive or commonsense violation event**. If the key elements constituting a violation of common sense cannot be directly observed and confirmed in the video, the violation of common sense is deemed not to have occurred.

Assistant: The prompt describes an event where a desk Lamp grows brighter after being unplugged, defying normal electrical behavior.

In the frames provided, the lamp is indeed glowing and appears to get brighter, but there is no visible indication of the lamp being unplugged or without a power source. This crucial aspect of the counter-intuitive event that the lamp remains illuminated without being connected to power is not evident in the images.

Without visual confirmation of the disconnection and the continued operation of the lamp, the video does not faithfully follow the commonsense-defying scenario described in the text prompt.

The answer is: [No]

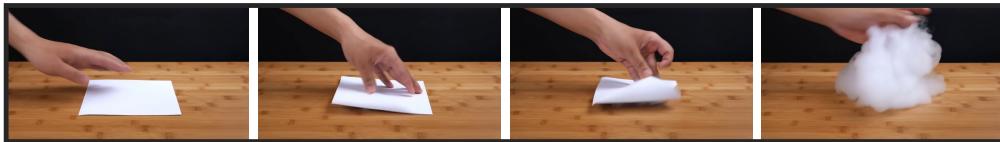
User: [New video frames] + [New prompt]

Here is a video generated by AI from the provided text prompt. Please check whether the counter-intuitive or commonsense violation event in the text prompt is faithfully followed by the generated video and explain the reason. You should first identify and brief the counter-intuitive or commonsense violation event from the prompt. Then, carefully check if this event occurs in the video. Finally, summarize the reason and conclude your answer with: [Yes] or [No]. **Do not analyze other factors** such as video quality, etc. **Only focus on the counter-intuitive or commonsense violation event**. If the key elements constituting a violation of common sense cannot be directly observed and confirmed in the video, the violation of common sense is deemed not to have occurred.

Figure 12: CoT examples we use to prompt GPT-4o for impossible prompt following evaluation. Videos may contain impossible events that are outside the scope of the prompt. We ignore such events when evaluating impossible prompt following.

Impossible Videos

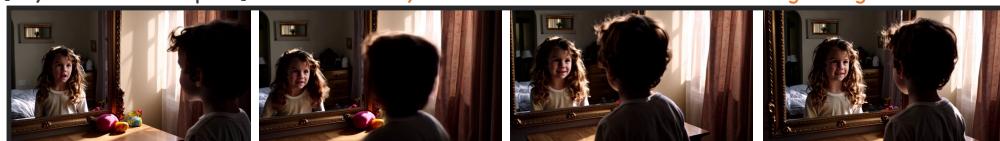
[Physical Laws → Material Properties] A paper **transforms into smoke** when being touched.



[Physical Laws → Mechanics] A car **flies** up into the sky, defying **gravity**.



[Physical Laws → Optics] A **short-hair boy's** reflection in the mirror shows a **long-hair girl**.



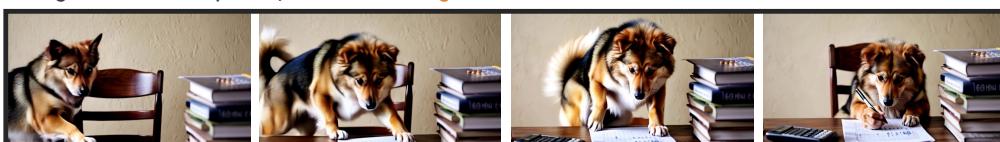
[Physical Laws → Material Properties] **White** chalk writes **pink** words on the blackboard.



[Biological Laws → Anthropomorphism] A **toy dog** **come to life**, playing with the toy bone.



[Biological Laws → Capability Violation] A **dog** **writes** on the table.



[Biological Laws → Morphology] A **crab** on the beach **grows two wings**.



[Social Laws → Magical Effects] A man **takes out food** from of a computer screen.



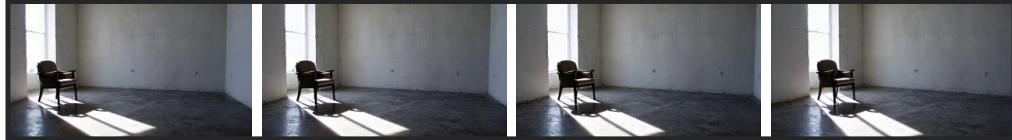
[Geographical Laws → Terrain and Environment] A **fountain** **appears** in the **desert**.



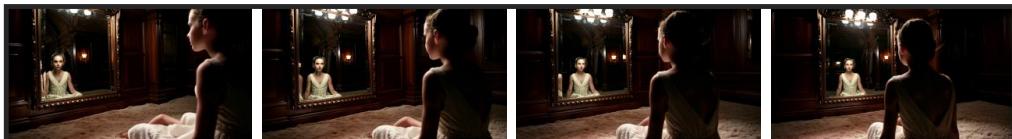
Figure 13: More examples of impossible videos.

Impossible Videos

[Physical Laws -> Conservation Laws] A solitary chair suddenly and inexplicably **duplicates itself**.



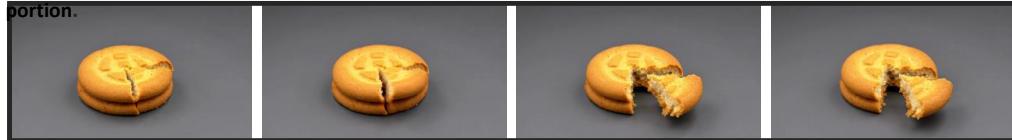
[Physical Laws -> Optics] A girl sits **motionless** in front of a mirror, her mirror reflection **moves independently**.



[Biological Laws -> Anthropomorphism] The rigid metal wings of a commercial airliner **flap** like a bird's wings.



[Physical Laws -> Conservation Laws] A partially eaten cookie magically **regenerates** its missing portion.



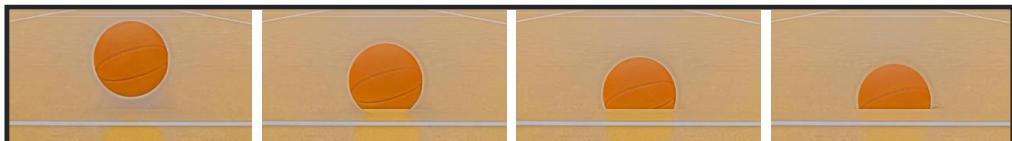
[Physical Laws-> Optics] A **yellow** car changes its color to **green**.



[Physical Laws -> Material Properties] A commercial **aircraft** mysteriously **transforms** into a luxury **car**.



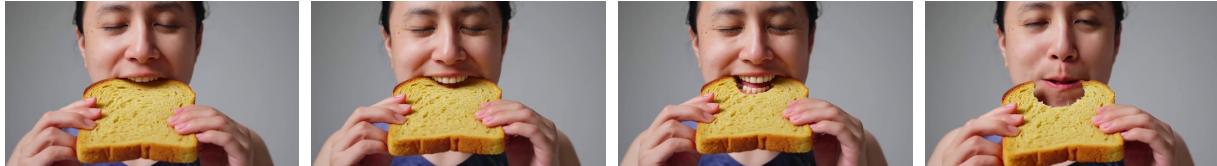
[Physical Laws-> Material Properties] A ball drops to the floor and **sinks** halfway into the ground.



[Social Laws -> Magical Transformations] Massive buildings suddenly begin **rotating**.



Figure 14: Failures in Generating Impossible Videos.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video depicts a woman eating a piece of bread, which is a common and normal activity. However, **the unusual aspect of the video is that the woman is eating the bread with her eyes closed**, which is not a typical way to eat. This could be interpreted as a playful or humorous gesture, but it is not a common way to eat bread.

Intern-VL-2.5: The video depicts a person eating a slice of bread, which is a common and everyday activity. **There is nothing in the video that makes the content impossible or unusual in the real world.** The actions shown are consistent with normal human behavior and the environment is typical for such an activity.

LongVU: The video content or event is impossible or unusual in real-world because the person is holding a piece of bread and **taking a bite out of it**.

Qwen2-VL: The video depicts a woman biting into a piece of bread and then immediately spitting it out. **This action is unusual in real-world scenarios because biting into a piece of bread typically involves chewing and swallowing, not spitting it out immediately.** The woman's reaction suggests that the bread might have a bitter or unpleasant taste, which is not a common occurrence with bread.

LLaVA-Next: The video is impossible in real-world because the person is holding **a slice of bread with no visible filling or topping**, and there is no context provided for why they are eating it.

GPT-4o: The sequence of images shows a person taking a bite of bread, **but the unusual aspect is that the bite mark appears before the person actually bites into the bread.** This suggests that the video might be edited or reversed, making it impossible in a real-world scenario where the bite mark should only appear after the person bites the bread.

Human Annotation: **A piece of toast mysteriously disappears as if bitten, despite no one actually taking a bite of it.**

Figure 15: Case study of impossible video understanding.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video is impossible or unusual in real-world because it shows two water bottles with the same amount of water, **but one bottle is frozen while the other is not**.

Intern-VL-2.5: The video is impossible in the real world because **it shows two plastic bottles filled with water that are frozen solid, which is not a common or physically possible state for water in a bottle**. Water freezes at 0 degrees Celsius, and if the bottles were filled with water, they would not remain solid without being in a controlled environment where the temperature is kept below freezing for an extended period. The video suggests that the water has frozen, but this is not a typical or realistic scenario for a simple demonstration.

LongVU: The video shows two identical plastic bottles of water placed side by side, **but one of them is frozen while the other is not**. This is an unusual event in the real world.

Qwen2-VL: The video shows two plastic bottles filled with water shaking violently on a table. **This is unusual because water is a liquid and should not shake violently unless there is an external force acting on it**, such as an earthquake or a strong wind. Additionally, the bottles are not attached to anything, so there is no reason for them to shake violently on their own. This makes the video impossible, unusual, or counterintuitive in real-world scenarios.

LLaVA-Next: The video is impossible in real-world because **the water level in both bottles rises simultaneously, which is not possible in reality**.

GPT-4o: The images depict **two water bottles with liquid inside that appears to be frozen or turning into ice**. This phenomenon is known as “supercooling,” where water remains in a liquid state below its freezing point and then rapidly crystallizes when disturbed. In a real-world context, this is unusual but not impossible. Supercooled water can instantly freeze when shaken or disturbed, which might be what’s happening here. This effect can be counterintuitive because we expect water to freeze at 0°C (32°F), but under certain conditions, it can remain liquid below this temperature until a disturbance triggers freezing.

Human Annotation: **The water level in the bottles spontaneously increases without any water source, which violates the law of conservation of mass.**

Figure 16: Case study of impossible video understanding.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video depicts a person pouring water into a container filled with sand. **This is an unusual event because sand is a dry substance and does not absorb water.**

Intern-VL-2.5: The video depicts a hand pouring a liquid into a container filled with sand, which is a common and expected action. However, **the unusual aspect is the presence of holes in the sand, which suggests that the sand might have been pre-drilled or has some other unusual property.** This could be counterfactual in a real-world scenario where sand is typically poured into a container without pre-existing holes. The video does not provide an explanation for the holes, which adds an element of mystery or could imply a specific purpose or experiment that is not immediately clear.

LongVU: The video shows a person pouring a clear liquid into a container of sand, which is an unusual event in the real world.

Qwen2-VL: The video depicts a scene where a person is pouring water into a container filled with sand. **This action is unusual in real-world scenarios because sand is not a suitable medium for holding water.** Sand is composed of small particles that are not dense enough to retain water, and the water would simply seep through the gaps between the particles. Additionally, the container appears to be a plastic container, which is not typically used for storing or pouring water due to the risk of water damage.

LLaVA-Next: The video is impossible in real-world because the liquid poured into the sand creates a swirling motion and forms a vortex, which is not possible in real-world physics.

GPT-4o: The images depict a scenario where a liquid is being poured onto sand, **but the sand remains dry and unaffected by the liquid. In the real world, pouring a liquid onto sand would typically result in the sand becoming wet and clumping together.** This unusual behavior suggests that the sand might be treated with a hydrophobic substance, making it repel water. This is counterintuitive because, under normal circumstances, sand would absorb the liquid.

Human Annotation: **The water completely disappears into the sand without leaving any wet marks or darkening the surface, which defies how water normally interacts with sand.**

Figure 17: Case study of impossible video understanding.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video depicts a tomato being cut in half, **but the tomato is not cut in half**. This is an impossible and unusual event in the real world.

Intern-VL-2.5: The video depicts a tomato being sliced in half, which is a common and expected action. However, **the unusual aspect is that the tomato is already sliced in half, and the video shows the process of slicing it again**. This repetition of the same action might be considered counterintuitive or unusual in a real-world context, as it does not add new information or progress the scene.

LongVU: The video shows a tomato being cut in half, **but the tomato is already sliced in half before the knife is used**. This creates a paradoxical situation where the knife is used to cut a tomato that is already cut in half.

Qwen2-VL: The video shows a tomato being sliced in half with a knife, **but the tomato is already cut in half**. This is an impossible event because the tomato cannot be cut in half while it is still whole.

LLaVA-Next: The video shows a tomato being cut in half, **but the tomato is not actually being cut**. This creates an illusion of the tomato being cut, which is unusual in real-world scenarios.

GPT-4o: The video depicts a knife cutting into a tomato, **but the tomato appears to be unusually large, almost the size of a small watermelon**. In reality, tomatoes do not grow to this size, making the scenario impossible or counterintuitive in the real world.

Human Annotation: **The tomato splits horizontally despite being cut vertically with a knife, creating an impossible mismatch between the cutting direction and splitting result.**

Figure 18: Case study of impossible video understanding.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video depicts a book that is impossible to open, unusual in that it is not a real book, and counterintuitive in that it is not a real book.

Intern-VL-2.5: The video is impossible in real-world terms because it shows a book opening and closing on its own without any visible human interaction or external force. This defies the laws of physics and the natural behavior of objects, as books do not have the ability to open and close autonomously.

LongVU: The video shows a book being placed on top of another book, which is a normal action.

Qwen2-VL: The video depicts a sequence of events that are physically impossible in the real world. Specifically, the video shows a book being stacked on top of another book, which then causes the first book to flip over and land on the second book. This sequence of events is impossible because books are not capable of flipping over in this manner. In reality, if a book is placed on top of another book, the top book would simply rest on the bottom book without flipping over. The video's depiction of the books flipping over is therefore a violation of the laws of physics and is not possible in the real world.

LLaVA-Next: The video shows a book being placed on top of another book, which is an impossible event in real-world physics.

GPT-4o: The sequence of images shows books appearing to multiply or move without any visible external force or interaction, which is impossible in the real world. In reality, books cannot move or duplicate themselves without someone or something physically interacting with them. This scenario defies the laws of physics, making it unusual and counterintuitive.

Human Annotation: The book spontaneously splits and replicates itself into three copies, violating the laws of physics and conservation of matter.

Figure 19: Case study of impossible video understanding.



Question: Based on your observation of the video, what content or event that makes the video impossible or unusual in real-world?

NVILA: The video depicts a person holding a bowl of food that is on fire, which is impossible in real-world scenarios.

Intern-VL-2.5: The video presents a counterfactual scenario where a laptop screen displays a bowl of food that is steaming, and the person interacts with the steaming food as if it were a physical object. In the real world, a laptop screen is a flat, two-dimensional display and cannot physically hold or emit steam from a bowl of food. The interaction with the steaming food on the screen is an unusual and impossible activity in reality.

LongVU: The video shows a laptop screen displaying a bowl of food that is on fire, which is an unusual and potentially dangerous situation.

Qwen2-VL: The video depicts a person interacting with a laptop that displays a bowl of food on fire. This scenario is impossible in the real world because it is not physically possible for a laptop screen to emit flames or smoke. The image on the screen is likely a digital or virtual representation, but the interaction with the laptop and the bowl of food on fire is a surreal and unrealistic concept.

LLaVA-Next: The video shows a bowl of food on the laptop screen with flames and smoke, which is not possible in real life.

GPT-4o: The video depicts a scenario where food and steam appear to be coming out of a laptop screen, which is impossible in the real world. Screens cannot produce physical objects or substances like food or steam. This is a digitally manipulated or staged effect, making it unusual and counterintuitive.

Human Annotation: A person reaches through a computer screen to grab food from a video, defying physical boundaries between digital and real worlds.

Figure 20: Case study of impossible video understanding.