A Dataset for Deepfake Videos Detection

Tianbo Yang, Keyush Shah, Viola Hu, Chung Un Lee, Christopher Shen, Taran Anantasagar, Yikai Mao, Yuanzhe Liu, Yinuo Xu, Xingyu Fu
University of Pennsylvania
{yangti, keyush06, chrshen, violahu, culee, tars, yikaimao, yuanzhel, xuyinuo, xingyuf2}@seas.upenn.edu

Figure 1: **Examples from the dataset**. Representative frames showcasing three types of visual artifacts in AI-generated videos: *distortion*, *merging*, and *blurring*. Each artifact is highlighted with human-annotated bounding boxes and accompanied by detailed explanations.

Abstract

This paper explores the detection of deepfake videos, focusing on the analysis of visual artifacts like deformation, blurring, and distortion, prevalent in AI-generated content. We employ a comprehensive ontology for artifact classification across different AI models, highlighting the challenges and limitations in current deepfake detection methodologies. Our findings from four annotated video batches demonstrate the significance of deformation as a dominant error type, especially in dynamic scenes, thereby advancing the strategies for robust deepfake recognition.

1 Introduction

Deepfake technology has profound implications for the integrity of visual media, influencing sectors ranging from news and journalism to entertainment and social media. As AI-generated content becomes increasingly indistinguishable from authentic recordings, the potential for misuse escalates, necessitating the development of robust detection and mitigation techniques. These challenges are underscored by the rapid advancements in generative AI models, which can produce highly convincing synthetic videos, posing a threat to public trust and digital content authenticity.

Significant progress has been made in detecting deepfake videos, particularly those involving human faces (Gu et al., 2022; Xu et al., 2024; Gu et al., 2021). Techniques leveraging spatiotemporal inconsistencies and hierarchical feature analysis have demonstrated promising results in distinguishing fake from real. However, these methods predominantly target facial manipulation, leaving a critical gap in addressing the broader spectrum of deepfake video content. Modern video generation techniques extend far beyond facial imagery, encompassing general video synthesis that includes complex scenes, dynamic motions, and diverse object interactions.

To tackle the challenge of AI-generated general video detection, this work introduces the following contributions:

A video dataset generated by State-of-the-Art Text-to-Video (T2V) models We curated a dataset that includes a variety of scenarios and actions to ensure comprehensive coverage of generative capabilities and limitations.

In-depth analysis of failure categories We systematically examine the types of visual artifacts present in AI-generated videos, providing a taxonomy of common issues such as deformation, distortion, object merging or splitting.

A high-quality annotated dataset We enriched the video dataset with detailed annotations and explanations that highlight prominent visual artifacts, facilitating targeted analysis and training of detection models.

This work aims to bridge the gap in general video detection by providing a robust framework for analyzing and addressing the challenges posed by advanced video generation techniques. The proposed dataset and analysis serve as critical resources for developing next-generation deepfake detection methods.

Dataset	Data	Classify	Explain
Gu et al.	Only Facial	✓	✗
Xu et al.	Only Facial	✓	✗
Chen et al.	General	✓	✗
Ours	General	✓	✓

Table 1: Comparison of Different Deepfake Video Detection Datasets

2 Related Work

2.1 Text-to-Video (T2V) Generation Models

The task of Text-to-Video (T2V) generation focuses on producing videos from textual prompts, leveraging advancements in Transformer architectures and diffusion models (Vaswani et al., 2023; Ho et al., 2020). Closed-source models like OpenAI Sora, Meta Movie Gen, and Pika have demonstrated remarkable capability in generating coherent and visually compelling video content from descriptive prompts (OpenAI, 2024; Polyak et al., 2024; Pika, 2024). Meanwhile, recent advancements in foundational models, such as Diffusion Transformers (DiT), have propelled open-source models like Mochi and CogVideoX to exhibit competitive performance in video generation tasks (Peebles and Xie, 2023; Team, 2024; Yang et al., 2024).

2.2 AI-generated Video Detection

Previously, people have put attention to machine generated text detection (Dugan et al., 2024; Ippolito et al., 2020) and AI generated images detection (Guo et al., 2023; Lorenz et al., 2023; Wu et al., 2023; Wang et al., 2023). Since Deepfake came out, people developed several techniques to detect fake videos (Gu et al., 2022; Xu et al., 2024; Gu et al., 2021). However, they focus on human faces, and current video generation go far beyond human faces. Detailed Mamba (DeMamba) (Chen et al., 2024) is one of the most recent effort to identify AI-generated videos by analyzing inconsistencies in both temporal and spatial dimensions.

3 Methods

To tackle the challenges of detecting artifacts in AI-generated videos, we developed a systematic pipeline for data curation and annotation, followed by a detailed analysis of the collected annotations.

3.1 Data Curation Pipeline

Our data curation pipeline follows a two-step process: prompt generation and video collection. This approach combines generative models with human filtering to create a high-quality dataset.

3.1.1 Prompt Collection

We utilize GPT-4o (OpenAI et al., 2024) for generating prompts in two stages. First, it enriches video generation prompts from existing benchmarks (Huang et al., 2023) by adding contextual detail to improve downstream video generation. Second, it generates new prompts with a focus on dynamic actions and objects in a realistic setting. Lastly, a human review process ensures these prompts are precise, relevant, and aligned with dataset objectives.

3.1.2 Video Collection

To ensure robust experimental outcomes, we curate a diverse collection of high-quality, realistic, and artifact-rich videos. The primary sources include state-of-the-art commercial models, such as Kling and Pika (Kling, 2024; Pika, 2024), known for their ability to generate coherent and visually compelling videos. Additionally, we incorporate videos from existing video repositories like VBench (Huang et al., 2023) to further enhance content diversity. All videos undergo a rigorous filtering process to remove irrelevant or inconsistent samples, ensuring alignment with predefined standards of coherence and activity relevance.. As of this report, our dataset comprises a total of 3,253 videos.

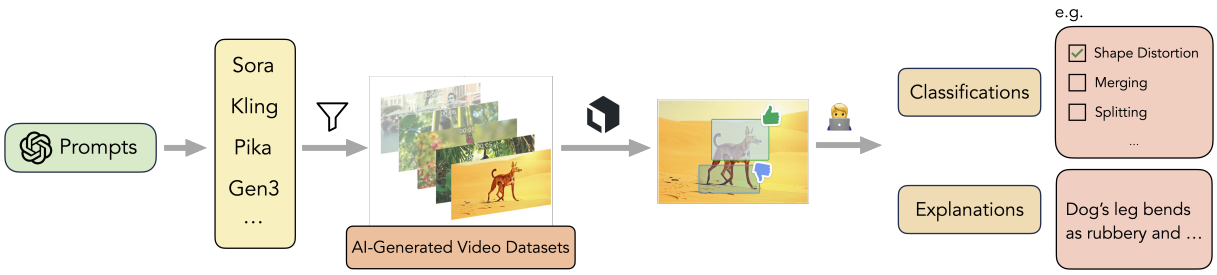

Figure 2: **Data curation pipeline.** GPT-4o (OpenAI et al., 2024) generates natural and realistic prompts, which are then fed into text-to-video (T2V) models to synthesize video datasets. The generated videos undergo a filtering process to exclude content depicting unrealistic or hypothetical scenes or topics. Subsequently, the curated videos are uploaded to LabelBox (LabelBox, 2024), where they are annotated with bounding boxes, classified into predefined categories, and supplemented with detailed textual explanations.

3.2 Annotation Design

For the collected AI-generated videos, we iteratively develop and refine an annotation framework to provide structured information and identify features indicative of AI-generated content. Initially, we explore a wide variety of categories to capture all potential visual clues, but many of these categories prove to be infrequent or irrelevant in most videos (see section 4.1).

Based on this insight, we narrow our focus to *deformation* artifacts, which emerges as the most prominent indicators of AI generation. Using LabelBox (LabelBox, 2024) (shown in Figure 3, we implement a structured annotation interface designed to capture specific visual inconsistencies, with an emphasis on deformation-related errors such as shape distortion, unnatural motion, and object merging.

3.3 Data Analysis

For data analysis, statistical measures were applied to identify patterns in artifact distribution, focusing on deformation's dominance as a marker of AI generation. Comparisons were drawn across batches and models to evaluate the consistency of artifact trends. Additionally, the annotated data was utilized to train and validate machine learning models, enabling an improved understanding of artifact detection capabilities. The results were analyzed to validate hypotheses around spatial-temporal inconsistencies in AI-generated content.

4 Results

4.1 Pilot Study Analysis

To explore the range of potential artifacts in AI-generated videos, we began with a pilot study on an initial batch of 1,097 videos, collected from Kling and Pika (Kling, 2024; Pika, 2024). This dataset served as the foundation for identifying recurring issues and iteratively developing a categorization schema (see Appendix A.1 for details of the schema). The count of each artifact type is illustrated in Table 2.

Label	Count
Deformation and Structural Issues	
Deformation	375
Blurring	203
Distortions	189
Object Merging or Overlapping	99
Disappearance	60
Appearance	54
Consistency	26
Unnatural Transitions	26
Temporal and Motion Issues	
Temporal Problem	144
Unsmooth Motion	81
Liquid Motion	16
Motion	7
Temporal Artifacts	1
Color and Lighting Issues	
Color Problem	19
Shadowing	15
Lighting	14
Other Issues	
I Can't Tell it's AI	91
Others	16

Table 2: Annotation stats of the first batch (1097 total) with higher-level categories

The analysis revealed deformation artifacts as the most prevalent, comprising approximately 24%

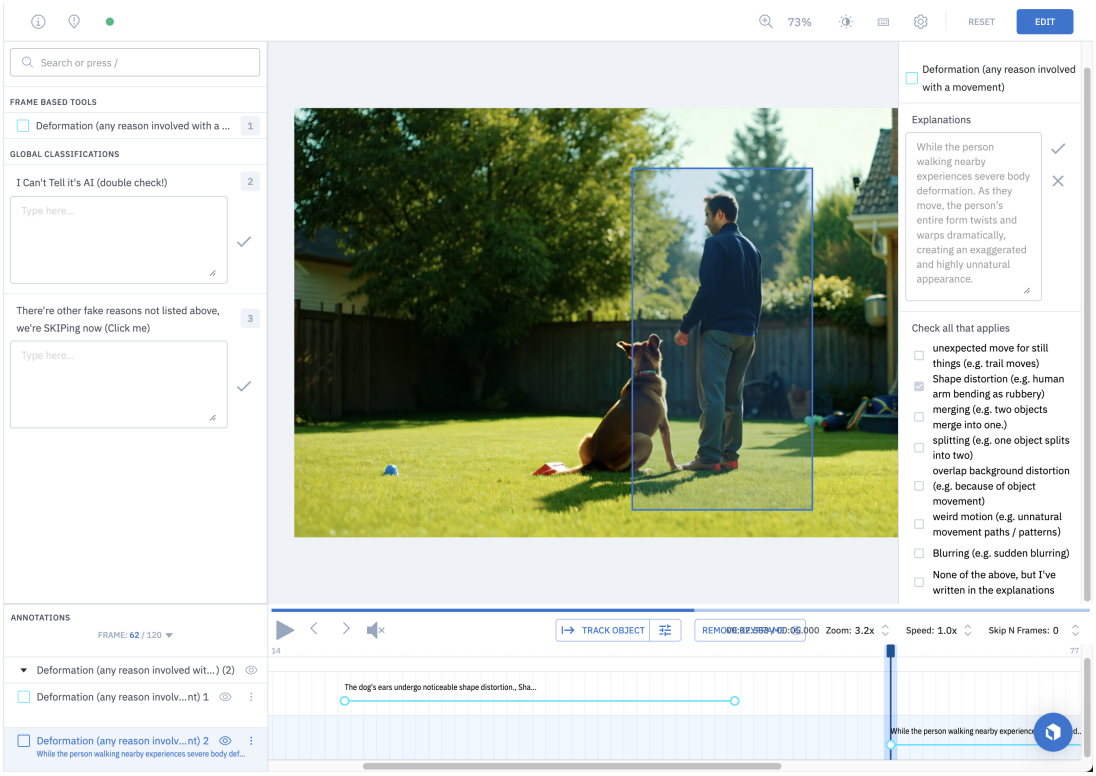

Figure 3: **Labelbox Annotation Interface.** Each video is annotated with bounding boxes highlighting specific sections across frames based on predefined categories, accompanied by qualitative natural language explanations.

of all detected issues. These deformations often involved unnatural bending or shape distortions, particularly during scenes with significant movement, such as dynamic human actions or crowded environments (see top of Figure 4). Blurring artifacts, the second most frequent category, accounted for about 13% of the annotations. Unlike motion-related blur seen in natural videos, these artifacts included abnormal blurring effects on static objects, indicating inconsistencies in rendering (see bottom of Figure 4).

These findings emphasize that deformation and blurring are frequent issues, particularly in dynamic scenes. This observation informed the subsequent refinement of our annotation framework for more in-detailed analysis.

4.2 Deformation-related Inconsistencies

To better understand deformation artifacts, we analyzed three batches of videos generated using Pika (Pika, 2024) under diverse scenarios, prompts, and model configurations. This analysis focused on deformation-related inconsistencies, including shape distortion, motion anomalies, and object interactions. Table 3 summarizes the distribution of annotated artifacts across these batches.

Our findings highlight systematic weaknesses in generative models, particularly in scenarios involving rapid motion, crowded environments, and visually complex scenes.

4.2.1 Scenario-Based Artifact Distribution

Across all batches, shape distortion consistently emerged as the most prevalent artifact, affecting 58%, 70%, and 52% of videos in the second, third, and fourth batches, respectively. These distortions were characterized by unnatural bending of limbs, elongation of body parts, or irregular morphologies. For example, the martial artist in the middle row of Figure 4 demonstrates severe limb distortions during dynamic movement, with unnaturally bent arms and legs. Such artifacts were particularly noticeable in action scenes involving rapid motion, where models struggled to maintain spatial consistency across frames.

Motion-related issues, including weird motion and unexpected movements, were prominent in fast-paced or erratic scenarios. Weird motion, observed in 48% of the second batch and 38% of the fourth batch, often manifested as jittering, unnatural acceleration, or erratic movement paths. An example is shown in the bottom row of Figure 4, where the legs of a walking subject display

Figure 4: **Example Frames of Deformation Artifacts.** Top row: *Merging/Overlapping*—instances of people's bodies blending together during dynamic motion in a crowded scene. Second row: *Shape Distortion*—unnatural elongation and bending of the martial artist's limbs during action. Third row: *Weird Motion*—awkward and inconsistent leg movement while walking, resulting in physically implausible dynamics. Bottom row: *Blurring*—loss of detail and smearing of textures, particularly noticeable in the wrestler's body during motion.

awkward and physically implausible movement over a lava-like surface, disrupting temporal coherence. Unexpected movements, though less frequent (5–9%), often appeared in scenes with sudden or complex actions, exacerbating temporal inconsistencies.

Merging and overlapping deformations, affecting 17–22% of cases, were particularly common in crowded scenes with multiple interacting objects. In the top row of Figure 4, individuals performing acrobatics blend unnaturally into each other, making their boundaries indistinct. Overlapping deformations, where objects merge with complex backgrounds, were often observed in visually intricate environments, such as crowded or heavily textured settings. These artifacts highlight the generative models' limitations in disentangling overlapping entities and maintaining object fidelity.

Blurring, affecting 20–37% of videos, occurred when models failed to render fine details, particularly during rapid movements. As shown in the bottom row of Figure 4, blurring artifacts degrade visual clarity, with the wrestler's body appearing smeared and lacking sharpness during motion. Such artifacts are especially prominent in high-speed actions like sprinting, martial arts, or dynamic transitions, where textures and object edges become indistinct due to poor temporal consistency.

4.2.2 Implications for Model Development and Annotation

These observations underscore the challenges that deformation artifacts pose to generative models:

1. Shape distortion and motion-related artifacts highlight deficiencies in both spatial consistency and temporal coherence, which are critical for realistic video generation.

2. Scene complexity, as seen in crowded or intricate environments, points to the need for better object disentanglement and rendering techniques.

3. Rapid motion artifacts reveal limitations in

Label	Second Batch		Third Batch		Fourth Batch	
	Count	% Share	Count	% Share	Count	% Share
Deformation - Shape Inconsistency						
Shape Distortion	370	58.08	543	70.43	320	51.86
Merging	142	22.29	134	17.38	108	17.50
Splitting	48	7.54	61	7.91	70	11.35
Blurring	128	20.09	286	37.09	154	24.96
Overlap with Background	37	5.81	37	4.80	16	2.59
Deformation - Motion-related Issues						
Weird Motion	308	48.35	250	32.43	232	37.60
Unexpected Movement	31	4.87	69	8.95	26	4.21
Overall						
Deformation	687	93.47	805	97.69	655	92.51
Reasons Except Deformation	40	5.44	7	0.85	37	5.23
Cannot Tell	8	1.09	12	1.46	16	2.26

Table 3: Merged annotations stats for the Second batch (735 total), Third batch (824 total), and Fourth batch (674 total). The table is organized into three sections: (1) **Deformation - Shape Inconsistency**, which includes specific issues like Shape Distortion, Merging, Splitting, Blurring, and Overlap with Background; (2) **Deformation - Motion-related Issues**, which covers motion-specific problems such as Weird Motion and Unexpected Movement; and (3) **Overall**, summarizing higher-level classifications including total Deformation, Reasons Except Deformation, and ambiguous cases categorized as Cannot Tell.

current temporal modeling, emphasizing the importance of capturing smooth transitions and natural dynamics.

5 Discussion

5.1 Findings and Implications

Our analysis highlights deformation as the most prevalent artifact in AI-generated videos, particularly in dynamic scenes with complex motion or interactions. This finding underscores significant weaknesses in current generative models, such as their inability to maintain temporal and spatial coherence. The inclusion of detailed artifact explanations and reasoning in our annotations enhances not only detection accuracy, but also the interpretability and reasoning capability of detection models. This structured approach has the potential to inform the design of future detection frameworks that makes them more robust and generalizable.

5.2 Alternative Explanations and Limitations

While our findings demonstrate the effectiveness of our annotation framework, some limitations must be acknowledged. First, our dataset is generated from a specific set of T2V models, which may not fully represent all generative techniques. Second, the focus on deformation and a narrowed set of artifact categories, though practical, may overlook less frequent but critical issues. Finally, inter-annotator agreement was not explicitly analyzed, which may introduce subjective bias in the annotations.

5.3 Future Work and Areas for Improvement

To address these limitations, future work should include expanding the dataset to incorporate videos from a wider range of generative models and scenarios. Incorporating inter-annotator agreement metrics can further validate annotation consistency. Additionally, integrating multi-modal analysis, combining visual and textual cues, could improve detection capabilities. Finally, developing models that leverage our annotated dataset for self-supervised learning could help improve artifact reasoning and detection accuracy in general video content.

By addressing these areas, this research can contribute to the development of more effective and scalable solutions for deepfake detection, ensuring the integrity of digital media in an increasingly AI-driven landscape.

6 Conclusion

This study advances deepfake detection by addressing general video synthesis beyond facial manipu-

lations. Our key contributions include:

1. Dataset Creation: A diverse dataset of 3,253 videos from state-of-the-art T2V models, ensuring broad coverage of generative capabilities and artifacts.

2. Annotation Framework: A systematic annotation approach with detailed explanations and reasoning for each identified artifact, enhancing model reasoning capabilities during detection.

3. Artifact Analysis: Iterative analysis identified deformation as the dominant artifact, particularly in dynamic scenes, highlighting limitations of current generative models.

By providing detailed artifact reasoning, our framework not only improves artifact detection but also supports enhanced interpretability and robustness in detection models. Future work will focus on leveraging this enriched dataset for advanced detection algorithms, exploring cross-modal techniques, and addressing evolving AI challenges.

References

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. 2024. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *Preprint*, arXiv:2405.19707.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *Preprint*, arXiv:2405.07940.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal inconsistency learning for deepfake video detection. *Preprint*, arXiv:2109.01860.

Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. 2022. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pages 596–613. Springer.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. *Preprint*, arXiv:2303.17111.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. Vbench: Comprehensive benchmark suite for video generative models. *Preprint*, arXiv:2311.17982.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. *Preprint*, arXiv:1911.00650.

Kling. 2024. Kling. https://kling.kuaishou.com.

LabelBox. 2024. Labelbox. https://labelbox.com.

Peter Lorenz, Ricard Durall, and Janis Keuper. 2023. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. *Preprint*, arXiv:2307.02347.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt,

Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Sora. https://openai.com/index/sora.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Pika. 2024. Pika. https://pika.art.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Tao Xu, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, and Zijian He. 2024. Movie gen: A cast of media foundation models.

Genmo Team. 2024. Mochi 1. https://github.com/genmoai/models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. *Preprint*, arXiv:2303.09295.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. 2023. Generalizable synthetic image detection via language-guided contrastive learning. *Preprint*, arXiv:2305.13800.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. 2024. Tall: Thumbnail layout for deepfake video detection. *Preprint*, arXiv:2307.07494.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

A Appendix: Annotation Design Details

A.1 Annotation Design for Pilot Study

For the first batch of videos, our categorization schema is divided into the following groups:

Consistency. The category of temporal and subject consistency encompasses metrics such as *Distortion*, *Deformation*, *Appearance*, and *Disappearance*, assessing whether objects maintain continuity and realism across frames. This category identifies issues such as unexpected shifts, abrupt disappearances, or inconsistencies in object appearance, signaling potential temporal instability. Ensuring subject consistency is critical for realistic AI-generated videos, as minor inconsistencies can disrupt the viewer's perception of authenticity (see Figure 4).

Dynamic Motion To evaluate the physical realism and continuity of motion, we included metrics such as *Motion Smoothness*, *Liquid Motions*, and *Temporal Logical Problems*. These metrics assess the fluidity of motion, detecting abrupt changes or unnatural pauses in the movement of humans, animals, and other objects. These evaluations are essential as AI models often struggle to reproduce seamless and natural movement. By analyzing the degree of dynamic motion, we can distinguish between static scenes and those with complex interactions, determining whether a video meets expectations for natural physical dynamics or exhibits anomalies that suggest manipulation.

Aesthetic Quality. This category entails a qualitative assessment of the visual appeal of each video, focusing on *Colors*, *Lighting*, *Shadowing*, and *Blurring*. These metrics capture nuanced aspects of visual composition that automated assessments may overlook. Aesthetic inconsistencies such as unnatural lighting, poor color balance, or unrealistic texturing can subtly signal a video's artificial nature. Evaluating aesthetic quality helps to identify less overt artifacts that impact the overall realism of the generated content (see Figure 4).