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

HOIGen-1M: A Large-scale Dataset for Human-Object Interaction Video Generation

Kun Liu[1] Qi Liu[2] Xinchen Liu[1] Jie Li[1]
Yongdong Zhang[3] Jiebo Luo[4] Xiaodong He[1] Wu Liu[3] †
[1]JD Explore Academy [2]UCAS [3]USTC [4]University of Rochester

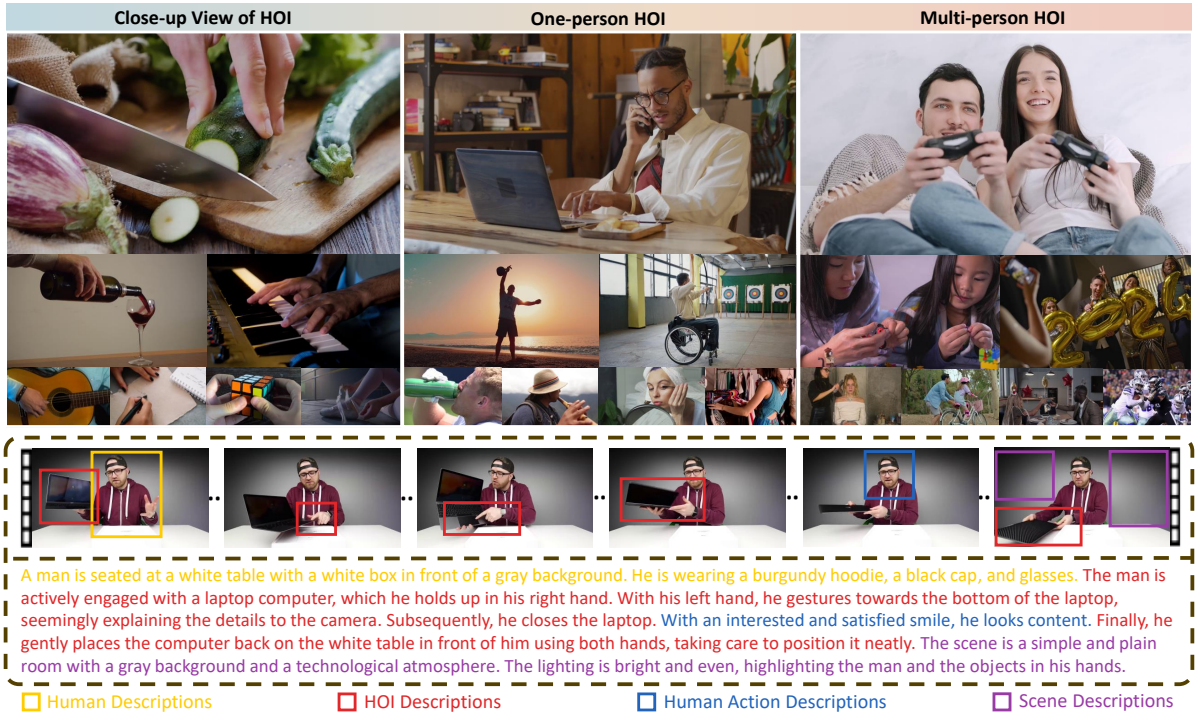


Figure 1. Overview of HOIGen-1M. HOIGen-1M contains over one million video clips for HOI video generation with multiple types of HOI videos, diverse scenarios ($15,000+$ objects and $7,000+$ interaction types), and expressive captions.

Abstract

*Text-to-video (T2V) generation has made tremendous progress in generating complicated scenes based on texts. However, human-object interaction (HOI) often cannot be precisely generated by current T2V models due to the lack of large-scale videos with accurate captions for HOI. To address this issue, we introduce **HOIGen-1M**, the first large-scale dataset for **HOI Gen**eration, consisting of over one million high-quality videos collected from diverse sources. In particular, to guarantee the high quality of videos, we first design an efficient framework to automatically curate HOI videos using the powerful multimodal large language models (MLLMs), and then the videos are further cleaned by human annotators. Moreover, to obtain accurate textual captions for HOI videos, we design a novel video description method based on a Mixture-of-Multimodal-Experts (MoME) strategy that not only generates expressive captions but also eliminates the hallucination by individual MLLM. Furthermore, due to the lack of an evaluation framework for generated HOI videos, we propose two new metrics to assess the quality of generated videos in a coarse-to-fine manner. Extensive experiments reveal that current T2V models struggle to generate high-quality HOI videos and confirm that our HOIGen-1M dataset is instrumental for improving HOI video generation.*

†Corresponding author.

1. Introduction

Text-to-video (T2V) generation, which focuses on creating video sequences from descriptive text, is receiving increasing attention in computer vision. Due to the significant strides of large multi-modal models like Sora [3], T2V generation has recently been viewed as a promising route to building general simulators of the real world, offering vast potential for video creation, self-driving technology, and the creation of physical agents. Nevertheless, as a fundamental aspect of the physical world, human-object interaction (HOI) is challenging to generate accurately for current T2V models because large-scale datasets of videos with precise captions are absent. Therefore, this work offers a dataset, a captioning strategy, and automatic metrics for HOI video generation.

As shown in Figure 2, the current T2V models, even with over 10B parameters, still struggles to produce simple HOI videos. Generally speaking, the videos' quality and the captions' accuracy are crucial to the performance of T2V models [6, 47]. However, current well-known video generation datasets, WebVid-10M [1] and Panda-70M [7], either have inferior quality or contain many videos without HOI. On the other hand, these video benchmarks for perceiving HOI are too small in scale to support the training of T2V models. Besides, existing captioning methods either produce descriptions that are too brief, restricting the in-depth video understanding, or they are not specifically designed HOI, resulting in the loss of many visual interaction details. Therefore, we aim to address the above issues by constructing a large-scale, high-quality training dataset for HOI video generation.

The construction of a large-scale HOI dataset faces two main challenges. The first challenge is acquiring high-quality and extensive video data that includes HOI. This involves accurately sourcing videos that capture these interactions. The second challenge is obtaining high-quality captions that precisely describe the people, objects, and scenes involved. This requires accurate and detailed captions to convey the complexity of the interactions and settings depicted in the videos. To address the aforementioned challenges, we build the first large-scale and high-quality dataset for HOI video generation named **HOIGen-1M** (see Figure 1). It exhibits three main features: 1) Large scale: HOIGen-1M curates over 1M video clips and all videos contain manually verified HOI, which is sufficient for training T2V models. 2) High quality: HOIGen-1M is strictly selected from the aspects of mete attribute, aesthetics, temporal consistency, motion difference, and MLLM assessment. 3) Expressive captions: the captions in HOIGen-1M are precise because a Mixture-of-Multimodal-Experts (MoME) strategy is employed to detect and eliminate hallucinations via cross-verification among multiple MLLMs.

To guarantee the above features of the dataset, we first design an efficient data pipeline, including multimodal large language model (MLLM) evaluation and human verification, to curate over 1 million videos showcasing HOI from a collection of 80 million videos. Next, we first propose a prompt engineering technique to drive MLLM to concentrate on describing human motion, object state changes, and HOI. More importantly, to mitigate hallucinations in video captioning by MLLMs, we design a Mixture-of-Multimodal-Experts (MoME) strategy to mutually verify the outputs among multiple MLLMs. Additionally, to accurately align human preference on the HOI video generation, we introduce two new automatic metrics, CoarseHOIScore and FineHOIScore, to assess the visual quality of interaction. We hope this project can act as a crucial resource for HOI video generation.



Figure 2. Video frames uniformly sampled from the video generated by several current T2V models. Most videos do not adequately follow the prompt of loading a suitcase onto a bus, showing that current T2V models struggle to generate videos that align with HOI. A dashed box means the content does not match the prompt, and a solid box indicates it does.

In summary, our work makes the following contributions:

- To our knowledge, we propose the first large-scale dataset for HOI video generation, comprising over one million high-quality and human-verified video clips.
- To guarantee the quality of HOIGen-1M, we first design an efficient data pipeline to curate videos. Moreover, we develop to detect hallucinations and correct the errors in the captions via a Mixture-of-Multimodal-Experts (MoME) strategy.
- We introduce two new automatic metrics to evaluate the quality of HOI in the generated videos.

Based on our dataset and evaluation framework, we compare ten T2V models with comprehensive analysis. The experiments show that the performance of the open-source

Table 1. A Comparison of the existing datasets for the T2V task and Our HOIGen-1M.

Dataset	#Videos	Duration(Hours)	Resolution	Scenario	Caption	Manually Filtering
CelebV-HQ [51]	35K	68	512p	Portrait	N/A	✗
CelebV-Text [44]	7K	279	512p	Face	Short	✗
ChronoMagic [46]	2K	7	720p	Time-lapse	Short	✗
ChronoMagic-Pro [47]	460K	30K	720p	Time-lapse	Short	✗
HowTo100M [25]	136M	134.5K	240p	Instructional	Short	✗
MSR-VTT [37]	10K	40	240p	General	Short	✗
Panda-70M [7]	70M	166.8K	720p	General	Short	✗
WebVid-10M [1]	10M	52K	360p	General	Short	✗
HD-VILA-100M [40]	103M	371.5K	720p	General	Short	✗
OpenVid-1M [27]	1M	2.0K	720p+	General	Long	✗
HOIGen-1M	1M	2.2K	720p+	HOI	Long	✓

T2V model and commercial software in generating HOI videos is far from real-world applications. Furthermore, all open-source models exhibit improved performance after fine-tuning on the constructed HOIGen-1M benchmark, which validates the value of this dataset for improving HOI video generation. Specifically, an open-sourced model (CogVideoX-5B) fine-tuned HOIGen-1M achieves competitive performance to the most recently available commercial software Kling 1.5 in terms of CoarseHOIScore.

2. Related Work

Our work touches two threads: text-to-video generation and HOI, which will be discussed respectively.

Text-to-video Generation. With the development of the diffusion model and the Transformer architecture, Text-to-video generation has made tremendous progress in visualizing complicated scenes based on textual description [3, 14, 16, 42, 50]. For example, Sora [3] is the first method to produce minute-long videos based on user prompts. Very recently, Kling 1.5 [14] is an advanced video generation software that enables users to create high-quality videos through intuitive tools and customizable templates. Meanwhile, OpenSora [50], OpenSoraPlan [16], and CogVideoX [42] leverage the million-scale video-text training samples to obtain the impressive results in the open-source community. While these models perform well in generating videos with general text, the quality of the generated HOI videos is far from satisfactory. Thus, it is necessary to construct a dedicated and large-scale dataset to generate HOI videos.

Recent advancements in the T2V field have predominantly focused on data quality, leading to the emergence of numerous video and caption datasets within the research community [1, 7, 31, 43, 45]. For instance, Taichi-HD [31] includes 2,668 videos of persons performing Taichi. ChronaMagic [45] consists of 2,265 high-quality time-lapse videos with corresponding detailed captions. CelebV-Text [43] comprises 70,000 in-the-wild video clips with diverse visual content for facial video generation. In contrast, HOIGen-1M offers a million-level, high-quality collection for an unexplored domain. As for general video generation datasets, both WebVid-10M [1] and Panda-70M [7] compile million-scale high-resolution and semantically consistent video samples. However, WebVid-10M [1] includes low-quality videos with watermarks and Panda-70M [7] contains numerous static, flickering, low-clarity videos along with short captions, while our HOIGen-1M stands out as an precise T2V dataset tailored for HOI with excellent video quality and detailed captions. Table 1 compares HOIGen-1M with existing datasets for the T2V task.

HOI Detection. Numerous video-based datasets have been developed for HOI perception, such as HOI detection [12, 13, 18, 23, 28]. For instance, CAD-120 [13] is designed for HOI detection and comprises 120 RGB-D videos, limited to 10 high-level activities. MPHOI-72 [28] focuses on multi-person activities, consisting of 72 videos captured from three distinct angles. The PVSG [41] dataset contains 400 videos of both third-person and egocentric perspectives.

HOI Reconstruction. As the performance of HOI detection improves, HOI reconstruction datasets has also attracted extensive attention [2, 8, 22, 30, 32, 39, 48, 49]. For instance, the GRAB [32] dataset comprises 1,334 full 3D shapes and pose sequences featuring 10 subjects engaging with 51 common objects. HOI4D [22] is a substantial 4D egocentric dataset comprising over 4,000 videos of nine participants interacting with 800 distinct object instances. BEHAVE [2] is the first full-body HOI dataset to offer multi-view RGBD frames alongside corresponding 3D SMPL models, totaling 15.2K video frames.

Even though these above datasets provide benchmarks for HOI perception, they cannot be directly utilized for video generation. A training set of at least a million is typically necessary for T2V task, yet the current HOI benchmark comprises no more than tens of thousands of videos. Besides, existing HOI benchmarks lack high-quality videos and expressive textual descriptions, which are of great importance for the training of T2V models. In contrast, our HOIGen-1M dataset contains more than 1 million high-quality HOI videos with detailed captions.

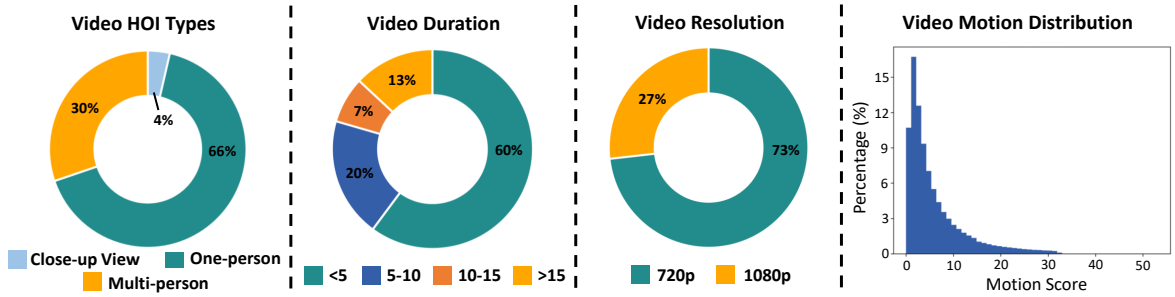


Figure 3. Statistics of video clips in HOIGen-1M. The dataset includes multiple types of HOI and spans a range of clip durations. All videos have a resolution of at least 720p and include significant motions.

3. The HOIGen-1M Dataset

3.1. Video Curation and Processing

Data Collection. Initially, we select three public datasets designed for HOI perception: BEHAVE [2], InterCap [10], and HOI4D [22], which have high resolution, and each clip showcases at least one HOI instance. However, these datasets only contain 22K valid videos, which is far from sufficient to support the training of large video generation models.

To collect more HOI videos, we expand our dataset selection from HOI perception data to general videos. We incorporate five large-scale datasets for T2V generation: Panda-70M [7], ViSR [21], and Mixkit, Pixabay, and Pexels introduced in OpenSoraPlan [16]. By this means, we obtain about eighty million raw videos. However, these videos are of varying quality and most of them contain non-HOI content. Therefore, to filter high-quality HOI videos from such a vast corpus, we adopt a data processing pipeline that is introduced as follows.

Metadata. We initially analyze video metadata, including duration, resolution, and frames per second (FPS). Videos are filtered to retain those exceeding one second in duration, with a resolution $\geq$ 720p and a frame rate $\geq$ 20 FPS.

Optical Character Recognition. The DBNet++ [19] model is employed for optical character recognition to exclude video samples with over one text, which are considered as noises for the training of video generation models.

Aesthetics Score. Video aesthetic scores are a key factor in determining the visual quality of videos. To ensure high-quality videos, we utilize the Laion Aesthetic Predictor [*] to assess the visual appeal of videos. Only those videos that achieve high aesthetic scores are retained.

Motion Score. To further ensure video quality, we enhance video quality by computing smooth motion. Specifically, we employ UniMatch [36] to calculate the optical flow scores of the videos. Videos with excessively high or low optical flow scores are excluded, while those with moderate scores are retained.

[*]https://github.com/christophschuhmann/improved-aesthetic-predictor

LLM Evaluation. To efficiently identify videos containing HOIs, we first employ a multi-modal model, i.e., PLLaVA [38], to annotate each video with a detailed caption. Then, we adopt a powerful LLM, i.e., Qwen2.5 [34] to judge whether there is an interaction in the caption using a predefined prompt. After the above processing steps, we obtain about 1.5 million videos in total.

Human Verification. To guarantee the high quality of the dataset, we employ seven annotators to manually check whether the remaining videos contain HOIs. In particular, the annotators are asked to scan videos to remove those featuring inconspicuous HOI and minimal object visibility. Finally, we obtain about 1.1 million videos from 80 million raw video clips. The entire annotation process requires approximately eight weeks.

Video Analysis. As illustrated in Table 1, our HOIGen-1M is a million-level, high-resolution, and manually verified video dataset for HOI generation. Compared to previous datasets [44, 46, 47, 51] that are usually constructed for specific domains with limited videos, our HOIGen-1M is a large-scale dataset for an unexplored field. As shown in Figure 3, all videos feature resolutions of at least 720p and encompass three types of fine-grained interactions, including close-up view interaction, one-person interaction, and multiple-person interaction. In addition, compared to previous million-level datasets, WebVid-10M [1] includes low-grade videos with watermarks and Panda-70M [7] comprises numerous static, flickering, or unclear videos. In contrast, our HOIGen-1M includes high-quality and clean videos without watermarks.

3.2. Video Captioning with MoME

The rich and detailed descriptions of videos are essential for training video generation models and the quality of generated videos. To obtain the expressive caption, we adopt the powerful MLLMs to generate the textual description for the one million videos. Specifically, we design a prompt engineering technique to adopt the MLLM for HOI description via a structured prompt, which encourages the model to focus on object attributes, human movements, and detailed

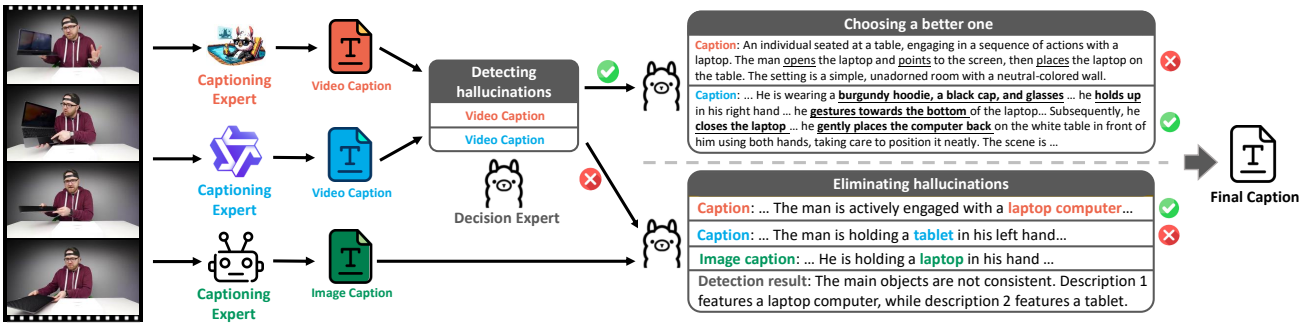


Figure 4. An illustration of the Mixture-of-Multimodal-Experts (MoME) strategy-based caption method.

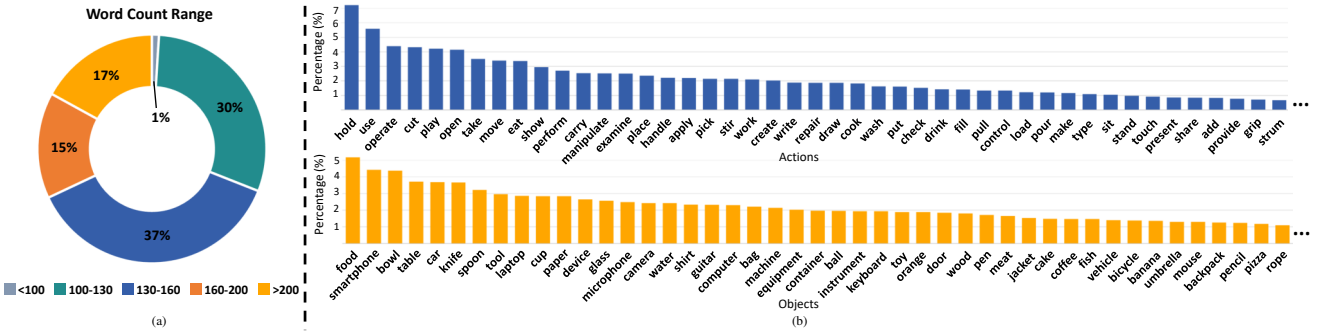


Figure 5. Caption words statistics in HOIGen-1M. The distribution of word numbers shows the captions are high-quality and fine-grained, with an average length of 152 words. The distribution of actions and objects in the captions further demonstrates the diversity of the dataset. There are over 15,000 objects and over 7,000 interaction action types, making it possible to train a T2V model to simulate the real world. For clarity, we have only listed the categories with the highest frequency.

temporal descriptions of interaction. However, the introduction of MLLMs may result in the hallucination phenomenon that the model generates plausible-sounding but factually incorrect or nonsensical texts. If we feed the video generation models with descriptions that are inconsistent with the visual content, the quality of the generated videos will be significantly degraded.

To eliminate hallucinations generated by large models in video descriptions, we propose a Mixture-of-Multimodal-Experts Strategy (MoME) to detect hallucinations and then fuse the characteristics of different MLLM to correct them. As shown in Figure 4, we first adopt two MLLMs, i.e., PLLaVa [38] and Qwen2-VL [35] as caption experts to generate multiple captions for one video. Then, we adopt Llama3.1 [9] as a decision expert to judge if both captions describe the same video, like the same object, interaction, and scene. A negative answer from the decision expert indicates a hallucination has been detected. Upon detecting hallucination, we encourage the third expert [20] to focus on the areas where the first two experts have disagreements. At this point, the reason for the detected hallucination and the third caption are fed into the decision expert [9] to regenerate the new caption, thus correcting the hallucination problem of the individual MLLM. If there is no hallucination, a decision expert is employed to evaluate which is superior, based on the richness of interactions, actions, and scene descriptions. In summary, MoME first adopts two captions and one decision expert to detect the hallucination. Then, an additional set of decision experts and caption experts is introduced to eliminate these hallucinations. Besides the textual description, we also annotate each video with HOI categories by extracting the core information from the caption. Thus, our dataset can also be utilized for HOI perception tasks.

Text Analysis. The captions in HOIGen-1M are longer and richer than those in WebVid-10M [1] and Panda-70M [7], where the average text lengths for WebVid-10M, Panda-70M, and HOIGen-1M are 12.0, 13.2, and 153.8, respectively. Besides, as illustrated in Figure 5, HOIGen-1M presents a wide variety of objects and interaction actions. There are over 15,000 objects and more than 7,000 interaction action types, which significantly exceeds the categories of HOI datasets. Thus, HOIGen-1M is a natural testbed for open-vocabulary HOI detection. Compared with previous video-text datasets, the captions in HOIGen-1M are precise because the MoME is employed to eliminate hallucinations via cross-verifications among multiple MLLM. Moreover, the interaction diversity within HOIGen-1M makes it possible to train a T2V model to simulate the real world.

Table 2. Evaluation results with proposed HOIScores and VBench in HOI video generation.

Method	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Temporal Flickering	Overall Consistency	Coarse HOIScore	Fine HOIScore
OpenSoraPlan [16]	**97.06%**	**96.83%**	**99.56%**	55.43%	59.62%	**99.09%**	14.12%	8.80%	91.54%
Mochi-10B [33]	87.27%	92.68%	98.90%	54.22%	54.31%	97.94%	19.83%	28.78%	**96.15%**
CogVideoX-2B [42]	91.31%	95.17%	98.08%	53.48%	60.94%	96.13%	20.29%	31.34%	92.06%
OpenSora [50]	94.33%	96.05%	99.05%	53.79%	**64.84%**	98.52%	18.94%	31.86%	91.86%
CogVideoX-5B [42]	90.29%	94.96%	97.68%	**54.84%**	62.27%	95.41%	**20.54%**	**32.84%**	94.25%
Gen-3 [29]	90.37%	95.31%	99.24%	59.22%	64.56%	97.95%	17.48%	23.98%	96.79%
Pika [15]	96.10%	**97.22%**	**99.51%**	60.55%	66.77%	**98.86%**	17.29%	29.37%	97.44%
Dreamina [4]	**96.46%**	97.08%	98.85%	**66.07%**	67.99%	97.46%	18.02%	36.36%	94.18%
Hailuo [26]	92.61%	95.45%	99.16%	58.49%	**70.07%**	97.73%	**19.55%**	39.56%	97.22%
Kling 1.5 [14]	92.55%	95.42%	99.24%	61.09%	69.28%	96.70%	18.74%	**42.72%**	**97.76%**

4. Experiments

In this section, we introduce the designed evaluation metrics and evaluate ten popular T2V models for HOI video generation. Moreover, we conduct an ablation study to highlight the effectiveness of HOIGen-1M and the proposed caption method. For implementation details, please refer to the **Supplementary File**.

4.1. Evaluation Protocol and Metrics

To make the experiment more convincing, we first employ the widely used evaluation metric VBench [11]. Because some of the 16 evaluation dimensions, such as color, are related to the prompts, we can adopt seven available metrics from VBench to assess the quality of the generated videos. Although the existing evaluation benchmarks are comprehensive, there are no specific metrics to assess HOI. Therefore, we first build a prompt set and then propose two evaluation metrics to assess the quality of generated interactive video clips in a coarse-to-fine manner.

Prompt Suite. The first step in evaluating T2V models is to construct a comprehensive and diverse prompt suite. Inspired by these previous works [11, 24], we select the object and interaction types from a large-scale HOI detection dataset [5]. Then, we manually filter out rare and unreasonable interactions and get a total of 306 prompts, including interaction with musical instruments, transportation, kitchen tools, etc. Only interactive-type phrases are not sufficient for the model to generate promising videos. Thus, LLM is utilized to extend prompts longer and more detailed without changing their original meaning.

The Proposed CoarseHOIScore. An ideal interaction should at least involve people, objects, and corresponding actions. Therefore, we adopt a HOI detector $d(\cdot)$ [17] to predict possible HOI triplets in the generated videos. Specifically, for each video, the CoarseHOIScore is calculated as:

$$S_{coarse} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1 & \text{if } d(f_i) > \theta, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where S_{coarse} is the CoarseHOIScore of one video, N is the number of uniformly adopted video frames, f_i denotes the i_{th} frame, and θ is the threshold for the confidence score. We average the CoarseHOIScore for all the videos generated by one model as the final score of this model.

The Proposed FineHOIScore. Although the above operation can roughly determine interactions, it does not capture the details well, such as the presence of contact between the human and the object, and the stability of the interaction action. To solve this problem, we propose a fine-grained, pixel-level approach to evaluate the generated videos. Specifically, for each video, the FineHOIScore is calculated as:

$$S_{Fine} = 1 - \frac{1}{T-1}\sum_{i=2}^{T}(|\langle M_i, K_i\rangle - \langle M_{i-1}, K_{i-1}\rangle|), \quad (2)$$

where S_{Fine} is the FineHOIScore of one video, T is the number of video frames, M_i and K_i denote the object mask and human keypoints in the i^{th} frame, and $\langle\cdot\rangle$ represents the pixel distance that satisfies the requirement of the contact between people and objects. We average the FineHOIScore for all the videos generated by one model as the final score of this model.

4.2. Main Results of HOI Video Generation

Despite the recent significant attention on T2V benchmarks, systematic evaluation of these models on HOI video generation is still lacking. To solve this problem, we evaluate five popular commercial software Kling 1.5 [14], Pika [15], Hailuo [26], Dreamina [4], and Gen-3 [29], as well as five representative open-source methods including OpenSora [50], OpenSoraPlan [16], CogVideoX-2B [42], CogVideoX-5B [42], and Mochi-10B [33].

As illustrated in Table 2, even the recent advanced commercial software, Kling 1.5 [14], only attains a score of 42.72% in terms of CoarseHOIScore. This indicates that even when trained on extremely large-scale open-domain datasets, existing T2V models still cannot produce clips that perfectly align with HOI. Consistent with the above observation, Michi-10B [33], one of the largest current text-to-video models with its 10 billion parameters, does not achieve an

Table 3. The effect of the proposed captioning method.

Method	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Temporal Flickering	Overall Consistency	Coarse HOIScore	Fine HOIScore
Fine-tuned with PLLaVA [38]	90.53%	94.29%	96.51%	52.91%	62.63%	95.24%	**19.93%**	33.17%	92.78%
Fine-tuned with our Caption	**94.98%**	**96.63%**	**98.16%**	**54.90%**	**65.85%**	**97.94%**	19.55%	**39.13%**	**93.82%**

Table 4. The effect of fine-tuning using HOIGen-1M.

Method	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Temporal Flickering	Overall Consistency	Coarse HOIScore	Fine HOIScore
OpenSora [50]	94.33%	96.05%	99.05%	53.79%	**64.84%**	98.52%	18.94%	31.86%	91.86%
Fine-tuned	**96.94%**	**97.21%**	**99.28%**	**54.29%**	64.74%	**99.13%**	**19.36%**	**35.38%**	**94.91%**
CogVideoX-2B [42]	91.31%	95.17%	98.08%	53.48%	60.94%	96.13%	**20.29%**	31.34%	92.06%
Fine-tuned	**94.98%**	**96.63%**	**98.16%**	**54.90%**	**65.85%**	**97.94%**	19.55%	**39.13%**	**93.82%**
CogVideoX-5B [42]	90.29%	94.96%	97.68%	54.84%	62.27%	95.41%	**20.54%**	32.84%	94.25%
Fine-tuned	**94.70%**	**96.42%**	**98.29%**	**56.14%**	**66.46%**	**98.01%**	19.88%	**44.04%**	**96.04%**

impressive HOI score for video generation. This indirectly reflects the necessity of constructing a dedicated and large-scale dataset to generate HOI videos.

In addition, we identify the following important observations: 1) While VBench provides a comprehensive evaluation standard, it does not accurately reflect the quality of the generated interactions. For instance, the score of motion smoothness over all models are more than 98% but the generated interactions in videos are clearly unreasonable. It indicates the necessity of designing a standard specifically for evaluating the quality of interaction generation; 2) Among open-source approaches, CogVideoX-5B [42] and Mochi-10B [33] perform comparatively well, both exceeding the performance level of OpenSoraPlan [16] and OpenSora [50] in terms of HOIScore. In relation to the scaling law of model parameters, the CogVideoX-5B model exhibits superior performance compared to the CogVideoX-2B model; 3) All commercial software performs well in the VBench evaluation. However, there is significant variation in the HOI evaluation, with the best software achieving a score of 42.72% and the worst only 23.98%; 4) Overall, under evaluation metrics VBench and HOIScore, commercial software tends to perform better than open-source software, possibly because commercial companies have more computational resources and easier access to large-scale video data for training models. For human evaluation on massive generated videos, please refer to the **Supplementary File**.

4.3. Ablation Study

We conduct a thorough effectiveness analysis of the proposed captioning method to demonstrate its ability to generate high-quality textual descriptions, which in turn contributes to generating videos that align with HOI. We also demonstrate the advantage of our HOIGen-1M by fine-tuning T2V models on this dataset. We observe a significant improvement in the quality of HOI generation. This improvement highlights the dataset's comprehensive coverage and its ability to enhance model performance by providing rich and diverse examples. The combination of our captioning method and the HOIGen-1M dataset sets a new benchmark in the field, offering a powerful tool for future research and applications in HOI modeling.

The effectiveness of the proposed captioning method. We adopt the recently proposed PLLaVA [38] as the baseline model due to its expressive performance on video captioning. Based on this model, we employ the proposed MoME to detect and eliminate hallucinations of individual MLLM, yielding a precise textual description. We train the T2V model separately using PLLaVA and our caption, and the evaluation results of the generated videos are listed in Table 3. We can observe that the T2V model trained with our caption achieves significant improvement in both VBench and HOI evaluation. Notably, our caption increases the CoarseHOIScore from 33.17% to 39.13%.

The effectiveness of HOIGen-1M dataset. The quality of the generated interactive videos significantly improves after fine-tuning the model on HOIGen-1M. Table 4 outlines the performance of the original model compared to its results after fine-tuning. Specifically, we first evaluate three models pre-trained on large-scale but general video samples. Then, we fine-tune all models on the HOIGen-1M using the LORA mechanism. A significant increase can be observed after fine-tuning in the HOI scores of all three models, which directly demonstrates the effectiveness of our dataset in generating HOI videos. In addition, we hope that our dataset can serve as a benchmark for interactive video generation.

Besides quantitative results, we would like to obtain further insight into the learned T2V models. In this sense, we provide visual comparisons between CogVideoX and its fine-tuned version on HOIGen-1M in Figure 6. According to the left example, we can see that the hands in the second row

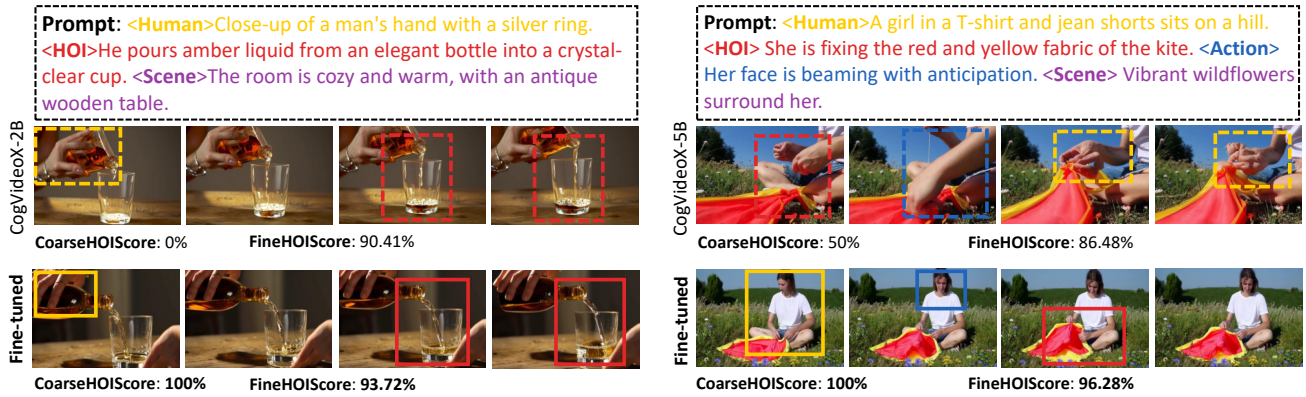


Figure 6. Visual comparison between CogVideoX and its fine-tuned version on HOIGen-1M. Please zoom in for more details.

appear more natural, the cup is more aesthetically pleasing, and the water increases as the pouring progresses. The right example also demonstrates that our model generates clearer hands and more reasonable interaction with kites after fine-tuning. The CoarseHOIScores of the original CogVideoX model in both examples are low (0% and 50%) because the generated hands are too terrible to be detected. Both HOI scores demonstrate substantial improvement after fine-tuning. The above observations further directly demonstrate the effectiveness of our dataset.

Best vs. Worst Generated Videos. Based on the HOIScore, we select the high-scoring interaction types and the low-scoring interaction types from these generated videos. For example, the interaction carrying a chair achieves a score of 87.5%, which far exceeds the average score of all generated videos. In contrast, the interaction brushing receives a very low score, possibly because the detailed body movements required by this action are difficult to model. We can observe that simple actions and interactions with rigid objects generally achieve better generation results. Adopting prior knowledge (e.g., human keypoints) as a prerequisite for generation can help the model better understand and capture complex body movements, thereby improving the accuracy and reliability of the generated content. More visual results can be found in the **Supplementary File**.

Out-of-Distribution vs. In-Distribution Video Generation. Although we aim to build a dataset that encompasses as many HOI as possible to simulate the physical world, there are still some interactions that rarely appear in HOIGen-1M. This provides us with an opportunity to test the emergent capabilities of the T2V model. Specifically, we select 20 prompts from all prompts that have not appeared in the captions of HOIGen-1M as the out-of-distribution samples. Additionally, we choose 20 prompts that appear with the highest frequency in the HOIGen-1M as in distribution samples. We then calculate the HOIScores of the videos generated from these two groups of prompts. The close scores (63.20% vs. 61.20%) demonstrate that whether a prompt has been seen in the training set does not significantly influence the generation results. This might indicate that the model is beginning to demonstrate emergent capabilities, as it can generate interactions it has not encountered before.

5. Discussion

While this work offers a high-quality dataset and an evaluation framework for HOI video generation, it still has two limitations. First, million-level videos may also be insufficient to support the training of T2V models as model sizes continue to scale up. Hence, we will further extend HOIGen-1M by collecting more high-quality videos and annotating more precise captions. Second, although two new evaluation metrics, i.e., CoarseHOIScore and FineHOIScore, are introduced to assess the quality of generated HOI videos, there is still a disparity between these metrics and human preferences. In future work, we will design more comprehensive metrics aligned with human judgment to evaluate T2V models.

6. Conclusion

This paper proposes HOIGen-1M, a large-scale and high-quality dataset for HOI generation. To ensure video quality, we develop an efficient data selection pipeline to guarantee that each video clip contains a single clear HOI through MLLM evaluation and human verification. Besides, we meticulously design a Mixture-of-Multimodal-Experts (MoME) strategy to describe the HOI in detail, which can detect and then correct the hallucinations in captions generated by individual MLLM. Due to the lack of evaluation metrics for HOI video generation, we propose a novel assessment framework to assess the generated videos in coarse and fine manners. The experimental results and in-depth analysis demonstrate the effectiveness of HOIGen-1M and the proposed caption method. With its data, annotations, and evaluation metrics, we believe HOIGen-1M can fuel future studies for HOI video generation.

7. Acknowledge

This work is supported by National Key Research and Development Program of China (NO. 2024YFE0203200), the National Nature Science Foundation of China (NO. U24A20329)

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1708–1718. IEEE, 2021. 2, 3, 4, 5

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *CVPR*, pages 15914–15925. IEEE, 2022. 3, 4

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 2, 3

[4] ByteDance. Dreamina ai. https://jimeng.jianying.com, 2024. 6

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, pages 1017–1025. IEEE, 2015. 6

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *CoRR*, abs/2406.04325, 2024. 2

[7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331. IEEE, 2024. 2, 3, 4, 5

[8] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, pages 19888–19901. IEEE, 2024. 3

[9] Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 5

[10] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view RGB-D images. *IJCV*, 132(7):2551–2566, 2024. 4

[11] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818. IEEE, 2024. 6

[12] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10233–10244. IEEE, 2020. 3

[13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from RGB-D videos. *IJRR*, 32(8):951–970, 2013. 3

[14] Kwai. Keling. https://klingai.kuaishou.com, 2024. 3, 6

[15] Pika Lab. Pika-2.0 lab discord server. https://pika.art, 2024. 6

[16] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan. https://github.com/PKU-YuanGroup/Open-Sora-Plan, 2024. 3, 4, 6, 7

[17] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *ICCV*, pages 6457–6467. IEEE, 2023. 6

[18] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, and Bryan C. Russell. Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *ICCV*, pages 1825–1835. IEEE, 2021. 3

[19] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *TPAMI*, 45(1):919–931, 2023. 4

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 5

[21] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *CVPR*, pages 3566–3574. IEEE, 2019. 4

[22] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, pages 20981–20990. IEEE, 2022. 3, 4

[23] Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. MOMA-LRG: language-refined graphs for multi-object multi-actor activity parsing. In *NeurIPS*, 2022. 3

[24] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *CoRR*, abs/2410.05363, 2024. 6

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640. IEEE, 2019. 3

[26] MiniMax. Hailuoai. https://hailuoai.com/video, 2024. 6

[27] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *CoRR*, abs/2407.02371, 2024. 3

[28] Tanqiu Qiao, Qianhui Men, Frederick W. B. Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert P. H. Shum. Geometric features informed multi-person human-object interaction

recognition in videos. In *ECCV*, pages 474–491. Springer, 2022. 3

[29] Runway. Gen-3 alpha. https://runwayml.com, 2024. 6

[30] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3d human motion capture with pose manifold sampling and dense contact guidance. In *ECCV*, pages 516–533. Springer, 2022. 3

[31] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, pages 7135–7145, 2019. 3

[32] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 3

[33] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 6, 7

[34] Qwen Team. Qwen2.5: A party of foundation models, 2024. 4

[35] Peng Wang, Shuai Bai, and Sinan Tan et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024. 5

[36] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *TPAMI*, 45(11):13941–13958, 2023. 4

[37] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296. IEEE, 2016. 3

[38] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See-Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *CoRR*, abs/2404.16994, 2024. 4, 5, 7

[39] Zhu Xu, Qingchao Chen, Yuxin Peng, and Yang Liu. Semantic-aware human object interaction image generation. In *ICML*, 2024. 3

[40] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045. IEEE, 2022. 3

[41] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *CVPR*, pages 18675–18685. IEEE, 2023. 3

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *CoRR*, abs/2408.06072, 2024. 3, 6, 7

[43] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *CVPR*, pages 14805–14814. IEEE, 2023. 3

[44] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *CVPR*, pages 14805–14814. IEEE, 2023. 3, 4

[45] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *CoRR*, abs/2404.05014, 2024. 3

[46] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *CoRR*, abs/2404.05014, 2024. 3, 4

[47] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *CoRR*, abs/2406.18522, 2024. 2, 3, 4

[48] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: towards controllable human-chair interactions. In *ECCV*, pages 518–535. Springer, 2022. 3

[49] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, pages 311–327. Springer, 2022. 3

[50] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora, 2024. 3, 6, 7

[51] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *ECCV*, pages 650–667. Springer, 2022. 3, 4