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

Supplementary Material

1. Overview

In this supplementary file, we provide more implementation details about the dataset curation and the proposed caption method in Section 2. We also present the human evaluation on massive generated videos to verify the proposed metrics' alignment with human perception in Section 3. Additionally, in Section 4, we discuss the limitations and societal impact of our dataset. Finally, we offer more visual results in Section 5 to support the conclusion and insights in the main paper.

2. Implementation Details

2.1. Dataset Curation

In this section, we present more details of all process in the dataset construction. Table 1 presents our data processing pipeline, explaining how we identify one million high-quality HOI videos from an initial collection of eight million.

Data Collection. Initially, we select three public datasets designed for HOI perception: BEHAVE [1], InterCap [5], and HOI4D [12], 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 [2], ViSR [11], and Mixkit, Pixabay, and Pexels introduced in OpenSoraPlan [8]. By this means, we obtain about eight 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++ [9] 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 [1] to assess the visual appeal of videos. Only those videos that achieve high aesthetic scores are retained.

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

Table 1. The pipeline of constructing HOIGen-1M.

Pipeline	Tool
Optical Character Recognition	DBNet++ [9]
Aesthetics Score	LAION Aesthetics Predictor
Motion Score	UniMatch [16]
Human Body Detection	RTMPose Model [6]
Cartoon Style Detection	ResNet-50 Binary Classifier [4]
LLM Evaluation	PLLaVA [17] and Qwen2.5 [14]
Human Verification	Volunteers

Motion Score. To further ensure video quality, we enhance video quality by computing smooth motion. Specially, we employ UniMatch [16] 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.

Human Body Detection. After the previous processing steps, we have obtained high-quality video data, next we introduce how to filter the data for human-object interaction. The goal of this step is to exclude videos that lack human body and show bodies that are too small. We first evenly sampled five frames from each video and then employed RTMPose model [6] to identify the whole body 2D keypoints (133 keypoints) of human figures. Furthermore, we utilize the detected results to classified the whole dataset into three HOI categories: close-up hand interactions, single-person, and multi-person.

Cartoon Style Detection. Despite previous filtering, we observed some cartoon-style video clips remaining in the dataset. These clips, being fictional, often depict interactions that defy real-world physical laws. To exclude such videos, we curated a dataset comprising 100 cartoon-style videos and 125 realistic videos, segmented them into frames, and employed these to train a ResNet-50 [4] as a binary classifier. We then extracted two frames from each video and discarded any video where both frames were recognized as cartoon-style.

LLM Evaluation. To rapidly identify videos containing HOI from a vast dataset, we initially employ PLLaVA [17] to annotate videos and produce detailed captions. Subsequently, we retain captions that include terms indicative of human presence to filter out the videos without human. Then, we ask the Qwen2.5 [14] whether there is interaction in these captions using a predefined question and require the LLM to explain the reason behind its answer. The advantage of

determining whether there is HOI from captions is that captions are an essential part of the generation process, and text tokens are fewer than video tokens, thereby reducing the overall computational cost.

Human Verification. Ultimately, to ensure the high quality of the dataset, we enlisted seven volunteers to manually check the remaining videos that showcases HOI. Specially, volunteers were asked to scan videos to remove these featuring inconspicuous human-object interactions and minimal object visibility.

2.2. Video Captioning with MoME

We design the Mixture-of-Multimodal-Experts (MoME) to eliminate the hallucination by individual MLLM. 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.

The complete prompt for three caption experts (e.g., PLLaVa [17], Qwen2-VL [15]), and LLaVA [10] is shown as followings:

> Please describe the content of this video in as much detail as possible, including the people, objects, colors, scenery, environment, and camera movements within the video. Focus on describing the interaction between people and objects in the video, such as what actions people take, changes in the position or shape of objects, etc. Please describe the content of the video and the changes that occur, in chronological order. The description should be useful for AI to re-generate the video. The description should not be less than six sentences.

We adopt Llama3.1 [3] to act as a decision expert, where the prompt for determining hallucinations is as follows:

> You are an AI assistant to check whether the given descriptions are about the same video. You will be provided with a description <A> and another one <B>.
>
> Please pay attention to the objects, attributes, and other relationships mentioned in these descriptions. Finally, make a judgment on whether these two descriptions are from the same video. Please ignore some of the very detailed differences. Please answer Yes or No first, and then explain the reason.
>
> For example:
>
> No, The main object in description A is an apple, but the object in description B is a pear.
>
> Yes, The main objects in these two descriptions are consistent, both depicting scenes where an apple is placed on a table.

If the hallucination exists, the prompt used for the decision expert to eliminate the hallucination is as follows:

> You are an AI assistant and need to choose a description of a video. You will be given a description <A>, a description <B>, a description <C> and a difference <D>.
>
> Descriptions A, B, and C are detailed descriptions of the same video. Difference D explains the differences between the objects, properties, and other relationships mentioned in description A and descrip tion B. Description C gives a more accurate description of the objects in the video. Your task is to choose from Description A and Description B the one that more closely matches the description of the object and property in Description C.
>
> You only need to answer "A" or "B". If descriptions A, B, and C are all completely different (ignoring very detailed points), answer "None".

If there is no hallucination, the prompt used for decision expert to choose a better caption is as follows:

> You are an AI assistant that needs to pick the description of one video. You will be given a description <A> and a description <B>.
>
> Descriptions A and B are detailed descriptions of the same video. Your task is to select a description from description A and description B. The requirement is for a more detailed and accurate description of the video, e.g., a more detailed and specific description of people, objects, scenes, colors, environments, state changes, human actions, and human-object interactions.
>
> You only need to answer "A" or "B".

3. Human Evaluation on Generated Videos

To validate the proposed evaluation metrics' alignment with human perception, we conduct human preference annotations on massive generated videos.

Data Preparation. Given a prompt p_i, and six models to be evaluated {A,B,C,A',B',C'}, including three open-sourced T2V models and corresponding models fine-tuned on our dataset, we employ each model to produce a video, constructing a set of videos $S_i = \{V_{i,A}, V_{i,B}, V_{i,C}, V_{i,A'}, V_{i,B'}, V_{i,C'}\}$. For every prompt p_i, we create pairs of videos in pairwise combinations, resulting in three pairs: $(V_{i,A}, V_{i,A'})$, $(V_{i,B}, V_{i,B'})$, $(V_{i,C}, V_{i,C'})$, and ask human annotators to judge their preferred video for each pair. Within the our proposed framework, a prompt suite of N prompts produces N×3 pairwise video comparisons. The sequence of videos in each pair is shuffled to guarantee unbiased labeling.

Table 2. The result of human preference annotation.

Methods	Coarse HOIScore	Fine HOIScore	Win Ratio
OpenSora [19]	31.86%	91.86%	30.32%
Fine-tuned	35.38%	94.91%	69.68%
CogVideoX-2B [18]	31.34%	92.06%	35.23%
Fine-tuned	39.13%	93.82%	64.77%
CogVideoX-5B [18]	32.84%	94.25%	32.86%
Fine-tuned	44.04%	96.04%	67.14%

Human Annotation Rules. Specifically, human annotators are instructed to focus solely on the particular evaluation dimension of interest and choose the video they prefer. For instance, regarding the CoarseHOIScore, the question is "Does the video exhibit the people, objects, and corresponding interaction in the prompts?". Annotators are asked to concentrate only on whether the generated video's content aligns with the interaction, disregarding other quality factors such as background consistency. For each metric, we meticulously prepare guidelines and train the human evaluators to comprehend the definition of the metric. For example, CoarseHOIScore pay more attention on the rough exist of the interaction while FineHOIScore concentrates on the details of interaction, such as the contact distance between the humans and objects. We also conduct two rounds of post-labeling checks to ensure the annotation quality.

Human Annotation Results. After performing a large-scale human annotation, we calculate the win ratio between the original model and the fine-tuend models. In pairwise comparisons, if a model's video is chosen as superior, the model receives a score of 1, while the other model receives a score of 0. In the event of a tie, both models earn a score of 0.5. The win ratio for each model is determined by dividing the total score by the total number of pairwise comparisons in which it participated. We present the result of human preference annotations (e.g., win ratio) and in Table 2. We can see that all fine-tuned models achieve a higher win ratio, which is consistent with achieving a higher HOI score. The above observation validate that our evaluation metrics can faithfully reflect human perception.

4. Discussion

4.1. Data Availability Statement

We are committed to maintaining transparency and compliance in our data collection and sharing methods. According to this principle, we have adhered to the following rules:

Public Data Sources: The data utilized in our work is publicly available. We do not rely on any exclusive or confidential data sources.

Data Distribution Policy: Our policy on data distribution builds upon the precedent set by previous works like Kinetics, InternVid, and others. Rather than offering the original raw data, we only provide the YouTube video IDs necessary for downloading the relevant content.

Usage Rights: The data released by us is solely for academic and research purposes. This agreement does not authorize any commercial use.

Compliance with YouTube Policies: Our methods for collecting and disseminating data are fully compliant with YouTube's privacy policies. We ensure that no personal data or privacy rights are compromised during the process.

Data Licensing: We employ the protocol of CC BY 4.0.

4.2. The Potential of HOIGen-1M

While HOIGen-1M is designed for video generation, it still has great potential in HOI perception. For example, HOIGen-1M contains over 15,000 objects and more than 7,000 interaction action types, which significantly extends the categories of existing HOI detection datasets. Consequently, it is a golden testbed for HOI perception, such as open-vocabulary HOI detection. Besides, This dataset also has the potential to be utilized for training robotic arms to interact with objects, facilitating the development of embodied AI.

4.3. Ethical Issues

There are three main ethical issues of this paper: 1) privacy, 2) data bias, and 3) human annotation bias. For the first issue, the dataset will be distributed only for research purposes and we do not plan to provide the original raw data. To eliminate data bias, we select the videos from multiple sources, including YouTube, movies, sports competitions, public area surveillance, etc. To reduce human annotation bias, we first recruit annotators from diverse backgrounds, including academia and industry, to ensure the diversity of the annotation team. Moreover, we provide a clear annotation guide and conduct regular calibration tests to ensure consistent understanding of the annotation standards.

5. Visual Results

In this section, more visual results are provided in order to verify the effectiveness of HOIGen-1M. We show the results of CogVideoX [18] and its fine-tuned version on HOIGen-1M, as well as two commercial models: Gen3 [13] and Kling1.5 [7]. For video comparisons of more models please watch our **supplementary video**.

Dashed boxes of different colors in the below figures indicate that the content does not match the corresponding cue, and solid boxes indicate that the content matches.

CVPR
#4635

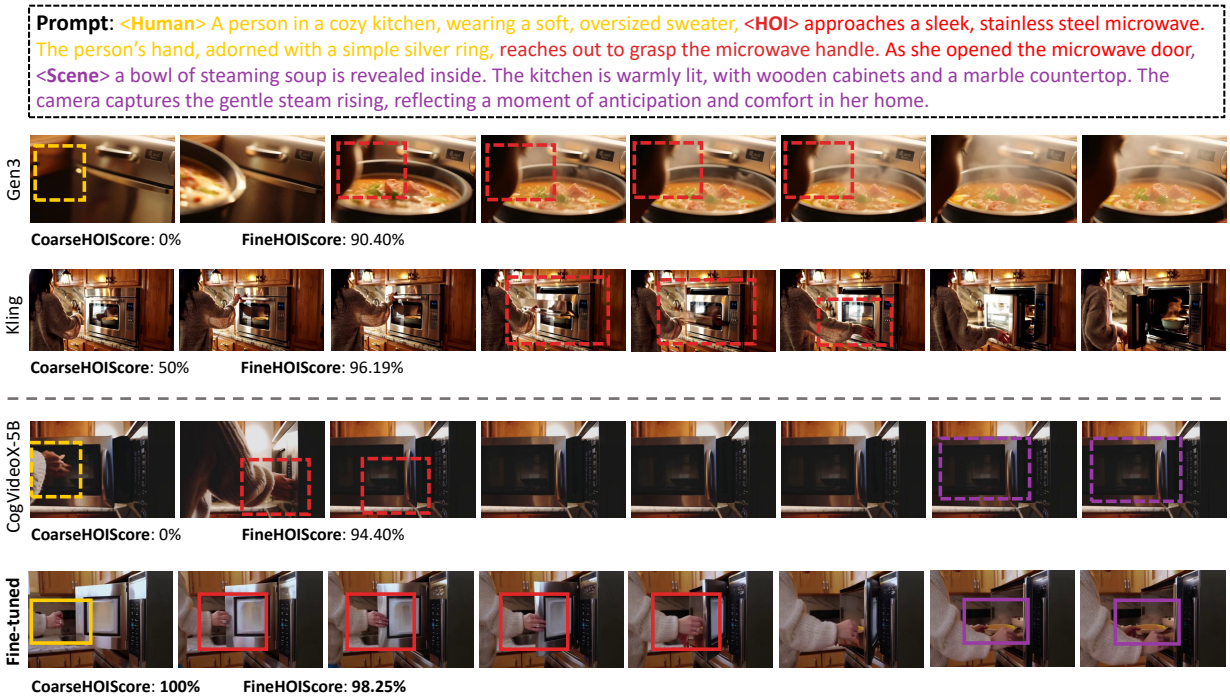

Figure 1. A person is opening a microwave door. Our generated video is the only one that generates the smooth interaction, which follows the physical law of the real world. Details in the prompt, such as the hand with a ring and the steaming soup, are also shown in full.

Figure 2. A man lifts and places a box in the trunk of the car. Our method generates the whole actions, including lifting and placing in, as well as detailed scenarios, such as a blanket and basket inside the trunk.

CVPR
#4635

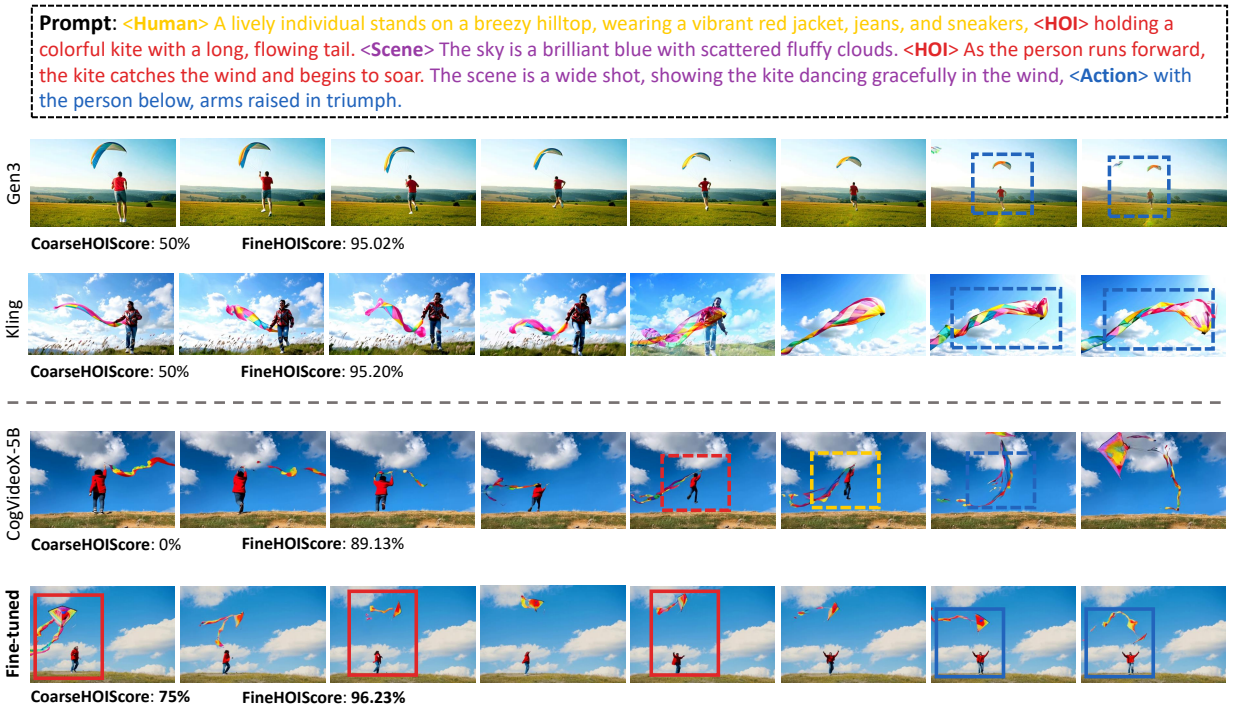

Figure 3. A person is running and flying a kite. Our approach successfully generates not only the interaction between the person and the kite, but also the wide shot as well as the action of the person raising his arms high.

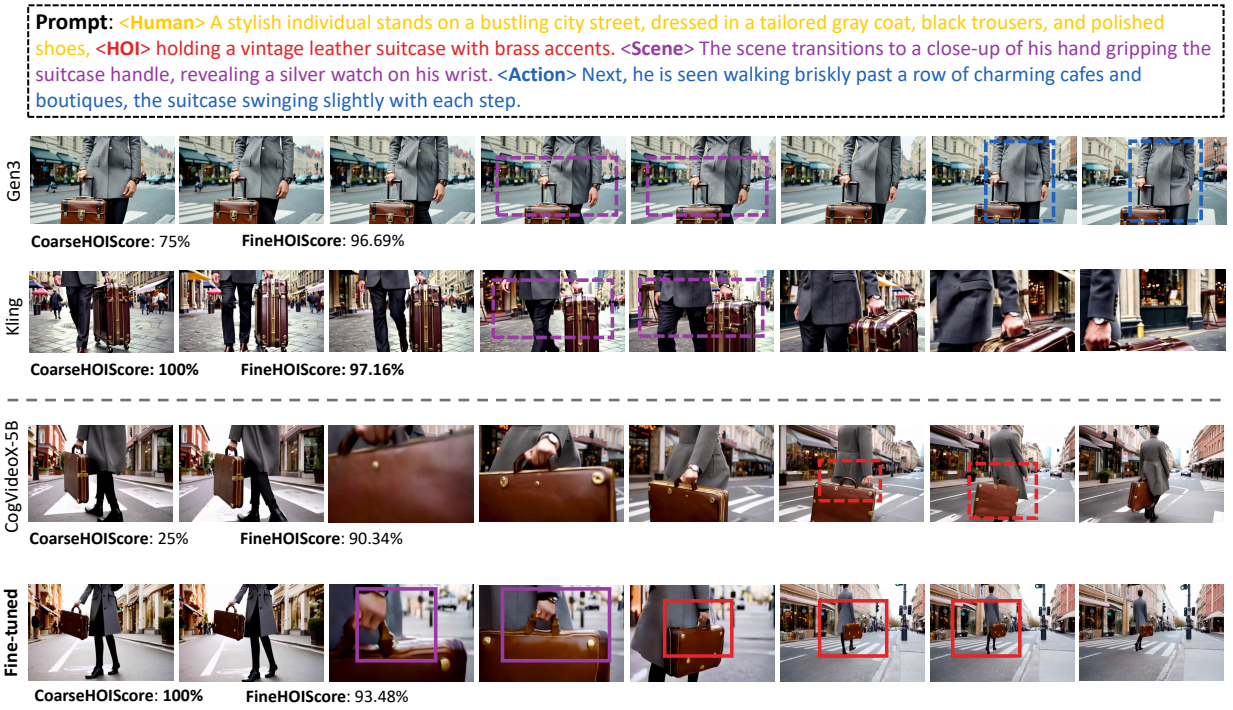

Figure 4. A man is carrying a suitcase and walking on the street. Our method successfully follows the scene transitions in the prompt: "close-up of his hand gripping the suitcase handle", and eliminates the unreasonable interactions of the CogvideoX-5B.

CVPR
#4635

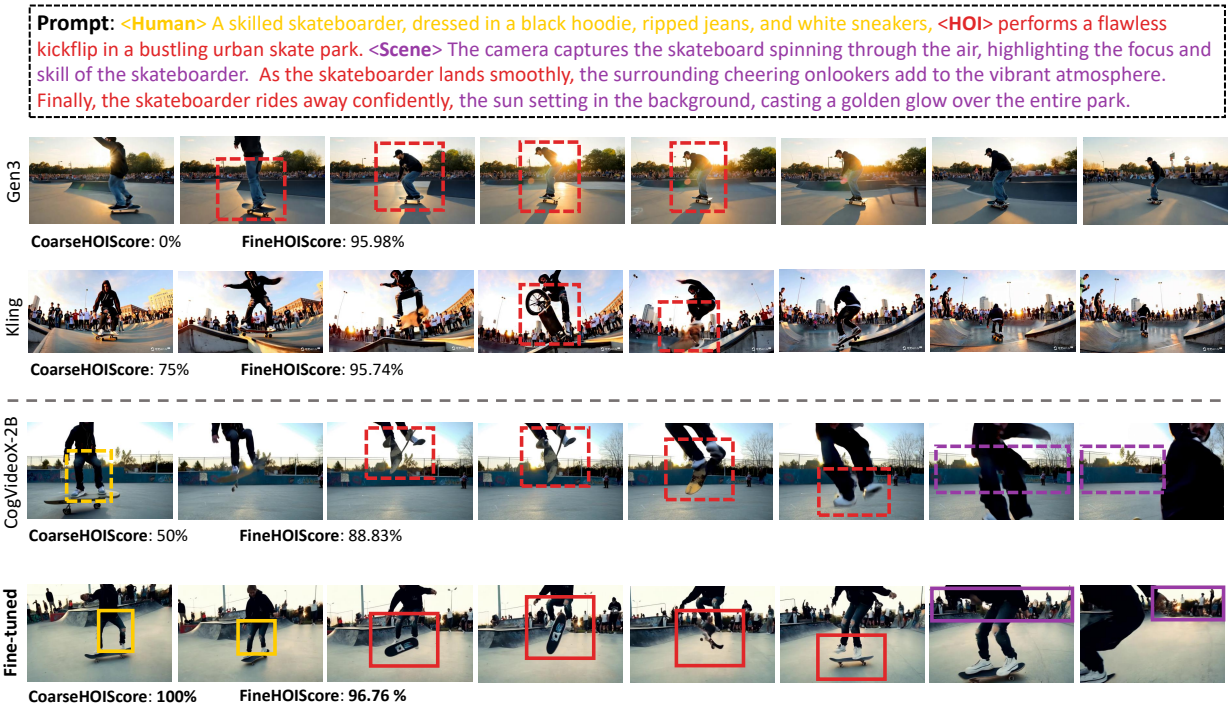

Figure 5. A complex interaction: kickflip with a skateboard. Our generated video maintains the best consistency in skateboarding. In addition, compared to CogvideoX-2B, our video also enhances the details of human and environments.

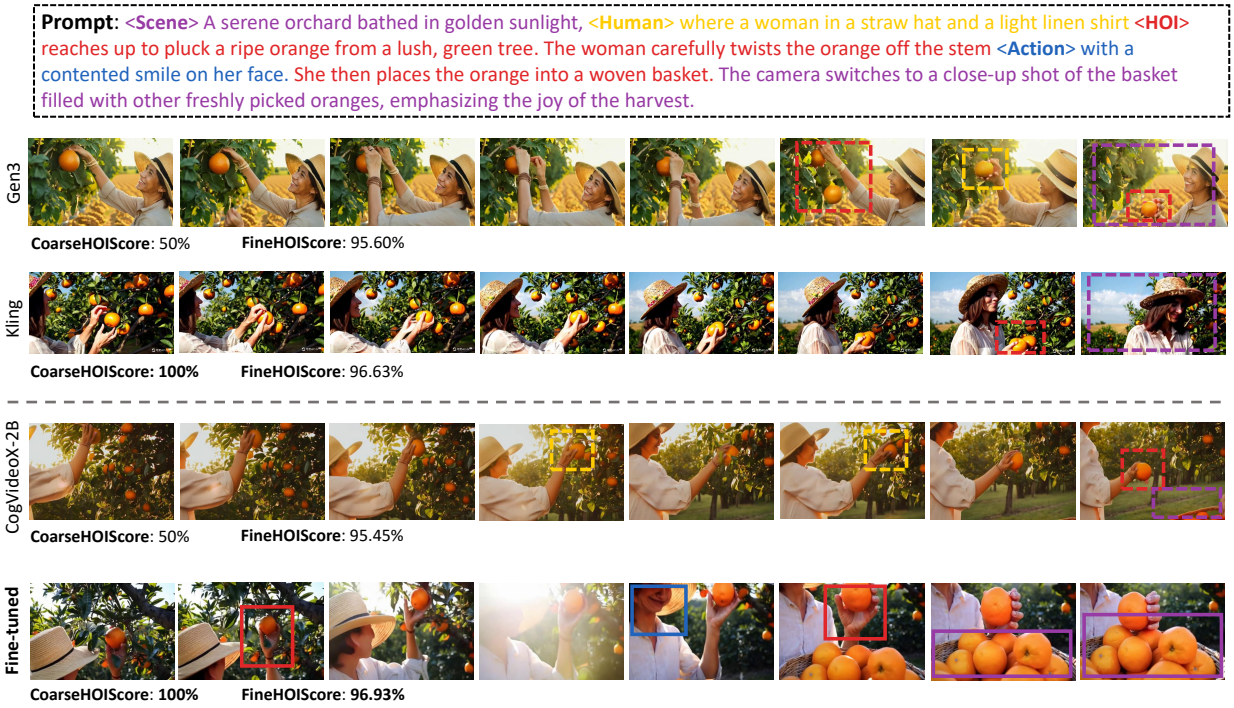

Figure 6. A woman is twisting the orange off the stem and placing the orange into a woven basket. Our method performs well on twisting, placing, and shot switching.

CVPR
#4635

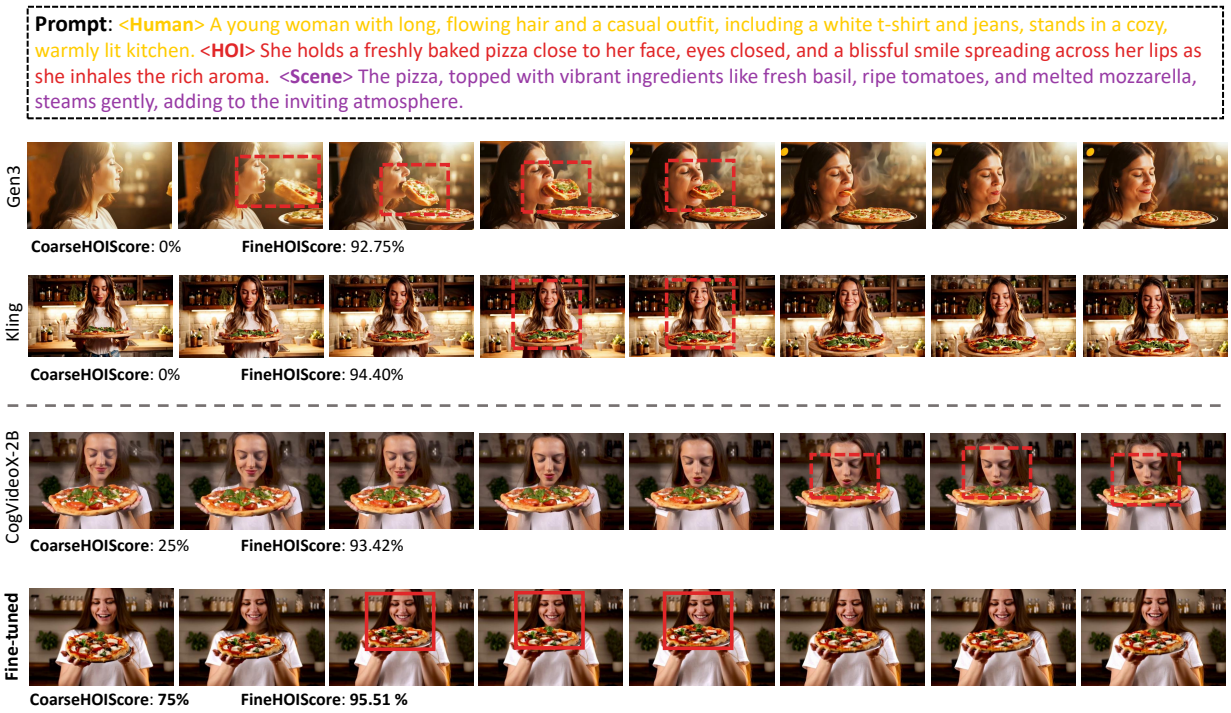

Figure 7. A woman is smelling Pizza. Other methods do not generate the action correctly, generating eating pizza or showing pizza to the camera.

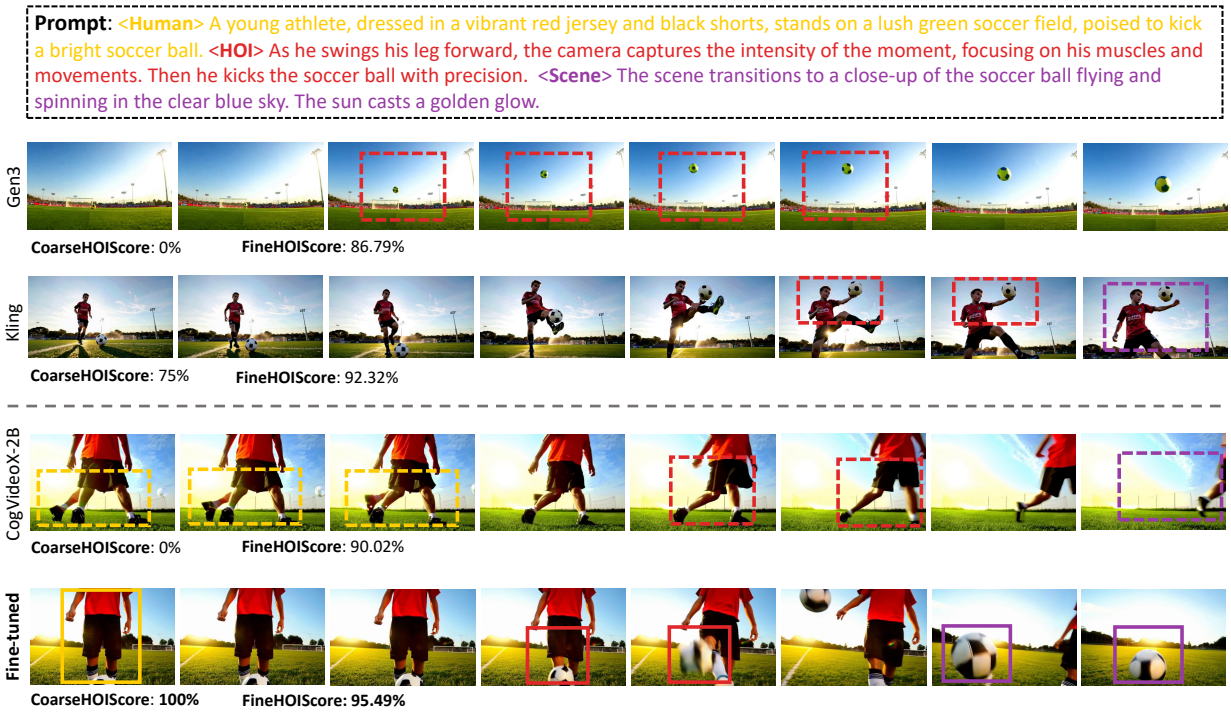

Figure 8. A young athlete is kicking a bright soccer. The video we generated not only correctly shows the full action of the kick, but also shows the precise scene transitions from the young athlete to the close-up of the soccer ball.

References

[1] 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. 1

[2] 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. 1

[3] Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016. 1

[5] 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. 1

[6] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *CoRR*, abs/2303.07399, 2023. 1

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

[8] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan. https://github.com/PKU-YuanGroup/Open-Sora-Plan, 2024. 1

[9] 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. 1

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2

[11] 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. 1

[12] 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. 1

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

[14] Qwen Team. Qwen2.5: A party of foundation models, 2024. 1

[15] 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. 2

[16] 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. 1

[17] 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. 1, 2

[18] 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

[19] 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