Open-world Hand-Object Interaction Video Generation Based on Structure and Contact-aware Representation

Supplementary Material

Abstract

In this supplementary material, we provide the following additional details and experimental comparisons:

- *The detailed Chain-of-Thought (CoT) prompt for hand and object grounding (see Sec. 1).*
- *Additional implementation details of network architecture and training protocols (see Sec. 2).*
- *Additional framework design analysis, including the qualitative comparison of HOI representations discussed in the main paper (see Sec. 3).*
- *Additional qualitative and quantitative comparisons on Taco [5], Taste-Rob [8], and open-world scenarios (see Sec. 4).*

1. VLM Grounding via Chain-of-Thought (CoT) Prompt

1.1. Detailed CoT Prompt

Our hand-object segmentation relies on a VLM-based grounding pipeline to identify hand and object regions. To achieve this in complex and open-vocabulary scenes, we employ the CoT prompt strategy detailed in Fig. S1. While we illustrate a single-hand case for clarity, our pipeline readily extends to double-hand scenarios with only minimal adaptation. Unlike a direct query, this CoT process decomposes the grounding task into a sequential cross-validation of three evidence sources: the **Task Description**, the **Visual Interaction**, and the **Temporal Object Motion**. The VLM first uses the task description to identify candidate objects in the first frame. Next, it analyzes the middle frame to localize the interacting hand (including the arm) and simultaneously verifies the object using the visual interaction cue. Finally, it confirms the object's identity by observing its temporal object motion in the last frame. This structured and multi-evidence reasoning improves grounding robustness, proving especially effective for open-vocabulary objects and in complex scenes involving distractors.

1.2. Qualitative Grounding Comparison

To further demonstrate the effectiveness of our CoT prompt, we compare three grounding settings: 1) Grounding DINO [4], 2) Qwen2.5-VL [1] without CoT reasoning, and 3) Qwen2.5-VL with our CoT prompt. As shown in Fig. S2, Grounding DINO and the VLM without CoT prompt often mislocalize the manipulated object or confuse visually similar items under cluttered conditions, whereas the CoT-guided model performs more consistent and precise grounding. This empirical observation demonstrates that the CoT-guided VLM more reliably handles open-vocabulary objects and complex scenes because its stepwise reasoning effectively reduces the contextual ambiguity that typically hinders direct grounding approaches.

2. Additional Implementation Details

2.1. Additional Architecture Details of Baselines

As introduced in the main paper, our proposed framework utilizes two mainstream pre-trained video diffusion models (VDMs): CogVideoXI2V-5B [7] and Wan2.1I2V-14B [6], denoted as SCAR_C and SCAR_W respectively, to showcase versatility. For the corresponding general-purpose baselines (CogVideoX [7] and Wan2.1 [6]), we employ the identical base models and apply LoRA [2] fine-tuning to ensure that the performance differences are attributable solely to our proposed methodology. We fine-tune CogVideoX using a LoRA dimension of 128 and Wan2.1 using a LoRA dimension of 256. For FLOVD [3], we adopt its CogVideoX-based instantiation and use the official architecture hyperparameters to ensure a fair comparison.

2.2. Training Details

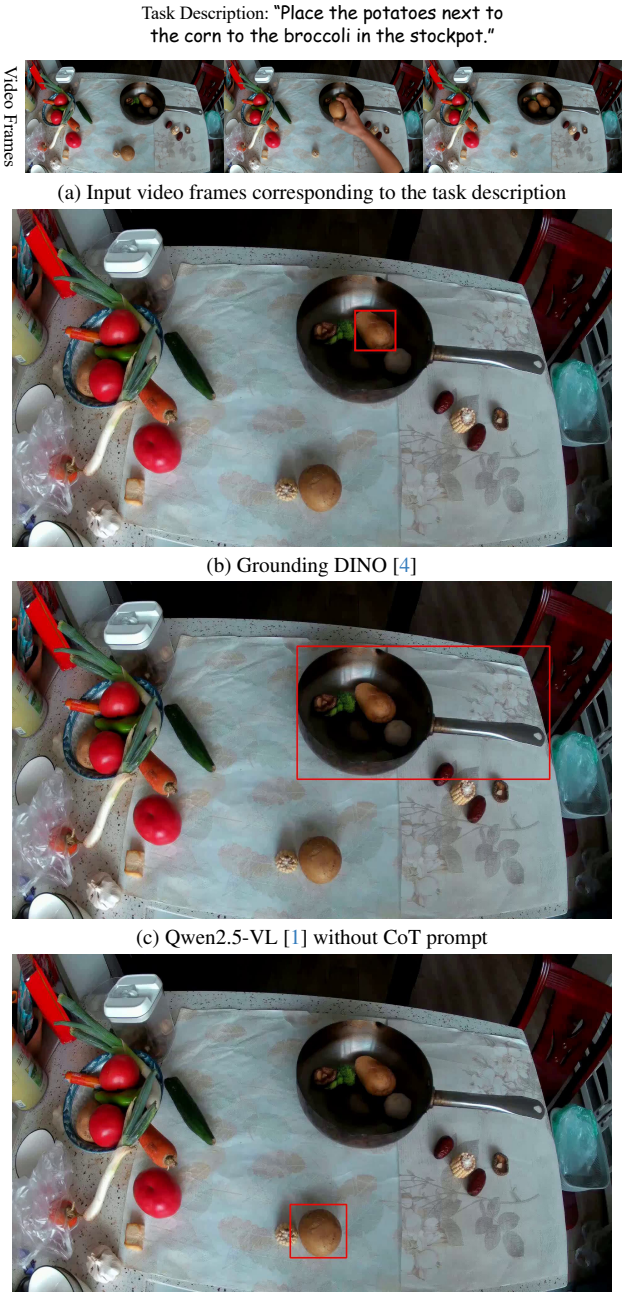
All models are trained on 8 NVIDIA A100-80G GPUs with a per-GPU batch size of 2. We train for 50k total steps on the Taste-Rob [8] dataset and 10k steps on the Taco [5] dataset. These training settings are kept consistent across all methods (our SCAR_C and SCAR_W, and all baselines) to ensure a fair comparison. We adhere to the

```
You are an advanced video analysis VLM. You will receive two types of inputs:
1. Task Description: A natural language command (e.g., "pick up the black spray bottle").
2. Image Frames: A sequence of three frames:
   • first_frame (index 0): The scene before the action begins.
   • middle_frame (index 1): The action in progress, where hand(s) appear.
   • last_frame (index 2): The state as the action is ending or completed.
You must output your reasoning by strictly following the Chain-of-Thought (CoT) Reasoning
Instructions provided below.
Your Objectives:
1. Hand Grounding: Localize the hand with arm (from shoulder → elbow → wrist →
   fingertips) in the middle_frame (index 1).
2. Object Grounding:
   • Localize the object interacted with hand and mentioned in the Task Description in the
     first_frame. (index 0).
Important Rules:
• Hand bounding box MUST be on middle_frame (index 1).
• Object bounding box MUST be on first_frame (index 0).
• The object box should not include the hand.
• If the object has a handle (e.g., cup, pot, tool), the handle must be fully enclosed.
• The hand bounding box must localize the hand with its arm (from shoulder → elbow → wrist
  → fingertips) and must fully include the arm.
Chain-of-Thought (CoT) Reasoning Instructions:
You must strictly follow this reasoning path and output your thoughts step-by-step.
1. Scene Analysis & Candidate Filtering (First Frame):
   Analyze the static scene in first_frame (0). Use the Task Description (e.g., "black
   spray bottle") to identify and pre-filter candidate objects that match the description.
2. Action Identification (Middle Frame):
   Analyze middle_frame (1). Detect the hand with the arm. Identify which object the
   hands are interacting with, confirming it is the same candidate object identified from
   the Task Description in Step 1.
3. Hand Grounding (Middle Frame):
   Using the hand with arm analysis from Step 2, localize the hand with arm in
   middle_frame (1). Output the bounding box(es) and justification.
4. Target Confirmation & Trace-back (Frames 0, 1, 2):
   Validate the object's identity (identified in Step 1 & 2) by testing for movement:
   compare its static location in first_frame(0) with its new location in last_frame
   (2). If movement is detected, this analysis confirms its identity and re-locates its
   original position in frame 0. If no movement is detected (e.g., a "wipe" task), this
   test fails; return to Step 1 to re-evaluate candidates, now informed that the target
   object is stationary.
5. Object Grounding (First Frame):
   Output the bounding box for the object's original location (identified in Step 4) onto
   first_frame (0).
Final Output Format:
[Fill in the complete reasoning process from Steps 1-5 here]

hand with arm bounding box (frame 1): [x1, y1, x2, y2]
Object bounding box (frame 0): [x1, y1, x2, y2]
```

Figure S1. Detailed CoT prompt for hand and object grounding.

original resolution and sampling configurations of the respective base models. The CogVideoX-based methods (our SCAR_C, CogVideoX and FLOVD) process videos at a resolution of 720×480 using 50 sampling steps, whereas the Wan2.1-based methods (our SCAR_W, Wan2.1) operate at 832×480 using 30 sampling steps. Note that while input videos are resized to these target resolutions during training, all generated videos are resized back to the original resolution for quantitative evaluation to ensure a fair comparison.



(a) Input video frames corresponding to the task description

(b) Grounding DINO [4]

(c) Qwen2.5-VL [1] without CoT prompt

(d) Qwen2.5-VL [1] with CoT prompt

Figure S2. Qualitative comparison of grounding methods. Given the input video frames and task description in (a), the CoT-guided model (d) achieves more accurate localization than both the specialized detector (Grounding DINO [4]) in (b) and the VLM baseline without reasoning steps in (c).

3. Additional Framework Design Analysis

3.1. Qualitative Comparison of Different HOI Representations

As shown in Fig. S3, we provide the qualitative comparison of different HOI representations. Optical flow (OF) and depth map (DM) representations fail to maintain spatiotemporal coherence, leading to visual artifacts such as the "blue ruler" disappearing. While maintaining object shape consistency to some extent, the hand-object mask (HOM) variant lacks explicit contact cues and thus frequently fails to model successful grasps. Removing hand-object contours (w/o HOC) or depth map (w/o DM) results in significant object inconsistency, as these components provide precise spatial localization and holistic structure, respectively. Furthermore, the w/o CG variant (removing the contact region) fails to execute fine-grained interactions, such as in the "measuring cup" scenario. Finally, adding 2D keypoints (+KP) also degrades visual quality, confirming that the overly complex auxiliary generative target hinders optimization rather than facilitating it. In contrast, the model with our full representation generates more physics-realistic videos. This demonstrates that our comprehensive and interaction-oriented HOI representation effectively guides the model in learning fine-grained interaction physics.

3.2. Effectiveness of Joint-generation Paradigm

Recall that we adopt a joint-generation paradigm that predicts the HOI video and its corresponding HOI representation simultaneously. To validate its effectiveness, we compare our framework to a two-stage baseline. This baseline adopts the network architecture of FLOVD [3] but predicts our proposed HOI representation rather than optical flow in the first stage. The second stage then uses this predicted representation as an explicit condition to generate the video. As shown in Fig. S4, this two-stage approach primarily suffers from error accumulation. The suboptimal HOI predictions from the first stage are propagated without refinement and compounded during video generation, resulting in a degraded final video. This is reflected in its lower performance across all metrics in Tab. S1. In contrast, our SCAR framework avoids this through mutual refinement, simultaneously yielding a more accurate HOI representation and a more physics-realistic video, thus achieving better quantitative performance.

3.3. Ablation Study of Share-and-specialization Generation Strategy

Recall that we propose a share-and-specialization strategy in our denoising network. To validate the superiority of this strategy, we compare it against two variants: 1) the specialization-only variant using an interaction embedding for the interaction tokens in the input layer but

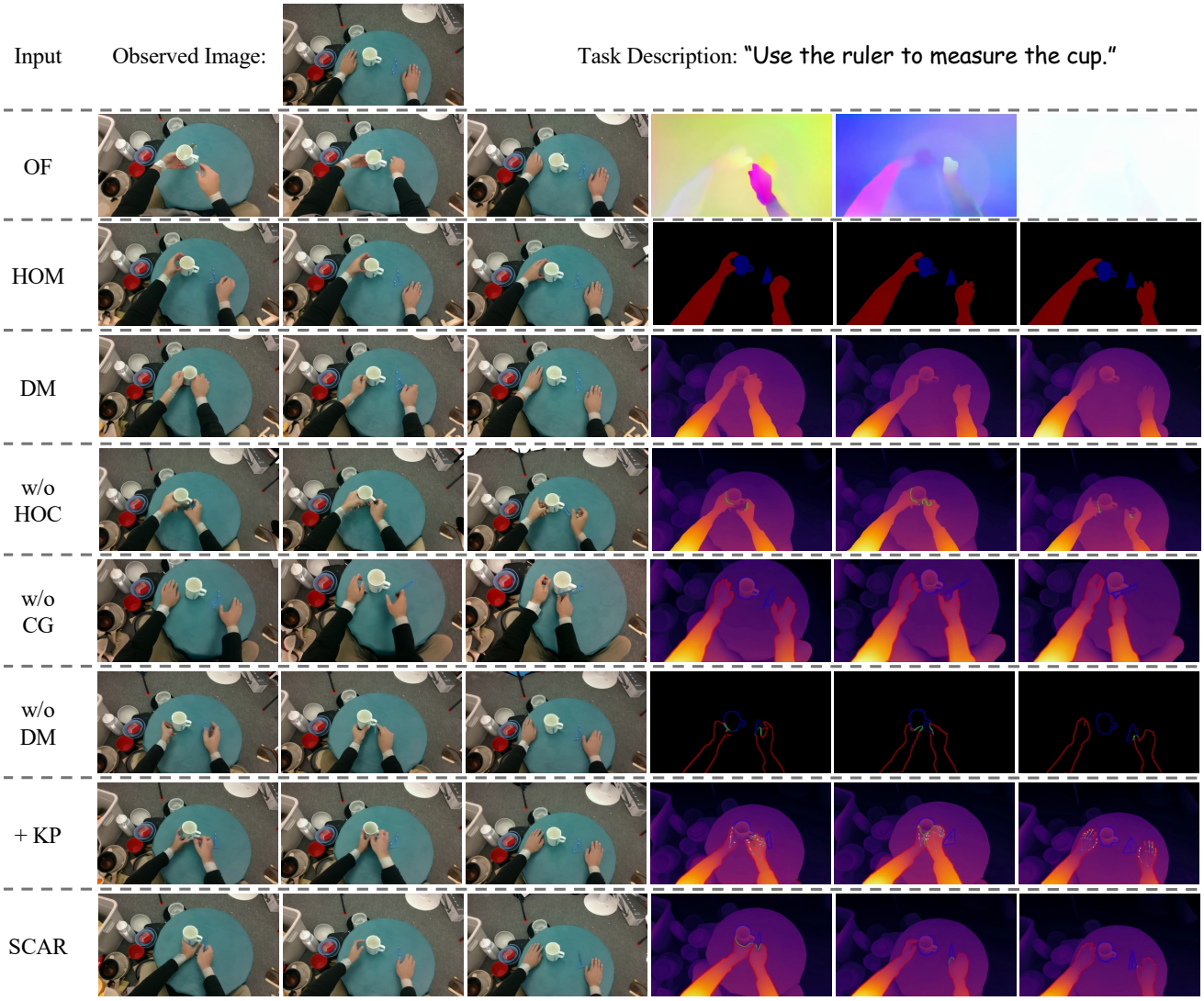


Figure S3. Qualitative comparison of different HOI representations. Baselines (OF, DM, HOM) and ablated variants (w/o HOC, w/o DM, w/o CG, +KP) all fail critical interaction aspects like object consistency or precise contact. In contrast, our full model generates physics-realistic and temporally coherent HOI videos.

Table S1. Quantitative comparison of our joint-generation paradigm against a two-stage variant. For all metrics, higher is better (↑). The best results are in **bold**.

Paradigms	Video Quality		Image Align.		Text Align.
	SC ↑	IQ ↑	ISC ↑	IBC ↑	VCS ↑
Two-stage	0.893	0.676	0.943	0.947	0.182
Joint-generation (our)	**0.916**	**0.698**	**0.951**	**0.954**	**0.187**

Table S2. Ablation study of our share-and-specialization generation strategy. For all metrics, higher is better (↑). The best results are in **bold**.

Strategy	Video Quality		Image Align.		Text Align.
	SC ↑	IQ ↑	ISC ↑	IBC ↑	VCS ↑
Spec.-only	0.908	0.689	0.944	**0.956**	0.184
Share-only	0.854	0.639	0.915	0.917	0.170
Share & Spec. (our)	**0.916**	**0.698**	**0.951**	0.954	**0.187**

removing the alignment loss (L_{align}), and 2) the share-only variant removing the interaction embedding but retaining the alignment loss (applied at the final DiT layer) to model the shared semantics between modalities. As demonstrated in Tab. S2, the share-only variant, which lacks

Figure S4. Qualitative comparison of our joint-generation paradigm against a two-stage variant. The two-stage variant suffers from error accumulation, where suboptimal HOI predictions from the first stage are compounded during video generation, culminating in a degraded final video. Our joint-generation framework avoids this via mutual refinement, simultaneously yielding a more accurate HOI representation and a more physics-realistic final video.

modality-specific modeling, leads to a suboptimal result. This is quantitatively confirmed by its significant performance degradation across all categories, especially in video quality (SC/IQ) and image-to-video alignment (ISC/IBC). The specialization-only variant (lacking shared semantics) also shows a degradation in key areas like video quality and text-to-video alignment.

In contrast, our share-and-specialization strategy achieves the best overall performance. This demonstrates the beneficial inductive bias of our strategy, which aligns with the modalities' dual nature by first building a robust understanding of their inherent semantic coupling and then refining their distinct and modality-specific characteristics.

3.4. Ablation Study on Alignment Loss Configuration

We conduct a comprehensive ablation study on the key hyperparameters of our alignment loss: the alignment layer (k^*) at which the loss is applied and its corresponding weight (λ_{align}). As shown in Tab. S3, applying the alignment loss at a shallow-middle layer ($k^* = 12$) yields the best performance. Enforcing alignment too early ($k^* = 4$) is suboptimal, as the hidden states have not yet captured rich enough semantics. Conversely, applying the loss too late ($k^* = 20$ or $k^* = 32$) begins to interfere with the model's ability to learn modality-specific details, degrading performance. With the optimal layer fixed at $k^* = 12$, we then find that a weight of $\lambda_{\text{align}} = 0.1$ provides the best balance. A weaker weight (0.01) provides an insufficient signal, and a stronger weight (0.5) imposes an overly strict constraint that interferes with the primary denoising task.

Table S3. Ablation study on the alignment loss configuration. For all metrics, higher is better (↑). The best results are in **bold**.

Configuration	Video Quality		Image Align.		Text Align.
	SC ↑	IQ ↑	ISC ↑	IBC ↑	VCS ↑
1. Ablating Alignment Layer k^* (fixing $\lambda_{align} = 0.1$)					
$k^* = 4$	**0.919**	0.685	0.941	0.946	0.183
$k^* = 12$ (our)	0.916	**0.698**	**0.951**	**0.954**	**0.187**
$k^* = 20$	0.910	0.693	0.948	0.951	0.186
$k^* = 32$	0.902	0.687	0.943	0.947	0.175
2. Ablating Alignment Weight λ_{align} (fixing $k^* = 12$)					
$\lambda_{\text{align}} = 0.01$	0.909	0.692	**0.953**	0.950	0.184
$\lambda_{\text{align}} = 0.1$ (our)	**0.916**	**0.698**	0.951	**0.954**	**0.187**
$\lambda_{\text{align}} = 0.5$	0.899	0.690	0.942	0.947	0.179

Table S4. Quantitative comparison in open-world scenarios. For all metrics, higher is better (↑). The best results are in **bold**, while the second best are underlined.

Method	Video Quality		Image Align.		Text Align.
	SC↑	IQ↑	ISC↑	IBC↑	VCS↑
CogVideoX [7]	**0.909**	0.709	0.921	0.929	0.239
Wan2.1 [6]	0.892	0.719	0.911	0.921	0.234
FLOVD [3]	0.903	0.711	0.916	0.924	0.238
SCAR (our)	0.907	**0.730**	**0.927**	**0.931**	**0.256**

4. Additional Qualitative and Quantitative Comparisons

4.1. Results on Taco Dataset

We illustrate additional qualitative comparisons on Taco [5] in Figs. S5 and S7. Correspondingly, the HOI representations generated by our SCAR and FLOVD [3] are presented in Figs. S6 and S8. Unlike existing methods that struggle with physical realism, our SCAR generates physics-realistic and temporally coherent videos by jointly generating our proposed HOI representation.

4.2. Results on Taste-Rob Dataset

We illustrate qualitative comparisons on Taste-Rob [8] in Figs. S9 and S11. Correspondingly, the HOI representations generated by our SCAR and FLOVD [3] are presented in Figs. S10 and S12. Existing methods struggle with physical realism, whereas our SCAR generates physics-realistic and temporally coherent videos through jointly generating our proposed HOI representation.

4.3. Results on Open-world Scenarios

We report the quantitative comparison on our open-world benchmark in Tab. S4. Compared to state-of-the-art methods, our SCAR consistently outperforms all baselines. We provide additional qualitative comparisons illustrated in Figs. S13, S14, S15, and S16. All baseline methods exhibit significant hand-object distortion, temporal inconsistencies, and physically implausible contact, while also frequently failing to align with the task description. In contrast, across diverse open-world scenarios, our SCAR consistently generates physics-realistic and temporally coherent HOI videos.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[3] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025. 1, 3, 6, 7, 8, 9, 10, 11, 12

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1, 3

[5] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 1, 6, 7, 8

[6] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 6, 7, 8, 9, 10, 11, 12

[7] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 6, 7, 8, 9, 10, 11, 12

[8] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27683–27693, 2025. 1, 6, 9, 10

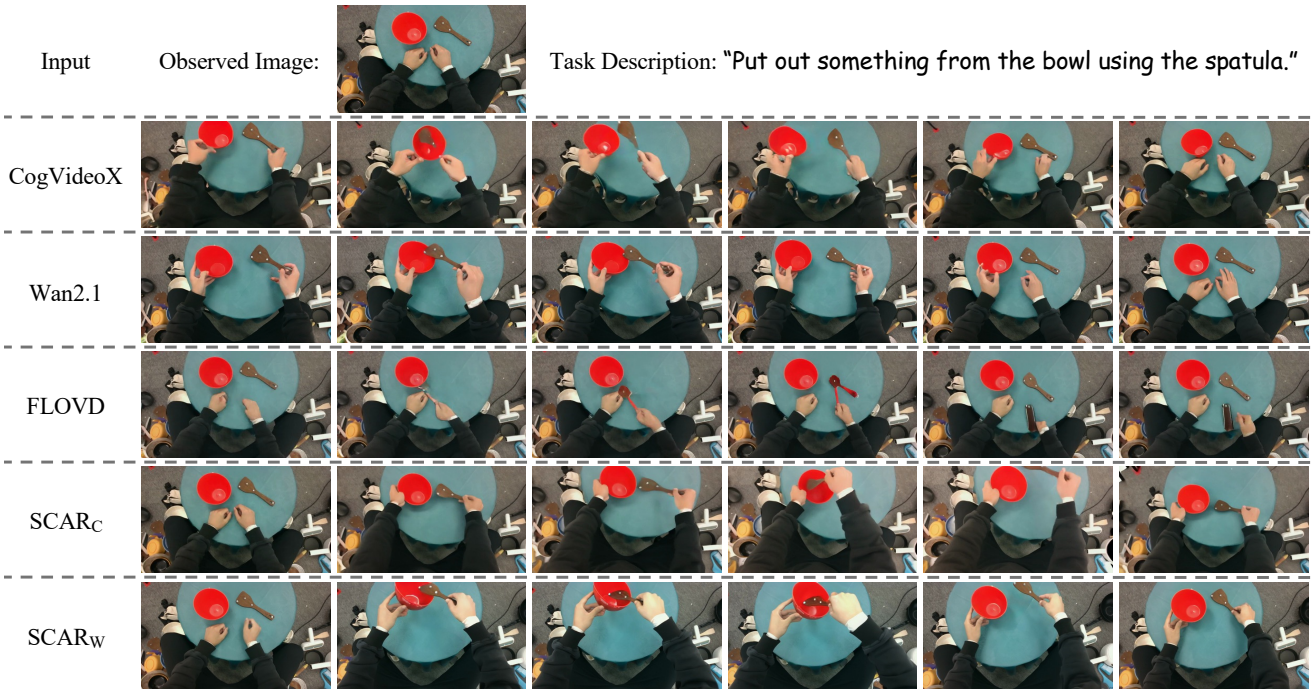


Figure S5. Qualitative comparison with state-of-the-art methods on Taco [5] dataset. CogVideoX [7] and Wan2.1 [6] suffer from significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In addition to these visual artifacts, they struggle to model the correct relative depth, resulting in a physically implausible interaction where the spatula fails to reach inside the bowl. The two-stage FLOVD [3] suffers from error propagation, leading to severe temporal inconsistency of the object. In contrast, our SCAR generates physics-realistic and temporally coherent videos by jointly generating our proposed HOI representation. Please refer to the supplementary video for better illustration.

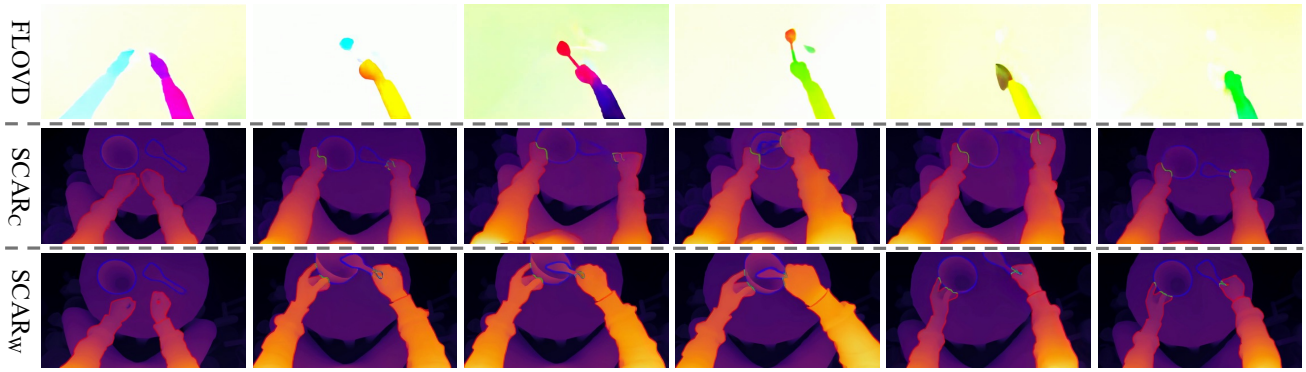


Figure S6. Qualitative comparison of generated HOI representations corresponding to Fig. S5. The optical flow generated by FLOVD [3] is noisy and inaccurate, which leads to the error propagation seen in the final video. In contrast, our jointly generated representation embodies consistent structural and contact cues, indicating that the model captures physical interaction patterns.

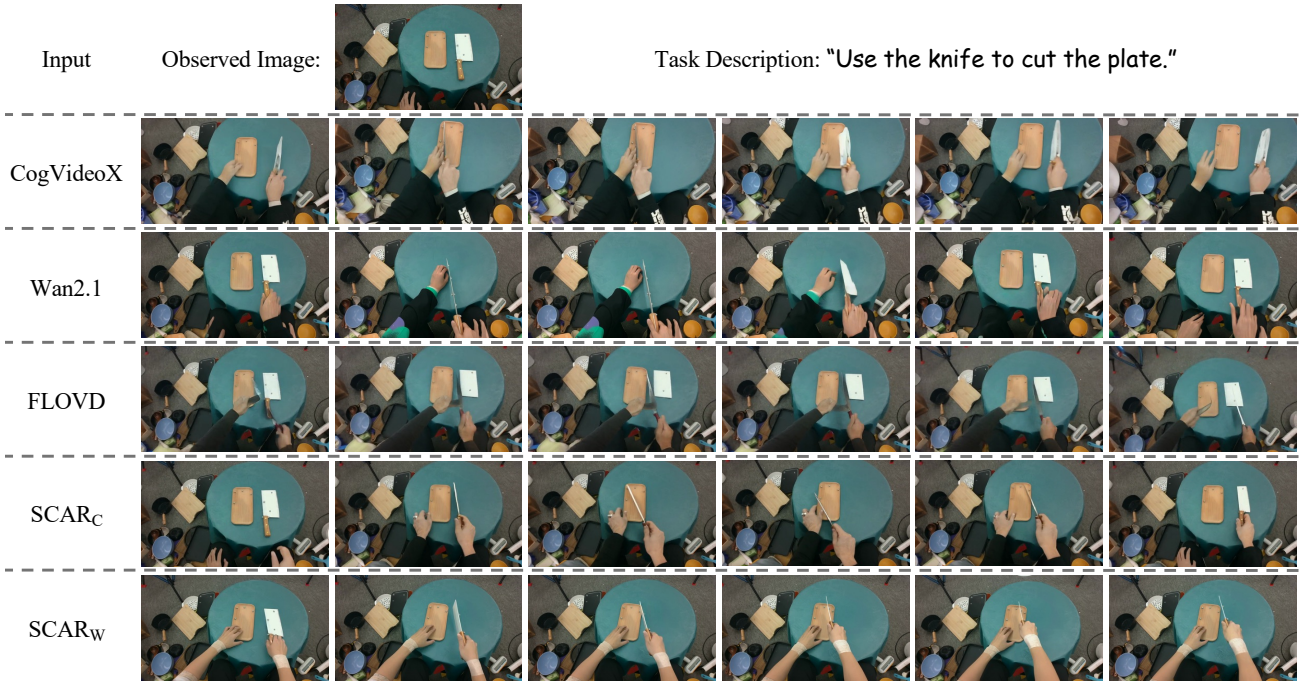


Figure S7. Qualitative comparison on the Taco [5] dataset. CogVideoX [7] and Wan2.1 [6] struggle with significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In addition to these visual artifacts, the two-stage FLOVD [3] suffers from error propagation, leading to object hallucinations (e.g., a new knife appearing). In contrast, our SCAR generates physics-realistic and temporally coherent videos by jointly generating our proposed HOI representation. Please refer to the supplementary video for better illustration.

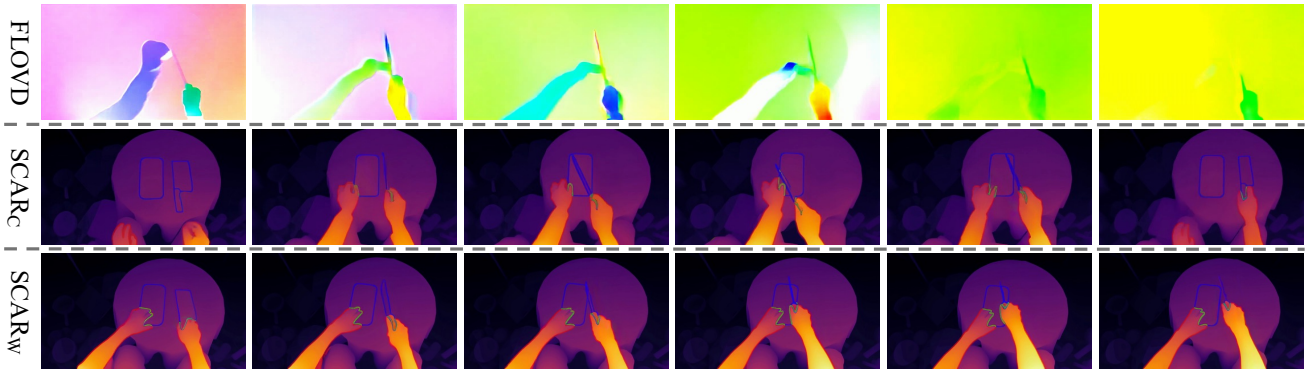


Figure S8. Qualitative comparison of generated HOI representations corresponding to Fig. S7. The optical flow generated by FLOVD [3] is noisy and inaccurate, which leads to the error propagation seen in the final video. In contrast, our jointly generated representation embodies consistent structural and contact cues, indicating that the model captures physical interaction patterns.

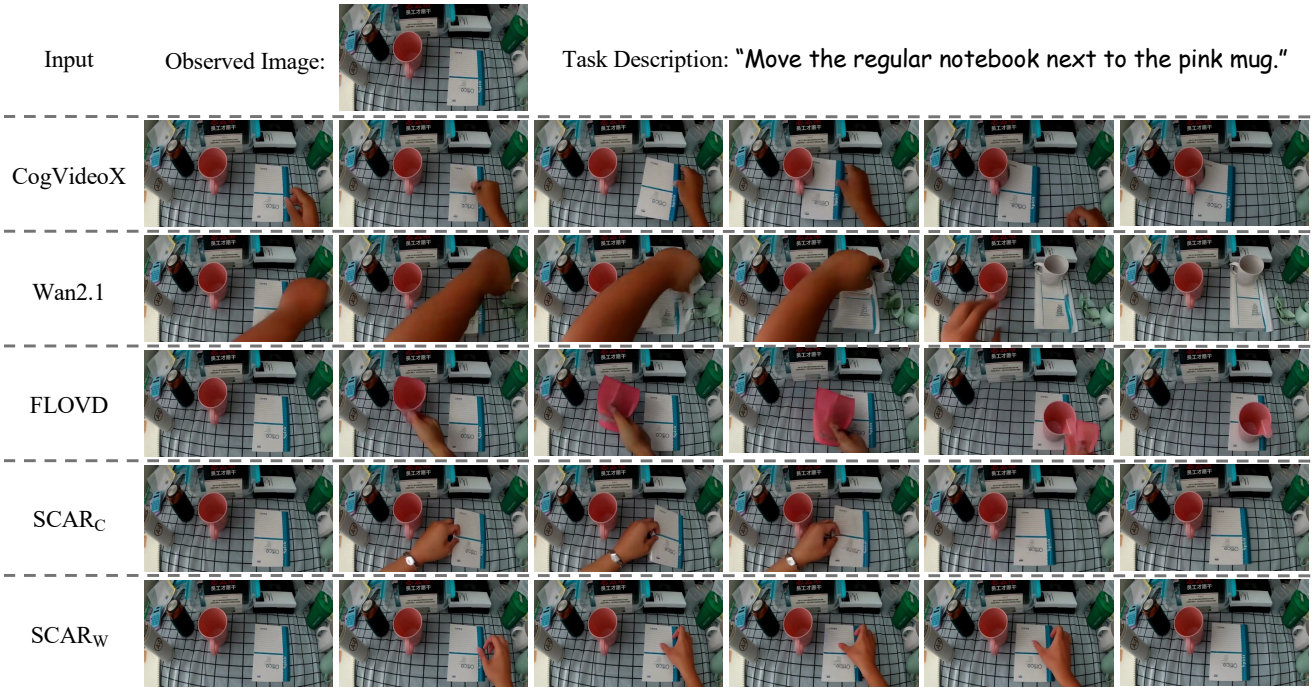


Figure S9. Qualitative comparison on the Taste-Rob [8] dataset. CogVideoX [7] fails to model relative depth, causing the notebook to slide under the pink mug. Wan2.1 [6] and FLOVD [3] exhibit severe object distortion and physically implausible contact. In addition to these visual artifacts, Wan2.1 and FLOVD also fail to adhere to the action specified in the task description (e.g., Wan2.1 interacting with a white mug and FLOVD moving a pink mug, instead of the target notebook). In contrast, our SCAR generates physics-realistic and temporally coherent videos by jointly generating our proposed HOI representation. Please refer to the supplementary video for better illustration.

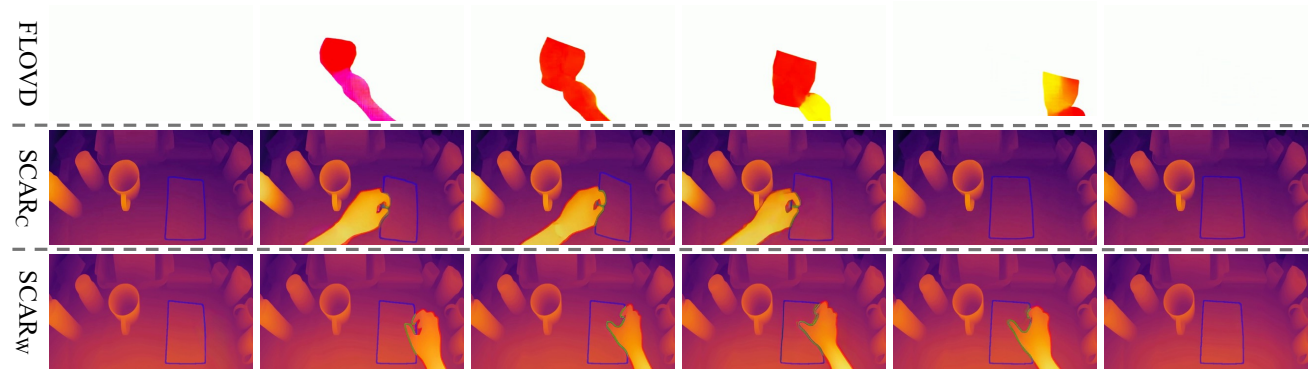


Figure S10. Qualitative comparison of generated HOI representations corresponding to Fig. S9. The optical flow generated by FLOVD [3] is noisy and inaccurate, which leads to the error propagation seen in the final video. In contrast, our jointly generated representation embodies consistent structural and contact cues, indicating that the model captures physical interaction patterns.

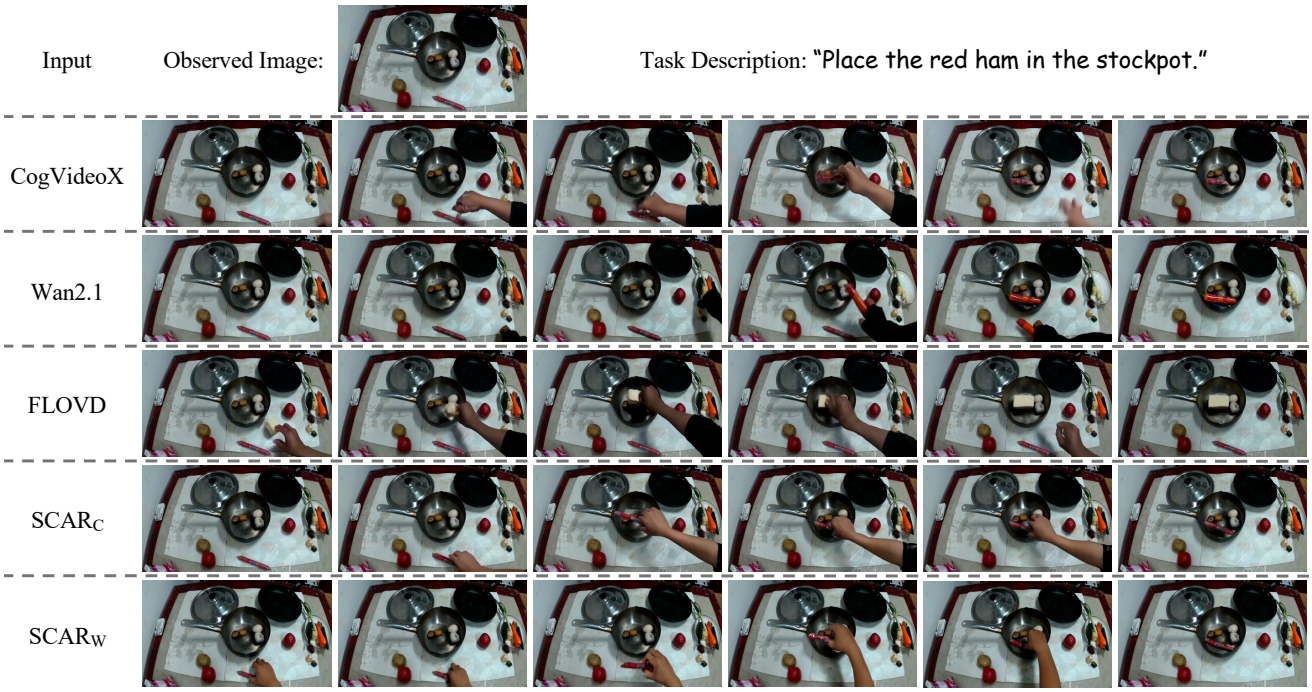


Figure S11. Qualitative comparison on the Taste-Rob [8] dataset. CogVideoX [7] struggles from significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In addition to these visual artifacts, Wan2.1 [6] and FLOVD [3] also fail to adhere to the action specified in the task description (e.g., Wan2.1 picking up a green onion and FLOVD interacting with tofu, instead of the target red ham). In contrast, our SCAR generates physics-realistic and temporally coherent videos by jointly generating our proposed HOI representation. Please refer to the supplementary video for better illustration.

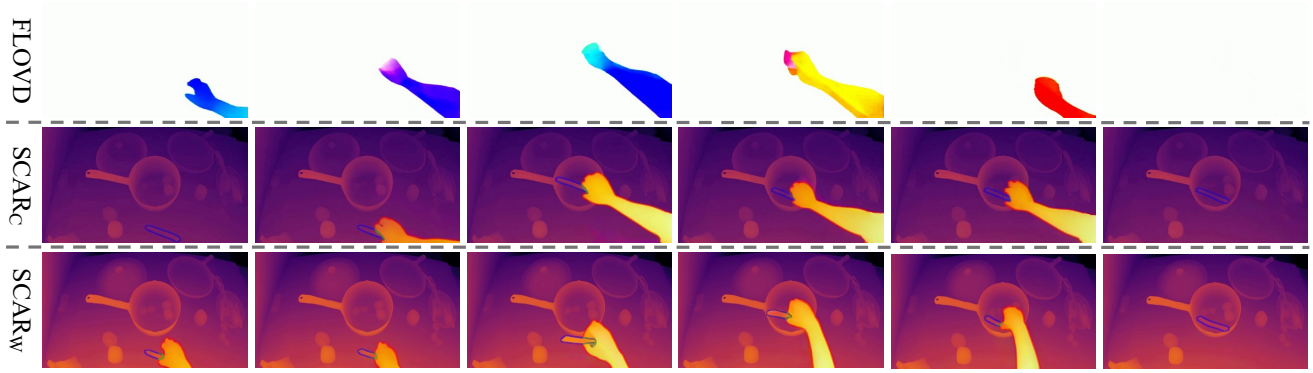


Figure S12. Qualitative comparison of generated HOI representations corresponding to Fig. S11. The optical flow generated by FLOVD [3] is noisy and inaccurate, which leads to the error propagation seen in the final video. In contrast, our jointly generated representation embodies consistent structural and contact cues, indicating that the model captures physical interaction patterns.

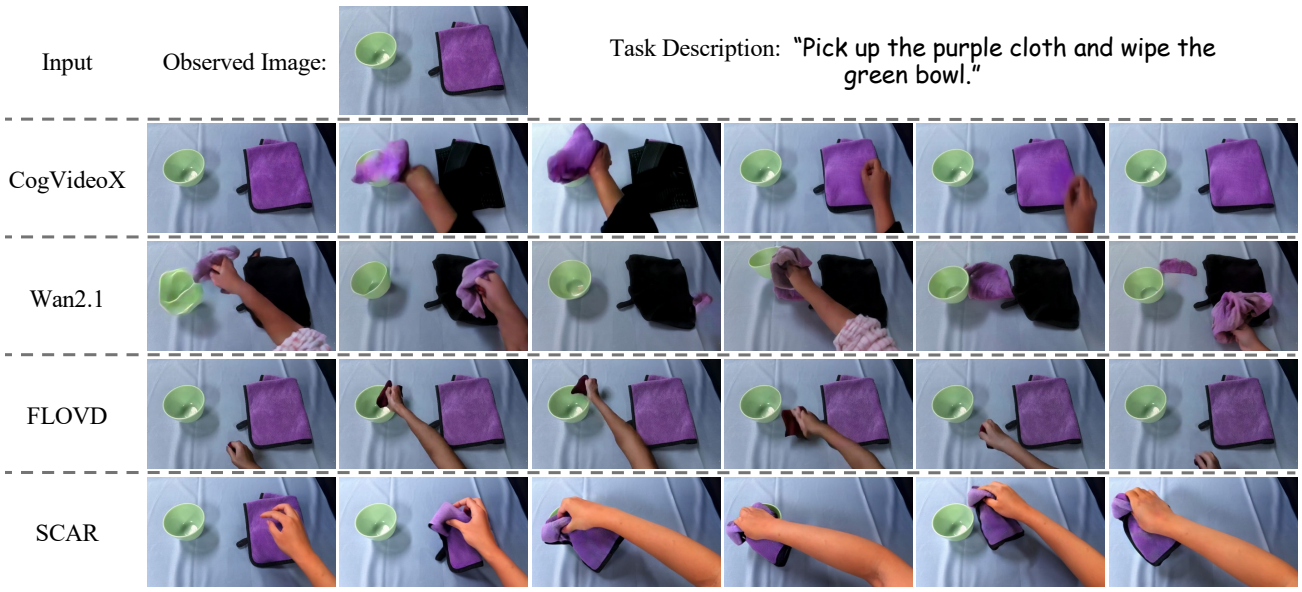


Figure S13. Qualitative comparison on a challenging open-world task. CogVideoX [7], Wan2.1 [6], and FLOVD [3] struggle with significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In contrast, our SCAR generates a physics-realistic and coherent video that correctly executes the challenging task involving unseen target objects. Please refer to the supplementary video for better illustration.

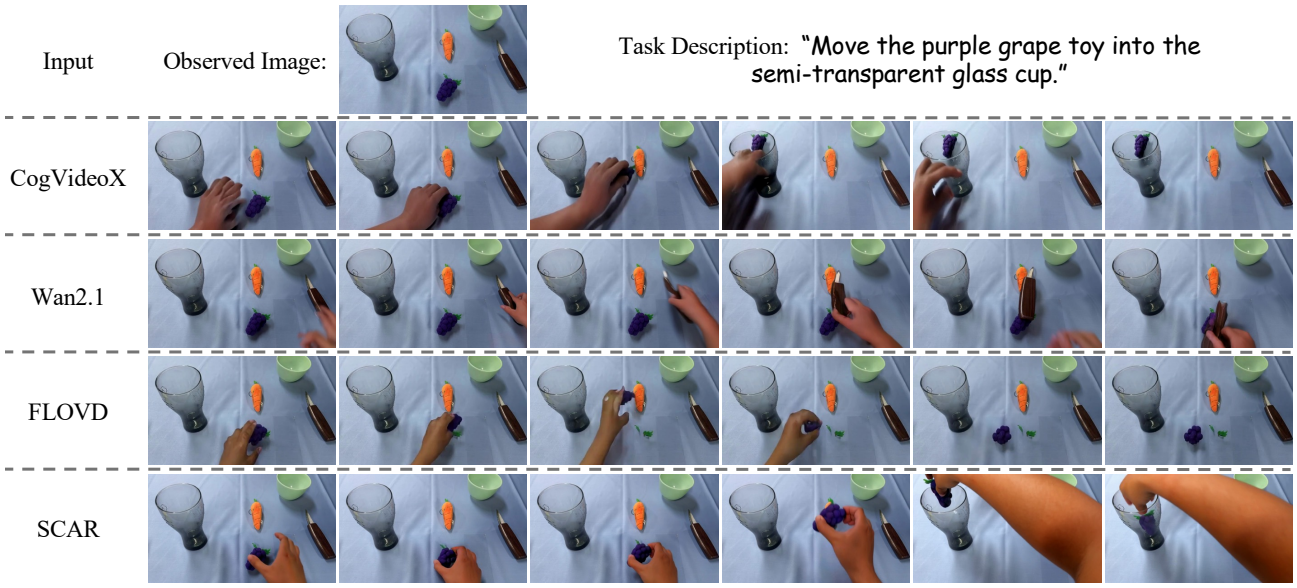


Figure S14. Qualitative comparison on a challenging open-world task. CogVideoX [7], Wan2.1 [6], and FLOVD [3] struggle with significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In contrast, our SCAR generates a physics-realistic and coherent video that correctly executes the challenging task involving unseen target objects. Please refer to the supplementary video for better illustration.

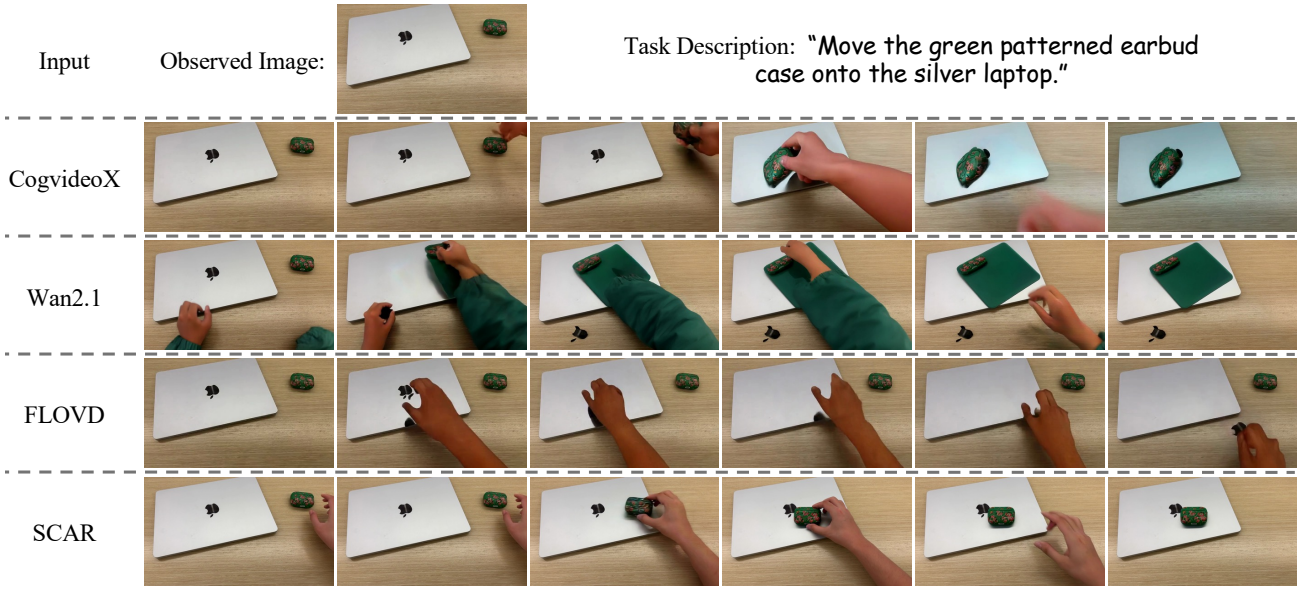


Figure S15. Qualitative comparison on a challenging open-world task. CogVideoX [7], Wan2.1 [6], and FLOVD [3] struggle with significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In addition to these visual artifacts, FLOVD also fails to adhere to the action specified in the task description (e.g., erroneously moving the Apple logo). In contrast, our SCAR generates a physics-realistic and coherent video that correctly executes the challenging task involving unseen target objects. Please refer to the supplementary video for better illustration.

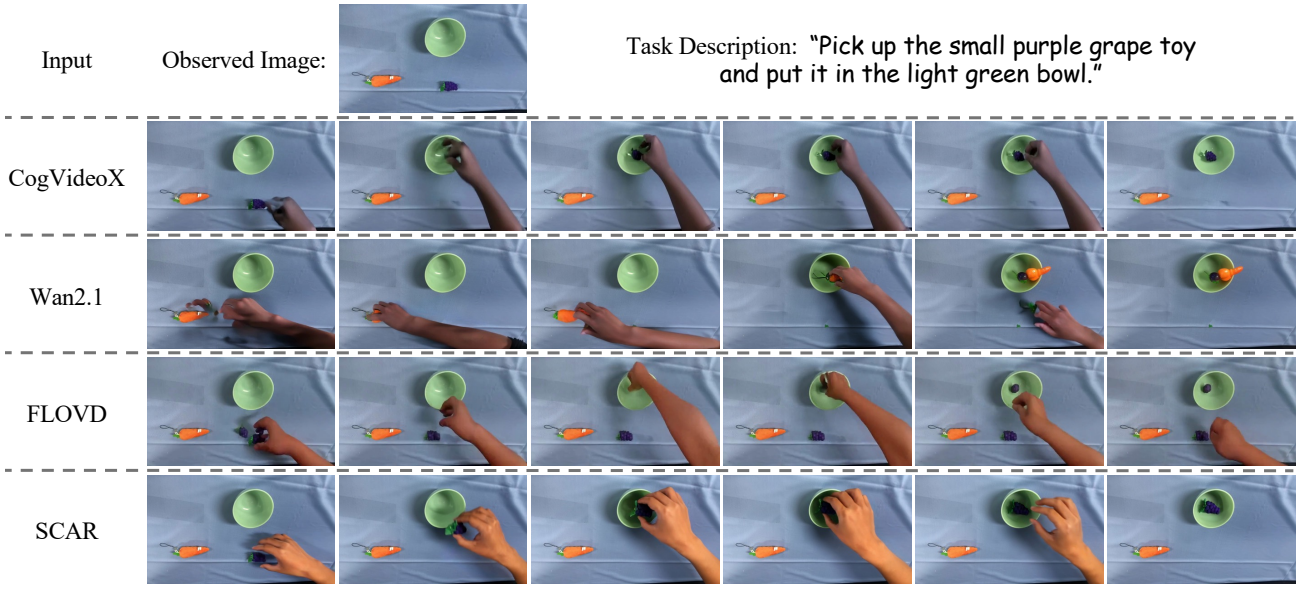


Figure S16. Qualitative comparison on a challenging open-world task. CogVideoX [7], Wan2.1 [6], and FLOVD [3] struggle with significant hand-object distortion, temporal inconsistencies, and physically implausible contact. In addition to these visual artifacts, FLOVD also fails to adhere to the action specified in the task description (e.g., picking up an orange carrot toy instead of a purple grape toy). In contrast, our SCAR generates a physics-realistic and coherent video that correctly executes the challenging task involving unseen target objects. Please refer to the supplementary video for better illustration.