

1 A Broader Impacts

2 Our physics-informed world model significantly advances robotic learning by generating high-
3 fidelity synthetic data with inherent physical plausibility, reducing reliance on costly real-world data
4 collection while improving simulation-to-reality transfer. This technology enables safer and more
5 efficient training of assistive robots for healthcare, disaster response, and industrial applications,
6 while its computational efficiency lowers barriers for broader research participation. By embedding
7 physical constraints during generation, we enhance the reliability of robotic data generation, though
8 future work should further address considerations in synthetic data diversity and establish governance
9 frameworks for responsible deployment in safety-critical domains.

10 B Limitations

11 While our physics-informed world model demonstrates significant improvements in physical consis-
12 tency and generation fidelity, we identify three key areas for future enhancement. First, the current
13 physical constraints focus primarily on geometric and kinematic properties, leaving opportunities to
14 incorporate additional material characteristics and dynamic force interactions for more comprehensive
15 physical understanding. Second, exploring emerging diffusion-autoregressive hybrid frameworks
16 is expected to achieve better trade-offs between generation quality (particularly for long-horizon
17 tasks) and computational efficiency. Third, we plan to extend to more diverse embodiments and larger
18 multi-domain datasets to further validate our model’s generalization capabilities.

19 C Code Reproducibility

20 For reproducibility, we provide codes in the following anonymous repository: [https://anonymous.](https://anonymous.4open.science/r/RoboScape-3652)
21 [4open.science/r/RoboScape-3652](https://anonymous.4open.science/r/RoboScape-3652).

22 D Baseline Details

23 We provide details of the compared baselines as follows:

- 24 • **IRASim**: A DiT-based robotic video generation model, capable of generating videos conditioned
25 on robot actions and trajectories.
- 26 • **iVideoGPT**: An auto-regressive interactive world model that takes the current video frame observa-
27 tion and action as input to predict the next frame while simultaneously estimating the reward signal
28 for robotic operations.
- 29 • **Genie**: A foundation world model trained through unsupervised learning on massive video data.
30 We implement it with a reproduced open-source repository ¹.
- 31 • **CogVideoX**: An advanced DiT-based text-to-video generation framework, with superior perfor-
32 mance in prompt-driven video generation.

33 E Scaling Behavior of RoboScape

34 We investigate the scaling behavior of RoboScape in terms of both model and data scales. As
35 shown in Figure 1, we evaluate three model variants—RoboScape-S (34M), RoboScape-M (131M),
36 and RoboScape-L (544M)—and observe a clear scaling law: all six evaluation metrics improve
37 significantly as model capacity increases.

38 In addition, we study the impact of data scale by training RoboScape-S on 1,000K, 3,000K, and
39 6,000K clips (Figure 2). While increasing data size consistently enhances visual quality and action
40 controllability, geometric accuracy exhibits marginal improvement or even slight degradation. We
41 find that this is because smaller datasets encourage overfitting to the final frame of conditional inputs,
42 artificially inflating geometric metrics without generating meaningful temporal dynamics. Despite
43 this, the overall trend confirms that more training data leads to better model performance.

¹<https://github.com/1x-technologies/1xgpt>

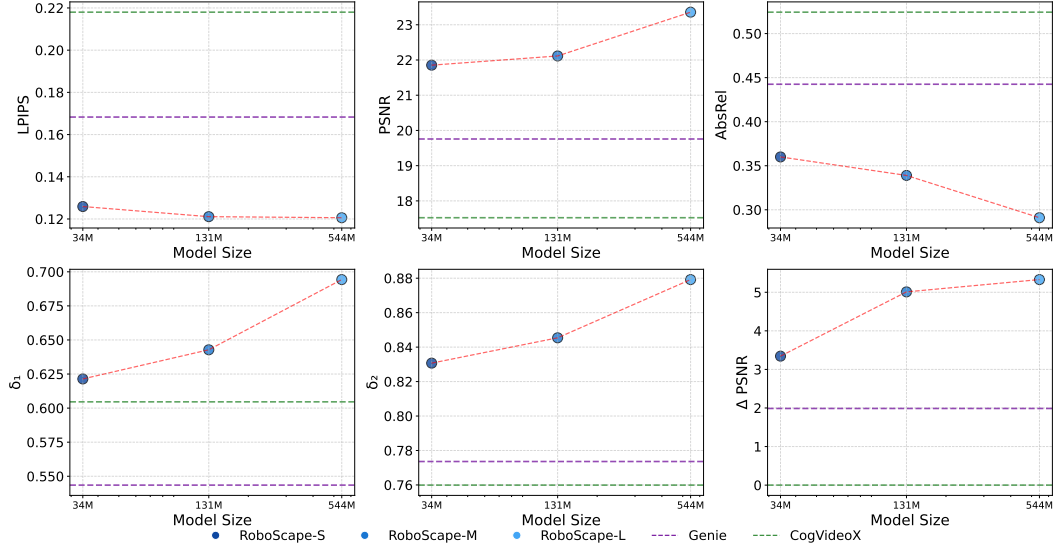


Figure 1: Model scaling law of RoboScape.

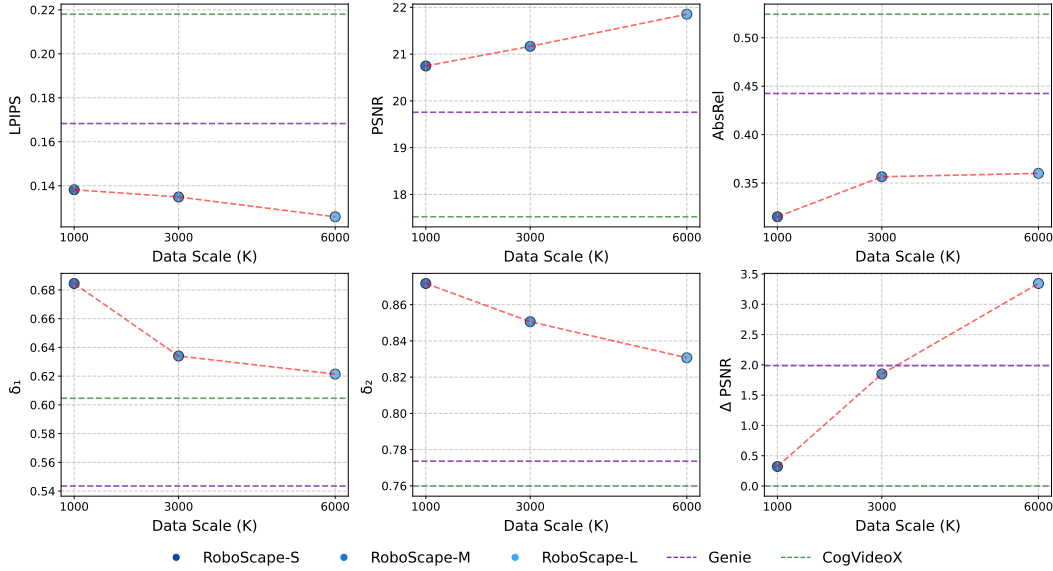


Figure 2: Data scaling law of RoboScape-S.

44 These findings highlight the importance of both model and data scaling in advancing robotic video
 45 generation, with larger models and datasets yielding better results.

46 F More Visualization Results

47 F.1 Video Generation Results

48 We provide more visualization results of generated videos using our model, as illustrated in Figure 3.

49 F.2 Robotic Policy Learning

50 We provide some visualization results of generated data on Robomimic and LIBERO using our model,
 51 which are shown in Figure 4 and Figure 5.

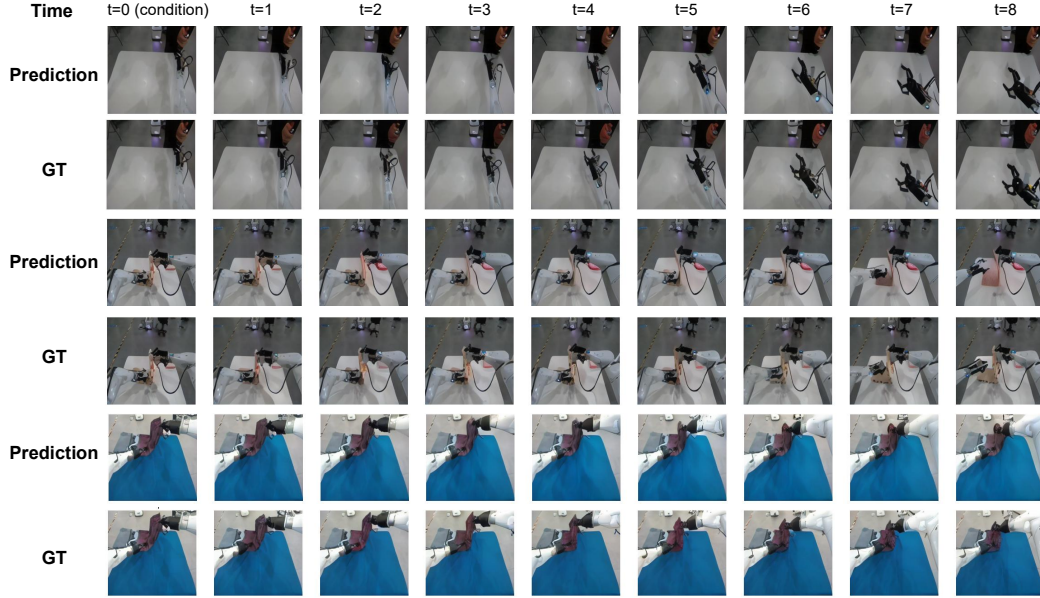


Figure 3: Supplemented visualization results from our model (only the subsequent 8 frames are shown).

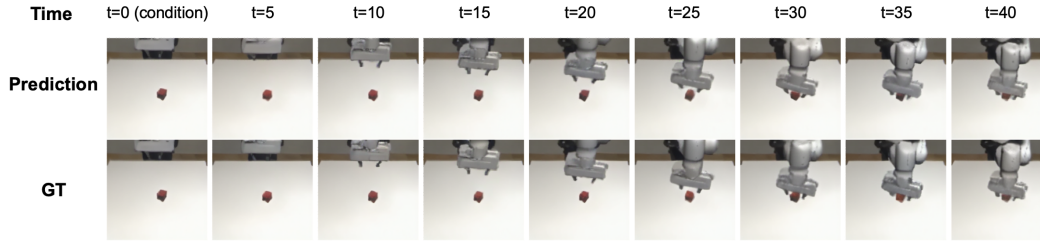


Figure 4: Supplemented visualization results on Robomimic (displaying every 5th frame; 8 frames shown from t=0 to t=40).

52 F.3 Robotic Policy Evaluation (add visualization results of our model and baselines)

53 In this part, we provide visualization results of RoboScape and other baselines in policy evaluation.
 54 The failure cases are presented in Figure 6 while the successful cases are shown in Figure 7.

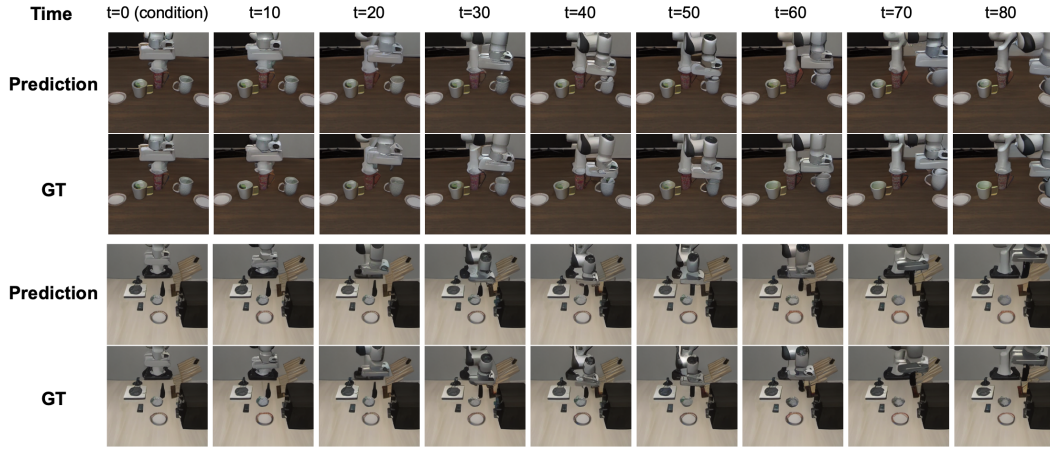


Figure 5: Supplemented visualization results on LIBERO (displaying every 10th frame; 8 frames shown from t=0 to t=80).

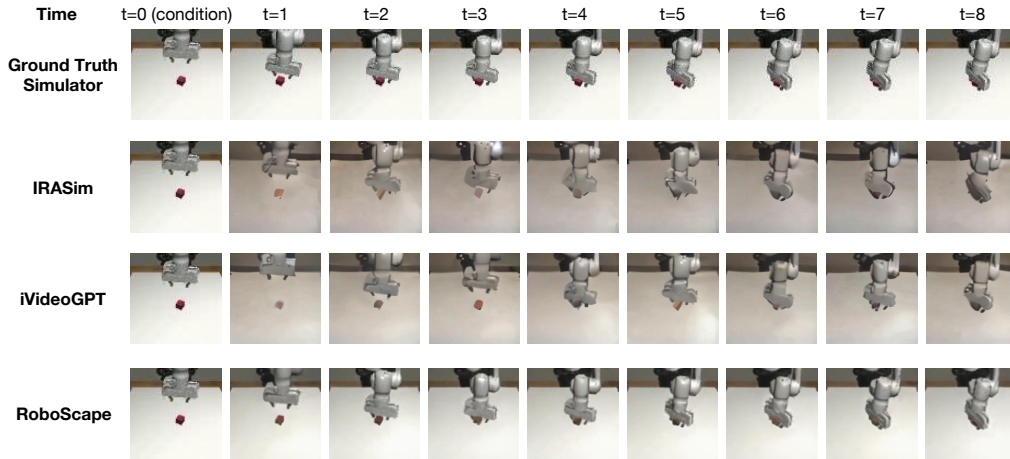


Figure 6: Supplemented visualization results of failure cases in policy evaluation.

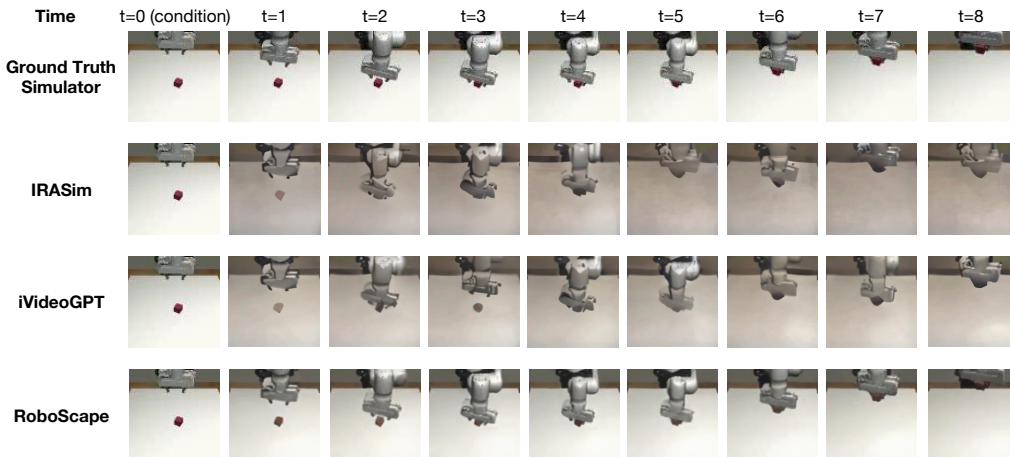


Figure 7: Supplemented visualization results of successful cases in policy evaluation.