PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation

Qiyao Xue, Xiangyu Yin, Boyuan Yang, Wei Gao

University of Pittsburgh

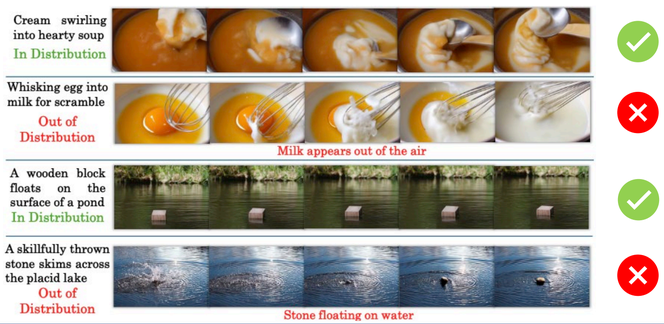
[Paper] SCAN ME

[Code] SCAN ME

Problem Statement

❖ **Physical illusions on text-to-video (T2V) generation**:

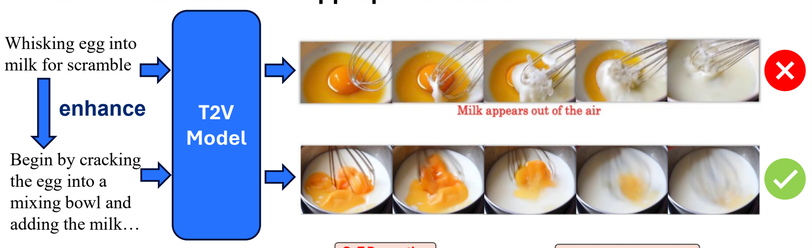
When given **out-of-distribution** prompts, the generated video often contain **physical illusions** or **artifacts**, reflecting the T2V model's limitations in generating realistic and coherent video contents under unfamiliar conditions



Motivation

❖ **Improving the model performance without training efforts**:

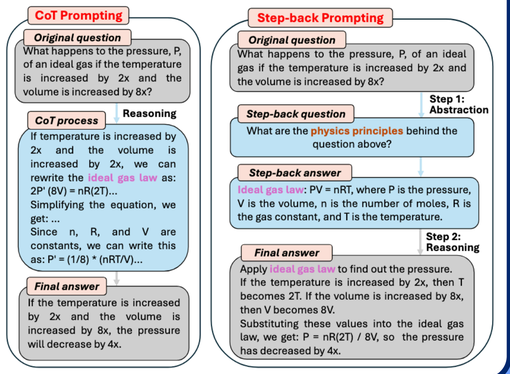
An **out-of-distribution** prompt can be improved by refining the prompt itself with **sufficient** and **appropriate** details



❖ **Reasoning in LLM**:

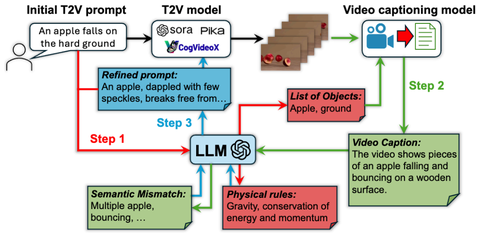
CoT prompting: emphasize **linear decomposition** and step-by-step reasoning

Step-back prompting: derive the step-back question at a **higher level of abstraction** and avoiding confusions and vagueness



PhyT2V Design

❖ **Iterative self-refinement of prompt and generated video**:

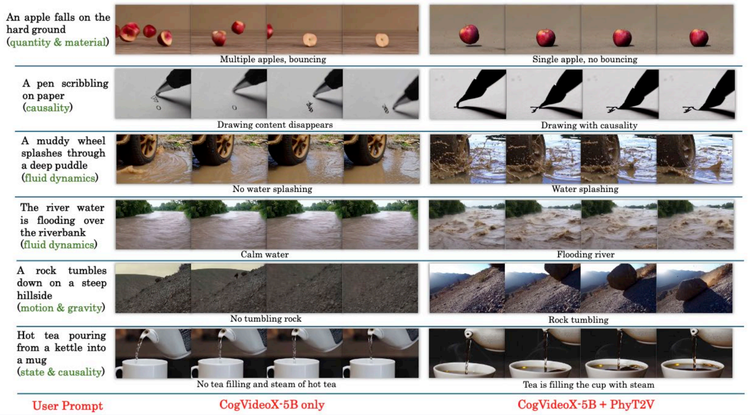


Step 1: Semantic alignment awareness

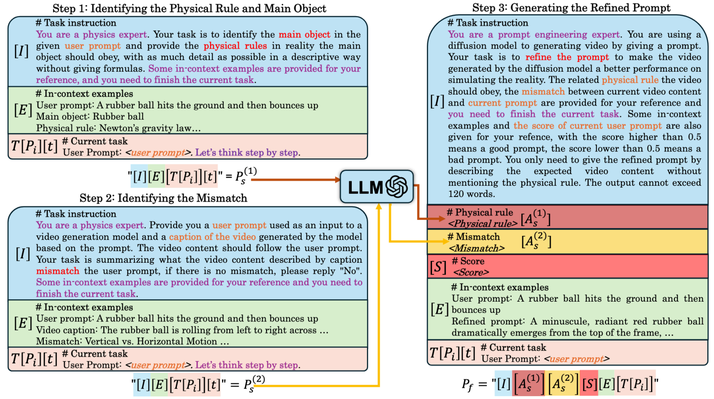
Step 2: Physical knowledge injection

- Feedback based
- Fully automated
- Training-free
- Data independent

❖ **Examples of improved T2V generation**:



Prompt Template for LLM Reasoning

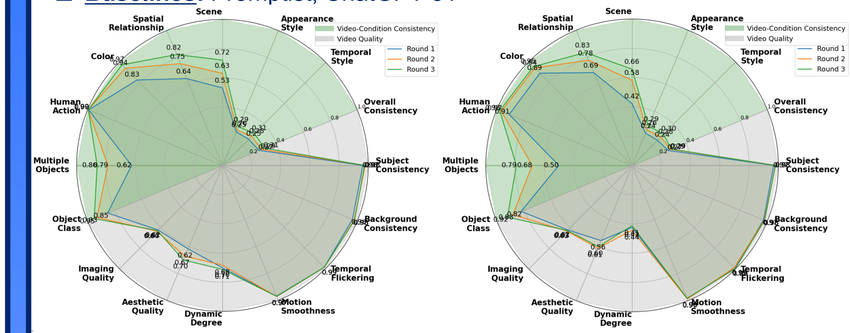


Evaluation Results

❑ **T2V models**: CogVideoX-5B&2B, OpenSora, VideoCrafter

❑ **Datasets**: VBench, VideoPhy, PhyGenBench

❑ **Baselines**: Promptist, ChatGPT-o1



VBench evaluation results with CogVideoX-5B (left) and OpenSora (right)

		CogVideoX-5B				CogVideoX-2B				OpenSora				VideoCrafter			
Round		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Solid-Solid	PC	0.21	0.28	0.34	0.32	0.09	0.13	0.14	0.22	0.12	0.27	0.29	0.30	0.19	0.22	0.27	0.28
	SA	0.24	0.48	0.49	0.47	0.18	0.25	0.36	0.33	0.16	0.34	0.37	0.35	0.24	0.40	0.45	0.47
Solid-Fluid	PC	0.22	0.27	0.28	0.30	0.11	0.18	0.28	0.27	0.17	0.21	0.24	0.25	0.18	0.24	0.25	0.26
	SA	0.39	0.54	0.60	0.61	0.29	0.43	0.44	0.43	0.16	0.40	0.41	0.36	0.34	0.43	0.48	0.52
Fluid-Fluid	PC	0.57	0.59	0.63	0.62	0.34	0.38	0.35	0.36	0.15	0.32	0.29	0.31	0.33	0.41	0.53	0.51
	SA	0.41	0.57	0.59	0.67	0.27	0.42	0.39	0.44	0.31	0.44	0.45	0.46	0.32	0.42	0.49	0.51

Improvement in different categories of physical rules in the VideoPhy dataset

		CogVideoX-5B				CogVideoX-2B				OpenSora				VideoCrafter			
Round		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Mechanics	PC	0.19	0.25	0.34	0.35	0.12	0.16	0.18	0.24	0.11	0.13	0.17	0.22	0.14	0.23	0.29	0.28
	SA	0.21	0.28	0.29	0.32	0.11	0.18	0.19	0.22	0.19	0.21	0.27	0.32	0.20	0.24	0.28	0.35
Optics	PC	0.22	0.35	0.41	0.39	0.22	0.25	0.29	0.28	0.24	0.26	0.25	0.25	0.22	0.21	0.27	0.32
	SA	0.27	0.42	0.39	0.44	0.23	0.34	0.37	0.39	0.26	0.31	0.29	0.30	0.22	0.28	0.35	0.39
Thermal	PC	0.33	0.35	0.35	0.35	0.13	0.15	0.15	0.14	0.27	0.30	0.31	0.33	0.25	0.28	0.26	0.28
	SA	0.22	0.36	0.43	0.45	0.12	0.16	0.24	0.27	0.23	0.25	0.37	0.36	0.25	0.37	0.41	0.43

Improvement in different categories of physical rules in the PhyGenBench dataset

		CogVideoX-5B	OpenSora
ChatGPT 4 [24]	PC	0.33	0.21
	SA	0.41	0.32
Promptist [17]	PC	0.25	0.19
	SA	0.39	0.33

		CogVideoX-5B	OpenSora
ChatGPT 4 [24]	PC	0.27	0.20
	SA	0.23	0.23
Promptist [17]	PC	0.32	0.19
	SA	0.24	0.21

Different prompt enhancers on the VideoPhy(left) and PhyGenBench(right) dataset

Empirical evaluations indicate that PhyT2V achieves a **2.3×** enhancement in physical realism compared to baseline T2V models and outperforms state-of-the-art T2V prompt enhancers by **35%**