

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



Qiyao Xue, Xiangyu Yin, Boyuan Yang, Wei Gao

University of Pittsburgh

PhyT2V Design

Step 1: Semantic alignment awareness Step 2: Physical knowledge injection

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

CogVideoX-5B + PhyT2V

Prompt Template for LLM Reasoning

-8		
rt. Your task is to identify the main object in the d provide the physical rules in reality the main h as much detail as possible in a descriptive way s. Some in-context examples are provided for your to finish the current task.		# Ta You diff You gen sim
ball hits the ground and then bounces up ll gravity law		[I] sho and you exa
$[E][T[P_i]][t]" = P_s^{(1)}$	M®	give mea bad dese men
tifying the Mismatch		120
t Provide you a user prompt used as an input to a	↑[L	→ # Pl <pl< td=""></pl<>
and a caption of the video generated by the model The video content should follow the user prompt. ting what the video content described by caption mpt, if there is no mismatch, please reply "No". les are provided for your reference and you need to		# M < <u>M</u> [S] # S <s # I</s
and a caption of the video generated by the model The video content should follow the user prompt. ting what the video content described by caption ompt, if there is no mismatch, please reply "No". les are provided for your reference and you need to ball hits the ground and then bounces up her ball is rolling from left to right across Horizontal Motion		# M < <u>/</u> [S] # § <

Step 3: Generating the Refined Prompt ask instruction are a prompt engineering expert. You are using a

ision model to generating video by giving a prompt. ir task is to refine the prompt to make the video nerated by the diffusion model a better performance or ulating the reality. The related physical rule the video uld obey, the mismatch between current video conten current prompt are provided for your reference and need to finish the current task. Some in-contex mples and the score of current user prompt are also ven for your refence, with the score higher than 0.5 eans a good prompt, the score lower than 0.5 means a prompt. You only need to give the refined prompt by cribing the expected video content without ntioning the physical rule. The output cannot exceed words.

$\stackrel{\text{\# Physical rule}}{<\!\!Physical rule\!\!>} [A_s^{(1)}]$
[S] #Score <score></score>
 # In-context examples User prompt: A rubber ball hits the ground and then [E] bounces up Refined prompt: A minuscule, radiant red rubber ball dramatically emerges from the top of the frame,
T[P _i][t] #Current task User Prompt: < <u>user prompt</u> >

 $P_f = "[I] \begin{bmatrix} A_s^{(1)} \end{bmatrix} \begin{bmatrix} A_s^{(2)} \end{bmatrix} \begin{bmatrix} S \end{bmatrix} \begin{bmatrix} E \end{bmatrix} \begin{bmatrix} T \begin{bmatrix} P_i \end{bmatrix} \end{bmatrix} "$



[Paper]

Evaluation Results

- □ <u>T2V models</u>: CogVideoX-5B&2B, OpenSora, VideoCrafter
- Datasets: VBench, VideoPhy, PhyGenBench
- **Baselines:** Promptist, ChatGPT-o1



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

		CogVid	eoX-5B		CogVid	eoX-2B		Oper	Sora		Video	Crafter	
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
Sona Sona	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													

improvement in unerent categories of physical rules in the videor

		CogVid	eoX-5B		CogVid	eoX-2B		Open	Sora		Video(Crafter	
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
- F	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
			e	1 1		f			41 I		L-	1 - 1	1

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

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-theart T2V prompt enhancers by 35%





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