CVPR 2026

NEURAI

PhysVid: Physics Aware Local Conditioning for Generative Video Models

Saurabh Pathak

Elahe Arani

Mykola Pechenizkiy

Bahram Zonooz

TU/e

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

PROBLEM

Modern text-to-video models look realistic, but often break basic physics.

WHY IT MATTERS

Lack of physics awareness in T2V generation produces videos far to reality, affecting potential for downstream use cases.

WHAT WE DO DIFFERENTLY?

Existing approaches take a purely global approach, our is local.

KEY CLAIM

With 1.7B parameters, PhysVid surpasses Wan-14B in Physical Consistency on two benchmarks.

A wine bottle pours a red blend into a glass.

A blender spins, mixing squeezed juice within it.

Physical realism failures in text-to-video generation

Looks realistic ≠ behaves realistically

Observation 1

The failure mode is local: short events such as pouring, splashing, sliding, and flow often break first.

Honey pours into a cup of tea.

Observation 2

These examples span multiple physical domains such as fluids, surfaces, optics and motion, so the issue is not tied to one single scene type.

A car glides over a road slick with rainwater.

Indication

We need conditioning that can guide local physical events over time, not only a single global caption.

Water gushes from a green garden hose.

Why global prompts fail

One caption can describe a scene, but not every short-horizon event inside it.

SINGLE GLOBAL PROMPT

"A wine bottle pours a red blend into a glass."

01 stream begins

the liquid stream has just started to come out of the bottle

02 stream hits glass

the stream hits the glass bottom and clearly starts interacting with it

03 swirling in glass

As the stream continues, the liquid inside the glass shows visible swirling dynamics

04 glass filling up

toward the end of the clip, the glass is visibly getting fuller, the fluid less turbulent.

One global caption must describe stream emergence, glass contact, in-glass swirling, and fill-level change over multiple short local sub-sequences.

- The same global text signal is applied across frames.
- It may not cover all physically meaningful local interactions.
- Short local events need more specific guidance.
- That mismatch shows up as physically weak motion.

Core intuition: add local physics per time chunk

Training-time annotation at the chunk level instead of only at the whole-video level.

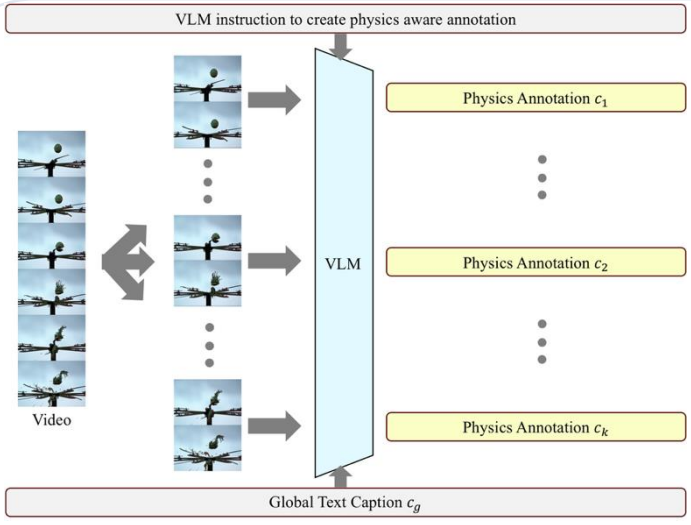
TRAINING-TIME ANNOTATION

- VLM analyzes one short chunk at a time.
- Prompts stay grounded in visible elements only.
- Focus stays on dynamics, shape, and optics.

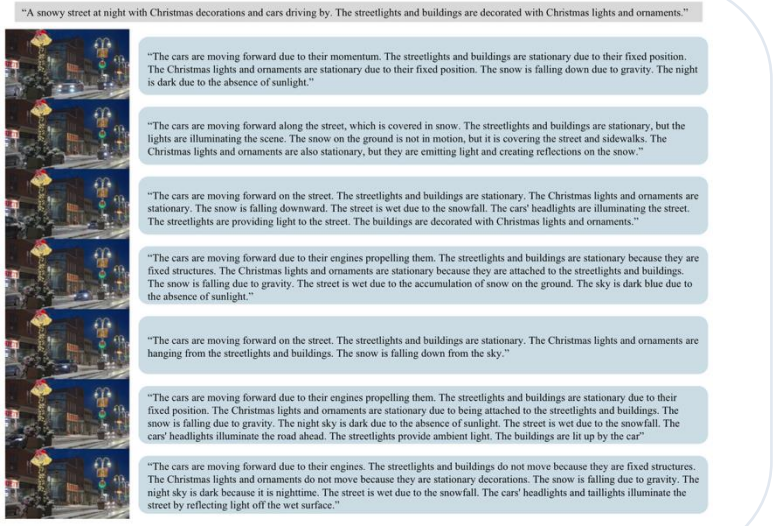
Split the video into short contiguous chunks.

Describe visible physics inside each chunk.

Condition generation with global + local text.



Each chunk is annotated with local physics-aware text.

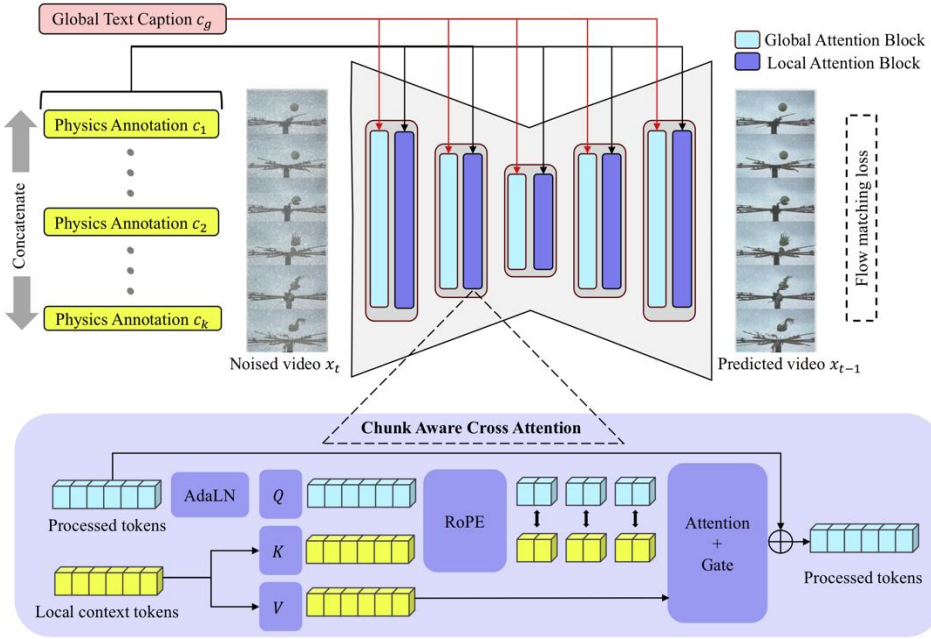


Example chunk-level prompts extracted from a training video.

Model architecture

Chunk-aware architecture

- Base: pretrained Wan-1.3B backbone
- Addition: local cross-attention blocks per transformer block
- Alignment: RoPE couples text chunks with video chunks
- Effect: different moments can attend to different local physics cues



Global pathway stays intact while local chunk-aware cross-attention is added.

Inference-time guidance with counterfactuals

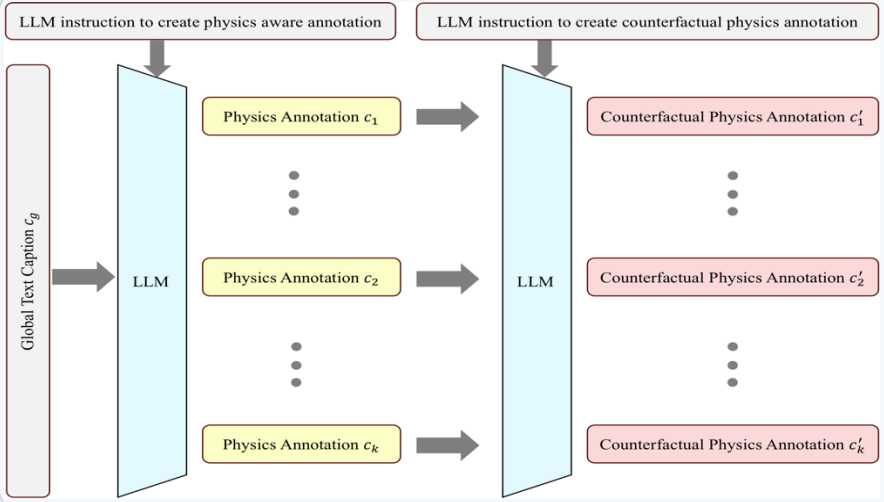
INFERENCE-TIME GUIDANCE

- Imagined locals: local prompts are synthesized from the global caption.
- Counterfactuals: wrong local behaviors are written explicitly.
- Guidance: sampling is pushed toward plausible motion and away from implausible motion.

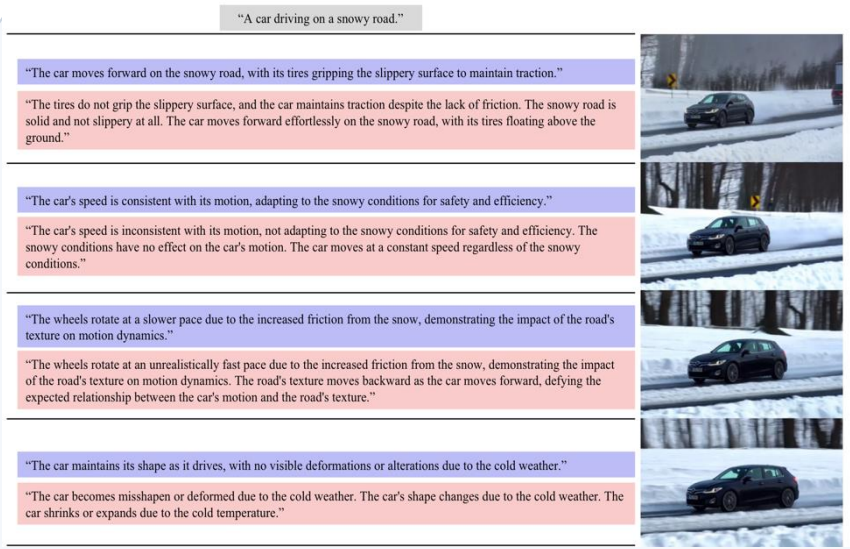
Derive local positive prompts from the caption alone.

Pair them with physically wrong counterfactual alternatives.

Use guidance to reward plausible dynamics and suppress failure modes.



Positive local prompts are paired with counterfactual wrong alternatives.



Positive and counterfactual prompt pairs for a single caption. Frames from the corresponding generated video samples on the right.

Setup

BACKBONE

Wan-1.3B

- 81 frames at 16 fps
- 832 × 480 resolution
- Full parameter finetuning

ADDED PARAMETERS

400M

TRAINING DATA

WISA-80k

- 5.06 second clips
- 53,000 videos
- 3000 training steps
- Effective batch size: 64

ANNOTATION

VideoLLama3-7B

TEMPORAL CHUNKS

7 chunks / clip

- 3 latent frames / chunk

BENCHMARKS

VideoPhy

VideoPhy2

Main quantitative results

≈ +33% VideoPhy PC

> +8% VideoPhy2 PC vs Wan-14B

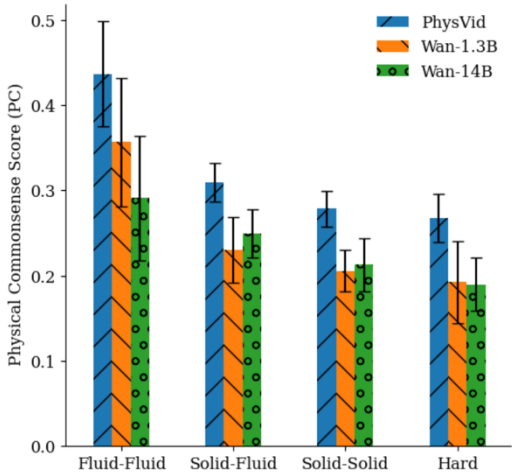
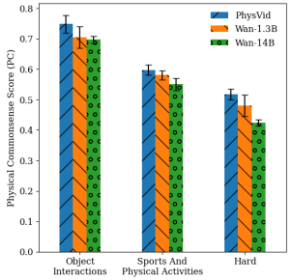


Figure 5. VideoPhy Physical Commonsense (PC) score by category.

Category-level physical commonsense improvements.



VideoPhy2 histogram: PhysVid stays ahead of the much larger Wan-14B baseline.

MAIN RESULTS (TABLE 1)

Model	Params	VideoPhy		VideoPhy2	
		SA	PC	SA	PC
Wan-1.3B	1.3B	0.46	0.24	0.28	0.61
Wan-14B	14B	**0.52**	0.24	**0.29**	0.59
PhysVid	1.7B	0.43	**0.32**	0.28	**0.64**

PhysVid leads the physical-commonsense columns while using only 1.7B parameters.

PAIRED METRICS · 2048 SAMPLE PAIRS (TABLE 4)

Minor fidelity trade-off, major physics gain.

Perceptual metrics stay close to the finetuned baseline while physical realism improves much more strongly.

Model	LPIPS ↓	FVD ↓	SSIM ↑	PSNR ↑
Wan-1.3B	0.703	417.352	0.217	8.625
Wan (finetuned)	**0.671**	**302.465**	0.239	**9.379**
PhysVid	0.679	318.087	**0.240**	9.234

Ablations

Local prompts > fine-tuning

Counterfactual guidance adds another boost

KEY POINTS

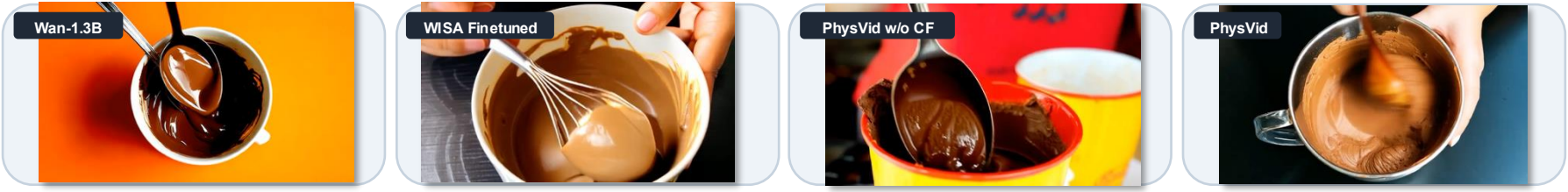
- Fine-tuning helps: VideoPhy PC rises from 0.2401 to 0.2866.
- Local prompts help again: without counterfactuals, PhysVid reaches 0.2924 / 0.6334.
- Full guidance wins: PhysVid reaches 0.3169 on VideoPhy PC and 0.6411 on VideoPhy2 PC.

ABLATIONS (TABLE 3)

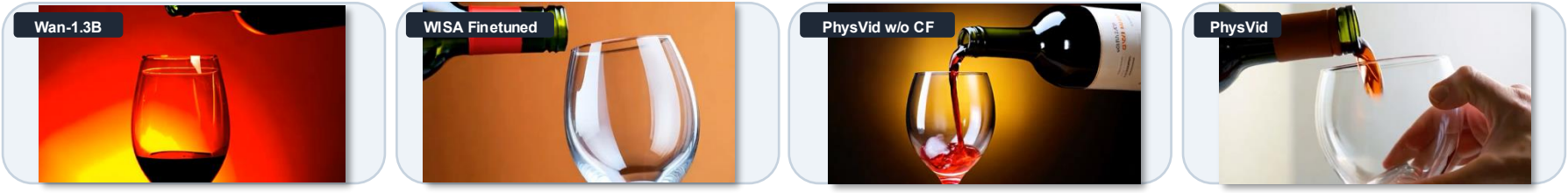
Method	VideoPhy		VideoPhy2	
	SA	PC	SA	PC
Baseline	0.4570	0.2401	0.2845	0.6144
Fine-tuning	0.4174	0.2866	0.2765	0.6261
PhysVid w/o CF	0.4355	0.2924	0.2791	0.6334
PhysVid	0.4302	**0.3169**	0.2775	**0.6411**

Each added ingredient improves physical commonsense, with the full model strongest on both benchmarks.

PROMPT A A mixing spoon stirring hot chocolate in a cup.



PROMPT B A wine bottle pours a red blend into a glass.



Comparison with prior work

KEY POINTS

- Compact model: PhysVid-1.7B stays competitive on VideoPhy while using far fewer parameters than many compared systems.
- Main in-house gain: relative to its baselines, PhysVid improves VideoPhy PC by **+33%,** which is highest gain among non-iterative methods.
- Mixed sources: symbols follow the respective reporting paper's original evaluation sources and reporting protocol.

General & baseline methods

Method	Reported in	SA	PC
Lavie	‡	0.49	0.28
Lavie	$	0.49	0.32
VideoCrafter2	†	0.47	0.36
VideoCrafter2	‡	0.49	0.35
VideoCrafter2	$	0.50	0.30
VideoCrafter2	~	0.24	0.15
OpenSora	#	0.38	0.43
OpenSora	‡	0.18	0.24
OpenSora	~	0.29	0.17
HunYuanVideo	†	0.46	0.28
HunYuanVideo	$	0.60	0.28
Cosmos-7B	†	0.57	0.18
Cosmos-7B	#	0.52	0.27
CogVideoX-2B	‡	0.47	0.34
CogVideoX-2B	$	0.52	0.26
CogVideoX-2B	~	0.22	0.13
CogVideoX-5B	†	0.60	0.33
CogVideoX-5B	‡	**0.63**	**0.53**
CogVideoX-5B	$	**0.63**	0.31
CogVideoX-5B	#	0.48	0.39
CogVideoX-5B	~	0.48	0.26
Wan2.1-1.3B	*	0.46	0.24
Wan2.1-14B	#	0.49	0.35
Wan2.1-14B	*	0.52	0.24

Physics-aware approaches

Method	SA	PC
PhyT2V [56]	0.59 (+23%)	0.42 **(+62%)**
PhyT2V †	0.61 (+2%)	0.37 (+12%)
WISA [49]	0.67 (+12%)	0.38 (+15%)
VideoREPA-5B [66]	0.72 **(+14%)**	0.40 (+29%)
Hao et al. [20]	0.49 (+0%)	0.40 (+14%)
PhysVid-1.7B	0.43 (−7%)	0.32 **(+33%)**

SYMBOL LEGEND

† WISA, 2025
$ VideoREPA, 2025
# Hao et al., 2025
~ PhyT2V, 2024
‡ VideoPhy,2025
* our baselines

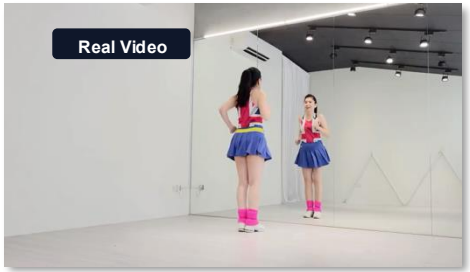
Conclusion

Local physics-aware guidance improves generative video realism

Why PhysVid?

- Scaling generative video models is expensive and still suffers from hard physics failures.

- PhysVid injects local conditioning and applies counterfactual physics guidance during inference, allowing it to significantly improve physical consistency without adding proportional scale.



TU/e

NEURAI

Please visit our project page

5aurabhpathak.github.io/PhysVid

THANKS

Arxiv paper

2603.26285

Contact

s.pathak@tue.nl

Acknowledgments

TU/e SCC

SYNERGIES

Funded by the European Union