Neuro-Symbolic Evaluation of Text-to-Video Models using Formal Verification

S P Sharan* Minkyu Choi* Sahil Shah Harsh Goel Mohammad Omama Sandeep Chinchali

The University of Texas at Austin, United States

Why Do We Care About Temporal Alignment? 💡

Synthetic videos are increasingly used to train systems like motion planners in AD. But if they fail to follow the prompt's temporal order, they can mislead downstream systems, compromising safety.

"A truck appears in frame 10 and veers in front of me onto my lane within 2 seconds."

(a) Autonomous Driving (b) Healthcare/Medical AI (c) Embodied Robotics/HRI

Figure 1. Misaligned videos can lead to unsafe behavior in high-stakes safety-critical settings.

NeuS-V–At a Glance ⚙️

Most T2V metrics focus on visual quality/semantic coverage. But they miss temporal fidelity, rely on black-box eval, and offer limited interpretability.

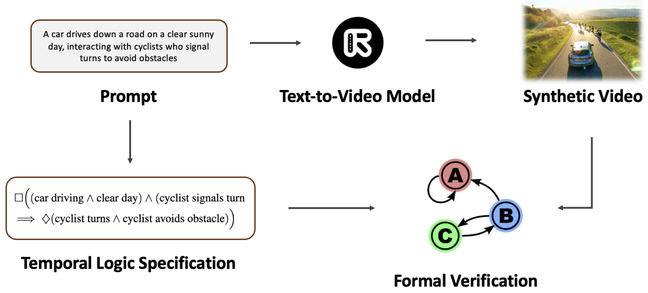
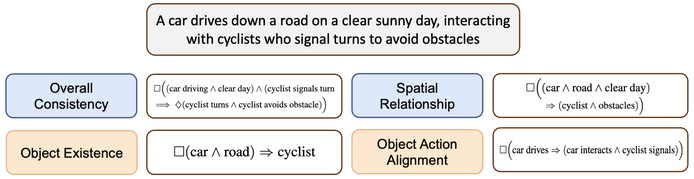


Figure 2. NeuS-V changes that with symbolic reasoning and structured temporal verification 📈.

Semantics + Temporal + Spatio → Video 🎥

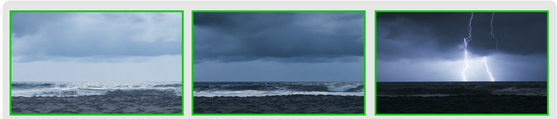
NeuS-V evaluates videos across four distinct modes to capture spatio-temporal semantics.



Key Takeaway 📖

Ocean waves gently lapping at the shore, suddenly a storm bursts in, and then lightning flashes across the sky

Prompt

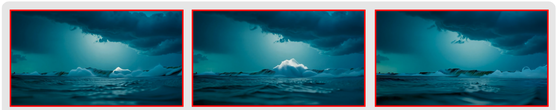


Video Frames (Temporally Aligned with Prompt)

VBench 0.707 ✓

NeuS-V (ours) 0.670 ✓

Video Score



Video Frames (Temporally Misaligned with Prompt)

VBench 0.752 ✗

NeuS-V (ours) 0.371 ✓

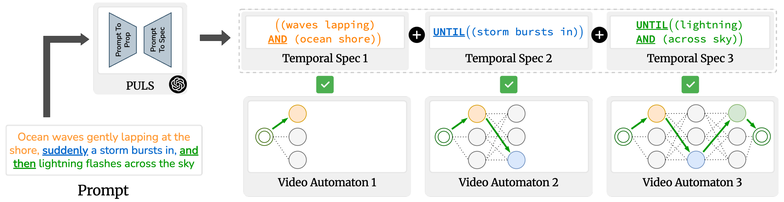
Video Score

Most T2V metrics evaluate appearance

We evaluate temporal alignment

Under the Hood–Construction of the Automata 🛠️

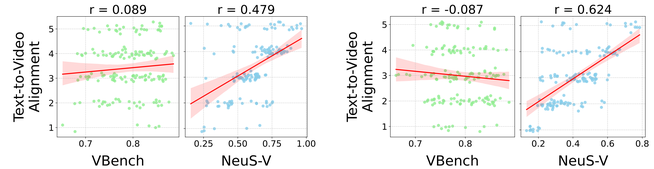
Prompts are decomposed into atomic propositions and structured into Temporal Logic (TL) formulas



Frame-windows are scored for semantics by a VLM and assembled into a probabilistic automaton over time. Each node represents frame-level semantics; colored nodes denote detected events, and green paths satisfy the spec.

Website 🌐 Paper 📜 HF Demo 🤗 GitHub 🚀

How well do we correlate with humans? 🎯



(a) Runway Gen-3 (b) Pika

Figure 4. Correlation with Human Annotators (Pearson coefficients at the top of each plot)

NeuS-V achieves up to 5× higher correlation ✨ with human ratings compared to VBench

Benchmarking Text-to-Video Models 💪

Our 360-prompt benchmark spans 4 semantic themes and 3 complexity levels.

	Prompts	Gen-3	Pika	T2V-Turbo-v2	CogVideoX-5B
By Theme	Nature	0.716 (0.47)	0.479 (0.70)	0.564 (0.46)	0.580 (0.53)
	Human & Animal Activities	0.752 (0.80)	0.531 (0.67)	0.564 (0.66)	0.623 (0.43)
	Object Interactions	0.710 (0.16)	0.500 (0.40)	0.553 (0.66)	0.573 (0.65)
	Driving Data	0.716 (0.48)	0.525 (0.66)	0.525 (0.30)	0.580 (0.52)
By Complexity	Basic (1 TL op.)	0.774 (0.60)	0.589 (0.70)	0.610 (0.58)	0.641 (0.65)
	Intermediate (2 TL ops.)	0.680 (0.27)	0.464 (0.44)	0.508 (0.38)	0.549 (0.28)
	Advanced (3 TL ops.)	0.692 (-0.01)	0.400 (0.33)	0.494 (0.42)	0.550 (0.78)
	Overall Score	0.723 (0.48)	0.508 (0.62)	0.552 (0.55)	0.589 (0.54)

Table 1. Gen-3 achieves the highest score across our suite. *NeuS-V* enjoys high correlation across all themes and complexities.

An Exciting (Upcoming) Follow-up! 🤩

Can this metric serve as feedback for zero-training refinement of black-box T2V models? YES!!! 🙀

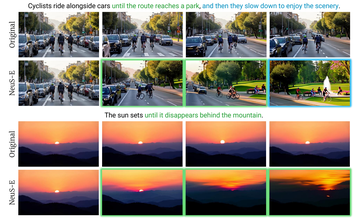


Figure 5. We formally identify the problematic video segments and surgically refine only those parts with targeted edits.

	Prompts	Pika-2.2	
		Original	Edited
By Theme	Nature	0.579	0.856 (+0.277)
	Human & Animal Activities	0.638	0.872 (+0.235)
	Object Interactions	0.420	0.707 (+0.287)
	Driving Data	0.676	0.810 (+0.134)
By Complexity	Basic (1 TL op.)	0.694	0.840 (+0.146)
	Intermediate (2 TL ops.)	0.480	0.795 (+0.315)
	Advanced (3 TL ops.)	0.373	0.729 (+0.356)
	Overall- Score	0.577	0.811 (+0.233)

Table 2. Comparison of Pika-2.2's original and edited videos—improvement is more prominent in complex prompts.

These scores guide video edits in our follow-up, improving temporal fidelity by up to 40%