ViRuleEval: Interpretable Neuro-Symbolic Evaluation of Text-to-Video Generative Models

Chufeng Jiang[1] Heng Li[1] Neng-Fa Zhou[1,2]

[1]City University of New York (CUNY) Graduate Center, USA [2]City University of New York (CUNY) Brooklyn College, USA

Motivation

Text-to-video (T2V) generative models have made rapid progress in visual realism and semantic alignment. However, evaluating their **temporal quality** remains challenging. Existing benchmarks, e.g., VBench (Han et al., 2025), primarily provide human annotated scores, offering limited insight into *why* models succeed or fail.

Key Problem

Current evaluation methods:

- Lack interpretability
- Provide limited diagnostic feedback
- Cannot explain temporal failure patterns

Our Idea

We propose **ViRuleEval**, an interpretable neuro-symbolic framework that:

- Learns explicit logical rules for temporal quality evaluation using FOLD-RM algorithm (Wang et al., 2022).
- Replaces black-box scores with transparent decision rules.
- Enables diagnostic analysis of generative model behavior.

From Features to Concepts

To improve interpretability, we further apply **prompt-based semantic labeling**:

- Map learned feature predicates to human concepts;
- Convert rules into human-readable explanations.

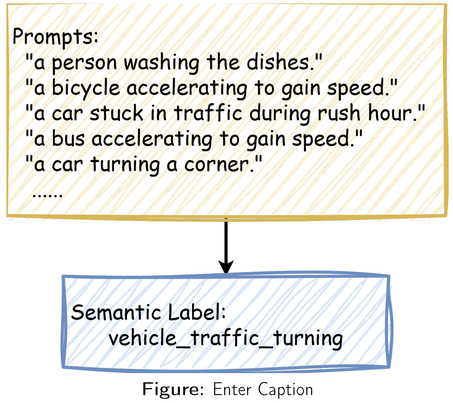

Figure: Enter Caption

```
label(X,poor) ?=> rule1(X).
label(X,good) ?=> rule2(X), not rule1(X).
label(X,good) ?=> rule3(X), not rule1(X), not ab2(X).
.....
rule1(X) ?=> vehicle_animal_accelerating(X,0), not ab1(X).
.....
ab1(X) ?=> animal_water_scene_drinking(X,0).
.....
```

Framework Overview

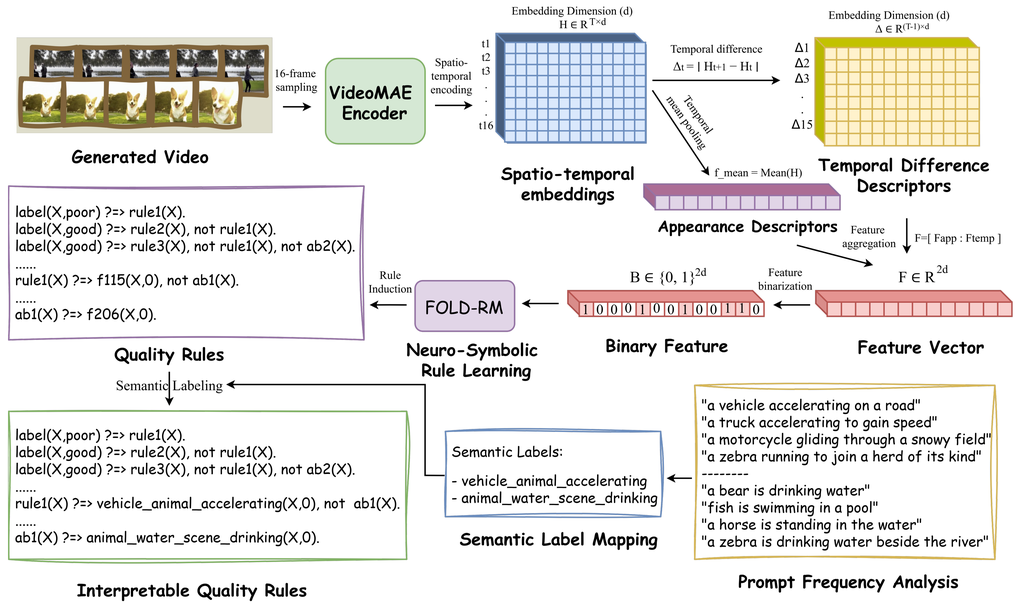

Figure: Overview of **ViRuleEval**, an interpretable neuro-symbolic framework for analyzing temporal quality in generated videos.

Early-Stage Results

Table: Comparison of ViRuleEval with baseline classifiers across four VBench temporal quality dimensions. Macro-F1 and per-class recall are reported on the held-out test set. † indicates interpretable methods.

Dimension	Method	Acc	Macro-F1	Good Recall	Poor Recall
Motion Smoothness	Majority Class	**0.761**	0.432	0.000	**1.000**
	Logistic Regression	0.669	0.615	**0.615**	0.686
	Random Forest	0.735	0.601	0.323	**0.865**
	MLP	**0.761**	0.680	0.538	0.831
	ViRuleEval (ours)†	0.750	**0.682**	0.600	0.797
Subject Consistency	Majority Class	0.621	0.383	0.000	**1.000**
	Logistic Regression	0.654	0.644	**0.641**	0.663
	Random Forest	0.669	0.648	0.563	0.734
	MLP	0.654	0.641	0.612	0.680
	ViRuleEval (ours)†	**0.710**	**0.693**	0.631	**0.757**
Dynamics Degree	Majority Class	**0.831**	0.454	0.000	**1.000**
	Logistic Regression	0.746	0.581	0.348	0.827
	Random Forest	0.824	0.507	0.065	**0.978**
	MLP	0.816	0.557	0.152	0.951
	ViRuleEval (ours)†	0.783	**0.630**	**0.413**	0.858
Temporal Flickering	Majority Class	0.592	0.372	0.000	**1.000**
	Logistic Regression	0.810	0.804	0.775	0.834
	Random Forest	0.835	0.830	0.806	0.856
	MLP	**0.851**	**0.846**	0.814	**0.877**
	ViRuleEval (ours)†	0.807	0.803	**0.814**	0.802

Cross-Dimension Analysis

A shared feature may behave differently across temporal quality dimensions.

Representative prompts include:

```
"a bird soaring gracefully in the sky."
"an airplane accelerating to gain speed."
"an airplane taking off."
"an airplane accelerating to gain speed."
"an airplane soaring through a clear blue sky."
```

Example: `vehicle_accelerating`

- → Improves **motion smoothness** (when inactive)
- → Degrades **subject consistency** (when active)

This reflects:

- Different evaluation criteria across dimensions
- Different feature activation regimes

Interpretation. Scenes with strong acceleration (e.g., airplanes, birds):

- ✓ Smooth motion trajectories
- × Hard to maintain object consistency

Takeaway. Temporal quality dimensions capture different aspects of video realism.

Work in Progress

This project is an **ongoing research effort**. The current results are:

- Preliminary experiments on selected VBench dimensions.
- Early-stage rule learning pipeline.
- Initial validation of interpretability-performance tradeoff.

Future work will expand datasets, improve feature learning, and refine symbolic reasoning.

Acknowledgment

This research was supported by the 2026 ERI/PS2 Summer Public Research Fellowship through the Early Research Initiative at the CUNY Graduate Center. The author would also like to thank Dr. Feng Gu from the Department of Computer Science at The College of Staten Island, CUNY, for valuable support.

References

Han, Hui, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. (2025). "Video-bench: Human-aligned video generation benchmark". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868.

Wang, Huaduo, Farhad Shakerin, and Gopal Gupta (2022). "FOLD-RM: a scalable, efficient, and explainable inductive learning algorithm for multi-category classification of mixed data". In: *Theory and Practice of Logic Programming* 22.5, pp. 658–677.