



Robust Video Diffusion via Structured Corruption

Chika Maduabuchi, Hao Chen, Yujin Han, Jindong Wang



CMU

The Fragility of Latent Video Diffusion

Problem:

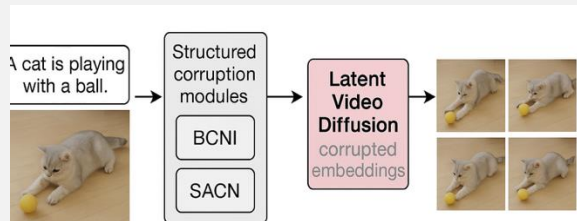
Latent video diffusion assumes perfectly clean text prompts or embeddings, yet actual captions are noisy.

Effect:

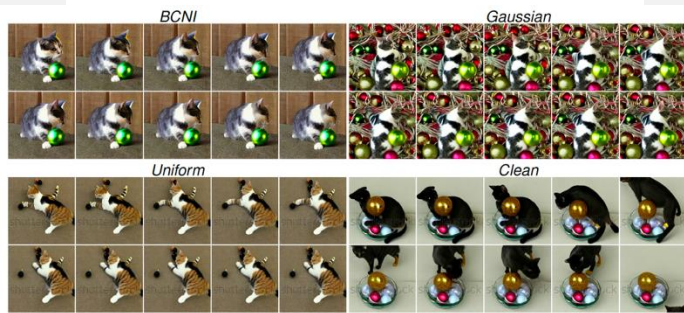
Structured corruption improves robustness, outperforming clean training across all metrics.

Fix:

Train with structured batch-centered noise (BCNI) to map noisy inputs to clean manifolds on WebVid, MSRVT, and MSVD. For short prompts, use spectrum-aware contextual noise (SACN) on the UCF-101 dataset.



CAT-LVDM improves video generation robustness via structured corruption-aware training.



Codes and pre-trained models are at
<https://github.com/chikap421/catlvdm>

Quantitative Results

BCNI & SACN beat all benchmarks

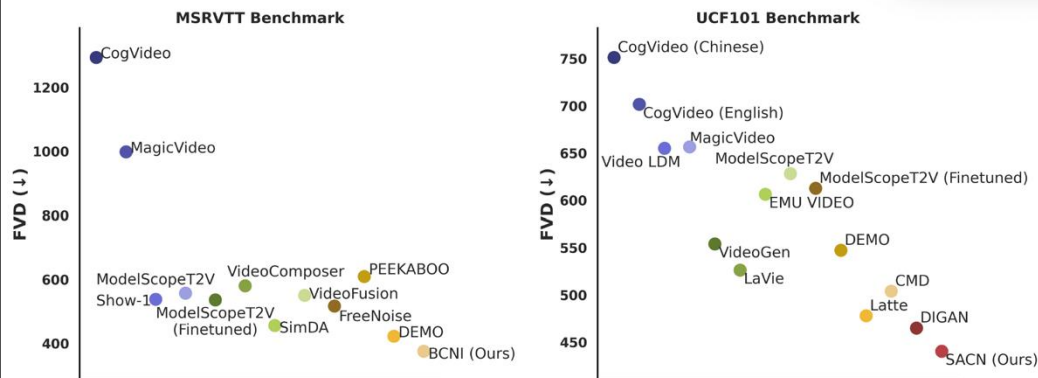
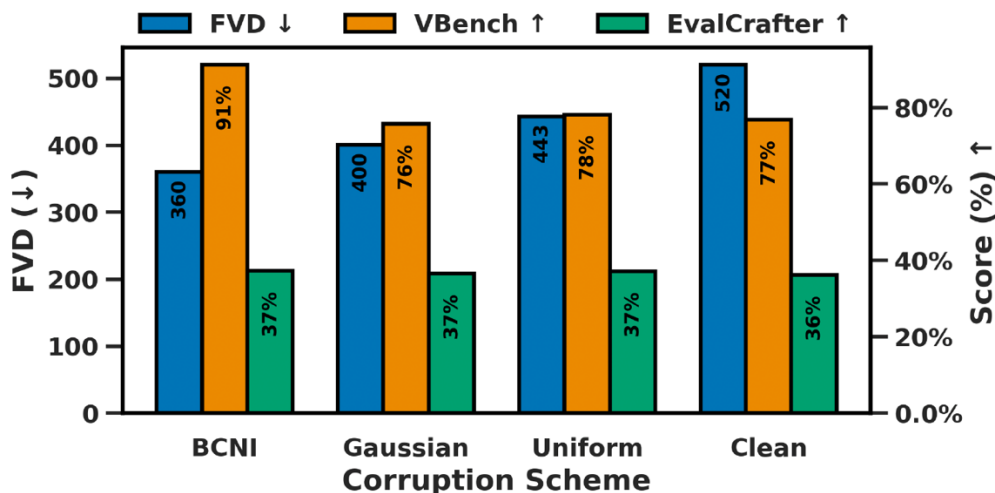


Table 1: **Model-Dataset Evaluations.** FVD comparisons across noise ratios. Lower is better.

Noise ratio (%)	WebVid-2M			MSRVT			MSVD			UCF101		
	BCNI	Gaussian	Uniform	BCNI	Gaussian	Uniform	BCNI	Gaussian	Uniform	SACN	Gaussian	Uniform
2.5	521.24	506.56	522.36	539.93	595.08	541.80	587.59	654.73	575.76	440.28	674.62	651.64
5	502.45	572.67	443.22	564.00	664.45	543.46	599.44	740.79	580.59	480.29	659.27	599.53
7.5	360.32	441.69	574.35	441.31	468.79	639.83	374.34	485.30	695.59	504.89	648.41	742.18
10	378.87	417.60	444.71	414.49	445.29	526.85	374.52	452.82	551.99	455.65	615.28	607.23
15	475.01	400.29	525.22	515.12	464.91	605.27	610.38	458.69	662.51	446.78	672.25	643.22
20	456.14	451.67	454.79	396.35	565.83	559.93	504.35	479.63	550.73	526.23	677.13	642.74
Uncorrupted	520.32			543.33			602.39			501.91		

Qualitative Results

Prompt: Rotation, close-up, falling drops of water on ripe cucumbers.

Observation: BCNI preserves fine water and cucumber detail and motion coherence better than Gaussian or Uncorrupted baselines.

