Dysen-VDM: Empowering Dynamics-aware Text-to-Video Diffusion with LLMs

Hao Fei[1], Shengqiong Wu[1], Wei Ji[1], Hanwang Zhang[2,3], and Tat-Seng Chua[1] 1. NUS 2. NTU. 3. Skywork AI

CVPR SEATTLE, WA JUNE 17-21, 2024

Project: https://haofei.vip/Dysen-VDM/

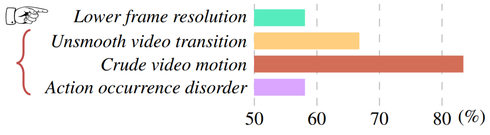
Paper: https://arxiv.org/abs/2308.13812

Code: https://github.com/scofield7419/Dysen

Background

➢ Common issues in Existing Text-to-Video (T2V) Diffusion

Easily solved issue:
resolution
Insufficient modeling of video temporal dynamics



➢ How we humans create a film from a given instruction?

☞ *We always first extract the key actions from the instruction into an event playlist with time order.*

☞ *We then enrich the simple events with more possible specific scenes, i.e., with our imagination.*

➢ Key key points of effective T2V modeling:

First, *sequential language mentions a set of movements that may not necessarily coincide with the physical order of occurrence, it is thus pivotal to properly organize the semantic chronological order of events.*

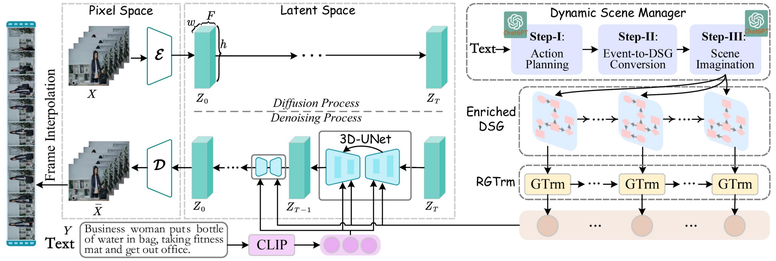
Second, *as prompt texts would not cover all action scenes, reasonable enrichment of video scenes is indispensable to produce delicate videos with detailed movements.*

Third, *the above processes should be carried out based on effective representations of structured semantics, to maintain the imagination of high-controllable dynamic scenes.*

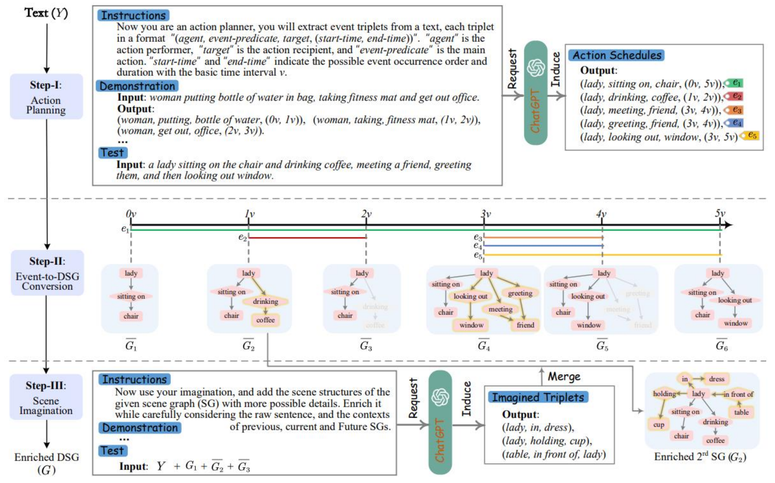
Finally, *fine-grained spatiotemporal features modeling should be realized for temporally coherent video generation.*

Methodology

➢ Overall Framework



➢ T2V Diffusion with Dynamic Scene Manager (Dysen)



Dysen carries out 3 steps to obtain enriched dynamic scene graph representations:

1) action planning
2) event-to-DSG conversion
3) scene imagination

Experiment

Table 1. Zero-shot results on UCF-101 and MSR-VTT data. The results of baselines are copied from their raw paper. The best scores are marked in bold.

Method	UCF-101		MSR-VTT	
	IS (↑)	FVD (↓)	FID (↓)	CLIPSIM (↑)
CogVideo [24]	25.27	701.59	23.59	0.2631
MagicVideo [91]	/	699.00	/	/
MakeVideo [55]	33.00	367.23	13.17	0.3049
AlignLatent [5]	33.45	550.61	/	0.2929
Latent-VDM [52]	/	/	14.25	0.2756
Latent-Shift [2]	/	/	15.23	0.2773
VideoFactory [70]	/	410.00	/	0.3005
InternVid [73]	21.04	616.51	/	0.2951
Dysen-VDM	**35.57**	**325.42**	**12.64**	**0.3204**

Table 2. Fine-tuning results on UCF-101 without pre-taining.

Method	IS (↑)	FVD (↓)
VideoGPT [82]	24.69	/
TGANv2 [53]	26.60	/
DIGAN [86]	32.70	577±22
MoCoGAN-HD [61]	33.95	700±24
VDM [23]	57.80	/
LVDM [18]	27.00	372±11
TATS [11]	79.28	278±11
PVDM [85]	74.40	343.60
ED-T2V [37]	83.36	320.00
VideoGen [33]	82.78	345.00
Latent-VDM [52]	90.74	358.34
Latent-Shift [2]	92.72	360.04
Dysen-VDM	**95.23**	**255.42**

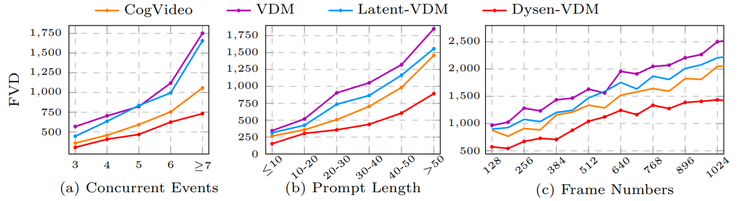


Figure 5: Performance on the action-complex scene video generation of ActivityNet data.

Table 3. Human evaluation on ActivityNet data.

	Action Faithfulness	Scene Richness	Movement Fluency
CogVideo [24]	67.5	75.0	81.5
VDM [23]	62.4	58.8	46.8
Latent-VDM [52]	70.7	66.7	60.1
Dysen-VDM	**86.6**	**92.4**	**87.3**

Table 4. Model ablation (fine-tuned results in FVD). 'w/o Dysen': degrading our system into the Latent-VDM model.

Item	UCF-101	ActivityNet
Dysen-VDM	**255.42**	**485.48**
w/o Dysen	346.40 (+90.98)	627.30 (+141.82)
w/o Scene Imagin.	332.92 (+77.50)	597.83 (+112.35)
w/o SWC	292.16 (+36.74)	533.22 (+47.74)
w/o RL-based ICL	319.01 (+63.59)	520.76 (+35.28)
RGTrm→RGNN [44]	299.44 (+44.02)	564.16 (+78.68)

Figure 6. Qualitative results on video generation with two pieces of examples. Visit the live demos at http://haofei.vip/Dysen-VDM/ for more cases.