Dysen-VDM

Empowering Dynamics-aware Text-to-Video Diffusion with LLMs

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

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

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

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

Motivation

Existing Diffusion-based Text-to-Video (T2V) Generation

➢ Common issues

- ~~*Lower frame resolution*~~ → **Easily solved**
- *Unsmooth video transition*
- *Crude video motion*
- *Action occurrence disorder*

→ **Insufficient modeling of video temporal dynamics**

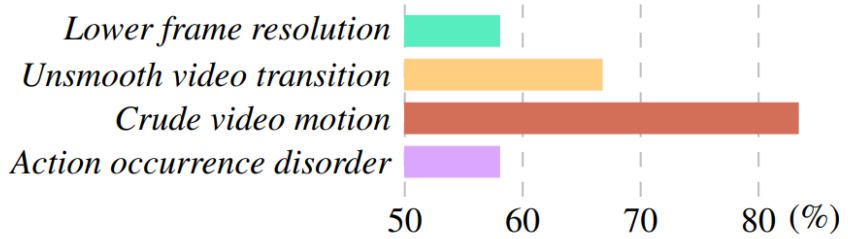


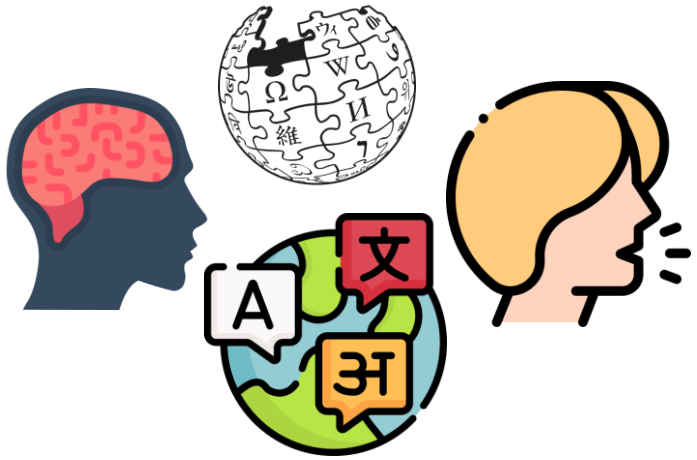
Figure 1: Common issues in the existing text-to-video (T2V) synthesis. We run the video diffusion model (VDM) [21] with random 100 prompts, and ask different users to summarize the problems.

✓ Real crux of high-quality video synthesis: modeling the intricate **video temporal dynamics**

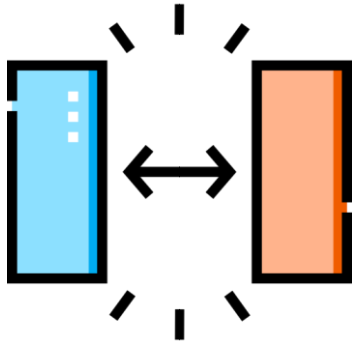
Motivation

The Gap between Language and Vision

Language



abstract & concise



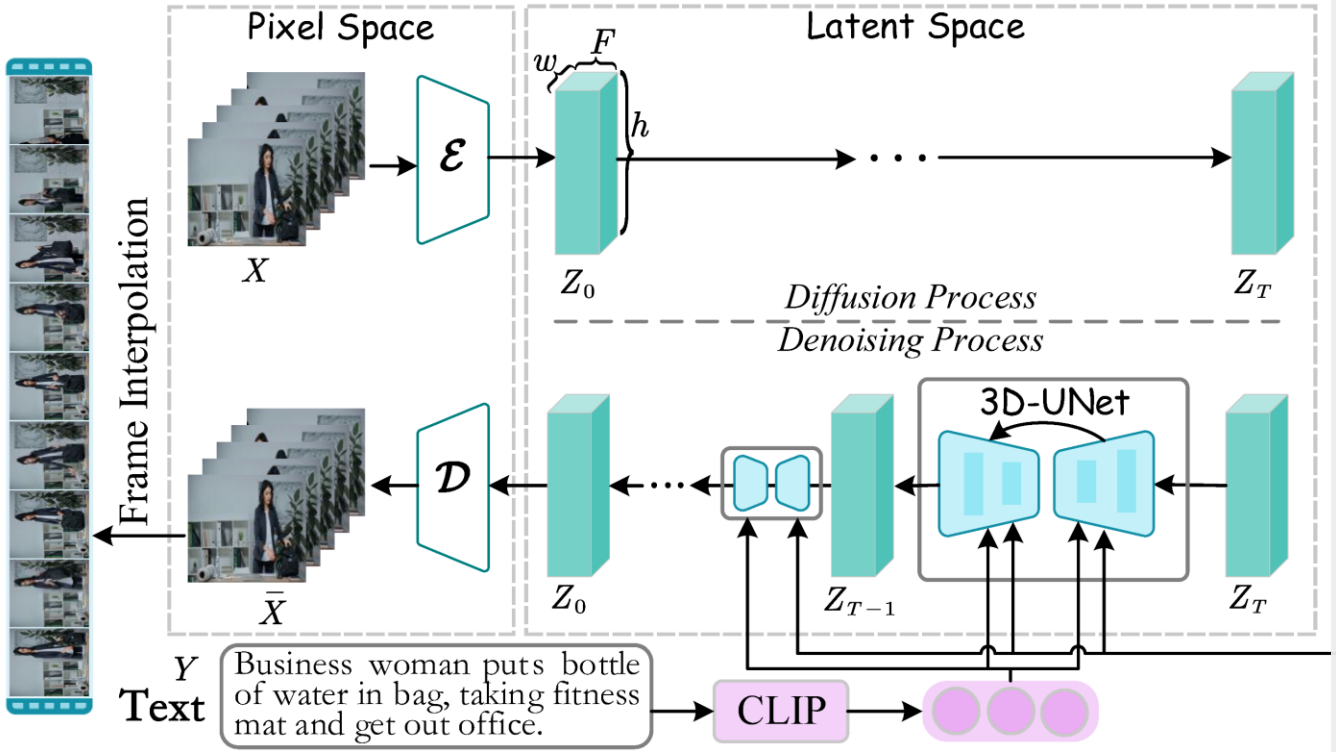
Gap to bridge

Vision

intricate & specific

Method

T2V Diffusion with Dynamic Scene Manager (Dysen)



Method

T2V Diffusion with Dynamic Scene Manager (Dysen)

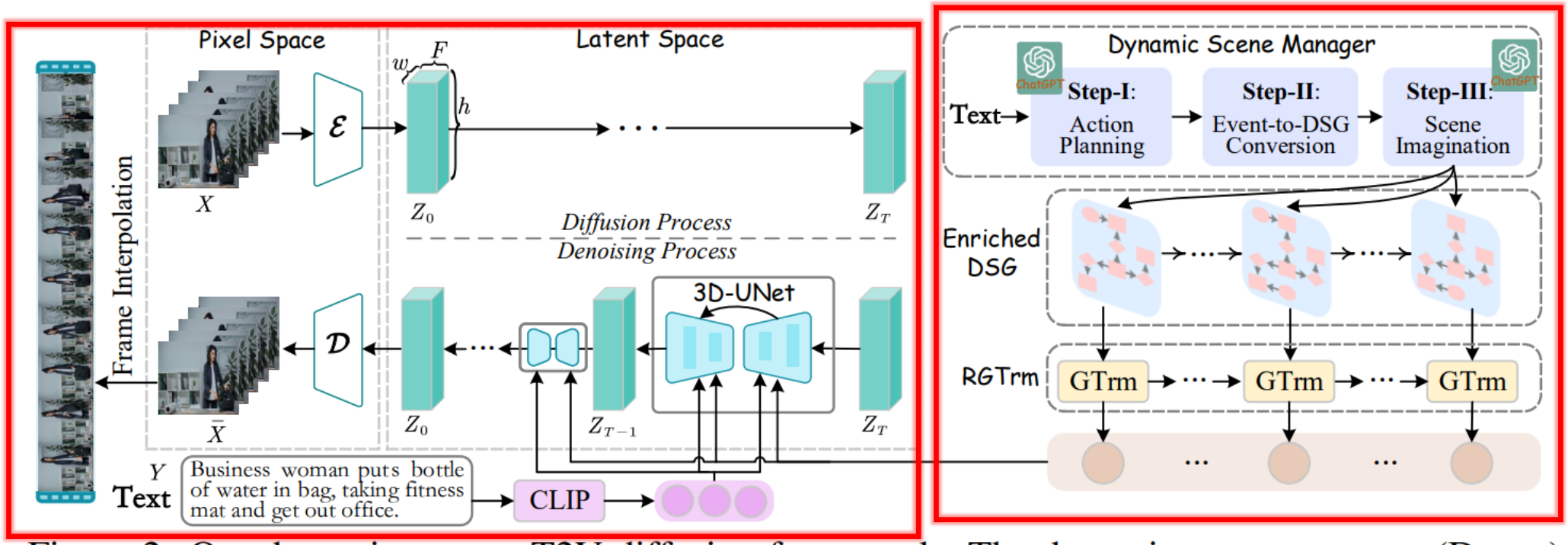


Figure 2: Our dynamics-aware T2V diffusion framework. The dynamic scene manager (Dysen) module operates over the input text prompt and produces the enriched dynamic scene graph (DSG), which is encoded by the recurrent graph Transformer (RGTrm), and the resulting fine-grained spatio-temporal scene features are integrated into the video generation (denoising) process.

Method

Dynamic Scene Manager (Dysen)

- Dynamic Scene Graph (DSG) Representation
  - Visual Scene Graph (VSG): *Representing visual content into semantic structured representation*
    - **Object Nodes:**

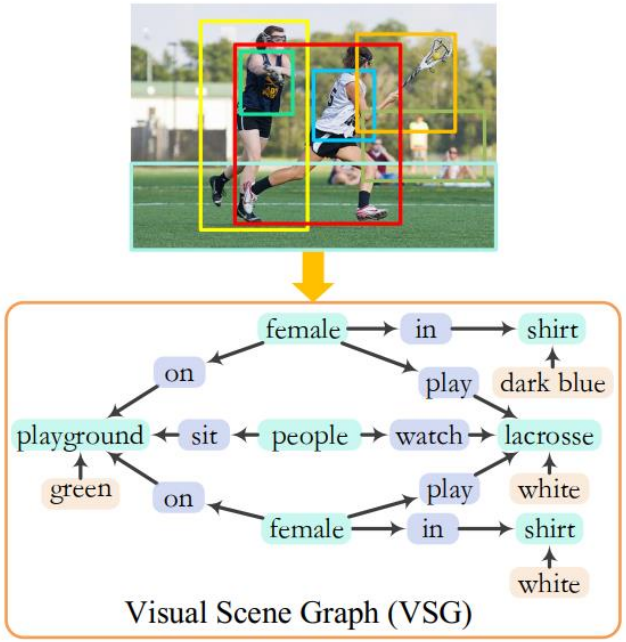
      Visually-seen entity objects

    - **Relation Nodes:**

      describing the semantic relations between objects

    - **Attribute Nodes**

      depicting the objects



Visual Scene Graph (VSG)

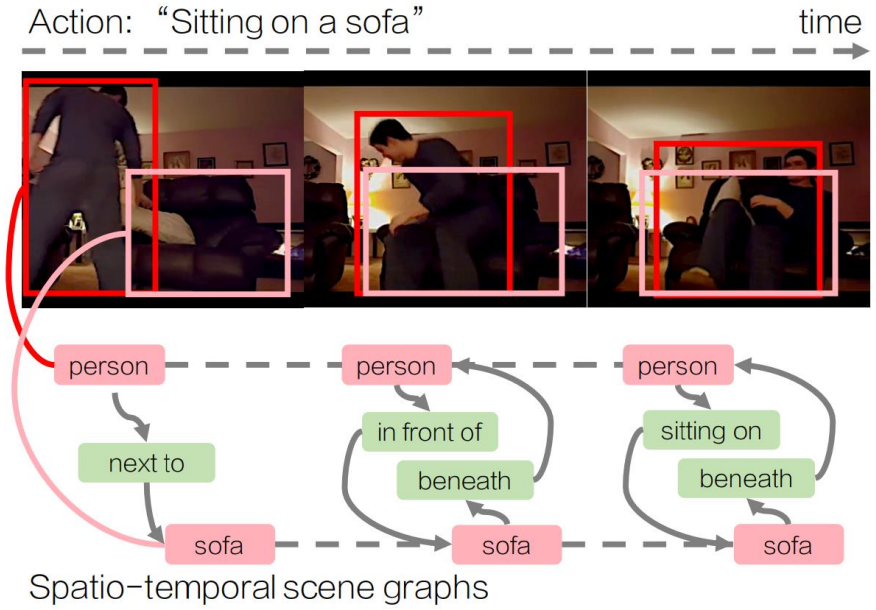
[1] Justin Johnson, etc, and Li Fei-Fei. Image retrieval using scene graphs. CVPR. 2015.

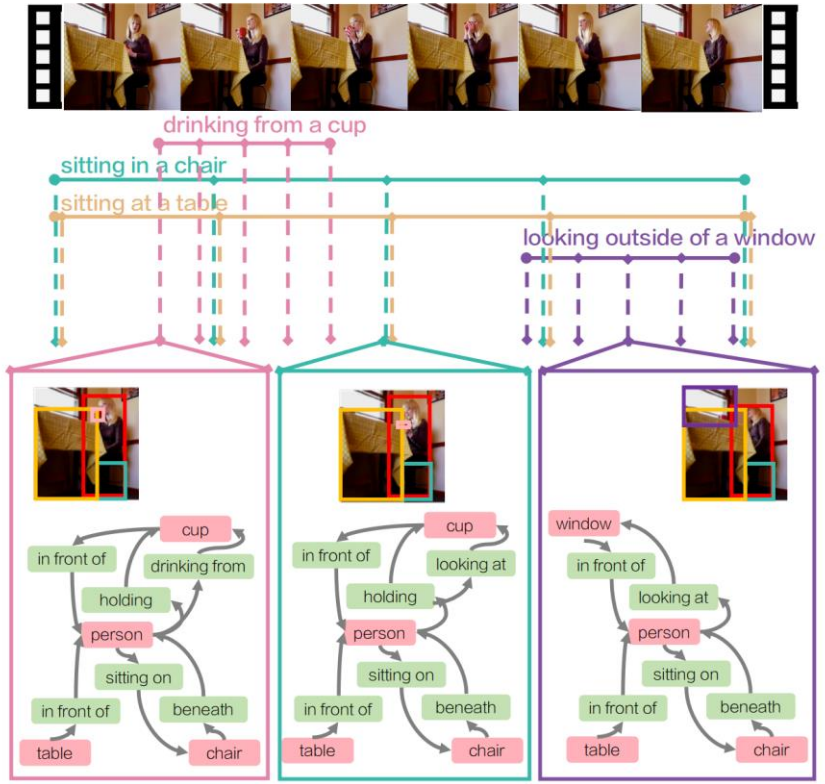
Method

Dynamic Scene Manager (Dysen)

- Dynamic Scene Graph (DSG) Representation

A sequence of VSG along time frames.





[1] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. CVPR, 2020.

Method

Dynamic Sc

- Process

- **Step-I:** occurri
- **Step-II** represe
- **Step-II** from Ch

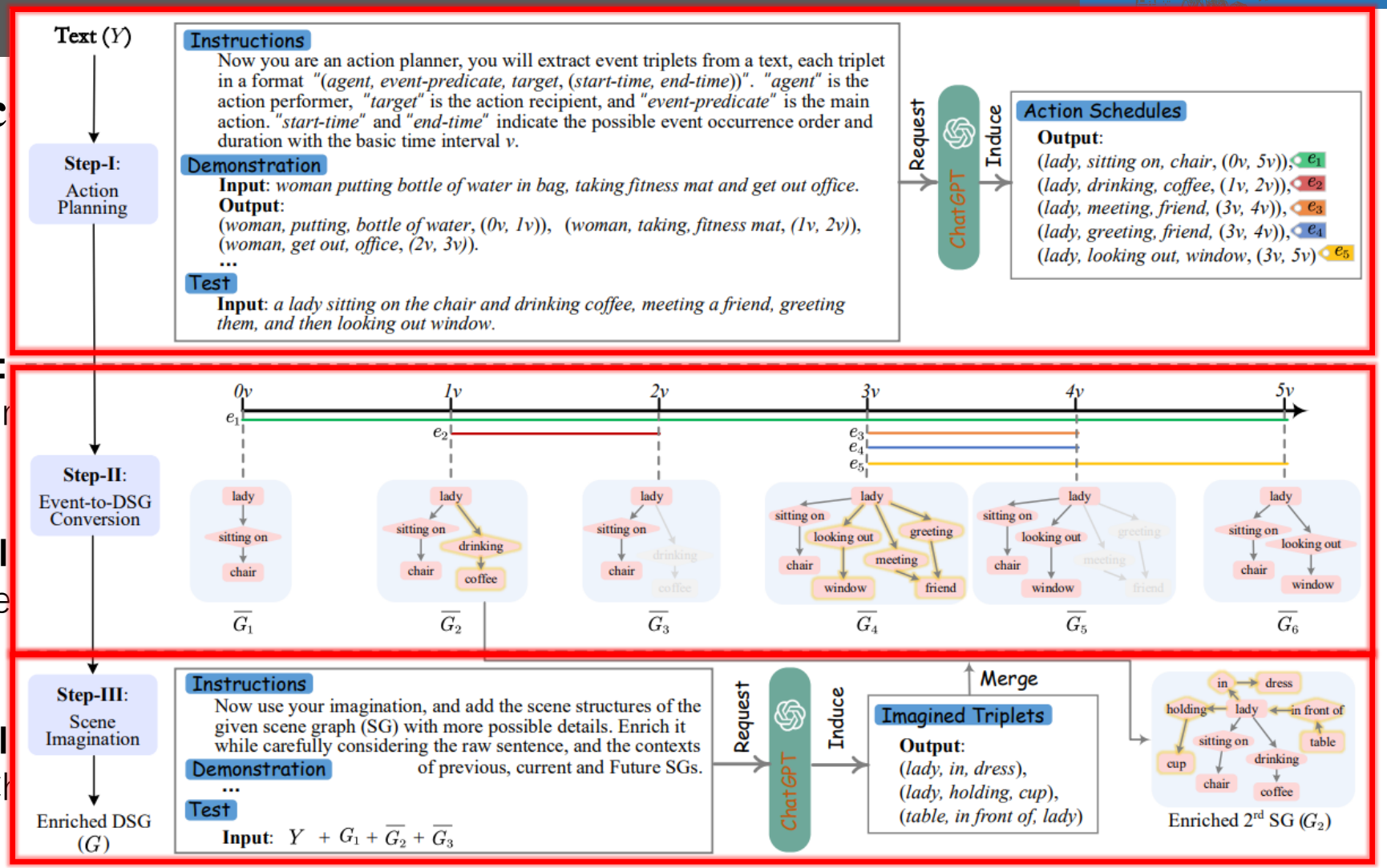
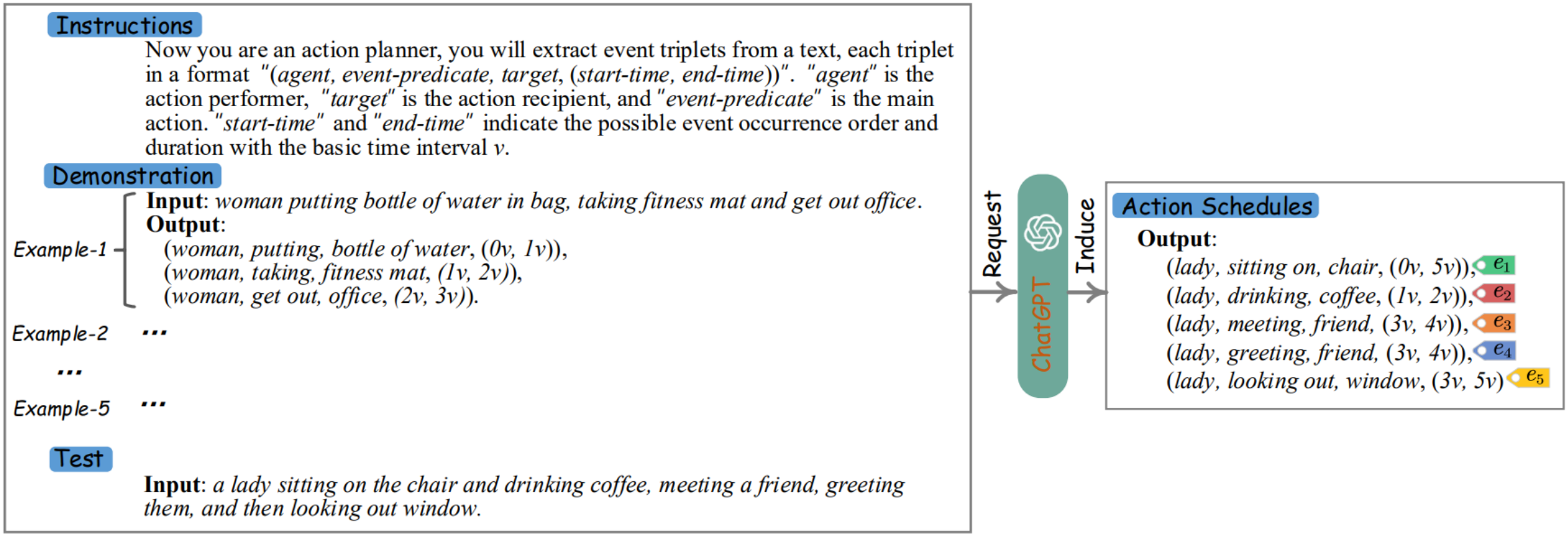


Figure 3: Based on the given text, Dysen module carries out three steps of operations to obtain the enriched DSG: 1) action planning, 2) event-to-DSG conversion, and 3) scene imagination, where we take advantage of the ChatGPT with in-context learning. Best viewed by zooming in.

Method

Dynamic Scene Manager (Dysen)

- Step-I, ICL for action planning



Method

Dynamic Scene Manager (Dysen)

- Step-II, ICL for step-wise scene imagination

Instructions

Now use your imagination, and add the scene structures of the given scene graph (SG) with more possible details. Enrich it while carefully considering the raw sentence, and the contexts of previous, current and Future SGs.

Demonstration

Example-1

Input:
Sentence: *woman putting bottle of water in bag, taking fitness mat and get out office.*
Enriched last SG: (*woman, holding, water*), (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, bag*), (*water, in, bottle*), (*bag, cloth*).
Current SG to enrich: (*woman, taking, fitness mat*).
Following SG: (*woman, get out of, office*).
Output:
(*woman, in, room*), (*woman, wearing, scarf*), (*lady, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*)

...

Example-5 ...

Test

Input:
Sentence: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*
Enriched last SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).
Current SG to enrich: (*lady, sitting on, chair*), (*lady, drinking, coffee*).
Following SG: (*lady, sitting on, chair*).

Request → ChatGPT → Induce

Imagined Triplets

Output:
(*lady, sitting on, chair*),
(*lady, in, suit*),
(*lady, in, glasses*),
(*lady, in, office*),
(*chair, white*),
(*chair, leather*),
(*lady, drinking, coffee*),
(*coffee, in, cup*),
(*cup, yellow*).

Instructions

Now imagine you are a film director, and polish the given sequentially-ordered scene graphs to make the dynamic scenes look more reasonable and temporally smooth in the global viewpoint. You may modify the input scene triplets for each SG and output more reasonable ones.

Demonstration

Example-1

Input:

Sentence: *woman putting bottle of water in bag, taking fitness mat and get out office.*

1st SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, water*), (*water, in, bottle*), (*bag, cloth*).

2nd SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*).

3rd SG: (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*), (*woman, get out of, office*).

...

Output:

1st SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, bag*), (*bottle, in, bag*), (*water, in, bottle*), (*bag, cloth*).

2nd SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*).

3rd SG: (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*), (*woman, get out of, office*).

...

...

Example-5 ...

Test

Input:

Sentence: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*

1st SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

2nd SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*), (*lady, drinking, coffee*), (*coffee, in, cup*), (*cup, yellow*).

3rd SG: (*lady, leaving, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

...

Request → ChatGPT → Induce

Polished Scene Graphs

Output:

1st SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

2nd SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*), (*lady, drinking, coffee*), (*coffee, in, cup*), (*cup, yellow*).

3rd SG: (*lady, leaving, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

...

Experiment

Result on Zero-shot T2V

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**

Experiment

- Result on Supervised Fine-tuned T2V

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**

Experiment

Results on Action-complex T2V Generation

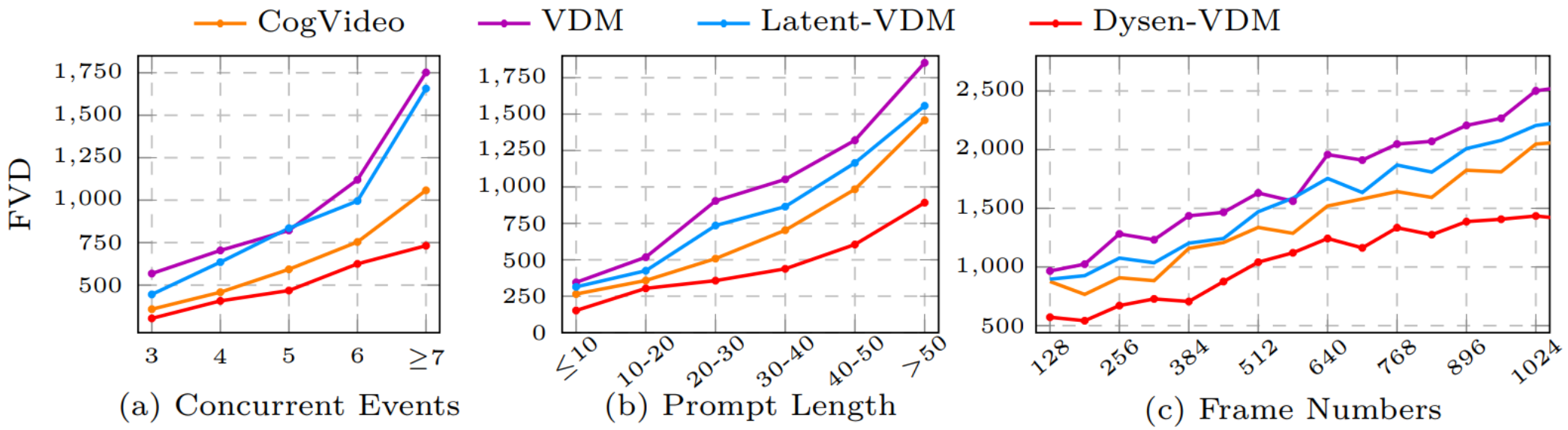


Figure 5: Performance on the action-complex scene video generation of ActivityNet data.

Experiment

In-depth Analyses

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)

Experiment

Visualizations

A clownfish swimming with elegance through coral reefs, presenting the beautiful scenery under the sea.

A woman is looking after the plant in her garden, and then she raises her head to observe the weather.

A man dressed as Santa Claus is riding a motorcycle on a big city road.

A horse in a blue cover walks at a fast pace, and then begins to slow down, taking a walk in the paddock.

Experiment

Visualizations

A person in a jacket riding a horse, is walking along the countryside road.

A cat eating food out of a bowl while looking around, then the camera moves away to a scene where another cat eats food.

A man and other man are standing together in the middle of a tennis court, and speaking to the camera.

A young violin player in a neat shirt with a collar, having a headphone on, is playing the violin.

Experiment

Visualizations

On a stage, a woman is rotating and waving her arms to show her belly dance.

A band composed of a group of young people is performing live music.

A woman hikes up the green mountain reaches the summit, and takes photos of the breathtaking view.

Two women sit on a park bench, reading books while chatting to each other.

Thanks

Q&A