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

Hao Fei[1] Shengqiong Wu[1] Wei Ji[1,*] Hanwang Zhang[2,3] Tat-Seng Chua[1]
[1]National University of Singapore [2]Skywork AI, Singapore [3]Nanyang Technological University
{haofei37,swu,jiwei,dcscts}@nus.edu.sg, hanwangzhang@ntu.edu.sg

Abstract

Text-to-video (T2V) synthesis has gained increasing attention in the community, in which the recently emerged diffusion models (DMs) have promisingly shown stronger performance than the past approaches. While existing state-of-the-art DMs are competent to achieve high-resolution video generation, they may largely suffer from key limitations (e.g., action occurrence disorders, crude video motions) with respect to the intricate temporal dynamics modeling*, one of the crux of video synthesis. In this work, we investigate strengthening the awareness of video dynamics for DMs, for high-quality T2V generation. Inspired by human intuition, we design an innovative dynamic scene manager (dubbed as* ***Dysen****) module, which includes (****step-1****) extracting from input text the key actions with proper time-order arrangement, (****step-2****) transforming the action schedules into the dynamic scene graph (DSG) representations, and (****step-3****) enriching the scenes in the DSG with sufficient and reasonable details. Taking advantage of the existing powerful LLMs (e.g., ChatGPT) via in-context learning, Dysen realizes (nearly) human-level temporal dynamics understanding. Finally, the resulting video DSG with rich action scene details is encoded as fine-grained spatio-temporal features, integrated into the backbone T2V DM for video generating. Experiments on popular T2V datasets suggest that our Dysen-VDM consistently outperforms prior arts with significant margins, especially in scenarios with complex actions. Codes at* http://haofei.vip/Dysen-VDM/.

1. Introduction

Recently, AI-Generated Content (AIGC) has witnessed thrilling advancements and remarkable progress, e.g., ChatGPT [46], DELLE-2 [49] and Stable Diffusion (SD) [51]. As one of the generative topics, text-to-video synthesis that generates video content complying with the provided textual description has received an increasing number of attention in the community. Prior researches develop a variety of methods for T2V, including generative adversarial networks (GANs) [1, 42, 53], variational autoencoders (VAEs) [8, 34, 82], flow-based models [4, 32], and auto-regressive models (ARMs) [11, 30, 75]. More recently, diffusion models (DMs) have emerged to provide a new paradigm of T2V. Compared with previous models, DMs advance in superior generation quality and scaling capability to large datasets [17, 25], and thus showing great potential on this track [39, 41, 43, 85].

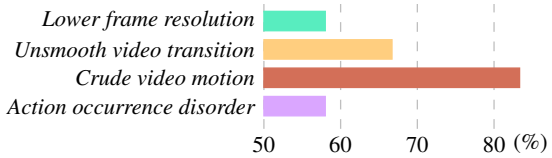

Figure 1. Common issues in the existing text-to-video (T2V) synthesis. We run the video diffusion model (VDM) [23] with random 100 prompts, and ask users to summarize the problems.

Although achieving the current state-of-the-art (SoTA) generative performance, DM-based T2V still faces several common yet non-negligible challenges. As summarized in Figure 1, four typical issues can be found in a diffusion-based T2V model, such as *lower frame resolution*, *unsmooth video transition*, *crude video motion* and *action occurrence disorder*. While the latest DM-based T2V explorations paid much effort into enhancing the quality of video frames, i.e., generating high-resolution images [43, 85, 91], they may largely overlook the modeling of the **intricate video temporal dynamics**, the real crux of high-quality video synthesis, i.e., for relieving the last three types of aforementioned issues. According to our observation, the key bottleneck is rooted in the nature of video-text modality heterogeneity: language can describe complex actions with few succinct and abstract words (e.g., predicates and modifiers), whereas video requires specific and often redundant frames to render an action.

Picturing that, whenever 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. With such integral *screenplay*, it can be effortless to project the whole video

*Corresponding Author.

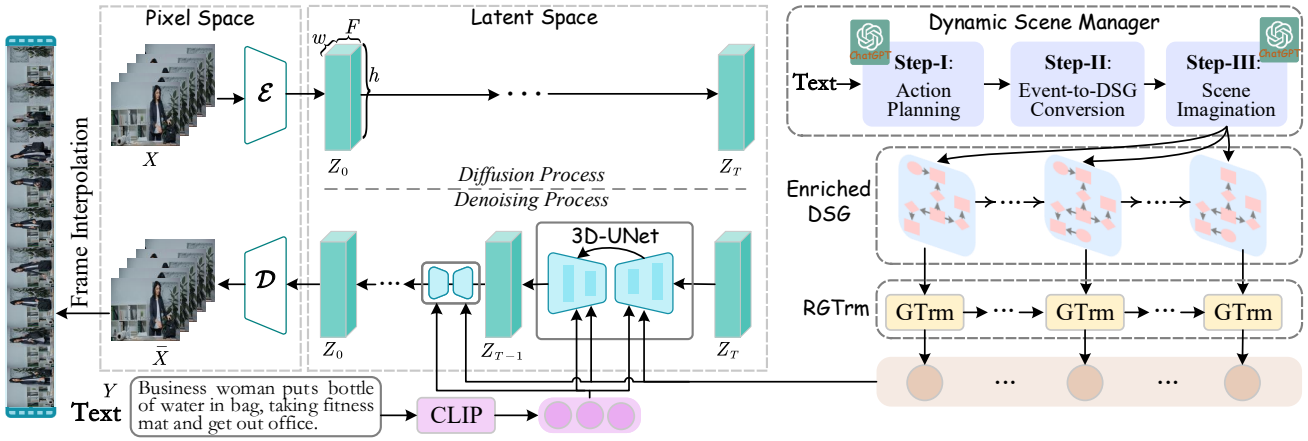

Figure 2. The architecture of the dynamics-aware T2V diffusion model, Dysen-VDM. The dynamic scene manager (Dysen) module operates over the input text prompt and produces the enriched dynamic scene graph (DSG), which is encoded into the resulting fine-grained spatio-temporal scene features are integrated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

successfully. Correspondingly, from the above intuition we can draw four key points of effective T2V modeling, especially for the scenario with complex dynamics. **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 spatio-temporal features modeling should be realized for temporally coherent video generation.

Based on the above observations, in this work we present a nichetargeting solution to achieve high-quality T2V generation by strengthening the awareness of video dynamics. We propose a dynamics-aware T2V diffusion model, as shown in Figure 2, in which we first employ the existing SoTA video DM (VDM) as the backbone T2V synthesis, and meanwhile devise an innovative **dy**namic **sce**ne manager (namely **Dysen**) module for video dynamics modeling. To realize the human-level temporal dynamics understanding of video, we take advantage of the current most powerful LLM, e.g., OpenAI ChatGPT (GPT3.5/GPT4); we treat ChatGPT as the consultant for action planning and scene imagination in Dysen. Specifically, in **step-I**, we extract the key actions from the input text, which are properly arranged in physically occurring orders. In **step-II**, we then convert these ordered actions into sequential dynamic scene graph (DSG) representations [26]. DSGs represent the intrinsic spatial&temporal characteristic of videos in semantic structures, allowing effective and controllable video scene management [35]. In **step-III**, we enrich the scenes in the DSG with sufficient and reasonable details. We elicit the knowledge from ChatGPT with the in-context learning [74]. At last, the resulting DSGs with well-enriched scene details are encoded with a novel recurrent graph Transformer, where the learned delicate fine-grained spatio-temporal features are integrated into the backbone T2V DM for generating high-quality fluent video.

We evaluate our framework on the popular T2V datasets, including UCF-101 [58], MSR-VTT [80], as well as the action-complex ActivityNet [31], where our model consistently outperforms existing SoTA methods on both the automatic and human evaluations with significant margins. We show that our Dysen-VDM system can generate videos in higher motion faithfulness, richer dynamic scenes, and more fluent video transitions, and especially improves on the scenarios with complicated actions. Further in-depth analyses are shown for a better understanding of how each part of our methods advances.

Overall, this paper addresses the crux of high-quality T2V synthesis by strengthening the motion dynamics modeling in diffusion models. We contribute in multiple aspects. ***(i)*** To our knowledge, this is the first attempt to leverage the LLMs for action planning and scene imagination, realizing the human-level temporal dynamics understanding for T2V generation. ***(ii)*** We enhance the dynamic scene controllability in diffusion-based T2V synthesis with the guidance of dynamic scene graph representations. ***(iii)*** Our system empirically pushes the current arts of T2V synthesis on benchmark datasets. Our codes will be open later to facilitate the community.

2. Related Work

Synthesizing videos from given textual instructions, i.e., T2V, has long been one of the key topics in generative AI. A sequence of prior works has proposed different genera-

tive neural models for T2V. Initially, many attempts extend the GANs [13] models from image generation [60, 81, 92] to video generation [7, 10, 14, 29, 56]. While GANs often suffer from the issue of mode collapse leading to hard scalability, other approaches have proposed learning the distribution with better mode coverage and video quality than GAN-based approaches, such as VAEs [8, 34, 82], flow-based models [4, 32] and ARMs [11, 30, 75]. Recently, diffusion models [22] have emerged, which learn a gradual iterative denoising process from the Gaussian distribution to the data distribution, generating high-quality samples with wide mode coverage. Diffusion-based T2V methods help bring better results with more stable training [17, 21, 25, 39, 41, 47, 78]. Further, latent diffusion models (LDMs) [52] have been proposed to learn the data distribution from low-dimensional latent space, which helps sufficiently reduce the computation costs, and thus receive increasing attention for T2V synthesis [18, 43, 85, 91]. In this work, we inherit the advance of LDMs, and adopt it as our backbone T2V synthesizer.

Compared with the text-to-image (T2I) generation [52, 57, 81, 92] that mainly focuses on producing static visions in high-fidelity resolutions, T2V further places the emphasis on the modeling both of spatial&temporal semantics, especially the scene dynamics. Previously, some T2V research explores video dynamics modeling for generating high-quality videos [9, 38, 67, 86, 87], i.e., higher temporal fluency, and complex motions, while they may largely be limited to the coarse-level operations, such as the spatio-temporal convolutions [67]. In the line of DM-based T2V [12, 36, 71, 72, 77, 79, 88, 90], most of the methods consider improving the video quality by enhancing the frame resolution [43, 85, 91], instead of the perception of dynamics. Most of the LDM-based T2V work also uses the spatio-temporal factorized convolutions in the 3D-UNet decoder [18, 69, 70, 91]. For example, [2] tries to strengthen motion awareness with a temporal shift operation. All of these attempts, unfortunately, can be seen as a type of coarse-grained modeling. In this work, we take fine-grained spatio-temporal feature modeling based on DSG representations. We propose a systematic solution to enhance the diffusion awareness of the action dynamics.

3. Preliminary

3.1. Text-to-video Latent Diffusion Model

We first formalize T2V task as generating an video X=$\{x_1, \cdots, x_F\} \in \mathbb{R}^{F\times H\times W\times C}$ that specifies the desired content in the input prompt text Y=$\{w_1, \cdots, w_S\}$. Here F, H, W, C are the frame length, height, width, and channel number of video, respectively. A latent diffusion model (LDM) is adopted for T2V which performs a forward (diffusion) process and a reverse (denoising) process in the video latent space. Firstly, an encoder $\mathcal{E}$ maps the video frames into the lower-dimension latent space, i.e., $Z_0 = \mathcal{E}(X)$, and later a decoder $\mathcal{D}$ re-maps the latent variable to the video, $X = \mathcal{D}(Z_0)$. Given the compressed latent code Z_0, LDM gradually corrupts it into a pure Gaussian noise $Z_T \sim \mathcal{N}(Z_T, 0, I)$ over T steps by increasingly adding noisy, formulated as $q(Z_{1:T}|Z_0) = \prod_{t=1}^{T} q(Z_t|Z_{t-1})$. and the learned reverse process $p_\theta(Z_{0:T}) = p(Z_T)\prod_{t=1}^{T} p_\theta(Z_{t-1}|Z_t, Y)$ gradually reduces the noise towards the data distribution conditioned on the text Y. T2V LDM is trained on video-text pairs (X, Y) to gradually estimate the noise ϵ added to the latent code given a noisy latent Z_t, timestep t, and conditioning text Y:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{Z\sim\mathcal{E}(X),Y,\epsilon,t}\left[\,||\epsilon - \epsilon_\theta(Z_t, t, \mathcal{C}(Y))||^2\,\right], \quad (1)$$

where $\mathcal{C}(Y)$ denotes a text encoder that models the conditional text, and the denoising network $\varepsilon_\theta(\cdot)$ is often implemented via a 3D-UNet [23], as illustrated in Figure 2.

3.2. Dynamic Scene Graph Representation

DSG [26] is a list of single visual SG of each video frame, organized in time-sequential order. We denote an DSG as G=$\{G_1, \cdots, G_M\}$, with each SG (G_m) corresponding to the frame (x_m). An SG contains three types of nodes, i.e., *object*, *attribute*, and *relation*, in which some scene objects are connected in certain relations, forming the spatially semantic triplets '*subject-predicate-object*'. Also, objects are directly linked with the attribute nodes as the modifiers. Besides, since a video comes with inherent continuity of actions, the SG structure in DSG is always temporal-consistent across frames. This characterizes DSGs with spatial&temporal modeling. Figure 2 (right part) simply visualizes a DSG.

4. Methodology

Overall Framework. The architecture of our proposed dynamics-aware T2V diffusion framework is shown in Figure 2. The backbone T2V synthesizer is an LDM (cf. §3.1). During the denoising, the dynamic scene manager (Dysen) module (cf. §4.1) effectively captures the intrinsic spatial-temporal characteristic of input texts to better guide the T2V generation in the mainstay LDM (cf. §4.2).

4.1. Dynamic Scene Manager

As cast earlier, the input language instruction can often be succinct and abstract, which causes trouble generating concrete dynamic visual scenes in videos, especially when the actions are semantically complex. To bridge the gap of dynamic scenes between texts and videos, here we propose a dynamic scene manager. With Dysen, we carry out three steps of operations: action planning, event-to-DSG conversion, and scene enrichment. Recently, the rise of LLMs has revealed the amazing potentials [6, 45, 62], among which, ChatGPT [46] is the most outstanding one in content un-

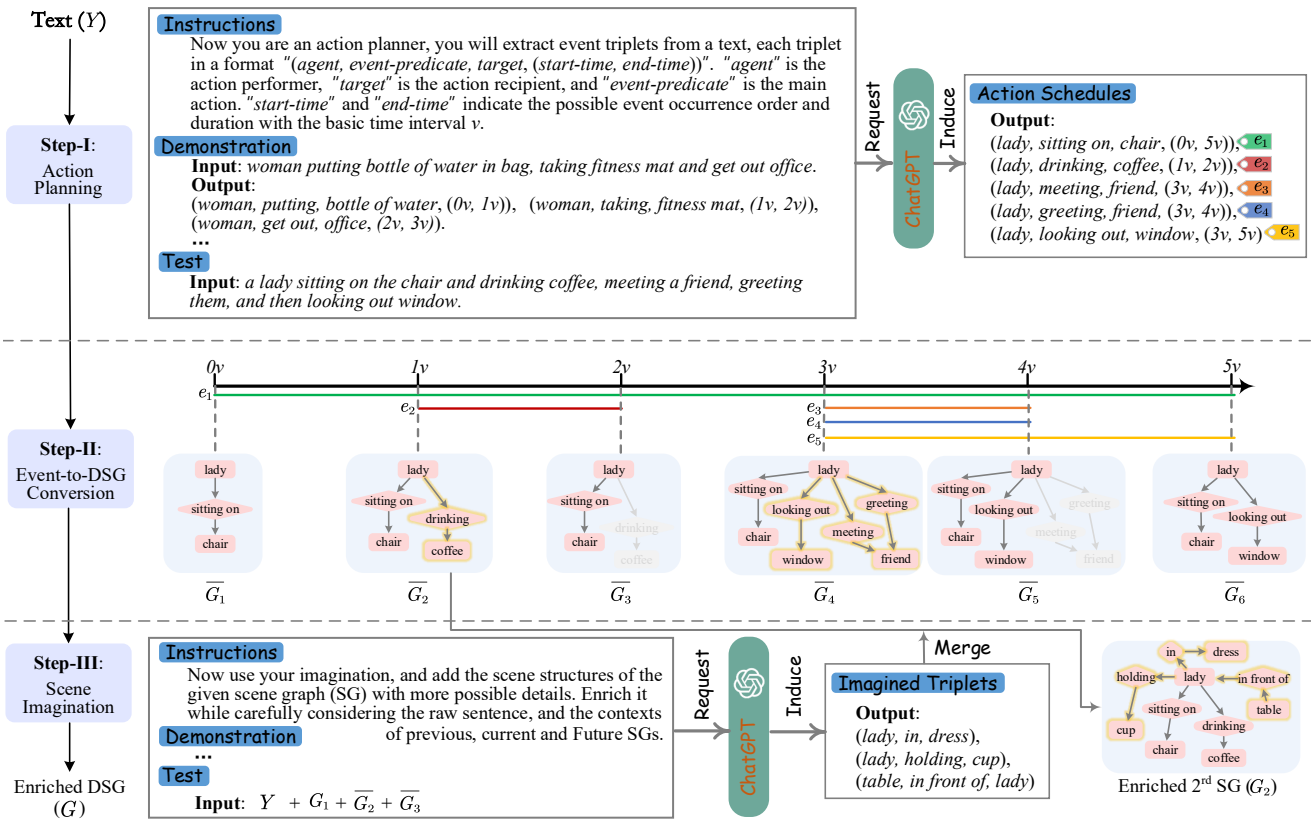

Figure 3. Based on the given text, Dysen module carries out three steps of operations to obtain enriched DSG: 1) action planning, 2) event-to-DSG conversion, and 3) scene imagination, where we take advantage of ChatGPT (i.e., GPT3.5 or GPT4) with in-context learning.

derstanding, and perceiving complex events from language and comprehending the dynamic scenes in the way humans do [68, 76]. Thus, we elicit such action planning and scene imagination abilities from ChatGPT.

Step-I: Action Planning. We first ask ChatGPT to extract the key actions from the prompt texts. Technically, we employ in-context learning (ICL) [74]. We write the prompts, which include 1) a job description (`Instruction`), 2) a few input-output in-context examples (`Demonstration`), and 3) the desired testing text (`Test`). Feeding the ICL prompts, we expect ChatGPT to return the desired action plans, as illustrated in Figure 3. Specifically, we represent an action scene as "(*agent, event-predicate, target*, (*start-time, end-time*))", in which '*agent, event-predicate, target*' is the event triplet corresponding to the relational triplets as described in DSG (cf. §3.2); '*start-time, end-time*' is the temporal interval of this event. Note that the atomic time interval is assumed as v, which is disentangled from a physical time duration. Both the event scene triplets and the time arrangements are decided via ChatGPT's understanding of the input. This way, even complex actions with multiple overlapped or concurrent events will be well supported.

Step-II: Event-to-DSG Conversion. With the event schedule at hand, we then transform it into a holistic DSG structure. Note that this DSG can be quite primitive, as each SG structure within DSG almost contains one triplet, which can be seen as the skeleton of the dynamic scenes. Specifically, we construct the DSG along with the time axis incrementally, i.e., with each frame step having a corresponding SG. According to the occurrence order and duration of events, in each frame we *add* or *remove* a triplet, until handling the last event. This also ensures the SG at each frame step is globally unique. The resulting DSG well represents the skeleton spatial-temporal feature of the events behind the input. Figure 3 illustrates the conversion process.

Step-III: Scene Imagination. Based on the above initial DSG (denoted as $\overline{G}$=$\{\overline{G}_1, \cdots, \overline{G}_M\}$), we finally enrich the scenes within each SG. For example, for each SG, there should be visually abundant scenes, e.g., objects will have various possible attributes, and different objects can be correlated with new feasible relations within the scene. Also, the temporal changes between SG frames should be reflected, instead of the constant SG across a period. This is intuitively important because all the events are continuous in time, and all the motions happen smoothly, e.g., in '*a person sitting on a chair*', the motion '*sitting*' can be broken down into a consecutive motion chain: '*approaching*'→'*near to*'→'*sitting*'. We again adopt the ChatGPT to

complete the job, as it is effective in offering rich and reasonable imagination [15, 16]. Concretely, the scene enrichment has two rounds. The first round preliminarily enriches each SG one by one, i.e., by either *adding* some new triplets or *changing* the existing triplets. As shown in Figure 3, the ICL technique is again used to prompt ChatGPT to yield the triplets to be added for the current SG, given the raw input text. To ensure the dynamic scene coherency, we consider a *sliding window context* (SWC) mechanism, when operating for the current SG, takes into account the current, previous (enriched), and following contexts of SGs (e.g., [G_{m-1}, $\overline{G}_m$, $\overline{G}_{m+1}$]). This way, $\overline{G}_m$ can inherit from the previous well-established scene G_{m-1}, and meanwhile decide what to add or change to better transit to the next scene $\overline{G}_{m+1}$. The second round further reviews and polishes the overall scenes of DSG from a global viewpoint, also via ChatGPT in another ICL prompting process. This ensures all the actions go more reasonably and consistently. The resulting final DSG is denoted as G=$\{G_1, \cdots, G_M\}$).

4.2. Scene Integration for T2V Generation

The enriched DSG (G) entails fine-grained spatial and temporal features. Instead of using the general graph neural networks to encode the DSG structure, e.g., GCN, GAT, and RGNN [40, 44, 66], we consider the Transformer architecture [65] that allows highly-parallel computation with the self-attention calculation. To further model the temporal dynamics of the graphs, we consider the recurrent graph Transformer (RGTrm). RGTrm has L stacked layers, with a total of M recurrent steps of propagation for each SG. The representation H_m^l of SG G_m of l-th layer is updated as:

$$H_m^{l+1} = O_k^l \,||_{k=1} \,(\sum_{j \in \mathcal{N}_i} w_{i,j,m}^{k,l} \, V_m^{k,l}) \,, \tag{2}$$

$$w_{i,j,m}^{k,l} = \text{Softmax}_j(\frac{\hat{Q}_m^{k,l} \cdot K_m^{k,l}}{\sqrt{d_k}}) \cdot E_m^{k,l} \,, \tag{3}$$

$$\hat{Q}_m^{k,l} = (1 - z_m) \cdot Q_{m-1}^{k,l} + z_m \cdot Q_m^{k,l} \,, \tag{4}$$

$$z_m = \sigma(W^z \cdot Q_m^{k,l} \cdot K_m^{k,l}) \,, \tag{5}$$

where k denotes the attention head number. O_k^l is the k-th attention head representation. $K_m^{k,l}$=$W^K H_m^l$, $Q_m^{k,l}$=$W^Q H_m^l$, $V_m^{k,l}$=$W^V H_m^l$ are the key, query and value representations in the Transformer. $E_m^{k,l}$=$W^E\{e_{i,j,m}\}$ is the embedding of edge $e_{i,j,m}$ in DSG. And $||$ is the concatenation operation. We denote the final DSG representation as $H^G = \{H_1^G, \cdots, H_M^G\}$, where H_m^G is the one of m-th SG (G_m).

Next, we integrate the fine-grained spatial-temporal DSG features (H^G) into the 3D-UNet decoder for enhanced T2V generation, i.e., denoising process. Although the original 3D-UNet [23] has a spatial-temporal feature modeling (as the green dotted box in Figure 4), it is limited by the coarse-grained operations, e.g., convolutions over 3D patches and attention over frames. Thus, we insert an additional Transformer block with cross-attention for fusing the fine-grained H^G representations, followed by another cross-attention to further fuse the raw text feature H^Y:

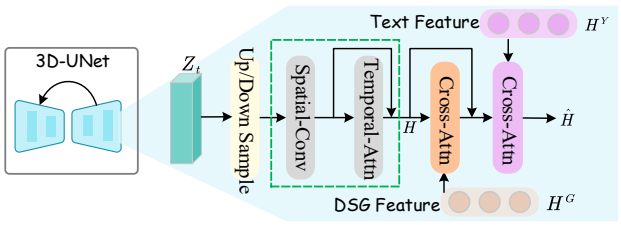

Figure 4. Illustration of the DSG integration.

$$\hat{H} = \text{Softmax}(\frac{H \cdot H^G}{\sqrt{d}}) \cdot H^G \,, \quad \hat{H} \leftarrow \text{Softmax}(\frac{\hat{H} \cdot H^Y}{\sqrt{d}}) \cdot H^Y \,, \tag{6}$$

where H is the coarse-grained spatio-temporal features. The text representation H^Y is encoded by CLIP [48].

4.3. Overall Training

The overall training of Dysen-VDM system entails three major steps.

- **Stage-I:** Pre-training backbone Latent VDM with autoencoder based on WebVid data [3].
- **Stage-II:** Further pre-training the backbone VDM for text-conditioned video generation, based on WebVid data. We update the backbone diffusion model of Dysen-VDM, where the 3D-UNet includes an RGTrm encoder, and they all will be updated. There we will use the DSG annotations generated from Dysen.
- **Stage-III:** Updating the overall Dysen-VDM with the dynamic scene managing (Dysen).

5. Experiments

5.1. Setups

We experiment on two popular T2V datasets, including the UCF-101 [58] and MSR-VTT [80]. In UCF-101, the given texts are the simple action labels. In MSR-VTT, there are integral video caption sentences as input prompts. To evaluate the action-complex scenario, we also adopt the ActivityNet data [31], where each video connects to the descriptions with multiple actions (at least 3 actions), and the average text length is 50.4. To relieve the computation burden, during the sampling phase in diffusion, we evenly sample 16 keyframes from a two-second clip, and then interpolate them twice with higher frame rates. We perform image resizing and center cropping with a spatial resolution of 256×256 for each input text. The latent space is 32×32×4. The denoising sampling step T is 1000. In default, we use the ChatGPT (*GPT-3.5* turbo) via OpenAI API.[1] For action planning and scene imagination, we sample D=5 in-context

[1] https://platform.openai.com

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**

demonstrations. RGTrm takes L=12 layers and k=8 attention heads. All dimensions are set as 768. Initial β is set 0.5, and then decays gradually.

Following previous works [2, 5, 18], we use the Inception Score (IS) and Fréchet Video Distance (FVD) for UCF-101, and Fréchet Image Distance (FID) and CLIP similarity (CLIPSIM) for MSR-VTT. We also use human evaluation for a more intuitive assessment of video quality. We consider two types of settings: 1) zero-shot, where our pre-trained model makes predictions without tuning on on-demand training data; 2) directly fine-tuned on training data without large pre-training. We consider several existing strong-performing T2V systems as our baselines, which are shown later. Also, we re-implement several open-sourced baselines for further customized evaluations, including CogVideo [24], VDM [23] and Latent-VDM [52]. Scores from our implementations are averaged in five runs with random seeds, and the results of other baselines are copied from the raw papers. All our training is conducted on 16 NVIDIA A100 GPUs.

5.2. Main Comparisons and Observations

Zero-shot Performance. We first present the comparison results on the zero-shot setting on UCF-101 and MSR-VTT datasets, respectively. As shown in Table 1, Dysen-VDM outperforms the baselines on both IS and FVD metrics with big margins on UCF-101 data, where the given texts are the simple action labels, and the dynamic scene imagination capability is especially needed. This shows the capability of our model. Note that on MSR-VTT data we calculate the frame-level metrics between the testing captions and video frames, and we see that our system secures the best results.

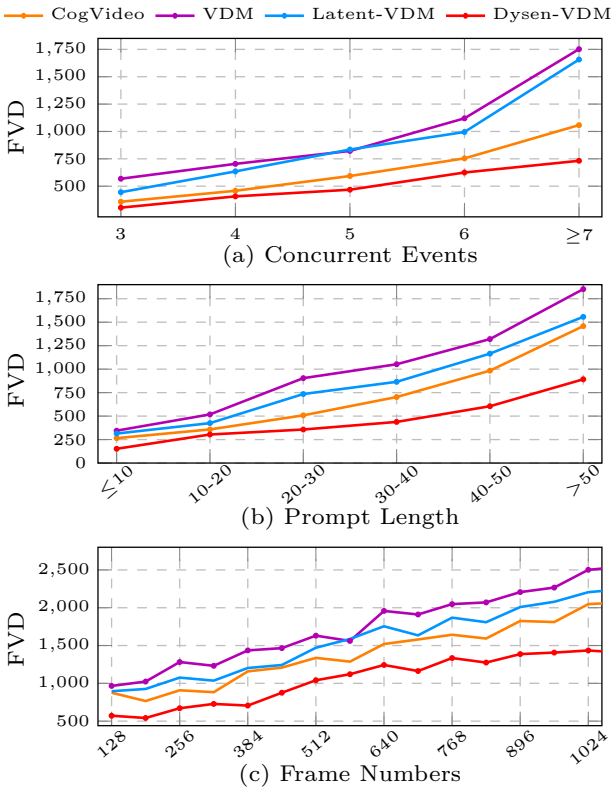

Figure 5. Performance on the action-complex scene video generation of ActivityNet data.

On-demand Fine-tuning Results. Table 2 further presents the results of the fine-tuned setting on UCF-101 data. We see that with the on-demand training annotations, the winning scores of our system over the baselines become more clear. In particular, Dysen-VDM model achieves 95.23 IS and 255.42 FVD scores, respectively, becoming a new state-of-the-art.

5.3. Results on Action-complex T2V Generation

Now we consider a more strict comparing setting of action-complex scenario. We use the ActivityNet data under the fine-tuning setup. We consider three different testing T2V scenarios: 1) the input texts containing multiple concurrent (or partially overlapped) actions, 2) the prompts having different lengths,[2] and 3) generating dif-

[2]As the text length in ActivityNet is no less than 20, we randomly add some test sets from MSR-VTT 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)

ferent lengths of video frames. We make comparisons with CogVideo, VDM and Latent-VDM, where the last two are diffusion-based T2V methods. As plotted in Figure 5, overall Dysen-VDM evidently shows stronger capability than the baseline methods, on all three tests of action-complex T2V generation. We also see that the superiority becomes more clear when the cases go harder, i.e., with more co-occurred events, longer input prompts and longer video generation. We note that CogVideo uses large pre-training, thus keeping comparatively better performance than the other two T2V diffusion models. In contrast, without additional pre-training, our system is enhanced with scene dynamics modeling can still outperform CogVideo significantly.

5.4. Human Evaluation

The standard automatic metrics could largely fail to fully assess the performance with respect to the temporal dynamics of generated videos. We further show the human evaluation results on the ActivityNet test set, in terms of action faithfulness, scene richness, and movement fluency, which correspond to the issues shown in Figure 1. We ask ten people who have been trained with rating guidelines, to rate a generated video from 0-10 scales, and we average the final scores into 100 scales. As seen in Table 3, overall, our system shows very exceptional capability on the complex-scene T2V generation than other comparing systems. In particular, Dysen-VDM receives a high 92.4 score on the scene richness, surpassing CogVideo by 17.4, and also wins over Latent-VDM on action failthfulness by 15.9. We can give the credit to the action planning and scene imagination mechanism in Dysen module.

5.5. System Ablations

We further conduct ablation studies to quantify the specific contribution of each design of our system. As shown in Table 4, we can find that, first of all, removing the whole

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.

Dynsen module (then equal to the Latent-VDM model) results in the most crucial performance loss, with +90.98 FVD on UCF-101 and +141.82 FVD on ActivityNet. This evidently verifies the efficacy of the Dynsen module and indirectly indicates that the core of high-quality T2V synthesis lies in modeling the motion dynamics. Further, removing step 3 of Dynsen, the scene imagination part, we see there are also significant drops, only second to the whole Dynsen. When only without the sliding window context (SWC) mechanism, the performance can be also hurt, indicating the importance of generating reasonable and fine scene details for T2V. Then, if canceling the RL optimization for ICL, the performance is hurt, especially on UCF-101 data, as the short labels require much more high-quality demonstrations to prompt ChatGPT for correct action planning and scene enrichment. Finally, the proposed RGTrm also serves irreplaceable roles for the fine-grained spatio-temporal feature encoding.

5.6. Qualitative Results

To gain a more direct understanding of how better our system succeeds in generating videos with smooth and complex movements, we present qualitative comparisons with the baseline models in Figure 6. As can be observed, Dysen-VLM has exhibited overall better performance. For

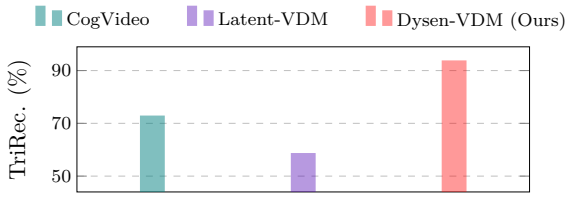

Figure 7. Aligning recall rate (TriRec.) of '*sub.-prdc.-obj.*' structures between prompt and generated video frames.

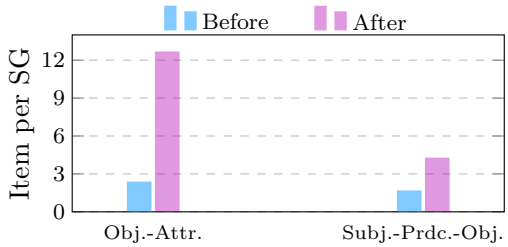

Figure 8. The number of SG structures ('*obj.-attr.*' and '*sub.-prdc-obj.*') before and after scene imagination.

both two prompts, our model shows smooth video frames with accurate motions occurring in order, while the videos by LVDM have quite jumpy transitions between different frames, and also the dynamic video scenes are not delicate, with some erroneous actions. The main reason largely lies in whether the T2V system models the intricate temporal dynamics. Also we see that the video by baseline may fail to faithfully reflect all the predicates mentioned in the input instructions; baseline missed certain actions. For example in prompt #1, '*man walks*' is missed by VDM.

5.7. In-depth Analyses

Controllability with DSG. SG has shown to have better semantic controllability, due to its semantically structured representations [28, 83, 84]. Here we examine such superiority of our system where our dynamic scene enhancement is built based on DSG. Following [78], we use the *Triplet Recall* (TriRec.) to measure the fine-grained '*subject-predicate-object*' structure recall rate between the SGs of input texts and video frames. Given a set of ground truth triplets, denoted G^{GT}, and TriRec. is computed as:

$$\text{TriRec.} = \frac{|G^{PT} \cap G^{GT}|}{|G^{GT}|}, \tag{7}$$

where G^{PT} are the relation triplets of the SG in the generated video DSG by a visual SG parser. As plotted in Figure 7, Dysen-VDM achieves the highest score than two baselines with clear margins.

Change of Scenes. We then make statistics of the structure changes before and after the scene imagination in Dysen module. We mainly observe the '*object-attribute*' and '*subject-predicate-object*' SG structures, where the former reflects the static contents, and the latter reflects the dynamic scenes. From Figure 8 we learn that both two types of SG structures are increased in numbers via scene imagination by LLM. This indicates a clear scene enrichment, leading to better video generation.

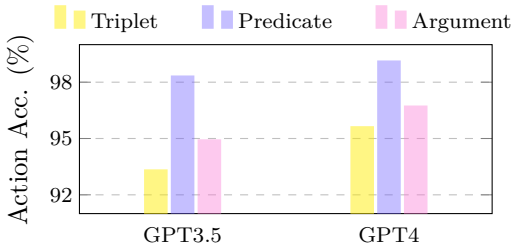

Figure 9. Accuracy of action annotation (overall triplet, predicate and argument (subject&object) using GPT3.5 and GPT4.

Action Parsing via ChatGPT. Action planning, as the first step, is pivotal to the overall following performance, where we employ the ChatGPT for inducing majorly-occurring events/actions. Here we analyze the quality of action parsing, and the influence of using GPT3.5 and GPT4. We randomly select 100 samples from the MSR-VTT data, and then compare between the ChatGPT-generated annotations and manually annotated ones. From Figure 9 we see that both GPT3.5 and GPT4 shows quite satisfied accurate induction, with GPT4 advancing more slightly.

6. Conclusion

In this work, we enhance the intricate temporal dynamics modeling of video diffusion models (VDMs) for text-to-video (T2V) synthesis. Inspired by human intuition of video filming, we design an innovative dynamic scene manager (Dysen) module, which performs three steps of temporal dynamics understanding: first extracting key actions with proper time-order arrangement; second, transforming the ordered actions into dynamic scene graph (DSG) representations; third, enriching the DSG scenes with sufficient reasonable details. We implement the Dysen based on ChatGPT, for human-level temporal dynamics understanding, where the in-context learning is optimized via reinforcement learning. Finally, we newly devise a recurrent graph Transformer to learn the fine-grained delicate spatio-temporal features from DSG, and then integrate them into the backbone T2V DM for video generation. Experiments on three T2V datasets show that our dynamics-aware video DM achieves new best results, especially performs stronger in scenarios with complex actions.

7. Acknowledgments

This research is supported by CCF-Baidu Open Fund and CCF-Baichuan Yingbo Innovation Research Funding.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *CoRR*, abs/1810.02419, 2018. 1

[2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *CoRR*, abs/2304.08477, 2023. 3, 6, 16, 18

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pages 1708–1718, 2021. 5

[4] Mohammad Bashiri, Edgar Y. Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas S. Tolias, and Fabian H. Sinz. A flow-based latent state generative model of neural population responses to natural images. In *Proceedings of the NeurIPS*, pages 15801–15815, 2021. 1, 3

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *CoRR*, abs/2304.08818, 2023. 6, 16, 18

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[7] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *CoRR*, abs/1907.06571, 2019. 3

[8] Yiping Duan, Mingzhe Li, Lijia Wen, Qianqian Yang, and Xiaoming Tao. From object-attribute-relation semantic representation to video generation: A multiple variational autoencoder approach. In *Proceedings of the MLSP*, pages 1–6, 2022. 1, 3

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[10] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *Proceedings of the BMVC*, pages 220–220, 2021. 3

[11] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *Proceedings of the ECCV*, pages 102–118, 2022. 1, 3, 6, 18

[12] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 3

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the NeurIPS*, pages 2672–2680, 2014. 3

[14] Cade Gordon and Natalie Parde. Latent neural differential equations for video generation. In *Proceedings of the NeurIPS*, pages 73–86, 2020. 3

[15] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023. 5

[16] Chao Guo, Yue Lu, Yong Dou, and Fei-Yue Wang. Can chatgpt boost artistic creation: The need of imaginative intelligence for parallel art. *IEEE/CAA Journal of Automatica Sinica*, 10(4):835–838, 2023. 5

[17] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Proceedings of the NeurIPS*, 2022. 1, 3

[18] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *CoRR*, abs/2211.13221, 2022. 3, 6, 16, 18

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the EMNLP*, pages 7514–7528, 2021. 17

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the NeurIPS*, pages 6626–6637, 2017. 17

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High defi-

nition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the NeurIPS*, pages 6840–6851, 2020. 3

[23] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Proceedings of the NeurIPS*, 2022. 1, 3, 5, 6, 7, 14, 18

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *CoRR*, abs/2205.15868, 2022. 6, 7, 18

[25] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *CoRR*, abs/2206.07696, 2022. 1, 3

[26] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the CVPR*, pages 10233–10244, 2020. 2, 3, 14, 16

[27] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the CVPR*, pages 1219–1228, 2018. 14

[28] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the CVPR*, pages 3668–3678, 2015. 8

[29] Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *Neural Networks*, 132:506–520, 2020. 3

[30] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the ICML*, pages 1771–1779, 2017. 1, 3

[31] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the ICCV*, pages 706–715, 2017. 2, 5

[32] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. In *Proceedings of the ICLR*, 2020. 1, 3

[33] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 6, 18

[34] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI*, pages 7065–7072, 2018. 1, 3

[35] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pretraining. In *Proceedings of the CVPR*, pages 13864–13873, 2022. 2

[36] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3

[37] Jiawei Liu, Weining Wang, Wei Liu, Qian He, and Jing Liu. Ed-t2v: An efficient training framework for diffusion-based text-to-video generation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 6, 18

[38] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tieniu Tan. Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 3

[39] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *CoRR*, abs/2303.08320, 2023. 1, 3

[40] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, 2017. 5

[41] Kangfu Mei and Vishal M. Patel. VIDM: video implicit diffusion models. *CoRR*, abs/2212.00235, 2022. 1, 3

[42] Andrés Muñoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift GAN for large scale video generation. In *Proceedings of the WACV*, pages 3178–3187, 2021. 1

[43] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. *CoRR*, abs/2303.13744, 2023. 1, 3

[44] Andrei Liviu Nicolicioiu, Iulia Duta, and Marius Leordeanu. Recurrent space-time graph neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Proceedings of the NeurIPS*, pages 12818–12830, 2019. 5, 7

[45] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 3

[46] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 1, 3

[47] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023. 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the ICML*, pages 8748–8763, 2021. 5, 17

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[50] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 91–99, 2015. 16

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the CVPR*, pages 10674–10685, 2022. 3, 6, 7, 18

[53] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *Int. J. Comput. Vis.*, 128(10):2586–2606, 2020. 1, 6, 18

[54] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the NeurIPS*, pages 2226–2234, 2016. 16

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022. 6, 18

[56] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the CVPR*, pages 3616–3626, 2022. 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the ICLR*, 2021. 3

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 2, 5

[59] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the CVPR*, pages 3713–3722, 2020. 16

[60] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, and Xiao-Yuan Jing. DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis. *CoRR*, abs/2008.05865, 2020. 3

[61] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *Proceedings of the ICLR*, 2021. 6, 18

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[63] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the ICCV*, pages 4489–4497, 2015. 16

[64] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. 17

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the NeurIPS*, pages 5998–6008, 2017. 5

[66] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua

Bengio. Graph attention networks. In *Proceedings of the ICLR*, 2018. 5

[67] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the NeurIPS*, pages 613–621, 2016. 3

[68] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Chatgpt empowered long-step robot control in various environments: A case application. *CoRR*, abs/2304.03893, 2023. 4

[69] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

[70] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 3, 6, 18

[71] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3

[72] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3

[73] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 6, 18

[74] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. 2, 4

[75] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *Proceedings of the ICLR*, 2020. 1, 3

[76] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. 4

[77] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[78] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 8

[79] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 3

[80] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, pages 5288–5296, 2016. 2, 5

[81] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the CVPR*, pages 1316–1324, 2018. 3

[82] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157, 2021. 1, 3, 6, 18

[83] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *CoRR*, abs/2211.11138, 2022. 8

[84] Sangwoong Yoon, Woo-Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI*, pages 10718–10726, 2021. 8

[85] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. *CoRR*, abs/2302.07685, 2023. 1, 3, 6

[86] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *Proceedings of the ICLR*, 2022. 3, 6, 18

[87] Vladyslav Yushchenko, Nikita Araslanov, and Stefan Roth. Markov decision process for video generation. In *Proceedings of the ICCV*, pages 1523–1532, 2019. 3

[88] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent

diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3

[89] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9073–9087, 2021. 15

[90] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 3

[91] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *CoRR*, abs/2211.11018, 2022. 1, 3, 6, 18

[92] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the CVPR*, pages 5802–5810, 2019. 3