DREAMRUNNER: Fine-Grained Compositional Story-to-Video Generation with Retrieval-Augmented Motion Adaptation

Zun Wang[1], **Jialu Li**[1], **Han Lin**[1], **Jaehong Yoon**[2], **Mohit Bansal**[1]
[1]UNC Chapel Hill
[2]Nanyang Technological University
{zunwang, jialuli, hanlincs, mbansal}@cs.unc.edu, jaehong.yoon@ntu.edu.sg

Abstract

Storytelling video generation (SVG) aims to produce coherent and visually rich multi-scene videos that follow a structured narrative. Existing methods primarily employ LLM for high-level planning to decompose a story into scene-level descriptions, which are then independently generated and stitched together. However, these approaches struggle with generating high-quality videos aligned with the complex single-scene description, as visualizing such complex description involves coherent composition of multiple objects/events, complex motion synthesis and character customization with sequential motions. To address these challenges, we propose **DREAMRUNNER**, a novel story-to-video generation method: First, we structure the input script using a large language model (LLM) to facilitate both coarse-grained scene planning as well as fine-grained object-level layout planning. Next, DREAMRUNNER presents retrieval-augmented test-time adaptation to capture target motion priors for objects in each scene, supporting diverse motion customization based on retrieved videos, thus facilitating the generation of new videos with complex, scripted motions. Lastly, we propose a novel spatial-temporal region-based 3D attention and prior injection module SR3AI for fine-grained object-motion binding and frame-by-frame spatial-temporal semantic control. We compare DREAMRUNNER with various SVG baselines, demonstrating state-of-the-art performance in character consistency, text alignment, and smooth transitions. Additionally, DREAMRUNNER exhibits strong fine-grained condition-following ability in compositional text-to-video generation, significantly outperforming baselines on T2V-ComBench. Finally, we demonstrate DREAMRUNNER's ability to generate multi-character interactions with qualitative examples.

Project Page — https://zunwang1.github.io/DreamRunner

1 Introduction

Advancing storytelling video generation (SVG) is crucial for real-world video generation applications, enabling the creation of rich, immersive narratives with multiple realistic scenes, characters, and interactive events. Unlike existing short-form video generation approaches (Wang et al. 2023a; Khachatryan et al. 2023; Qing et al. 2024; Bar-Tal et al. 2024), these models allow characters and objects to evolve across scenes, enhancing the coherence of generated content to align more closely with human storytelling. Such capabilities hold vast potential in media, gaming, etc.

Existing SVG methods (He et al. 2023; Oh et al. 2025; Zheng and Fu 2024; Zhao et al. 2024) primarily employ high-level planning with a large language model (LLM), breaking down a story into multiple key scene descriptions. Each scene is then generated independently as a separate video and later stitched together to form a complete long-form storytelling video. Generating high-quality single-scene video exhibits three key challenges: 1) *Coherent composition*: As a highly complex textual form, a story, even at the single-scene level (e.g. "*Lucy on the left and a man on the right is walking towards each other, they meet in the middle and start ballroom dancing*"), typically involves multiple objects/characters with distinct motion trajectories, attributes, and sequentially occurring events, all of which must be coherently composed in the generated video. 2) *Complex motion synthesis*: The complex scene descriptions often feature intricate character motions (e.g. "*ballroom dancing*") that are difficult to generate from the base text-to-video (T2V) models. 3) *Character customization with sequential events*: These descriptions usually involve characters with pre-defined reference images (e.g. *Lucy*), with sequential motions (e.g. *walking* to *ballroom dancing*), making it challenging to maintain both temporal and visual consistency with the character. However, recent methods often feed single-scene descriptions directly as textual conditions to the T2V model with limited constraints, resulting in sub-optimal fidelity, missed events/objects, unclear motions, etc.

To address the above challenges, we propose DREAMRUNNER, a novel SVG framework that enhances fine-grained alignment between scene descriptions and generated videos. Beyond high-level planning, DREAMRUNNER uses LLM-based compositional reasoning to decompose complex scenes into frame-by-frame layout plans of multiple entities with sequential motions/events, followed by region-based attention for *coherent composition*. For *complex motion synthesis*, we adopt retrieval-augmented prior learning, injecting priors only into relevant regions to support *character customization with sequential motions*. Specifically, DREAMRUNNER presents three essential processes in the framework: (1) *Dual-Level Video Plan Generation*, (2) *Motion Retrieval and Subject/Motion Prior Learning*,

and (3) *Spatial-Temporal Region-Based 3D Attention and Prior Injection (SR3AI)*. In **(1) plan generation stage**, given a user-provided story narration (e.g. "*write a story of the witch and her cat's one day*"), we employ an LLM for hierarchical planning: first generate a high-level plan with character-driven, motion-rich event descriptions across scenes, then decompose the scene descriptions into detailed, entity-specific frame-level layout plans within each scene. The generated frame-level plan serves as the fine-grained guidance for T2V. In **(2) prior learning stage**, we learn both subject and motion priors to enhance character consistency and motion fidelity. Subject priors are learned from character reference images using customization techniques (Ruiz et al. 2023) to adapt the model to specific appearances. Then we treat *complex motion synthesis* as a customization problem and learn motion priors to capture the visual patterns of target motions. To this end, we introduce an automatic retrieval pipeline that selects motion-relevant videos from a large-scale dataset (Wang et al. 2023b) as references. We then apply test-time fine-tuning (Zhao et al. 2023) to learn customized motion priors. We use per-video prompts—rather than a shared one as in prior methods—to improve motion specificity, and learn both priors via LoRA-based tuning (Hu et al. 2021) on specific layers of DiT (Peebles and Xie 2023). In **(3) video generation stage**, we introduce SR3AI, a novel spatial-temporal region-based 3D attention and prior injection module that enables fine-grained control without additional training. Unlike prior methods that support only spatial (Lian et al. 2024; Yang et al. 2024a; Jain et al. 2024) or temporal (Bansal et al. 2024) control, SR3AI leverages frame-level layout plans to enable spatial-temporal control over sequential events, object attributes, trajectories, and spatial relationships. We first encode multiple conditions from the fine-grained plan. SR3AI then computes visual latents for each condition based on its spatial-temporal layout and enforces attention masking so that each condition attends only to its designated region. This ensures precise control and coherent composition of multiple objects and motions. Moreover, we extend this region-based design to inject learned character and motion priors into their corresponding regions in the diffusion model, enabling coherent character and motion customization.

We validate the effectiveness of DREAMRUNNER on two tasks: story-to-video generation and compositional text-to-video generation. For SVG, we collect a story dataset, DreamStorySet, and compare DREAMRUNNER with SoTA methods (VideoDirectGPT (Lin et al. 2023) and VLogger (Zhuang et al. 2024)). DREAMRUNNER achieves a 13.1% relative improvement in character consistency score and an 8.56% gain in text alignment score. It also improves sequential event generation within a single scene, with a 27.2% boost for smoother multi-event transitions. Qualitative results further show strong generalization to multi-character settings. In compositional T2V generation, DREAMRUNNER outperforms baseline methods on T2V-CompBench (Sun et al. 2024) across all dimensions, demonstrating its strength in compositional generation. Notably, despite being based on open-source models (Yang et al. 2024b), DREAMRUNNER achieves the highest scores in dynamic attribute binding and object interaction, along with comparable results in spatial relationships and motion binding to closed-source models, showing our method's potential to bridge the performance gap between open- and closed-source models. In summary, our main contributions include:

- A retrieval-augmented prior learning approach to enhance the synthesis of complex motions.
- A spatiotemporal region-based attention module for coherent composition of multiple objects and sequential events, along with a region-based LoRA injection design for character and sequential motion customization.
- SoTA performance in both compositional T2V and SVG.

2 Related Work

Storytelling Video Generation. Storytelling video generation (Oh et al. 2025; Zheng and Fu 2024; Long et al. 2024) aims to produce multi-scene videos from input scripts. Existing approaches use either high-level LLM planning for story decomposition and generation (Zhuang et al. 2024; He et al. 2023; Lin et al. 2023) or text-to-image keyframe generation then video animation (He et al. 2024; Zhao et al. 2024; Chai et al. 2023). Reference-based customization methods (Ruiz et al. 2023; Gal et al. 2023; Kumari et al. 2023a; Sohn et al. 2023; Ye et al. 2023; Wei et al. 2023b; Li, Li, and Hoi 2023; Chen et al. 2025; Huang et al. 2025; Liu et al. 2025) is used to preserve character identity. Our work targets the video-centric challenge of generating multi-character, motion-rich videos with smooth, natural transitions.

Compositional Generation. Recent advances in diffusion models have enhanced compositional T2V generation by improving coherence, semantic alignment, and user control. Several methods leverage LLMs for scene planning (Lian et al. 2024; Lin et al. 2023; Qu et al. 2023), while others employ regional masks for multi-object control (Jain et al. 2024; Tian et al. 2024; Yu et al. 2024; Wei et al. 2024) or frame-level semantic conditioning (Bansal et al. 2024; Xing et al. 2024). Additionally, LoRA-based compositional techniques integrate diverse concepts within the generation process (Li et al. 2024; Zhong et al. 2024). However, these approaches do not explicitly bind objects to their corresponding actions/events spatial-temporally. Our method ensures fine-grained control over both objects and motions, maintaining a cohesive object-action link throughout the video.

Motion Customization. Motion customization remains a key challenge in video generation. Existing methods either learn pixel-level motions for editing (Zhang et al. 2023; Jeong, Park, and Ye 2023; Lin et al. 2024) or model high-level motion priors (e.g., human/camera) from curated datasets (Wu et al. 2023; Wei et al. 2023a), often requiring test-time LoRA or adapter tuning. Unlike these methods, we retrieve motion-relevant videos from large-scale databases to supply diverse, context-aware motion priors, and use per-video detailed prompts rather than a single prompt to improve motion fidelity and overall quality.

3 Methodology

Task Setup. Storytelling Video Generation focuses on creating multi-scene, character-driven videos based on a given

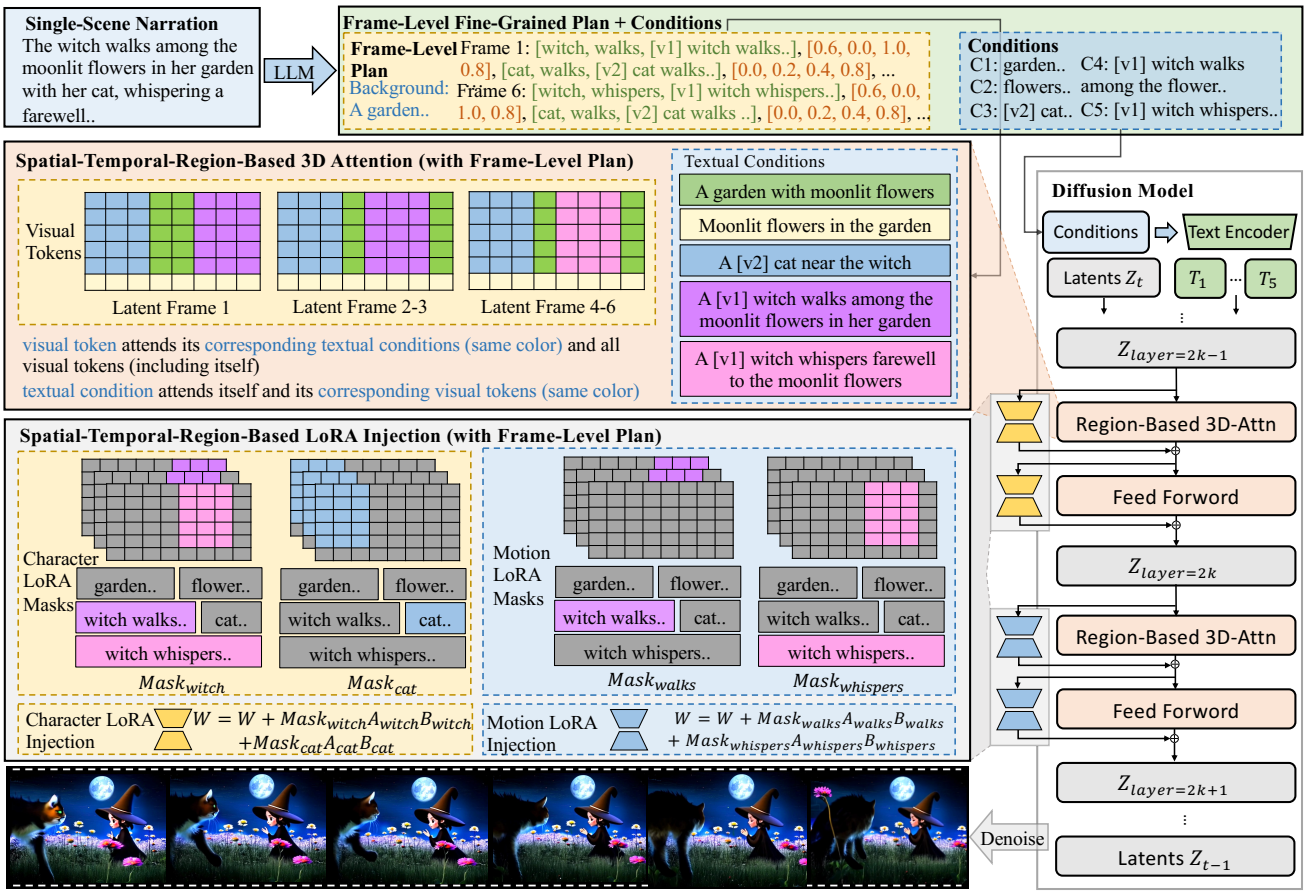


Figure 1: **Implementation details for region-based diffusion.** We extend the vanilla self-attention mechanism to *spatial-temporal-region-based 3D attention* (see upper orange part), which is capable of aligning different regions with their respective text descriptions via region-specific masks. The region-based character and motion LoRAs (see lower yellow and blue parts) are then injected interleavingly to the attention and FFN layers in each transformer block (see the right part). Note that though we resize the visual latents into sequential 2D latent frames for better visualization, they are flattened and concatenated with all conditions when performing region-based attention. Fig. 2 and Appendix A.3 provide example of the region-based attention mask and more details of region-based LoRA injection, respectively.

topic. The characters are defined by reference images (e.g. images of a witch), and the topic is presented as an instructional prompt (e.g. "witch's one day"). The generated videos should align with the given topic and accurately reflect the characteristics and behavior of the characters.

Method Overview. Our approach employs a hierarchical system where an LLM generates event-based scripts across multiple scenes, followed by detailed plans specifying the layout and motion transitions of key objects per scene (Section 3.1). A video diffusion model then synthesizes each scene step by step. We train motion priors from retrieval videos aligned with the LLM-generated plans, sourced from a large-scale video-language database, and character priors using the reference images (Section 3.2). Finally, we inject these priors and detailed plans into the video generation process in a zero-shot manner using our spatial-temporal regional diffusion module SR3AI (Section 3.3).

Base Generation Model. We leverages CogVideoX-2B (Yang et al. 2024b) as the base text-to-video model. CogVideoX-2B employs a DiT-based architecture that integrates full 3D attention, and generates 6-second videos at 8 fps conditioned on input text. In our method, we extend CogVideoX-2B by training character and motion priors in distinct layers (see Sec. 3.2) and by modifying its 3D attention (see Sec. 3.3) for better motion and character binding.

3.1 Generating Dual-Level Plans with LLMs

Story-Level Coarse-Grained Planning. We prompt an LLM (GPT-4o (OpenAI 2024)) to generate 6∼8 character-driven, motion-rich scene descriptions based on the story topic, task requirements, and a single in-context example. Each description follows a structured format: `scene`, `motions`, and `narrations`, where motions are defined first, followed by corresponding event narrations. This sequence forms a high-level plan that guides story progression across scenes, ensuring narrative coherence.

Scene-Level Fine-Grained Planning. After generating a list of single-scene descriptions with narrations, we create detailed, entity-level plans for each latent frame. Each plan consists of an overall background description followed by entity-specific details for each latent frame. As shown in the yellow *Frame-Level Plan* box at the top of Figure 1, the background provides a global scene description (e.g., "*A large garden*"), formatted as `Background: background description`. Entity-level details specify each entity's description, motion (e.g., "*A [v1] witch is walking among the moonlit flowers in her garden*"), and bounding box layout, formatted as: `Frame: [entity name, entity motion, entity description], [x0,y0,x1,y1]`. Here, `[x0,y0,x1,y1]` denotes the top-left and bottom-right corners of the bounding box, with coordinates normalized to $[0, 1]$. Entities without motion are labeled "`none`". When bounding boxes overlap, we prompt the LLM to generate a unified caption that integrates the descriptions of all entities within the overlapping region. Each scene includes plans for six key frames, with each frame guiding one second of video generation (we interpolate key frames to match the #frames of visual latents), resulting in a six-second output using CogVideoX. Detailed prompt templates for both levels' planning are in Appendix I.

3.2 Motion Retrieval and Prior Learning

Retrieving Motion-Related Videos from Database. We employ a retrieval-augmented approach to fine-tune motion priors at test time for complex motion synthesis. Based on motion descriptions generated from the LLM planning, we retrieve relevant videos from a large-scale video database (Wang et al. 2023b). Our retrieval pipeline first uses BM25 (Robertson, Zaragoza et al. 2009) for initial text-based retrieval, followed by attribute-based filtering and clip segmentation via object tracking (Jocher 2020). We then compute semantic similarity scores using CLIP (Radford et al. 2021) and ViCLIP (Wang et al. 2023b) to refine the selection, ensuring high-quality motion-aligned videos (see Appendix A.1 for details). By following this process, we retrieve $4 \sim 20$ video clips per motion, which are then used as reference videos for learning motion priors.

Motion Prior Training. We follow recent motion customization methods (Zhao et al. 2023) with test-time fine-tuning for learning motion priors. Reference videos are used to learn an appearance-debiased temporal LoRAs by injecting it into temporal attention layers with video-specific spatial LoRAs into spatial layers (Wang et al. 2023a; Zer 2023). Only temporal LoRAs are used during inference. Since our approach is based on CogVideoX (Yang et al. 2024b), which employs 3D full attention instead of separate spatial-temporal attention, we manually designate even layers as "spatial" and odd layers as "temporal" to disentangle the learning. We train LoRAs on top-ranked retrieved videos with the backbone frozen. Unlike prior methods that use a single prompt for all retrieved videos, we condition each video with its database caption, which implicitly reduces appearance and background bias and encourages motion-specific learning. More design details are in Appendix A.2.

Subject Prior Learning. We learn the subject's appearance by injecting LoRA modules into the spatial transformer layers. To train these LoRAs, we create videos by repeating reference images multiple times and focus on reconstructing the first frame of the video during training. Notably, the subject priors are learned within spatial LoRAs, while the motion priors are learned within temporal LoRAs. Since their injections target different layers, there is no overlap, effectively avoiding conflicts between multiple LoRAs.

3.3 Spatial-Temporal-Region-Based Diffusion

Region-Based 3D Attention. We build our model on CogVideoX (Yang et al. 2024b), a text-to-video generation model designed on top of a Diffusion Transformer (DiT). CogVideoX 3D full attention, integrating self-attention across all visual latents and the text condition embeddings. We extend this module to enable region-specific conditioning via masking, aligning different regions with their respective text descriptions. Specifically, given a fine-grained plan with N region-specific text descriptions $C_1, C_2, \ldots, C_N$ and corresponding layouts $L_1, L_2, \ldots, L_N$ across frames, we encode each text condition C_i to produce embeddings $T_1, T_2, \ldots, T_N$ (Figure 1 top right). At each attention layer, we identify the visual latents corresponding to each layout L_i in the latent space. We then perform masked self-attention on the concatenation of $T_1, T_2, \ldots, T_N$ and $L_1, L_2, \ldots, L_N$. The self-attention mask is defined as follows: for each region's visual latents L_i, attention is allowed to its corresponding text condition embeddings T_i and all visual latents $L_1, L_2, \ldots, L_N$. Conversely, for each condition embeddings T_i, attention is restricted to itself and its corresponding latents L_i. This design ensures each region is conditioned on its specific textual description while maintaining interactions among visual latents through unmasked attention among $L_1, L_2, \ldots, L_N$. No modifications are made to other modules in the base model, preserving the integrity of its original architecture. A visualization example of such masking strategy is contained in Fig. 2.

Region-Based LoRA Injection. We adopt a similar region-based strategy for injecting LoRA priors into diffusion models. For each LoRA, we identify the corresponding regions of latent tokens based on the associated text description and layout information. LoRA injection is applied exclusively to these regions, ensuring precise alignment between the priors and their designated areas. This enables handling multiple LoRAs simultaneously while avoiding conflicts, preserving the integrity of each injected prior. Appendix A.3 provides details of this with equation derivations, explanations, etc.

4 Experiments

In this section, we first introduce the evaluation datasets and evaluation metrics details in Section 4.1, then compare our DREAMRUNNER with prior methods on story-to-video generation in Section 4.2. Next, we present detailed ablation studies on the necessity of RAG and effectiveness of SR3AI in Section 4.3, and demonstrate the generalizability of our DREAMRUNNER to improve compositional text-to-video generation on T2V-CompBench (Sun et al. 2024) in

Method	Character		Fine-Grained Text		Full Text		Transition	Visual Quality		
	CLIP	DINO	CLIP	ViCLIP	CLIP	ViCLIP	DINO	Aesthetics	Imaging	Smoothness
VideoDirectorGPT (Lin et al. 2023)	54.3	9.5	23.7	21.7	22.4	22.5	63.5	42.3	60.3	94.3
VLogger (Zhuang et al. 2024)	62.5	41.3	23.5	23.1	22.5	22.2	73.6	43.4	61.2	96.2
DREAMRUNNER (Ours)	**70.7** (+13.1%)	**55.1** (+33.4%)	**24.7** (+5.11%)	**23.7** (+2.60%)	**24.2** (+7.56%)	**24.1** (+8.56%)	**93.6** (+27.2%)	**55.4** (+27.6%)	**62.1** (+1.47%)	**98.1** (+1.98%)

Table 1: **Evaluation of story-to-video generation on DreamStorySet.** We compare ours with VideoDirectorGPT and VLogger on character consistency (CLIP and DINO scores), text instructions following and full prompt adherence (CLIP and ViCLIP scores), and event transitions smoothness (DINO score). Our relative improvement over VLogger is highlighted in blue.

Model	Consist-attr	Dynamic-attr	Spatial	Motion	Action	Interaction
Gen-3 (gen 2024)	0.7045	0.2078	0.5533	*0.3111*	0.6280	0.7900
Dreamina (Dre 2024)	*0.8220*	0.2114	0.6083	0.2391	0.6660	0.8175
PixVerse (Pix 2024)	0.7370	0.1738	0.5874	0.2178	*0.6960*	0.8275
Kling (kli 2024)	0.8045	*0.2256*	*0.6150*	0.2448	0.6460	*0.8475*
Open-Sora-Plan v1.1.0 (Lab and etc. 2024)	0.7413	0.1770	0.5587	0.2187	**0.6780**	0.7275
VideoTetris (Tian et al. 2024)	0.7125	0.2066	0.5148	0.2204	0.5280	0.7600
CogVideoX-5B (Yang et al. 2024b)	0.7232	0.2250	0.5845	0.2551	0.6040	0.7995
CogVideoX-5B+SR3A (Ours)	**0.7650** (+5.8%)	**0.2832** (+25.9%)	**0.6875** (+17.5%)	**0.3041** (+19.2%)	0.6340 (+5.0%)	**0.8725** (+9.1%)

Table 2: **T2V-CompBench evaluation results**. Best/2nd best scores for open-source models are **bold**/underline. Gray : closed-source models; Yellow : best closed-source score.

RAG	SR3AI	Fine-Grained Text		Full Text		Trans.	Quality		
		CLIP	ViCLIP	CLIP	ViCLIP	DINO	Asth.	Img.	Smth.
×	×	23.8	22.5	22.2	22.1	87.1	54.3	61.3	94.3
×	✓	23.9	23.1	23.5	22.4	92.5	55.4	61.9	98.0
✓	×	24.7	23.5	23.9	24.0	84.6	**55.6**	61.9	98.1
✓	✓	**24.7**	**23.7**	**24.2**	**24.1**	**93.6**	55.4	**62.1**	**98.1**

Table 3: **Ablation studies for the effectiveness of RAG and SR3AI in DREAMRUNNER.** Our full model achieves the best text-following ability and event transition smoothness.

Section 4.5. We show the effectiveness of RAG for learning the motion prior on a comprehensive motion dataset in Section 4.6. Lastly, we present qualitative comparison between our DREAMRUNNER and previous methods in Section 4.4.

4.1 Experimental Setups

Evaluation Datasets. We evaluate DREAMRUNNER on two tasks: (1) story-to-video generation, and (2) compositional text-to-video generation. The first task focuses on the model's ability to follow the text closely while maintaining character and scene consistency throughout the story. The second task assesses various aspects of compositionality in video generation. For (1) story-to-video generation, we collect and introduce a new benchmark dataset, DreamStorySet. Specifically, we collect 10 characters, including 6 from existing customization datasets (CustomConcept101 (Kumari et al. 2023b) and Dreambooth (Ruiz et al. 2023)), and 4 with generation models (FLUX (flu 2024)). (featuring two motions per scene) and three multi-character stories (featuring two or three motions per scene). Each story comprises 5 to 8 scenes, incorporating a total of 64 diverse motions throughout. We focus on single-character stories for quantitative evaluation of SVG models and reserve multi-character stories for qualitative evaluation. For (2) compositional text-to-video generation, we use the T2V-CompBench (Sun et al. 2024) to benchmark the performance of DREAMRUNNER, where we select six dimensions except numeracy.

Evaluation Metrics. We evaluate storytelling videos on character consistency (Frame-to-Reference CLIP/DINO), narration- and scene-level text alignment (Image/Video-to-Text CLIP/ViCLIP), and transition smoothness (Frame-to-Frame DINO). For visual quality, we adopt three representative VBench metrics—aesthetic quality, imaging quality, and video smoothness—from VBench (Huang et al. 2023; Li et al. 2023), with full results and metric details in the Appendix. For compositional T2V, we follow the evaluation protocol of T2V-ComBench (Sun et al. 2024).

Implementations. We use CogVideoX-2B as our base model for SVG. Test-time-finetuning each prior requires 5min on a single A6000 GPU. For compositional T2V we evaluate with both CogVideoX-2B and CogVideoX-5B.

4.2 Story-To-Video Generation Evaluation

We compare DREAMRUNNER with prior SoTAs (VideoDirectorGPT (Lin et al. 2023) and VLogger (Zhuang et al. 2024)) on our DreamStorySet dataset for story-to-video generation. For fairness, each scene narration is split into two single-motion descriptions, with corresponding videos later merged into a single-scene video. As shown in Table 1, DREAMRUNNER improves CLIP/DINO scores by 13.1%/33.4% over VLogger, demonstrating the effectiveness of our learned subject prior and region-based

Figure 2: **Visualization of spatial-temporal region-based 3D attention mask**. Different text colors represent different conditions, while the white region indicates masked areas. For simplicity, we reduce each condition to two words, each frame to three segments, and display only three conditions and two frames in the figure. In practice, conditions can be longer and more numerous, frames can have more segments, and there are 12 latent frames in total.

LoRA injection for character consistency. To evaluate text-following capability, we assess both full-prompt adherence and fine-grained event alignment. DREAMRUNNER improves CLIP/ViCLIP scores consistently on both settings, showing superior alignment with both full-scene and fine-grained event descriptions. For transition quality, we compute the DINO-based transition score to measure scene and event consistency. DREAMRUNNER improves transitions by 27.2% over VLogger, highlighting the effectiveness of SR3AI in generating sequential events in a single scene. Lastly, we evaluate visual quality across aesthetic quality, imaging quality, and motion smoothness. DREAMRUNNER enhances aesthetics while slightly improving the other two, demonstrating its capability to generate high-quality videos adhere to complex scene descriptions with smooth event transitions. We provide three additional quality scores from VBench and qualitative examples in the Appendix.

4.3 Ablation Studies

In this section, we demonstrate the effectiveness of RAG-based video retrieval for motion prior learning and SR3AI for fine-grained object–motion control. As shown in Table 3, SR3AI (2nd row) substantially improves event-transition smoothness, visual quality, and text alignment, as its region-based decomposition enables more effective multi-object, multi-event binding. Adding retrieval-augmented motion priors (3rd row) further boosts video–text similarity for both fine-grained and full-prompt alignment. Combining both components (last row) achieves the best performance across transitions, alignment, and visual quality. Additional ablations on the RAG pipeline, layer-separation strategy, and computational cost are provided in Appendix B.

Method	CLIP	ViCLIP
CogVideoX-2B	23.39	20.84
CogVideoX-2B + RAG (w/ single prompt for all videos)	24.01	22.02
CogVideoX-2B + RAG (w/ per-video prompt)	**24.67**	**23.04**

Table 4: Effect of RAG and per-video prompt.

4.4 Qualitative Ablations and Comparisons

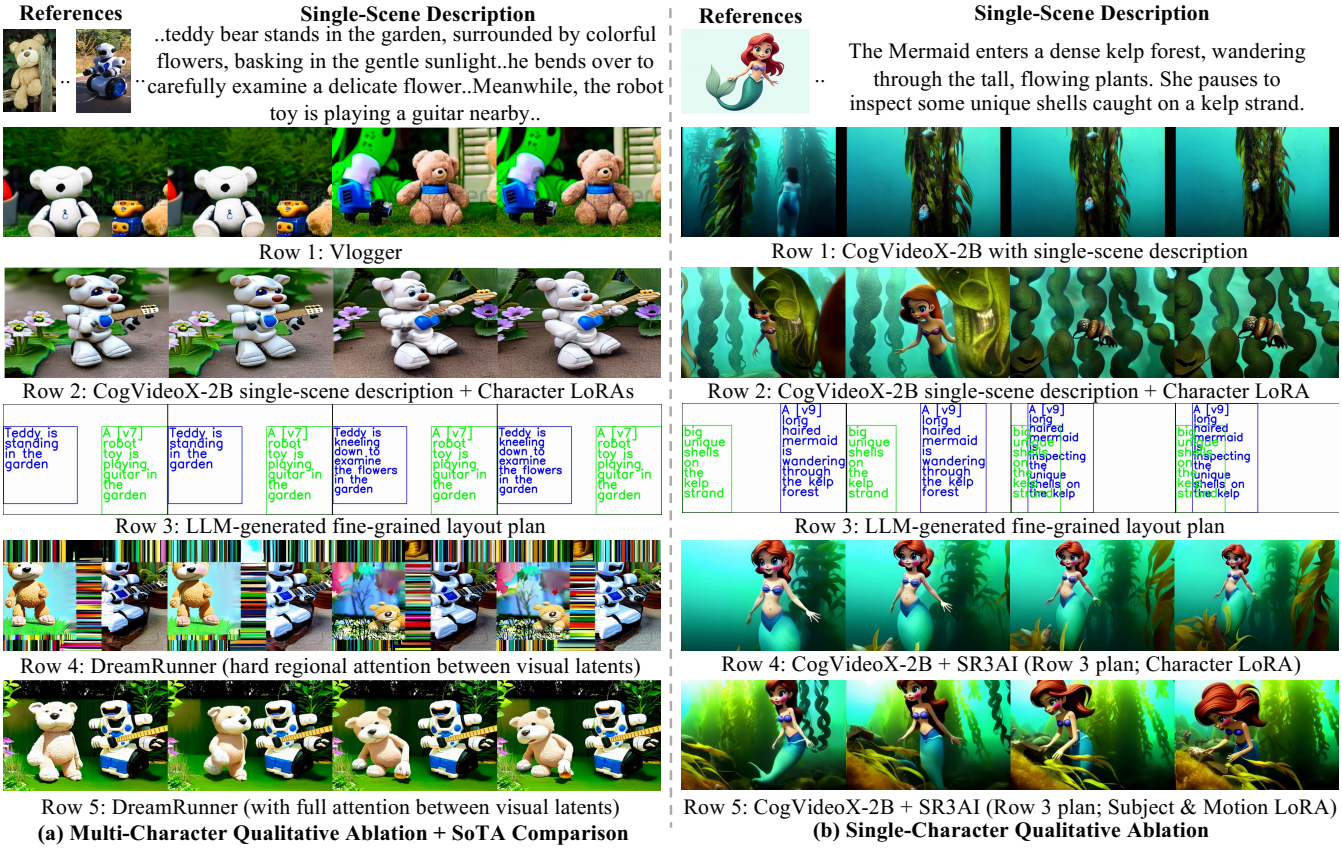
We provide qualitative comparisons and ablations in Fig. 3.

We compare DREAMRUNNER with VLogger(Zhuang et al. 2024) for multi-character generation and analyze the effects of region-based attention and LoRA injection (Fig. 3(a)). Our method (row 5) generates coherent multi-character composition and motions, outperforming VLogger (row 1). Using CogVideoX-2B with character LoRAs injected globally (row 2) results in interference, producing a robot-like teddy bear and blurry compositions, highlighting the necessity of region-based LoRA injection. To study attention design, we ablate our region-based attention by comparing hard-regional attention (row 4) and full attention (row 5). While hard attention strictly follows the layout plan (row 3), it limits spatial-temporal continuity due to lack of inter-region interaction. In contrast, full attention enables smooth transitions while maintaining spatial-region constraints, supporting high-quality multi-character customization.

In Fig. 3(b), we present single-character ablations. Using the base CogVideoX-2B with only scene-level text (row 1) leads to vague character/background and missing actions. Injecting global character LoRA (row 2) improves character appearance but still fails on action and transition quality, and degrades background fidelity (e.g., cartoonish kelp forest). Applying SR3AI with layout plans (row 4) improves trajectory control and preserves background fidelity through localized injection, but motion remains limited. Injecting RAG-learned motion priors (row 5) enables clear, fine-grained motion execution (e.g., the mermaid stooping and interacting with kelp), demonstrating the benefit of our motion prior learning and injection strategy. Overall, our full model combines coherent composition with strong motion quality, showing the effectiveness of SR3AI and retrieval-augmented prior learning for complex, multi-entity video generation.

4.5 Compositional T2V Generalization

In this section, we demonstrate how our spatial-temporal region-based attention module (SR3A) enhances compositional T2V, as evaluated on T2V-CompBench (Sun et al. 2024). We use SR3A (no LoRA injection) as no customization is required. Given a prompt, we use GPT-4o to generate layout plans, and SR3A ensures coherent composition of objects and events. As shown in Table 2, SR3A significantly improves both CogVideoX-2B and CogVideoX-



(a) Multi-Character Qualitative Ablation + SoTA Comparison **(b) Single-Character Qualitative Ablation**

Figure 3: **Qualitative comparison and ablations of DREAMRUNNER on SVG.** In (a) multi-character example, DREAMRUNNER produces significantly better character consistency compared to other strong baselines, while others fail to maintain object consistency (e.g., VLogger), or fail to generate multiple objects ((a) Row 2,4). In (b) single-character setting, integrating SR3AI and locally-injected priors consistently improve overall quality, complex motion synthesis and coherent composition. Note that in the overlapped regions in (b) row 3, the caption is a merge of the two. For cleaner visualization, we don't show it here.

5B (Yang et al. 2024b) across all categories. Specifically, it boosts dynamic attribute binding by over 25%, spatial binding by over 15%, and motion binding by at least 10%, highlighting SR3A's ability to maintain coherent multi-object compositions, trajectories, and sequential events. It also improves scores on other fine-grained aspects, demonstrating strong control capabilities. Notably, DREAMRUNNER built on CogVideoX-5B achieves SoTA results in five dimensions among open-source models, and surpasses all closed-source models in dynamic attribute binding, spatial binding, and object interactions, highlighting its ability to close the open-closed source model gap and adapt to stronger base models. Qualitative examples per dimension are in Appendix E.

4.6 Effect of RAG and Per-Caption Prompt

We investigate the effectiveness of retrieval-augmented test-time fine-tuning and our per-caption prompt design for learning an enhanced motion prior. Specifically, for each motion in the 64-motion set, we use GPT-4o to generate six prompts and evaluate the average CLIP/ViCLIP scores. As shown in Table 4, applying our approach to CogVideoX-2B improves both scores, with the significant ViCLIP gain indicating better story-video alignment and enhanced motion accuracy. Rows 2–3 further show that per-video prompts outperform single prompts, suggesting that video-specific conditioning helps the model ignore unrelated visual cues and better capture motion-specific patterns. These results confirm that RAG effectively retrieves motion-relevant videos and facilitates the learning of more accurate motion priors.

5 Conclusion

In this work, we present DREAMRUNNER, a novel framework for story-to-video generation. Specifically, DREAMRUNNER utilizes a LLM to structure a hierarchical video plan, then introduces retrieval-augmented test-time adaptation to capture target motion priors, and finally generates videos using a novel region-based 3D attention and prior injection module for coherent composition. Experiments on both story-to-video and compositional T2V generation benchmarks show that DREAMRUNNER outperforms strong baselines and SoTAs in tackling fine-grained complex motions, maintaining multi-scene consistency of multiple objects, and ensuring seamless scene transitions.

Acknowledgments

This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL2112635, DARPA Machine Commonsense (MCS) Grant N66001-19-2-4031, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, Accelerate Foundation Models Research program, and a Bloomberg Data Science PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

References

2023. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w.

2024. Dreamina. https://dreamina.capcut.com/ai-tool/platform.

2024. Flux. https://github.com/black-forest-labs/flux.

2024. Gen-3. https://runwayml.com/blog/introducing-gen-3-alpha/.

2024. Kling. https://kling.kuaishou.com/.

2024. PixVerse. https://app.pixverse.ai.

Bansal, H.; Bitton, Y.; Yarom, M.; Szpektor, I.; Grover, A.; and Chang, K.-W. 2024. TALC: Time-Aligned Captions for Multi-Scene Text-to-Video Generation. *arXiv preprint arXiv:2405.04682*.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Fang, Y.; Lee, K. S.; Skorokhodov, I.; Aberman, K.; Zhu, J.-Y.; Yang, M.-H.; and Tulyakov, S. 2025. Multi-subject open-set personalization in video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6099–6110.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

He, H.; Yang, H.; Tuo, Z.; Zhou, Y.; Wang, Q.; Zhang, Y.; Liu, Z.; Huang, W.; Chao, H.; and Yin, J. 2024. DreamStory: Open-Domain Story Visualization by LLM-Guided Multi-Subject Consistent Diffusion. *arXiv preprint arXiv:2407.12899*.

He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Yuan, Z.; Liu, Q.; Wang, Q.; Wang, X.; Zhang, R.; Wan, P.; Zhang, D.; and Gai, K. 2025. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2023. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*.

Jain, Y.; Nasery, A.; Vineet, V.; and Behl, H. 2024. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8079–8088.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2023. VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models. *arXiv preprint arXiv:2312.00845*.

Jocher, G. 2020. YOLOv5 by Ultralytics.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J. 2023a. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023b. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Lab, P.-Y.; and etc., T. A. 2024. Open-Sora-Plan.

Li, D.; Li, J.; and Hoi, S. C. H. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, J.; Cho, J.; Sung, Y.-L.; Yoon, J.; and Bansal, M. 2024. SELMA: Learning and Merging Skill-Specific Text-to-Image Experts with Auto-Generated Data. *arXiv preprint arXiv:2403.06952*.

Li, Z.; Zhu, Z.-L.; Han, L.-H.; Hou, Q.; Guo, C.-L.; and Cheng, M.-M. 2023. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9801–9810.

Lian, L.; Shi, B.; Yala, A.; Darrell, T.; and Li, B. 2024. LLM-grounded Video Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Lin, H.; Cho, J.; Zala, A.; and Bansal, M. 2024. Ctrl-Adapter: An Efficient and Versatile Framework for Adapting Diverse Controls to Any Diffusion Model. *arXiv preprint arXiv:2404.09967*.

Lin, H.; Zala, A.; Cho, J.; and Bansal, M. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*.

Liu, L.; Ma, T.; Li, B.; Chen, Z.; Liu, J.; Li, G.; Zhou, S.; He, Q.; and Wu, X. 2025. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*.

Long, F.; Qiu, Z.; Yao, T.; and Mei, T. 2024. VideoStudio: Generating Consistent-Content and Multi-Scene Videos. arXiv:2401.01256.

Oh, G.; Jeong, J.; Kim, S.; Byeon, W.; Kim, J.; Kim, S.; and Kim, S. 2025. MEVG: Multi-event Video Generation with Text-to-Video Models. In *European Conference on Computer Vision*, 401–418. Springer.

OpenAI. 2024. Hello, GPT-4 Turbo.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Qing, Z.; Zhang, S.; Wang, J.; Wang, X.; Wei, Y.; Zhang, Y.; Gao, C.; and Sang, N. 2024. Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qu, L.; Wu, S.; Fei, H.; Nie, L.; and Chua, T.-S. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 643–654.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Sohn, K.; Jiang, L.; Barber, J.; Lee, K.; Ruiz, N.; Krishnan, D.; Chang, H.; Li, Y.; Essa, I.; Rubinstein, M.; Hao, Y.; Entis, G.; Blok, I.; and Chin, D. C. 2023. StyleDrop: Text-to-Image Synthesis of Any Style. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; and Liu, X. 2024. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. *arXiv preprint arXiv:2407.14505*.

Tian, Y.; Yang, L.; Yang, H.; Gao, Y.; Deng, Y.; Chen, J.; Wang, X.; Yu, Z.; Tao, X.; Wan, P.; Zhang, D.; and Cui, B. 2024. VideoTetris: Towards Compositional Text-to-Video Generation. arXiv:2406.04277.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2023a. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*.

Wei, Y.; Zhang, S.; Yuan, H.; Wang, X.; Qiu, H.; Zhao, R.; Feng, Y.; Liu, F.; Huang, Z.; Ye, J.; et al. 2024. DreamVideo-2: Zero-Shot Subject-Driven Video Customization with Precise Motion Control. *arXiv preprint arXiv:2410.13830*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023b. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Wu, R.; Chen, L.; Yang, T.; Guo, C.; Li, C.; and Zhang, X. 2023. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*.

Xing, Z.; Dai, Q.; Weng, Z.; Wu, Z.; and Jiang, Y.-G. 2024. AID: Adapting Image2Video Diffusion Models for Instruction-guided Video Prediction. *arXiv preprint arXiv:2406.06465*.

Yang, S.; Hou, L.; Huang, H.; Ma, C.; Wan, P.; Zhang, D.; Chen, X.; and Liao, J. 2024a. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24)*, 12. New York, NY, USA: ACM.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR*, abs/2308.06721.

Yu, S.; Fang, J. Z.; Zheng, S.; Sigurdsson, G.; Ordonez, V.; Piramuthu, R.; and Bansal, M. 2024. Zero-shot controllable image-to-video animation via motion decomposition. *ACM, Multimedia*.

Zhang, Y.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. MotionCrafter: One-Shot Motion Customization of Diffusion Models. *arXiv preprint arXiv:2312.05288*.

Zhao, C.; Liu, M.; Wang, W.; Yuan, J.; Chen, H.; Zhang, B.; and Shen, C. 2024. MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequence. *arXiv preprint arXiv:2407.16655*.

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.; Wu, W.; Keppo, J.; and Shou, M. Z. 2023. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. *arXiv preprint arXiv:2310.08465*.

Zheng, S.; and Fu, Y. 2024. TemporalStory: Enhancing Consistency in Story Visualization using Spatial-Temporal Attention. *arXiv preprint arXiv:2407.09774*.

Zhong, M.; Shen, Y.; Wang, S.; Lu, Y.; Jiao, Y.; Ouyang, S.; Yu, D.; Han, J.; and Chen, W. 2024. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*.

Zhuang, S.; Li, K.; Chen, X.; Wang, Y.; Liu, Z.; Qiao, Y.; and Wang, Y. 2024. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8806–8817.