VIDEODIRECTORGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning

Han Lin **Abhay Zala** **Jaemin Cho** **Mohit Bansal**
UNC Chapel Hill
{hanlincs, aszala, jmincho, mbansal}@cs.unc.edu
videodirectorgpt.github.io

Abstract

In this paper, we propose VIDEODIRECTORGPT, a novel framework for consistent multi-scene video generation that uses the knowledge of LLMs for video content planning and grounded video generation. Specifically, given a single text prompt, we first ask our video planner LLM (GPT-4) to expand it into a '*video plan*', which includes the scene descriptions, the entities with their respective layouts, the background for each scene, and consistency groupings of the entities. Next, guided by this *video plan*, our video generator, named Layout2Vid, has explicit control over spatial layouts and can maintain temporal consistency of entities across multiple scenes, while being trained only with image-level annotations. Our experiments demonstrate that our proposed VIDEODIRECTORGPT framework substantially improves layout and movement control in both single- and multi-scene video generation and can generate multi-scene videos with consistency, while achieving competitive performance with SOTAs in open-domain single-scene T2V generation.

1 Introduction

Text-to-video (T2V) generation has achieved rapid progress following the success of text-to-image (T2I) generation. Most works in T2V generation focus on producing short videos from the given text prompts (Wang et al., 2023; He et al., 2022; Ho et al., 2022; Singer et al., 2023; Zhou et al., 2022). Recent studies on long video generation (Blattmann et al., 2023; Yin et al., 2023; Villegas et al., 2023; He et al., 2023) aim at generating long videos of a few minutes with holistic visual consistency. Although these works could generate longer videos, the generated videos often display the continuation or repetitive patterns of a single action (e.g., driving a car) instead of transitions and dynamics of multiple changing actions/events (e.g., five steps about how to make caraway cakes). Meanwhile, large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b; Chowdhery et al., 2022) have demonstrated their capability in generating layouts and programs to control visual modules (Dídac et al., 2023; Gupta and Kembhavi, 2023), especially image generation models (Cho et al., 2023; Feng et al., 2023). This raises an interesting question: *Can we leverage the knowledge embedded in these LLMs for planning consistent multi-scene video generation?*

In this work, we introduce VIDEODIRECTORGPT, a novel framework for consistent multi-scene video generation. VIDEODIRECTORGPT decomposes the T2V generation task into two stages: **video planning** and **video generation**. In the first stage, video planning, we employ an LLM (e.g., GPT-4 (OpenAI, 2023)) as a video planner to generate a ***video plan***, a multi-component video script with multiple scenes to guide the downstream video synthesis process. Our *video plan* consists of four components: (1) multi-scene descriptions, (2) entities (names and their 2D bounding boxes), (3) background, and (4) consistency groupings (scene indices for each entity indicating where they should remain visually consistent). We generate the *video plan* in two steps by prompting an LLM with different in-context examples. In the first step, we expand a single text prompt into multi-step scene descriptions with an LLM, where each scene is described with a text description, a list of entities, and a background (see Fig. 1 blue part for details). In the second step, we expand the detailed layouts of each scene with an LLM by generating the bounding boxes of the entities per frame, given the list of entities and scene description. This overall '*video plan*' guides the downstream video generation.

In the second stage, video generation, we introduce Layout2Vid, a grounded video generation module to render videos based on the generated *video plan* (see yellow part of Fig. 1). For the grounded video generation module, we build upon

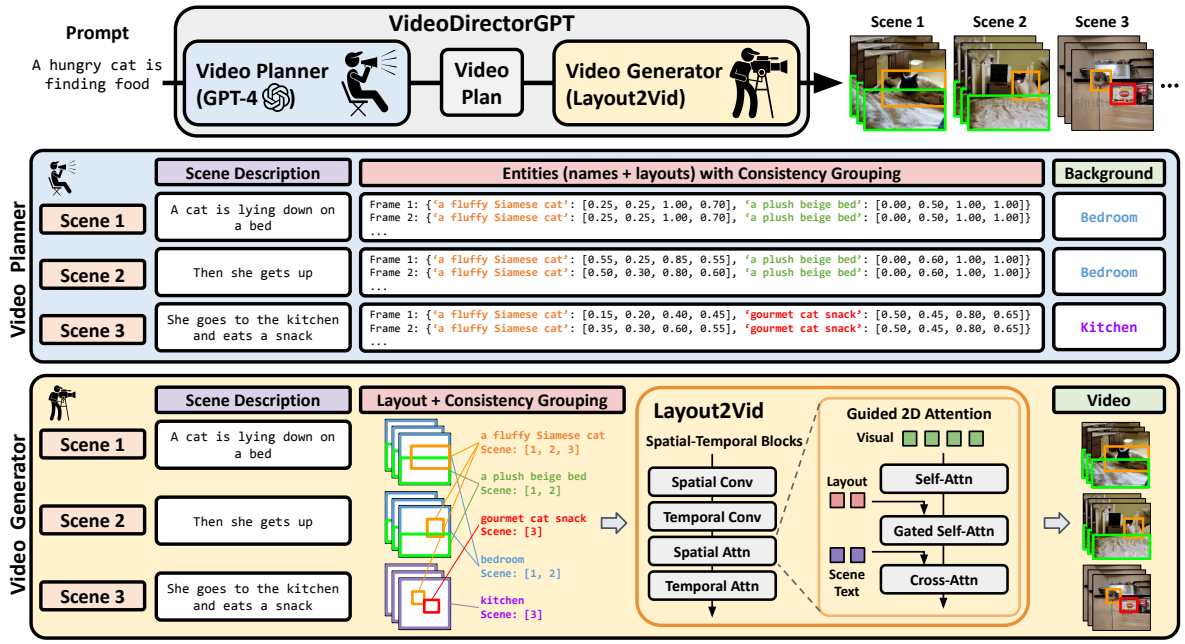

Figure 1: **Detailed illustration of VIDEODIRECTORGPT.** In the first stage, we employ the **LLM as a video planner** to craft a ***video plan***, which provides an overarching plot for multi-scene videos, which consists of scene descriptions, entities/background, and consistency groups. In the second stage, we utilize **Layout2Vid**, a grounded video generation module, to render videos based on the *video plan*. The guided 2D attention of Layout2Vid ensures spatial control, as well as temporal consistency by using identical embeddings to represent the same entities and backgrounds across scenes.

ModelScopeT2V (Wang et al., 2023), an off-the-shelf T2V generation model, by freezing its original parameters and adding spatial/consistency control of entities through a small set of trainable parameters (13% of total parameters) through the gated-attention module (Li et al., 2023). This enables our Layout2Vid to be trained solely on layout-annotated images, thus bypassing the need for expensive training on annotated video datasets. To preserve the identity of entities across scenes, we use shared representations for the entities within the same consistency group. We also propose to use a joint image+text embedding as entity grounding conditions which we find more effective than the existing text-only approaches (Li et al., 2023) in entity identity preservation (see appendix). Overall, our Layout2Vid avoids expensive video-level training, improves the object layout and movement control, and preserves objects temporal consistency.

We conduct experiments on both single-scene and multi-scene video generation. Experiments show that our VIDEODIRECTORGPT demonstrates better layout skills (object, count, spatial, scale) and object movement control, capable of generating multi-scene videos with visual consistency across scenes, and competitive with SOTAs on single-scene open-domain T2V generation. Detailed ablation studies, including dynamic adjustment of layout control strength and video generation with user-provided images, confirm the effectiveness and capacity of our framework.

Our main contributions are as follows:

- A new T2V generation framework VIDEODIRECTORGPT with two stages: video content planning and grounded video generation, which is capable of generating a multi-scene video from a single text prompt.
- We employ LLMs to generate a multi-component '*video plan*' which consists of detailed scene descriptions, entity layouts, and entity consistency groupings to guide downstream video generation.
- We introduce Layout2Vid, a novel grounded video generation module, which brings together image/text-based layout control ability and entity-level temporal consistency. Our Layout2Vid can be trained efficiently using only image-level layout annotations.
- Empirical results demonstrate that our framework can accurately control object layouts and movements, and generate temporally consistent multi-scene videos.

References

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Visual programming for text-to-image generation and evaluation. In *NeurIPS*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Surís Dídac, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

OpenAI. 2023. Gpt-4 technical report.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2023. Make-a-video: Text-to-video generation without text-video data. In *ICLR*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual description. In *ICLR*.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. 2023. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1309–1320, Toronto, Canada. Association for Computational Linguistics.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.