Multi-sentence Video Grounding for Long Video Generation

Wei Feng[1], Xin Wang[1,2,*], Hong Chen[1], Zeyang Zhang[1], Wenwu Zhu[1,2,*]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University
{fw22,h-chen20,zy-zhang20}@mails.tsinghua.edu.cn,{xin_wang,wwzhu}@tsinghua.edu.cn

***Abstract*—Video generation has witnessed great success recently, but its application in generating long videos still remains challenging due to the difficulty in maintaining the temporal consistency of generated videos and the high memory cost during generation. To tackle the problems, in this paper, we propose a brave and new idea of Multi-sentence Video Grounding for Long Video Generation, connecting the massive video moment retrieval to the video generation task for the first time, providing a new paradigm for long video generation. The method of our work can be summarized as three steps: (i) We design sequential scene text prompts as the queries for video grounding, utilizing the massive video moment retrieval to search for video moment segments that meet the text requirements in the video database. (ii) Based on the source frames of retrieved video moment segments, we adopt video editing methods to create new video content while preserving the temporal consistency of the retrieved video. Since editing can be conducted segment by segment and even frame by frame, it largely reduces memory costs. (iii) We also attempt video morphing and personalized generation methods to improve the subject consistency of long video generation, providing ablation experimental results for the subtasks of long video generation. Our approach seamlessly extends the development in image/video editing, video morphing and personalized generation, and video grounding to the long video generation, offering effective solutions for generating long videos at a low memory cost.**

***Index Terms*—Video Generation, Video Grounding**

I. Introduction

Video generation has made significant progress in recent years, demonstrating the incredible ability to generate multi-media content. The main existing works of video generation focus on developing generative models [1], which can be divided into diffusion-based models [2] and non-diffusion-based models, such as VQGAN [3] . Yin et al. proposed NUWA-XL [4] by applying Diffusion over Diffusion method for long video generation. In addition, some works strengthened temporal information by combining generative diffusion models or VQGAN models with the transformers to generate long videos of up to 1 minute [2].

However, there are still many limitations in the generation of long videos. The first issue is that the generated video content often overlooks some physical laws of real-world knowledge (such as chair running). In addition, the overall consistency of the generated video is lower than that of the real video due to unnatural transitions between frames. Last but not least, the longer the video is generated, the higher GPU memory cost would be required.

To address these challenges, we propose a brave and new idea named grounding-based video generation, which applies the multi-sentence video grounding method for long video generation. This idea shares a similar spirit with the retrieval-augmented generation [5] in large language models. To begin with, based on the video grounding technique, we can obtain several moments of different videos from our video database that match target text queries to provide video generation tasks with guided source video segments that follow physical rules and remain highly consistent. Subsequently, based on the retrieved video segments, we adopt the video editing method to create new content in the video segments, such as changing the subject or background. Additionally, we combine the retrieved video segments with a unified subject or style through video editing, and achieve long video generation while ensuring overall consistency and adherence to the physical laws of the generated video content. Meanwhile, since video editing can be conducted segment by segment and even frame by frame, our work maintains a relatively low level of GPU memory cost, making it possible for the public to generate long videos. Extensive experiments show that our proposed method can generate long videos with better consistency. With a larger video corpus and more advanced video grounding methods, our proposed method can work as a powerful tool for long video generation.

To summarize, we make the following contributions:

- To the best of our knowledge, this is the first work to study the feasibility of leveraging the multi-sentence video grounding for long video generation, which we believe will inspire a lot of future work.
- We propose the Multi-sentence Video Grounding-based Long Video Generation framework, consisting of i) a massive video moment retrieval model capable of locating suitable video segments for the text prompts, ii) a video editor that creates new content for the video segments while preserving temporal consistency and iii) a video personalization and morphing scheduler that enables customized video generation and smooth transition between generated videos.
- We conduct experiments on various video editing and

*Corresponding authors: Xin Wang and Wenwu Zhu.

video personalization methods, demonstrating the feasibility of retrieval augmentation to improve the continuity and diversity of generated long videos through the video grounding method.
- We conduct ablation analysis under different video editing methods and the application of video morphing and personalization, providing importing references for improving the performance of long video generation.

II. Related Work

A. Video Grounding

Video grounding aims to locate the starting and ending times of a given segment target from a video [6], which is a popular computer vision task and has drawn much attention over the past few years [7], [8]. The early-stage video grounding task mainly focused on searching for target segments from a single video, which limited its ability to obtain information from the entire video pool. Therefore, tasks such as video corpus moment retrieval (VCMR) [9], [10] and massive video moment retrieval (MVMR) [11] have emerged in the field of computer vision, which focuses on finding correct positive video segments associated with the VG query from the video pools. This task requires the model to distinguish positive from massive negative videos.

B. Long Video Generation and Video Editing

Existing work has demonstrated impressive capabilities in generating high-quality images and short videos [12]. By introducing the transformer architecture to enhance temporal understanding and reasoning ability, some work has been able to generate long videos [13]. With the rapid development of Diffusion Models (DM) such as stable diffusion, William et al. [2] proposed Diffusion Transformers (DiTs), a simple transformer-based backbone for diffusion models that outperform prior U-Net models. Given the promising scaling results of DiTs, OpenAI proposes Sora, presenting powerful capabilities to generate long videos and simulate the physical world. These works, however, all have limitations when addressing long video generation tasks. On the one hand, some generated videos may violate physical knowledge or have poor overall video consistency, which is caused by both inadequate prompt understanding of input and the limitations of the generation algorithm, leading to insufficient modeling of physical laws. On the other hand, to generate a single long video at once, the longer the video is generated, the more computational resources are needed for generated models to consider the overall temporal consistency.

Benefiting from the rapid development of diffusion models in image and video generation, many zero-shot video editing methods have been proposed [14], [15], [16], which apply the pre-trained image diffusion model to transform an input source video into a new video. The critical problem of video editing is to maintain the visual motion and temporal consistency between the generated video and the source video. To address these problems, some works introduce additional spatial conditioning controls or internal features to keep motion consistency between the generated video images and the source video images. Pix2Video [14] uses self-attention feature injection to propagate the changes to the future frames. LOVECon [15] applies ControlNet [17] using condition controls such as edges, depth, segmentation, and human pose for text-driven training-free long video editing. In addition, other methods are proposed to improve the temporal coherency in generation by considering the relationship between source video frames. VidToMe [16] unifies and compresses internal features of diffusion by merging tokens across video frames, balancing the performance of short-term video continuity and long-term consistency of video editing.

III. Method

Long video generation task requires generating a consistent and diverse video of at least one minute conditioned on a story or a sequence of prompts $(p_1, p_2, ..., p_n)$. As shown in Figure 1, the method of Multi-sentence Video Grounding for Long Video Generation can be decomposed into the following steps. Given a sequence with several target queries $(q_1, q_2, ..., q_n)$, each query in the format of *'A person does something...'*, our first step is to input them into the Massive Video Moment retrieval model to search for video time segments that meet the requirements in the video database and filter a sequence of videos $V_1, V_2, ..., V_n = Grounding(q_1, q_2, ..., q_n)$. Secondly, we will use their modified queries $(q'_1, q'_2, ..., q'_n)$ in the format of *'**Customized subject** does something...**in a customized scenario**'* and the video editing method to edit each grounding video segment into video content $V' = Editing(V, q')$ with a unified subject we want to customize and change the background as we expected, forming several story segments $V'_1, V'_2, ..., V'_n$ of a continuous long video. In addition, we also attempt video morphing approaches for the combination of generated video segments, and personalized generation methods to improve the subject consistency of long video generation.

A. Multi-sentence Video Moment Grounding

Given a sequence with several target queries $(q_1, q_2, ..., q_n)$ in the format of *'A person does something...'*, we first tokenize each query q and then perform global average pooling over all the tokens to transform them into text feature$f^q \in R^d$. We hope to find video clips from a video pool that match the text feature through the video grounding method.

Each video in the database would be divided into N video clips and used by a pre-trained visual model for feature extraction. Utilizing these features through FC layer dimensionality reduction and max pooling, we constructed a 2D temporary moment feature map $F \in R^{N \times N \times d}$. After each query and video moment is encoded to a joint visual-text space, we apply Reliable Mutual Matching Network (RMMN) [11] to solve the Multi-sentence Video grounding task. The matching score between the query and the video moment representations could be computed through cosine similarity between them: $f^q_{mm} = W_{mm} f^q + b_{mm}$ and $\forall f^v \in F, s^{mm} = f^{v^T}_{mm} f^q_{mm}$, where W_{mm} and b_{mm} are learnable and the embeddings $||f^v_{mm}|| = ||f^q_{mm}|| = 1$ through a l_2-normalization layer. The

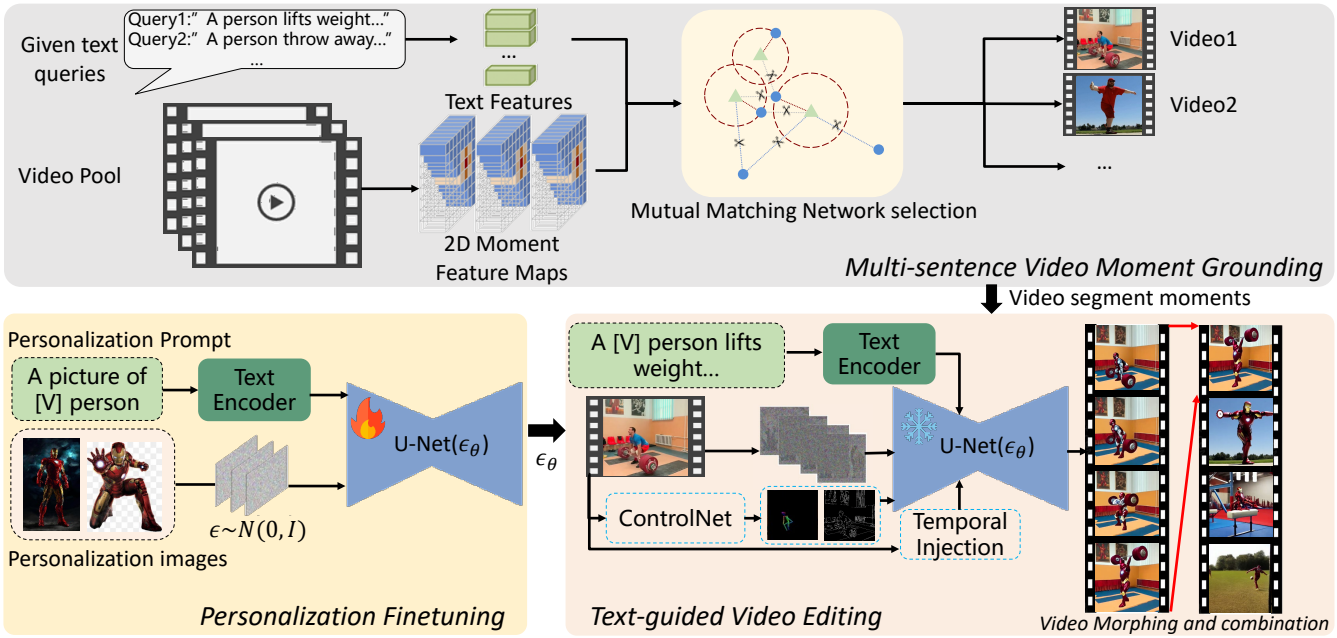

Fig. 1. Framework of Multi-sentence Video Grounding for Long Video Generation. In the stage of Multi-sentence Video Moment Grounding, we input a sentence of queries $(q_1, q_2, ..., q_n)$ and obtain their corresponding video segments $V_1, V_2, ..., V_n = Grounding(q_1, q_2, ..., q_n)$. In the stage of Text-guided Video Editing, each received video segment would go through video editing and form the generated video $V' = Editing(V, q')$ with a unified subject or scenario. The obtained edited videos $V'_1, V'_2, ..., V'_n$ would be smoothly combined into a long video using the Video Morphing method. The Personalization Finetuning is optional to replace the diffusion model for generating videos with customized subjects.

video grounding segments from videos with the highest match s_{mm} with each query would be selected for the subsequence.

B. Text guided Video Editing

After receiving video segments $V_1, V_2, ..., V_n = Grounding(q_1, q_2, ..., q_n)$, For each video clip V with n frames$(v^1, v^2, ..., v^n)$ corresponding to a text query q, each frame would be encoded into a low-dimensional latent representation $z_0^i = E(v^i)$, and go through DDIM Inversion, converting them back into noise latent in T reverse steps:

$$\begin{aligned} z_{t+1}^i = &\sqrt{\alpha_{t+1}} \frac{z_t^i - \sqrt{1-\alpha_t}\epsilon_\theta(z_t^i, t, c_q)}{\sqrt{\alpha_t}} \\ &+ \sqrt{1-\alpha_{t+1}}\epsilon_\theta(z_t^i, t, c_q), \\ &t = 0, ..., T-1, \end{aligned} \tag{1}$$

where ϵ_θ represents an image-to-image translation U-Net of the diffusion models we implement, α_t represents the noise variance based on the decreasing schedule and c_q is the text embedding encoded from the text query q.

Using the noise latent as a starting point for video editing, we modify each original video query q to a new query q', making different text queries have the same character subject or visual style. The basic video frame editing approach uses the following DDIM Sampling methods to directly edit each video frame:

$$\begin{aligned} z_{t-1}^i = &\sqrt{\alpha_{t-1}} \frac{z_t^i - \sqrt{1-\alpha_t}\epsilon_\theta(z_t^i, t, c_{q'})}{\sqrt{\alpha_t}} \\ &+ \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t^i, t, c_{q'}), \\ &t = 0, ..., T-1. \end{aligned} \tag{2}$$

Furthermore, we introduce additional methods that improve video editing and generation from two aspects. On the one hand, we use ControlNet, which adds conditional control such as edges, depth, segmentation, and human pose for better generation guidance. On the other hand, we modify the intermediate latent through video editing approaches such as pre-frame latent injection, cross-window attention, and global token merging. Therefore, each latent z_t^i would contain richer reference information from latent and source video images.

In the end, each final latent would be mapped back to an image frame through a decoder $f^i = D(z_0^i)$, forming the generated video $V' = Editing(V, q')$ with higher temporal consistency that more physically makes sense.

C. Video Morphing and Personalization

1) Video Morphing: Although we have obtained edited videos of several different scenes $V'_1, V'_2, ..., V'_n$ with consistent character subjects or styles through video grounding and generative methods, there are still significant differences between the videos based on different text queries, and they cannot be smoothly combined into a long video since the end frame of the previous video could not be not coherent with the starting frame of the next video. Therefore, we adopted the video morphing method to concatenate the start and end segments of all edited videos. In each transition task, we obtain the VAE encoded latent of the preceding video's last frame v^i and the following video's first frame v^j, represented as z_0^i and z_0^j, along with the modified queries q'_i and q'_j of the two videos. Inspired by the approach of DiffMorpher [18], we use two sets of latent-query pairs to relatively fine-tune the diffusion model and train two LoRAs [19] $\Delta\theta_i$ and $\Delta\theta_j$ on the SD UNet ϵ_θ according to the following learning objective:$L(\Delta\theta) = E_{\epsilon,t}||\epsilon -$

$\epsilon_{\theta+\Delta\theta}(\sqrt{\alpha_t}z_0 + \sqrt{1-\alpha_t}\epsilon, t, c_q)||^2$, where $\epsilon \sim N(0, I)$ is the random sampled Gaussian noise.

After fine-tuning, $\Delta\theta_i$ and $\Delta\theta_j$ are fixed and fused into a linear interpolation for the semantics of the input images $\Delta\theta_\alpha = (1-\alpha)\Delta\theta_i + \alpha\Delta\theta_j$, where $\alpha = \frac{k}{n}, k = 1, 2, ..., n-1$, representing generation process of the k-th transition image latent from q'_i to q'_j.

Using LoRA fine-tuned by the images, we utilize LoRA-integrated UNet $\epsilon_{\theta+\Delta\theta_k}(k = i, j)$ to relatively inverse image latent z_0^i and z_0^j to z_T^i and z_T^j, and obtain the intermediate latent noise z_T^α through spherical linear interpolation:

$$z_T^\alpha = \frac{sin((1-\alpha)\phi)}{sin\phi} z_T^i + \frac{sin(\alpha\phi)}{sin\phi} z_T^j,$$
$$\phi = arccos(\frac{z_T^i \cdot z_T^j}{||z_T^i||||z_T^j||}). \tag{3}$$

When generating the intermediate image latent z_0^α, we use UNet with interpolated LoRA $\epsilon_{\theta+\Delta\theta_\alpha}$ during DDIM Sampling steps, while the latent condition $c_\alpha = (1-\alpha)c_{q'_i} + \alpha c_{q'_j}$ is applied through the linear interpolation. By generating $z_0^\alpha, \alpha = \frac{k}{n}, k = 1, 2, ..., n-1$ and $v^\alpha = D(z_0^\alpha)$, we formed a transition video segment $V^{ij} = (v^{\frac{1}{n}}, v^{\frac{2}{n}}, ..., v^{\frac{n-1}{n}})$ from v^i to v^j.

2) Video Personalization: Video personalization is designed to generate characters from the real or virtual world that were not used for training the diffusion model [20], [21]. In our approach, we fine-tune the diffusion model using 3-5 images of a specialized character paired with the text prompt containing a rare token identifier [20] and the name of the character's class(e.g., "A [V] dog" or "A [sks] man"). After fine-tuning, we replace the diffusion model ϵ_θ used in the section of video editing and video morphing. This step is optional to generate customized subjects for the final generated videos.

IV. Experiments

In this section, we will present the results of our method on long video generation tasks based on multi-sentence video grounding and generative methods.

A. Setups

1) Datasets: We conduct massive video moment retrieval on the Charades [22], ActivityNet [23], and TACoS [24] datasets and mainly choose **Activitynet** as our video database for the next stage of video editing since it covers a wide range of complex human activities that are of interest to people in their daily living. We generate about 100 edited videos for evaluation, each containing frames ranging from 150 to 1000 (depending on the length of the video clip captured by video grounding) with a resolution of 512×512.

2) Models: We apply Stable Diffusion model with video editing approaches including **Pix2Video**, **LOVECon** and **Vid-ToMe** under their default setting for the video editing. These video editing methods apply different methods to enhance the temporal consistency of edited generated videos.

3) Baseline : We compare our results with the Stable Diffusion-based video generation method **Text2Video-Zero** [25] and **FreeNoise** [26]. **Text2Video-Zero** allows zero-shot text-to-video generation that encodes motion dynamics in the latent codes and reprograms cross-frame attention of frames. **FreeNoise** is a tuning-free paradigm to enhance the generative capabilities of pre-trained models while maintaining content consistency and being able to generate high-fidelity long videos conditioned on multiple texts.

4) Evaluation Metrics : We evaluate generated long videos using CLIP-T for similarity between the CLIP [27] feature of the video and the textual prompt, and CLIP-SMI for semantic consistency across adjacent frames. Additionally, we calculate FID and FVD between original images/short videos and longer video subsets and use VBench [28] metrics, including subject consistency, image quality, temporal style, and temporal flickering.

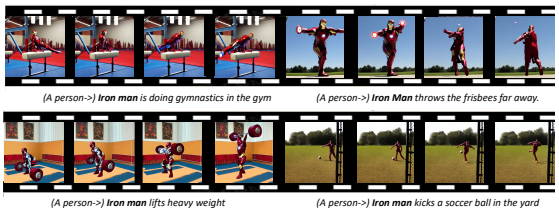

Fig. 2. Example combined video using multi-sentence video grounding for long video generation. The non-bold texts represent queries for video grounding, while **bold** text represents a portion of the content in the query being replaced in video editing to generate a customized subject.

B. Main results

Our main quantitative results for video grounding in long video generation are shown in Table I. The results show that: (i) Our method achieves higher scores in subject consistency and temporal flickering, demonstrating the effectiveness of retrieval augmentation for improving continuity and diversity through video grounding. (ii) Using VidToMe, our method outperforms the baseline in most of the other evaluation metrics, enhancing image quality in generated videos. An example of a generated long video using our method in Figure 2 shows strong subject consistency across frames.

C. Ablation Studies

1) Results with Different Editing Method: As shown in Table II, we compared the results of different video editing methods using different video editing approaches and different pre-trained weights of ControlNet.

The results show that (i) VidToMe outperforms the Stable Diffusion editing method and other video editing methods for long videos, highlighting the importance of balancing local continuity and global long-term consistency. (ii) Using a single ControlNet for editing different video segments does not guarantee quality improvement, as different video types benefit from different ControlNets. Methods incorporating multiple ControlNets achieve better performance compared to those using single or no ControlNet.

2) Results with Video Morphing and Personalization: As shown in Table III, using VidToMe under the depth Conttrol-Net setting, we conducted further evaluations under CLIP-I

TABLE I

COMPARISON BETWEEN OUR METHOD WITH BASELINE. THE BEST AVERAGE PERFORMANCE IS IN **BOLD** AND SECOND IS UNDERLINED. ↑ INDICATES HIGHER METRIC VALUE REPRESENTS BETTER PERFORMANCE AND VICE VERSA.

Model	CLIP-T↑	CLIP-SIM↑	FID↓	FVD↓	subject consistency↑	imaging quality↑	temporal style↑	temporal flickering↑
Text2Video-Zero	0.265	0.8676	87.91	104.02	68.76%	62.18%	10.46%	80.10%
FreeNoise	0.271	0.8996	84.18	86.60	75.49%	65.37%	9.57%	97.28%
Ours with Pix2Video	0.265	**0.9080**	74.91	81.02	73.07%	61.69%	10.16%	98.41%
Ours with LoveCon	0.267	0.8945	79.47	96.52	70.83%	54.94%	9.22%	98.48%
Ours with VidToMe	**0.283**	0.8733	**59.28**	**64.57**	**76.79%**	**69.50%**	**10.59%**	**98.67%**

TABLE II

RESULTS WITH DIFFERENT EDITING METHODS OF OUR APPROACH.

Editing Method	ControlNet	subject consistency	imaging quality	temporal style	temporal flickering
SD-1.5 only	None	68.24%	55.98%	8.16%	79.85%
Pix2Video	None	73.07%	61.69%	10.16%	98.41%
LoveCon	Canny	70.99%	53.65%	8.39%	98.49%
LoveCon	Depth	71.60%	54.61%	9.21%	98.18%
LoveCon	Hed	70.83%	54.94%	9.22%	98.48%
LoveCon	**Multi**	74.49%	56.37%	9.63%	98.45%
VidToMe	None	76.79%	69.50%	10.59%	98.67%
VidToMe	Depth	72.90%	63.33%	10.58%	98.43%
VidToMe	Canny	73.43%	66.17%	10.85%	97.59%
VidToMe	Softedge	72.19%	63.66%	10.08%	98.26%
VidToMe	**Multi**	**78.01%**	**70.54%**	**11.29%**	**99.13%**

TABLE III

ABLATION STUDY OF VIDEO MORPHING AND PERSONALIZATION.

Video Morphing	Personalization	CLIP-I	CLIP-SMI
✗	✗	0.593	0.873
✓	✗	0.691	0.875
✗	✓	0.590	0.899
✓	✓	0.676	0.902

to measure the similarity between the generated video and the personalization image and CLIP-SMI for frame-to-frame consistency in the morphing sections. Results demonstrate that our method effectively follows prompts and generates consistently customized frames.

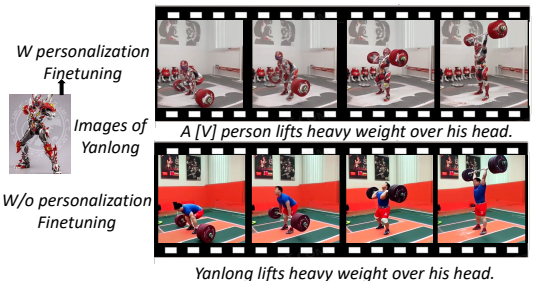

Fig. 3. Examples of personalization. As shown above, *Yanlong* is a subject that the pre-trained SD model could not generate in lack of personalization finetuning. Our personalization process through Dreambooth achieves exact and consistent subject generation for video.

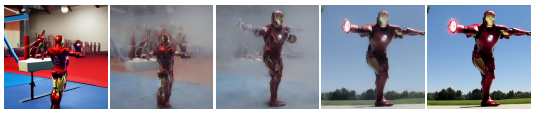

Fig. 4. Examples of video morphing.

In addition, considering that some subjects have distinctive features and are well remembered by the SD model, we tested pre-trained SD models based on our method specifically for subject roles that could not be generated before personalized customization and provide one of the examples shown in Figure 3, which includes *Yanlong*, a subject that the SD model we use could not generate directly, presenting the availability of our method to maintain exact subject consistency across clips through Dreambooth. For the video morphing part, we provide an example of the video morphing results shown in Figure 4. Since the current morphing is still imperfect, it is worth looking forward to better methods in this field to make video transitions more artistic and smooth.

D. Case Analysis

As shown in Figure 5, we select a few video samples generated using our method and baseline method Text2Video-Zero respectively for case analysis. Our method uses VidToMe as the video editing method and selects the depth as the type of ControlNet. We further visualize the X-T slice for each frame from the videos.

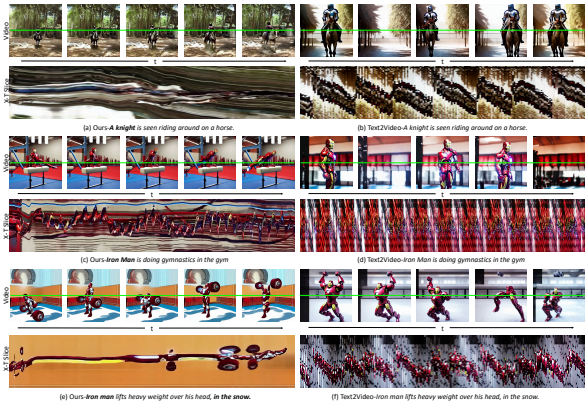

Fig. 5. Qualitative example results. (a),(c) and (e) are example videos generated through our method, including the customized subject or scenario in the text queries represented by the **bold** characters. While (b), (d), and (f) are videos generated through the baseline model.

The X-T slices of our generated videos demonstrate the effectiveness of our method. For instance, in Figure 5(a), the video *A knight is seen riding around on a horse'* shows continuous subject and background changes, highlighting the stability and variability of our approach. Figure 5(c), with *Iron Man is doing gymnastics in the gym'*, illustrates our ability to capture complex movements despite frequent subject motion. Figure 5(e) showcases our method's capability for multiple customizations, which includes both **customized subject** and **scenario**.

In contrast, Figures 5(b), (d), and (e) illustrate the limitations of the baseline method, Text2Video, which suffers from frame inconsistencies and subject omissions. The FreeNoise method, shown in Figure 6, maintains subject consistency but violates real-world physical laws without reference guidance. Overall, our multi-sentence video grounding-based approach leads to more consistent video generation.

Fig. 6. *'Iron Man lifts heavy weight over his head'* generated by **FreeNoise**. The video generated without reference guidance exhibits violations of real-world physical laws. (Heavy weight is lifted, but Iron Man's lifting action did not actually come into contact with heavy weight.)

E. Comparison about Computing Resources

Based on the Stable Diffusion model, our model maintains a comparable GPU memory cost (8-10GB) to the SD-only method, Text2Video-Zero, and FreeNoise, demonstrating better capabilities to generate long and consistent videos while maintaining better obedience to physical laws and the common sense of human life. Moreover, our method has a lower GPU memory cost compared to the high-performance long video generation methods such as Sora or Open-Sora, which normally cost GPU memory 20GB or even more.

V. Conclusion

In this paper, we study the problem of utilizing video grounding models to conduct data augmentation for long video generation. We propose the Multi-sentence Video Grounding for Long Video Generation framework, which consists of a multi-sentence video grounding model to retrieve different video moments matching text queries from the video database, and a data augmentation strategy to edit video contents into videos with unified subjects through the video editing method. Experiments demonstrate that our proposed framework outperforms baseline models for long video generation. Our approach seamlessly extends the development in image/video editing, video morphing, personalized generation, and video grounding to the long video generation, offering effective solutions for generating long videos at a low memory cost. Utilizing video grounding methods to enhance the long video generation could be a promising future research direction.

Acknowledgments

This work is supported by National Natural Science Foundation of China No.62222209, Beijing National Research Center for Information Science and Technology under Grant No.BNR2023TD03006.

References

[1] H. Chen, X. Wang, Y. Zhou, B. Huang, Y. Zhang, W. Feng, H. Chen, Z. Zhang, S. Tang, and W. Zhu, "Multi-modal generative ai: Multi-modal llm, diffusion and beyond," *arXiv preprint arXiv:2409.14993*, 2024.

[2] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[3] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[4] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation," in *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[6] X. Lan, Y. Yuan, X. Wang, Z. Wang, and W. Zhu, "A survey on temporal sentence grounding in videos," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2, pp. 1–33, 2023.

[7] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 271–14 280.

[8] X. Wang, Z. Wu, H. Chen, X. Lan, and W. Zhu, "Mixup-augmented temporally debiased video grounding with content-location disentanglement," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4450–4459.

[9] M. Jung, S. Choi, J. Kim, J.-H. Kim, and B.-T. Zhang, "Modal-specific pseudo query generation for video corpus moment retrieval," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7769–7781.

[10] H. Chen, X. Wang, H. Chen, Z. Zhang, W. Feng, B. Huang, J. Jia, and W. Zhu, "Verified: A video corpus moment retrieval benchmark for fine-grained video understanding," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[11] N. Yang, M. Kim, S. Yoon, J. Shin, and K. Jung, "Mvmr: Evaluating natural language video localization bias over multiple reliable videos pool," *arXiv preprint arXiv:2309.16701*, 2023.

[12] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2023.

[13] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2022.

[14] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 206–23 217.

[15] Z. Liao and Z. Deng, "Lovecon: Text-driven training-free long video editing with controlnet," *arXiv preprint arXiv:2310.09711*, 2023.

[16] X. Li, C. Ma, X. Yang, and M.-H. Yang, "Vidtome: Video token merging for zero-shot video editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7486–7495.

[17] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[18] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan, "Diffmorpher: Unleashing the capability of diffusion models for image morphing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7912–7921.

[19] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[20] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "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*, 2023, pp. 22 500–22 510.

[21] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation," in *The Twelfth International Conference on Learning Representations*.

[22] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[23] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[24] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[25] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[26] H. Qiu, M. Xia, Y. Zhang, Y. He, X. Wang, Y. Shan, and Z. Liu, "Freenoise: Tuning-free longer video diffusion via noise rescheduling," in *The Twelfth International Conference on Learning Representations*, 2024.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[28] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.