

Make Pixels Dance: High-Dynamic Video Generation

Yan Zeng* Guoqiang Wei* Jiani Zheng
 Jiaxin Zou Yang Wei Yuchen Zhang Hang Li

ByteDance Research

* Equal Contribution

{zengyan.yanne, weiguoqiang.9, lihang.lh}@bytedance.com

<https://makepixelsdance.github.io>



Figure 1. Generation results of **PixelDance** given text, first frame instruction highlighted in red box (and last frame instruction in green). Six frames sampled from a 16-frame clip are displayed. Human faces presented in this paper are synthesized using text-to-image models.

Abstract

Creating high-dynamic videos such as motion-rich ac-

tions and sophisticated visual effects poses a significant challenge in the field of artificial intelligence. Unfortunately, current state-of-the-art video generation methods,

*primarily focusing on text-to-video generation, tend to produce video clips with minimal motions despite maintaining high fidelity. We argue that relying solely on text instructions is insufficient and suboptimal for video generation. In this paper, we introduce **PixelDance**, a novel approach based on diffusion models that incorporates image instructions for both the first and last frames in conjunction with text instructions for video generation. Comprehensive experimental results demonstrate that PixelDance trained with public data exhibits significantly better proficiency in synthesizing videos with complex scenes and intricate motions, setting a new standard for video generation.*

1. Introduction

Generating high-dynamic videos with motion-rich actions, sophisticated visual effects, natural shot transitions, or complex camera movements, has long been a lofty yet challenging goal in the field of artificial intelligence. Unfortunately, most existing video generation approaches focusing on text-to-video generation [10, 22, 55] are still limited to synthesize simple scenes, and often falling short in terms of visual details and dynamic motions. Recent state-of-the-art models have significantly enhanced text-to-video quality by incorporating an image input [11, 24, 43], which provides finer visual details for video generation. Despite the advancements, the generated videos frequently exhibit limited motions as shown in Figure 2. This issue becomes particularly severe when the input images depict out-of-domain content unseen in training data, indicating a key limitation of current technologies.

To generate high-dynamic videos, we propose a novel video generation approach that incorporates image instructions for both the first and last frames of a video clip, in addition to text instruction. The image instruction for the first frame depicts the major scene of the video clip. The image instruction for the last frame, which is *optionally* used in training and inference, delineates the ending of the clip and provides additional control for generation. The image instructions enable the model to construct intricate scenes and actions. Moreover, our approach can create longer videos, in which case the model is applied multiple times and the last frame of the preceding clip serves as the first frame instruction for the subsequent clip.

The image instructions are more direct and accessible compared to text instructions. We use ground-truth video frames as the instructions for training, which is easy to obtain. In contrast, recent work has proposed the use of highly descriptive text annotations [4] to better follow text instructions. Providing detailed textual annotations to precisely describe both the frames and the motions of videos is not only costly to collect but also difficult to learn for the model. To understand and follow complex text instructions, the model

needs to significantly scale up. The use of image instructions overcome these challenges together with text instructions. Given the three instructions in training, the model is able to focus on learning the dynamics of video content, and in inference the model can better generalize the learned dynamics knowledge to out-of-domain instructions.

Specifically, we present **PixelDance**, a latent diffusion model based approach to video generation, conditioned on $\langle \text{text}, \text{first frame}, \text{last frame} \rangle$ instructions. The text instruction is encoded by a pre-trained text encoder and is integrated into the diffusion model with cross-attention. The image instructions are encoded with a pre-trained VAE encoder [32] and concatenated with either perturbed video latents or Gaussian noise as the input to the diffusion model. In training, we use the (ground-truth) first frame to enforce the model to strictly adhere to the instruction, maintaining continuity between consecutive video clips. In inference, this instruction can be conveniently obtained from T2I models [32] or directly provided by users.

Our approach is unique in its way of using the last frame instruction. We intentionally avoid encouraging the model to replicate the last frame instruction exactly since it is challenging to provide a perfect last frame in inference, and the model should accommodate user-provided coarse drafts for guidance. Such kind of instruction can be readily created by the user using basic image editing tools.

To this end, we develop three techniques. First, in training, the last frame instruction is randomly selected from the last three (ground-truth) frames of a video clip. Second, we introduce noise to the instruction to mitigate the reliance on the instruction and promote the robustness of model. Third, we randomly drop the last frame instruction with a certain probability, e.g. 25%, in training. Correspondingly, we propose a simple yet effective sampling strategy for inference. During the first τ denoising steps, the last frame instruction is utilized to guide video generation towards the desired ending status. Then, during the remaining steps, the instruction is dropped, allowing the model to generate more temporally coherent video. The impact of last frame instruction can be adjusted by τ .

Our model’s ability of leveraging image instructions enables more effective use of public video-text datasets, such as WebVid-10M [2] which only contains coarse-grained descriptions with loose correlation to videos [37], and lacks of content in diverse styles (*e.g.*, comics and cartoons). Our model with only 1.5B parameters, trained mainly on WebVid-10M, achieves state-of-the-art performance on multiple scenarios. First, given text instruction only, PixelDance leverages T2I models to obtain the first frame instruction to generate videos, reaching FVD scores of 381 and 242.8 on MSR-VTT [50] and UCF-101 [38] respectively. With the text and first frame instructions (the first

frame instruction can also be provided by users), PixelDance is able to generate more motion-rich videos compared to existing models. Second, PixelDance can generate continuous video clips, outperforming existing long video generation approaches [8, 16] in temporal consistency and video quality. Third, the last frame instructions are shown to be a critical component for creating intricate out-of-domain videos with complex scenes and/or actions, as shown in Figure 1. Overall, by actively interacting with PixelDance, we create the first three-minute video with a clear storyline at various complex scenes and characters hold consistent across scenes.

Our contributions can be summarized as follows:

- We propose a novel video generation approach based on diffusion model, **PixelDance**, which incorporates image instructions for both the first and last frames in conjunction with text instruction.
- We develop training and inference techniques for PixelDance, which not only effectively enhances the quality of generated videos, but also provides users with more control over the video generation process.
- Our model trained on public data demonstrates remarkable performance in high-dynamic video generation with complex scenes and actions, setting a new standard for video generation.

2. Related Work

2.1. Video Generation

Video generation has long been an attractive and essential research topic [8, 31, 42]. Previous studies have resorted to different types of generative models such as GANs [12, 25, 29, 39] and Transfomers with VQVAE [9, 23, 51]. Recently, diffusion models have significantly advanced the progress of photorealistic text-to-image generation [3, 34], which exhibit robustness superior to GANs and require fewer parameters compared to transformer-based counterparts. Latent diffusion models [32] are proposed to reduce the computational burden by training a diffusion model in a compressed lower-dimensional latent space. For video generation, previous studies typically add temporal convolutional layers and temporal attention layers to the 2D UNet of a pre-trained text-to-image diffusion models [10, 14, 16, 27, 37, 43, 45, 55]. Although these advancements have paved the way for the generation of high-resolution videos through the integration of super-resolution modules [26], the videos produced are characterized by simple, minimal motions as shown in Figure 2.

Recently, the field of video editing has witnessed remarkable progress [28, 52, 54], particularly in terms of content modification while preserving the original structure and motion of the video, for example, modifying a cattle to a cow [6, 48]. Despite these achievements, the neces-



Figure 2. Videos generated by state-of-the-art video generation model [11], compared with our results given the same text prompts and image conditions in Figure 1 and Figure 4.

sity to search for an appropriate reference video for editing is time-consuming. Furthermore, this approach inherently constrains the scope of creation, as it precludes the possibility of synthesizing entirely novel content (*e.g.*, a polar bear walking on the Great Wall) that may not exist in any reference video.

2.2. Long Video Generation

Long video generation is a more challenging task which requires seamless transitions between successive video clips and long-term consistency of the scene and characters. There are typically two approaches: 1) autoregressive methods [15, 22, 41] employ a sliding window to generate a new clip conditioned on the previous clip. 2) hierarchical methods [9, 15, 17, 53] generate sparse frames first, then interpolate intermediate frames. However, the autoregressive approach is susceptible to quality degradation due to error cumulation over time. As for the hierarchical method, it needs long videos for training, which are difficult to obtain due to frequent shot changes in online videos. Besides, generating temporally coherent frames across larger time interval exacerbates the challenges, which often leads to low-quality initial frames, making it hard to achieve good results in later stages of interpolation. In this paper, PixelDance generates continuous video clips in the auto-regressive way and exhibits superior performance in synthesizing long-term consistent frames compared to existing models. Concurrently, we advocate for active user engagement with the generation process, akin to a film director’s role, to ensure that the produced content closely aligns with the user’s expectation.

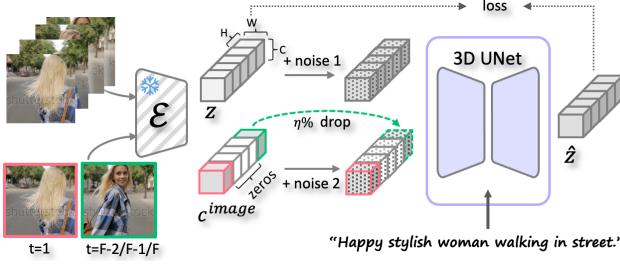


Figure 3. Illustration of the training procedure of PixelDance. The original video clip and image instructions (in red and green boxes) are encoded into \mathbf{z} and \mathbf{c}^{image} , which are then concatenated along the channel dimension after perturbed with different noises.

3. Method

Existing models in text-to-video [10, 22, 55] and image-to-video generation [11, 24, 43] often produce videos characterized by simple and limited movements. In this paper, we attempt to enable the model to focus on learning the dynamics of video contents, to generate videos with rich motions. We present a novel approach that integrates image instructions for both the first and last frames in conjunction with text instruction for video generation, and we effectively utilize public video data for training. We will elaborate on the model architecture in Sec. 3.1, and then introduce the training and inference techniques tailored for our approach in Sec. 3.2.

3.1. Model Architecture

Latent Diffusion Architecture We adopt latent diffusion model [32] for video generation. Latent diffusion model is trained to denoise from a perturbed input in the latent space of a pre-trained VAE, in order to reduce the computational burden. We take the widely used 2D UNet [33] as diffusion model, which is constructed with a series of spatial down-sampling layers followed by a series of spatial upsampling layers with inserted skip connections. Specifically, it is built with two basic blocks, *i.e.*, 2D convolution block and 2D attention block. We extend the 2D UNet to 3D variant with inserting temporal layers [22], where 1D convolution layer along temporal dimension after 2D convolution layer, and 1D attention layer along temporal dimension following 2D attention layer. The model can be trained jointly with images and videos to maintain high-fidelity generation ability on spatial dimension. The 1D temporal operations are disabled for image input. We use bi-directional self-attention in all temporal attention layers. We encode the text instruction using a pre-trained CLIP text encoder [30], and the embedding \mathbf{c}^{text} is injected through cross-attention layers in the UNet with hidden states as queries and \mathbf{c}^{text} as keys and values.

Image Instruction Injection We incorporate image instructions for both the first and last frames in conjunction with text instruction. We utilize ground-truth video frames as the instructions in training, which is easy to obtain. Given the image instructions on the first and last frame, denoted as $\{\mathbf{I}^{first}, \mathbf{I}^{last}\}$, we first encode them into the input space of diffusion models using VAE, result in $\{\mathbf{f}^{first}, \mathbf{f}^{last}\}$ where $\mathbf{f} \in \mathbb{R}^{C \times H \times W}$. To inject the instructions without loss of the temporal position information, the final image condition is then constructed as:

$$\mathbf{c}^{image} = [\mathbf{f}^{first}, \text{PADs}, \mathbf{f}^{last}] \in \mathbb{R}^{F \times C \times H \times W}, \quad (1)$$

where $\text{PADs} \in \mathbb{R}^{(F-2) \times C \times H \times W}$. The condition \mathbf{c}^{image} is then concatenated with noised latent \mathbf{z}_t along the channel dimension, which is taken as the input of diffusion models.

3.2. Training and Inference

The training procedure is illustrated in Figure 3. For the first frame instruction, we employ the ground-truth first frame for training, making the model adhere to the first frame instruction strictly in inference. In contrast, we intentionally avoid encouraging the model to replicate the last frame instruction exactly. During inference, the ground-truth last frame is unavailable in advance, the model needs to accommodate user-provided coarse drafts for guidance to generate temporally coherent videos. To this end, we introduce three techniques. First, we randomly select an image from the last three ground-truth frames of a clip to serve as the last frame instruction for training. Second, to promote robustness, we perturb the encoded latents \mathbf{c}^{image} of image instructions with noise.

Third, during training, we randomly drop the last frame instruction with probability η , by replacing the corresponding latent with zeros. Correspondingly, we propose a simple yet effective inference technique. During inference, in the first τ out of the total T denoising steps, the last frame instruction is applied to guide the video generation towards desired ending status, and it is dropped in the subsequent steps to generate more plausible and temporally consistent videos:

$$\tilde{\mathbf{x}}_\theta = \begin{cases} \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{f}^{first}, \mathbf{f}^{last}, \mathbf{c}^{text}), & \text{if } t < \tau \\ \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{f}^{first}, \mathbf{c}^{text}), & \text{if } \tau \leq t \leq T \end{cases}. \quad (2)$$

τ determines the strength of model dependency on last frame instruction, adjusting τ will enable various applications. For example, our model can generate high-dynamic videos without last frame instruction (*i.e.*, $\tau = 0$). Additionally, we apply the classifier-free guidance [19] in inference, which mixes the score estimates of the model conditioned on text prompts and without text prompts.

4. Experiments

4.1. Implementation Details

Following previous work, we train the video diffusion model on WebVid-10M [2], which contains about 10M short video clips with an average duration of 18 seconds, predominantly in the resolution of 336×596 . Each video is associated with a paired text which generally offers a coarse description weakly correlated with the video content. Another nuisance issue of WebVid-10M lies in the watermarks placed on all videos, which leads to the watermark’s presence in all generated videos. Thus, we expand our training data with other self-collected 500K watermark-free video clips depicting real-world entities such as humans, animals, objects, and landscapes, paired with coarse-grained textual descriptions. Despite comprising only a modest proportion, we surprisingly find that combining this dataset with WebVid-10M for training ensures that PixelDance is able to generate watermark-free videos if the image instructions are free of watermarks.

PixelDance is trained jointly on video-text dataset and image-text dataset. For video data, we randomly sample 16 consecutive frames with 4 fps per video. Following previous work [21], we adopt LAION-400M [36] as image-text dataset. Image-text data are utilized every 8 training iterations. The weights of pre-trained text encoder and VAE model are frozen during training. We employ DDPM [20] with $T = 1000$ time steps for training. A noise corresponding to 100 time steps is introduced to the image instructions \mathbf{c}^{image} . We first train the model at resolution of 256×256 , with batch size of 192 on 32 A100 GPUs for 200K iterations, which is utilized for quantitative evaluations. This model is then finetuned for another 50K iterations with higher resolution. We incorporate ϵ -prediction [20] as training objective.

Table 1. Zero-shot T2V performance comparison on MSR-VTT. All methods generate video with spatial resolution of 256×256 . Best in bold.

Method	#data	#params.	CLIPSIM(\uparrow)	FVD(\downarrow)
CogVideo (En) [23]	5.4M	15.5B	0.2631	1294
MagicVideo [55]	10M	-	-	1290
LVDM [16]	2M	1.2B	0.2381	742
Video-LDM [5]	10M	4.2B	0.2929	-
InternVid [46]	28M	-	0.2951	-
ModelScope [43]	10M	1.7B	0.2939	550
Make-A-Video [37]	20M	9.7B	0.3049	-
Latent-Shift [1]	10M	1.5B	0.2773	-
VideoFactory [44]	-	2.0B	0.3005	-
PixelDance	10M	1.5B	0.3125	381

Table 2. Zero-shot T2V performance comparison on UCF-101. All methods generate video with spatial resolution of 256×256 . Best in bold.

Method	#data	#params.	IS(\uparrow)	FID(\downarrow)	FVD(\downarrow)
CogVideo (En) [23]	5.4M	15.5B	25.27	179.00	701.59
MagicVideo [55]	10M	-	-	145.00	699.00
LVDM [16]	2M	1.2B	-	-	641.80
InternVid [46]	28M	-	21.04	60.25	616.51
Video-LDM [5]	10M	4.2B	33.45	-	550.61
ModelScope [43]	10M	1.7B	-	-	410.00
VideoFactory [44]	-	2.0B	-	-	410.00
Make-A-Video [37]	20M	9.7B	33.00	-	367.23
VidRD [13]	5.3M	-	39.37	-	363.19
Dysen-VDM [7]	10M	-	35.57	-	325.42
PixelDance	10M	1.5B	42.10	49.36	242.82



Figure 4. Illustration of video generation conditioned on the text and first frame instructions. Please refer to the Supplementary for more examples.

4.2. Video Generation

4.2.1 Quantitative Evaluation

We evaluate zero-shot video generation performance of our PixelDance on MSR-VTT [50] and UCF-101 [38] datasets, following previous work [5, 16, 23, 55]. MSR-VTT is a video retrieval dataset with descriptions for each video, while UCF-101 is an action recognition dataset with 101 action categories. *To make a comparison with previous T2V approaches which are conditioned on text prompts only, we also evaluate only with text instructions.* Specifically, we utilize off-the-shelf T2I Stable Diffusion V2.1



Figure 5. Illustration of complex video generation conditioned on the text, first frame and last frame instructions. Please refer to the Supplementary for more examples.

[32] to obtain the first frame instructions, and generate videos given the text and first frame instructions. Following previous work [7, 44], we randomly select one prompt per example to generate 2990 videos in total for evaluate, and report the Fréchet Video Distance (FVD) [40] and CLIP-similarity (CLIPSIM) [47] on MSR-VTT dataset. For UCF-101 dataset, we construct descriptive text prompts per category and generate about 10K videos, and compare with previous works in terms of the widely-used Inception Score (IS) [35], Fréchet Inception Distance (FID) [18] and FVD, following previous work [7, 44]. Both FID and FVD measure the distribution distance between generated videos and the ground-truth data. IS assesses the quality of generated videos and CLIPSIM estimates the similarity between the generated videos and corresponding texts.

Zero-short evaluation results on MSR-VTT and UCF-101 are presented in Table 1 and Table 2, respectively. Compared to other T2V approaches on the MSR-VTT, PixelDance achieves state-of-the-art result in terms of FVD and CLIPSIM, demonstrating its remarkable ability to generate high-quality videos with better alignment to text prompts. Notably, PixelDance achieves an FVD score of 381, which substantially surpasses the previous state-of-the-art ModelScope [43], with an FVD of 550. On UCF-101 benchmark, PixelDance outperforms other models across various metrics, including IS, FID and FVD.

4.2.2 Qualitative Analysis

Effectiveness of Each Instruction Our video generation approach incorporates three distinct instructions: text, first frame, and last frame instruction. In this section, we will probe deeply into the influence of each instruction on the quality of generated videos.

In PixelDance, the text instruction could be concise, considering that the first frame instruction has delivered the objects/characters and scenes, which are challenging to be



Figure 6. First two rows: text instruction helps enhance the cross-frame consistency of key elements like the *black hat* and *red bow tie* of the polar bear. Last row: natural shot transition.

described succinctly and precisely with text. Nonetheless, the text prompt plays a vital role of specifying various motions, including but not limited to body movements, facial expressions, object movements, and visual effects (first two rows of Figure 4). Besides, it allows for manipulating camera movements with specific prompts like "zoom in/out," "rotate," and "close-up," as demonstrated in the last row of Figure 4. Moreover, the text instruction helps to hold the cross-frame consistency of specified key elements, such as the detailed descriptions of characters (polar bear in Figure 6).

The first frame instruction significantly improves the video quality by providing finer visual details. Moreover, it is key to generate multiple consecutive video clips. With the text and first frame instructions, PixelDance is able to generate more motion-rich videos (Figure 4 and Figure 6) compared to existing models.

The last frame instruction, delineating the concluding status of a video clip, provides an additional control on video generation. This instruction is instrumental for synthesizing intricate motions, and becomes particularly crucial for out-of-domain video generation as depicted in the first two samples in Figure 1 and Figure 5. Furthermore, we can generate a natural shot transition using last frame instruction (last sample of Figure 6).

Strength of Last Frame Guidance To make the model work well with user-provided drafts, even if they are somewhat imprecise, we intentionally avoid encouraging the model replicate the last frame instruction exactly, with the proposed techniques detailed in Sec. 3. As shown in



Figure 7. Illustration of the effectiveness of the proposed techniques ($\tau = 25$) to avoid replicating the last frame instruction.

the Figure 7, without our techniques, the generated video abruptly ends in the given last frame instruction exactly. In contrast, with our proposed methods, the generated video is more fluent and temporally consistent.

Generalization to Out-of-Domain Image Instructions
 Despite the notable lack of training videos in non-realistic styles (e.g., science fictions, comics, and cartoons), PixelDance demonstrates a remarkable capability to generate high-quality videos in these out-of-domain categories. This generalizability can be attributed to that our approach focuses on learning dynamics and ensuring temporal consistency, given the image instructions. As PixelDance learns the underlying principles of motions in real world, it can generalize across various stylistic domains of image instructions.

4.3. Ablation Study

Table 3. Ablation study results on UCF-101.

Method	FID(\downarrow)	FVD(\downarrow)
① T2V baseline	59.35	450.58
② PixelDance	49.36	242.82
③ PixelDance w/o c^{text}	51.26	375.79
④ PixelDance w/o f^{last}	49.45	339.08

To evaluate the key components of PixelDance, we conduct a quantitative ablation study on the UCF-101 dataset following the zero-shot evaluation setting in Sec. 4.2.1.

First, we provide a T2V baseline (①) for comparison trained on the same dataset. We further analyze the effectiveness of instructions employed in our model. Given the indispensable nature of the first frame instruction for the generation of continuous video clips, our ablation focuses on the text instruction (③) and the last frame instruction (④). The experimental results indicate that omitting ei-

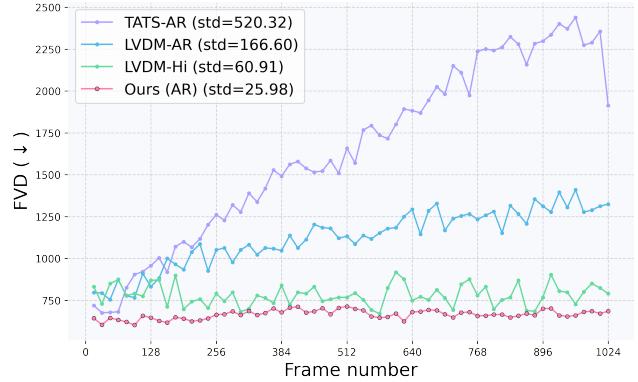


Figure 8. FVD comparison for long video generation (1024 frames) on UCF-101. AR: auto-regressive. Hi: hierarchical. The construction of long video with PixelDance is in an autoregressive manner.

ther instruction results in a significant deterioration in video quality. Notably, even though the evaluation does not incorporate the last frame instruction, model trained with this instruction (②) outperforms the model trained without it (④). This observation reveals that relying solely on the $\langle text, first frame \rangle$ for video generation poses substantial challenges due to the significant diversity of video content. In contrast, incorporating all three instructions enhances PixelDance’s capacity to model motion dynamics and hold temporal consistency.

4.4. Long Video Generation

4.4.1 Quantitative Evaluation

As aforementioned, PixelDance is trained to strictly adhere to the first frame instruction, in order to generate long videos, where the last frame from preceding clip is used as the first frame instruction for generating the subsequent clip. To evaluate PixelDance’s capability of long video generation, we follow the previous work [8, 16] and generate 512 videos with 1024 frames on UCF-101 datasets, under the zero-shot setting detailed in Sec. 4.2.1. We report the FVD of every 16 frames extracted side-by-side from the synthesized videos. The results, as shown in Figure 8, show that PixelDance demonstrates lower FVD scores and smoother temporal variations, compared with auto-regressive models, TATS-AR [8] and LVDM-AR [16], and the hierarchical approach LVDM-Hi. Please refer to the Supplementary for visual comparisons.

4.4.2 Qualitative Analysis

Recognizing that most real-world long videos (e.g., videos or films on YouTube) comprise numerous shots rather than a single continuous shot, this qualitative analysis focuses on

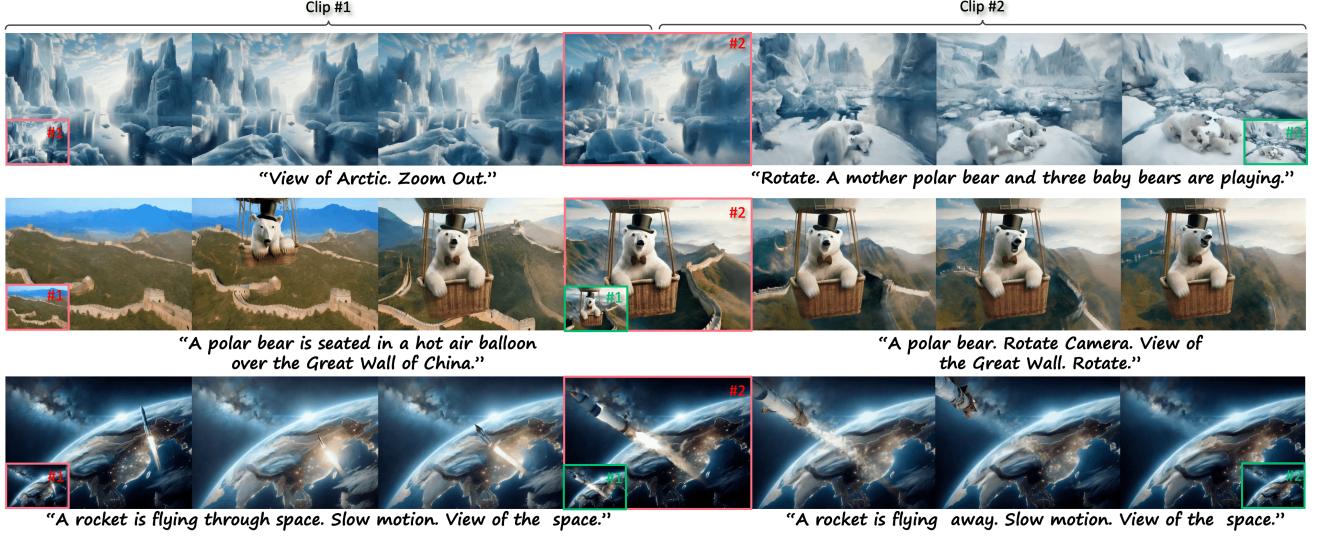


Figure 9. Illustration of PixelDance handling intricate shot compositions consisting of two continuous video clips, in which case the last frame of the Clip #1 serves as the first frame instruction for Clip #2.

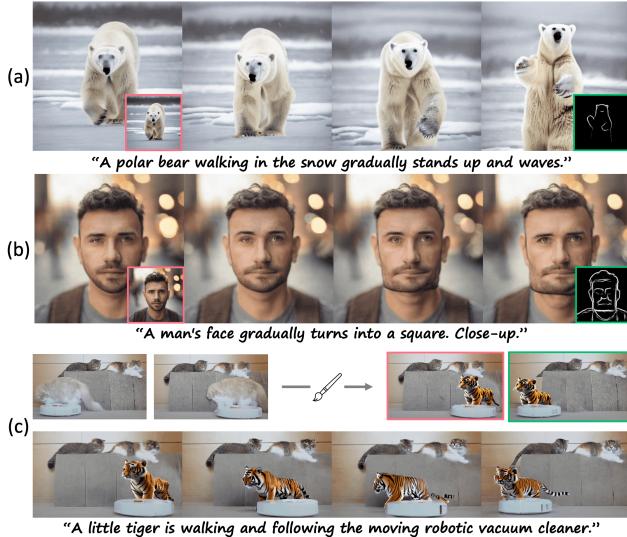


Figure 10. Illustration of video generation with sketch image as last frame instruction (first two examples), and PixelDance for zero-shot video editing (c).

PixelDance’s capabilities of generating a composite shot. This is formed by stringing together multiple continuous video clips that are temporally consistent. Figure 9 illustrates the capability of PixelDance to handle intricate shot compositions involving complex camera movements (in Arctic scenes), smooth animation effects (polar bear appears in a hot air balloon over the Great Wall), and precise control over the trajectory of a rocket. These instances exemplify how users interact with PixelDance to craft desired video sequences. Leveraging PixelDance’s advanced gener-

ation capabilities, we have successfully synthesized a three-minute video that not only tells a coherent story but also maintains a consistent portrayal of the main character.

4.5. More Applications

Sketch Instruction Our proposed approach can be extended to other types of image instructions, such as semantic maps, image sketches, human poses, and bounding boxes. To demonstrate this, we take the image sketch as an example and finetune PixelDance with image sketch [49] as the last frame instruction. The results are shown in the first two rows of Figure 10, exhibiting that a simple sketch image is able to guide the video generation process.

Zero-shot Video Editing PixelDance is able to perform video editing without any training, achieved by transforming the video editing task into an image editing task. As shown in the last example in Figure 10, by editing the first frame and the last frame of the provided video, PixelDance generates a temporally consistent video aligned with user expectation on video editing.

5. Conclusion

In this paper, we proposed a novel video generation approach based on diffusion models, PixelDance, which incorporates image instructions for both the first and last frames in conjunction with text instruction. We developed training and inference techniques tailored for this approach. PixelDance trained mainly on WebVid-10M exhibited exceptional proficiency in synthesizing videos with complex

scenes and actions, setting a new standard in video generation.

While our approach has achieved noteworthy results, there is potential for further advancements. First, the model can benefit from training with high-quality, open-domain video data. Second, fine-tuning the model within specific domains could further augment its capabilities. Third, incorporating annotated texts that outline key elements and motions of videos could improve the alignment to user’s instructions. Lastly, PixelDance currently consists of only 1.5B parameters, presenting an opportunity for future scaling up. Further investigation into these aspects will be explored in future work.

References

- [1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 5
- [2] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 5
- [3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3
- [4] James Betker, Gabriel Goh, Li Jiang, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image captioning with better captions. 2023. 2
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 5
- [6] Patrick Esser, Johnathan Chiu, Parmida Atighchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3
- [7] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023. 5, 6
- [8] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 3, 7
- [9] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 3
- [10] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2, 3, 4
- [11] Gen-2: The Next Step Forward for Generative AI. <https://research.runwayml.com/gen2>. 2, 3, 4
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3
- [13] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 5
- [14] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 3
- [15] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 3
- [16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3, 5, 7
- [17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3
- [18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5
- [21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 5
- [22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646. Curran Associates, Inc., 2022. 2, 3, 4
- [23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for

- text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3, 5
- [24] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2, 4
- [25] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3
- [26] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1557–1567, 2023. 3
- [27] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 3
- [28] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3
- [29] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions, 2018. 3
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4
- [31] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 3
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 6
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015. 4
- [34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3
- [35] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10–11):2586–2606, 2020. 6
- [36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarek, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5
- [37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3, 5
- [38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. 2, 5
- [39] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021. 3
- [40] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6
- [41] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. 3
- [42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 3
- [43] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3, 4, 5, 6
- [44] Wenjing Wang, Huan Yang, Zixi Tu, Huigu He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 5, 6
- [45] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3
- [46] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 5
- [47] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 6
- [48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yafei Shi, Wynne Hsu, Ying Shan, Xiaohu

- Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 3
- [49] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 8
- [50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 5
- [51] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3
- [52] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3
- [53] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 3
- [54] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. Magicavatar: Multimodal avatar generation and animation. *arXiv preprint arXiv:2308.14748*, 2023. 3
- [55] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3, 4, 5