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, weiguqiang.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 actions and sophisticated visual effects poses a significant challenge in the field of artificial intelligence. Unfortunately, current state-of-the-art video generation methods,
To significantly scale up, the image instructions overcome and follow complex text instructions, the model may need to not only be difficult to learn for the model. To understand the subsequent 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 may need to significantly scale up. The image instructions overcome these challenges together with text instructions. Given the three instructions in training, the model focuses on learning the dynamics of video content, and better generalizes the learned dynamics knowledge to out-of-domain instructions in inference.

Specifically, we present PixelDance, a latent diffusion model based approach to video generation, conditioned on <text, first frame, last frame> instructions. The text instruction is encoded by a pre-trained text encoder and is integrated into the diffusion model via cross-attention. The image instructions are encoded with a pre-trained VAE encoder 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 with T2I models or provided by users.

Our approach is unique in the 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 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 $\tau$ denoising steps, the last frame instruction is utilized to guide video generation towards the desired ending status. Then, in 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 $\tau$.

Our model’s ability of leveraging image instructions enables more effective use of public <text, video> datasets, such as WebVid which only contains coarse descriptions of videos, and lacks of content in diverse styles (e.g., comics and cartoons). Our model with only 1.5B parameters, trained mainly on WebVid, 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 and UCF-101 respectively. With the text and first frame instructions, 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 in tem-
poral 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.

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.
• PixelDance 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 [9, 34, 45]. Previous studies have resorted to different types of generative models such as GANs [13, 26, 30, 42] and Transformers with VQVAE [10, 24, 54]. Recently, diffusion models have significantly advanced the progress of photorealistic T2I generation [3, 37], which exhibit superior to GANs and are parameter-efficient compared to transformer-based counterparts. Latent diffusion models [35] 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 introduce temporal layers into the 2D UNet of a pre-trained T2I diffusion models [11, 15, 17, 28, 40, 46, 48, 58]. Although these advancements have paved the way for the generation of high-resolution videos through the integration of super-resolution modules [27], the videos produced are characterized by simple, minimal motions as shown in Figure 2.

Recently, the field of video editing has witnessed remarkable progress [29, 55, 57], particularly in terms of content modification while preserving the original structure and motion of the video [7, 51]. Despite these achievements, the necessity 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 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 [16, 23, 44] employ a sliding window to generate a new clip conditioned on the previous clip. 2) hierarchical methods [10, 16, 18, 56] generate sparse frames first, then interpolate intermediate frames. However, the autoregressive approach is susceptible to quality degradation due to error accumulation 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, limiting achieving good results in interpolation stage. In this paper, we generate continuous video clips in the auto-regressive way and exhibits superior performance in synthesizing long-term consistent frames. 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.

3. Method

In this section, 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 [35] for video generation. Latent diffusion model is trained to denoise from a perturbed input in the latent space.
The training procedure is illustrated in Figure 3. For the first 3.2. Training and Inference
then concatenated with noised latent $z$ padded where $f \in \mathbb{R}^{C \times H \times W}$, which are then concatenated along the channel dimension after perturbed with different noises.

of a pre-trained VAE, in order to reduce the computational burden. We take the widely used 2D UNet [36] 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 [23], 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 [33], and the embedding $c^{\text{text}}$ is injected through cross-attention layers in the UNet with hidden states as queries and $c^{\text{text}}$ 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 $\{I^{\text{first}}, I^{\text{last}}\}$, we first encode them into the input space of VAE, result in $\{f^{\text{first}}, f^{\text{last}}\}$ where $f \in \mathbb{R}^{C \times H \times W}$. To inject the instructions without loss of the temporal position information, the final image condition is then constructed as:

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

where $\text{PADs} \in \mathbb{R}^{(F-2) \times C \times H \times W}$. The condition $c^{\text{image}}$ is then concatenated with noised latent $z$ 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 $c^{\text{image}}$ of image instructions with noise.

Third, during training, we randomly drop the last frame instruction with probability $\eta$, by replacing the corresponding latent with zeros. Correspondingly, we propose a simple yet effective inference technique. During inference, in the first $\tau$ 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:

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

$\tau$ determines the strength of model dependency on last frame instruction, adjusting $\tau$ will enable various applications. For example, our model can generate high-dynamic videos without last frame instruction ($\tau = 0$). Additionally, we apply the classifier-free guidance [20] 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 [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 of video. Another nuisance issue of WebVid 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 for training ensures that PixelDance generates watermark-free videos if the image instructions are free of watermarks.

PixelDance is trained jointly on $<$text,video$>$ dataset and $<$text,image$>$ dataset. For video data, we
Table 1. Zero-shot T2V performance comparison on MSR-VTT. All methods generate video with spatial resolution of 256 x 256.

<table>
<thead>
<tr>
<th>Method</th>
<th>#data</th>
<th>#params.</th>
<th>CLIPSIM↑</th>
<th>FVD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CogVideo (En) [24]</td>
<td>5.4M</td>
<td>15.5B</td>
<td>0.2631</td>
<td>1294</td>
</tr>
<tr>
<td>MagicVideo [58]</td>
<td>10M</td>
<td>-</td>
<td>-</td>
<td>1290</td>
</tr>
<tr>
<td>LVDM [17]</td>
<td>2M</td>
<td>1.2B</td>
<td>0.2381</td>
<td>742</td>
</tr>
<tr>
<td>Video-LDM [6]</td>
<td>10M</td>
<td>4.2B</td>
<td>0.2929</td>
<td>-</td>
</tr>
<tr>
<td>InternVid [49]</td>
<td>28M</td>
<td>-</td>
<td>0.2951</td>
<td>-</td>
</tr>
<tr>
<td>ModelScope [46]</td>
<td>10M</td>
<td>1.7B</td>
<td>0.2939</td>
<td>550</td>
</tr>
<tr>
<td>Make-A-Video [40]</td>
<td>20M</td>
<td>9.7B</td>
<td>0.3049</td>
<td>-</td>
</tr>
<tr>
<td>Latent-Shift [1]</td>
<td>10M</td>
<td>1.5B</td>
<td>0.2773</td>
<td>-</td>
</tr>
<tr>
<td>VideoFactory [47]</td>
<td>-</td>
<td>2.0B</td>
<td>0.3005</td>
<td>-</td>
</tr>
<tr>
<td>PixelDance</td>
<td>10M</td>
<td>1.5B</td>
<td>0.3125</td>
<td>381</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>#data</th>
<th>#params.</th>
<th>IS↑</th>
<th>FID↓</th>
<th>FVD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CogVideo (En) [24]</td>
<td>5.4M</td>
<td>15.5B</td>
<td>25.27</td>
<td>179.00</td>
<td>701.59</td>
</tr>
<tr>
<td>MagicVideo [58]</td>
<td>10M</td>
<td>-</td>
<td>145.00</td>
<td>699.00</td>
<td></td>
</tr>
<tr>
<td>LVDM [17]</td>
<td>2M</td>
<td>1.2B</td>
<td>21.04</td>
<td>60.25</td>
<td>616.51</td>
</tr>
<tr>
<td>InternVid [49]</td>
<td>28M</td>
<td>-</td>
<td>21.04</td>
<td>60.25</td>
<td>616.51</td>
</tr>
<tr>
<td>ModelScope [46]</td>
<td>10M</td>
<td>1.7B</td>
<td>-</td>
<td>-</td>
<td>410.00</td>
</tr>
<tr>
<td>VideoFactory [47]</td>
<td>-</td>
<td>2.0B</td>
<td>-</td>
<td>-</td>
<td>410.00</td>
</tr>
<tr>
<td>Dysen-VDM [8]</td>
<td>10M</td>
<td>-</td>
<td>35.57</td>
<td>-</td>
<td>325.42</td>
</tr>
<tr>
<td>PixelDance</td>
<td>10M</td>
<td>1.5B</td>
<td>42.10</td>
<td>49.36</td>
<td>242.82</td>
</tr>
</tbody>
</table>

randomly sample 16 consecutive frames with 4 fps per video. Following previous work [22], we adopt LAION-400M [39] as <text, image> 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 [21] with $T = 1000$ time steps for training. A noise corresponding to 100 time steps is introduced to the image instructions $c_{image}$. We incorporate $\epsilon$-prediction [21] as training objective.

4.2. Video Generation

4.2.1 Quantitative Evaluation

We evaluate zero-shot video generation performance of our PixelDance on MSR-VTT [53] and UCF-101 [41] datasets, following previous work [6, 17, 24, 58]. 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 [35] to obtain the first frame instructions, and generate videos given the text and first frame instructions. Following previous work [8, 47], we randomly select one prompt per example to generate 2990 videos in total for evaluation, and report the Fréchet Video Distance (FVD) [43] and CLIP-similarity (CLIPSIM) [50] 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 Inception Score (IS) [38], Fréchet Inception Distance (FID) [19] and FVD, following previous work [8, 47].

Zero-shot 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 [46], with an FVD of 550. On UCF-101 benchmark, PixelDance outperforms other models across various metrics, including IS, FID and FVD.

4.2.2 Human Evaluation

Though above automatic evaluation metrics can assess the performance of models from various perspectives, they sometimes do not align well with human preference and fail to reflect the improvements in quality [22, 32, 40]. To make a more fair comparison, we use human evaluation to assess the performance of our PixelDance, Gen2 [12] and PiKa [31]. Specifically, we write 50 different prompts, covering diverse styles (real-world and cartoon, people and animal, wide-range landscapes and close-up), then we generate first frame with Stable Diffusion V2.1 [35] per prompt and generate one video for each method with text, first frame instructions. Users are asked to sort the randomly displayed videos per prompt (rating belongs to {3, 2, 1}, the higher, the better) w.r.t. two aspects: text faithfulness and video quality.

The human preference comparison result is demonstrated in Figure 4, where our PixelDance shows superior user preference in terms of both of text faithfulness and video quality, and outperforms Gen2 and PiKa with a clear margin.

4.2.3 Qualitative Analysis

Effectiveness of Each Instruction 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 described succinctly and precisely with text. Nonetheless, the text prompt plays a vital role of specifying various motions, including not limited to body movements, facial expressions, object movements, and visual effects (first two rows of Figure 5). Besides, it allows for manipulating camera movements with prompts like "zoom in/out" and "rotate", as shown in the last row of Figure 5. 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 7).

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 5 and Figure 7) 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 6. Furthermore, we can generate a natural shot transition using last frame instruction (last sample of Figure 7).

Strength of Last Frame Guidance With the proposed techniques detailed in Sec. 3, we intentionally avoid encouraging the model replicate the last frame instruction exactly. As shown in Figure 8, without our techniques, the generated video abruptly ends in the given last frame 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 PixelDance focuses on learning dynamics, 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
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, to validate the influence of our self-collected data which is used for watermark-free video generation, we train two T2V baselines, ① on WebVid and self-collected data, and ② only on WebVid data. These two baselines have similar performance, demonstrating that the self-collected data has negligible influence on improving video generation performance but generating watermark-free videos.

Based on the T2V baseline (③), we further analyze the effectiveness of different instructions employed in our model. Considering the indispensable nature of the first frame instruction for the generation of continuous video clips, our ablation focuses on the text instruction (②) and the last frame instruction (③). The experimental results indicate that omitting either 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 (③). This observation reveals that relying solely on the <text, first frame> 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 [9, 17] 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 9, show that PixelDance demonstrates lower FVD scores and smoother temporal variations, compared with auto-regressive models, TATS-AR [9] and LVDM-AR [17], and the hierarchical approach LVDM-Hi. Please refer to the Supplementary for visual comparisons.

4.4.2 Qualitative Analysis
This qualitative analysis focuses on PixelDance’s capabilities of generating a composite shot. This is formed by stringing together multiple continuous video clips that are temporally consistent. Figure 10 illustrates the capability of PixelDance to handle intricate shot compositions involv-
ing 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 generation capabilities, we have successfully synthesized a two-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 fine-tune PixelDance with image sketch [52] as the last frame instruction. The results are shown in the first two rows of Figure 11, 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 11, 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 exhibited exceptional proficiency in synthesizing videos with complex scenes and actions, setting a new standard in video generation.
References


[30] Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. 2018. 3

[31] PiKa. https://pika.art/. 2, 5, 6


