Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models

Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu*, Jiashi Li, Xin Xia, Xuefeng Xiao, Liang Lin

Sun Yat-Sen University

Project Page: https://controlavideo.github.io

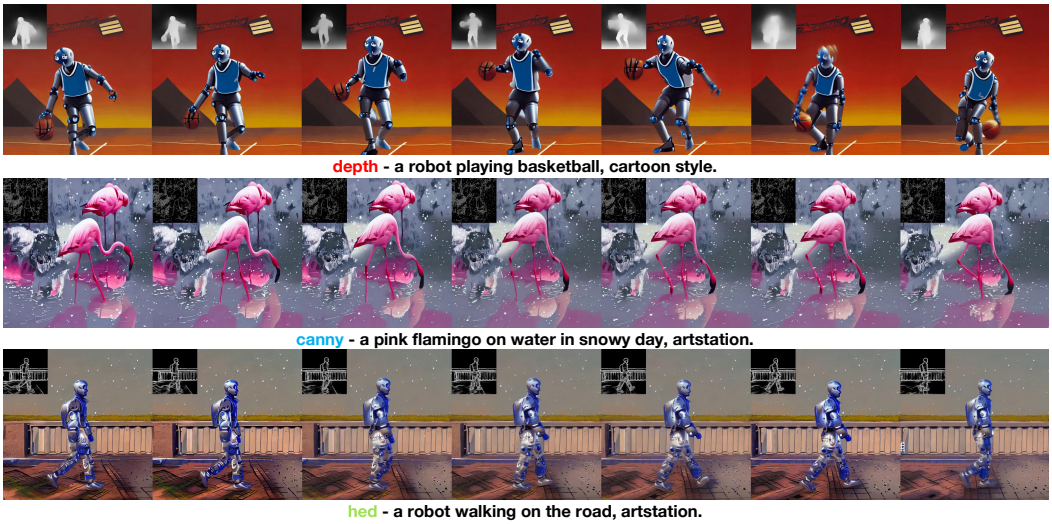

Figure 1: Our model generates high-quality and consistent videos conditioned on a text prompt and additional control maps, such as depth maps (first row), canny edge maps (second row), hed edge maps (third row).

Abstract

This paper presents a controllable text-to-video (T2V) diffusion model, named Video-ControlNet, that generates videos conditioned on a sequence of control signals, such as edge or depth maps. Video-ControlNet is built on a pre-trained conditional text-to-image (T2I) diffusion model by incorporating a spatial-temporal self-attention mechanism and trainable temporal layers for efficient cross-frame modeling. A first-frame conditioning strategy is proposed to facilitate the model to generate videos transferred from the image domain as well as arbitrary-length videos in an auto-regressive manner. Moreover, Video-ControlNet employs a novel residual-based noise initialization strategy to introduce motion prior from an input video, producing more coherent videos. With the proposed architecture and strategies, Video-ControlNet can achieve resource-efficient convergence and generate superior quality and consistent videos with fine-grained control. Extensive experiments demonstrate its success in various video generative tasks such as video editing and video style transfer, outperforming previous methods in terms of consistency and quality.

*Corresponding author.

Preprint. Under review.

1 Introduction

In recent years, there has been a rapid development in text-based visual content generation. Trained with large-scale image-text pairs, current T2I diffusion models [25, 29, 27] have demonstrated an impressive ability to generate high-quality images based on user-provided text prompts. Built on these pre-trained T2I models, personalizing generation [28, 5] and conditional generation [39, 18] provide more fine-grained control over the generated videos. Success in image generation has also been extended to video generation. Some methods leverage T2I models to generate videos in one-shot [37] or zero-shot [15, 22] manner, while videos generated from these models are still inconsistent or lack variety. Scaling up video data, text-to-video (T2V) diffusion models [31, 8] are able to generate consistent videos with text prompts, but these models generate videos in a fully automatic way, lacking control over the generated content. A recent study [4] propose a T2V diffusion model that allow for depth maps as control. However, a large-scale dataset is required to achieve consistency and high quality, which is resource-unfriendly. Additionally, it's still challenging for T2V diffusion models to generate videos of consistency, arbitrary length and diversity.

To address these issues, this paper presents a controllable T2V model, namely Video-ControlNet, which has the following advantages: (i) Improved consistency: Video-ControlNet employs motion prior and control maps to achieve better consistency. (ii) Generation of videos with arbitrary length: by adopting the proposed first-frame conditioning strategy, Video-ControlNet can auto-regressively generate videos of any length. (iii) Domain generalization: the model generates succeeding videos based on the initial frame, transferring knowledge from images to videos. (iv) Resource-efficient: our model is easy to converge with fewer training resources. e.g. , our batch size*steps are 16*10k compared to 1152*100k in Gen-1's [4].

Specifically, the Video-ControlNet is designed to generate videos based on text and reference control maps. In terms of model architecture, we develop our video generative model by reorganizing a pre-trained controllable T2I model [39], incorporating additional trainable temporal layers, and presenting a spatial-temporal self-attention mechanism that facilitates fine-grained interactions between frames. This approach allows for the creation of content-consistent videos, even without extensive training. To ensure video structure consistency, we propose a pioneering approach that incorporates the motion prior of the source video into the denoising process at the noise initialization stage. By leveraging motion prior and control maps, the Video-ControlNet is able to produce videos that are less flickering and closely resemble motion changes in the input video, while also avoid error propagation in other motion-based methods [14] due to the nature of multi-step denoising process.

Furthermore, instead of previous methods that train models to directly generate entire videos, we introduce an innovative training scheme that produces video predicated on the initial frame. With such a straightforward yet effective strategy, it becomes more manageable to disentangle content and temporal learning, since the former is presented in the first frame and the text prompt. Our model only needs to learn how to generate subsequent frames, inheriting generative capabilities from the image domain and easing the demand for video data. During inference, we generate the first frame conditioned on the control map of the first frame and a text prompt. Then, we generate subsequent frames conditioned on the first frame, text, and subsequent control maps. Meanwhile, another benefit of such strategy is that our model can auto-regressively generate an infinity-long video by treating the last frame of the previous iteration as the initial frame.

In summary, our contributions can be outlined as follows:

- We propose a controllable T2V diffusion model, named Video-ControlNet, by refactoring a controllable T2I model. Our model is able to generate text-guided videos conditioned on various types of control maps.
- We introduce a residual-based noise initialization strategy that incorporates motion from the input video into the diffusion process, resulting in the generation of videos that are less flickering and motion-aligned.
- We present a novel first-frame conditioning strategy that not only empowers our model to generate videos generalized from the image domain but also to generate arbitrary-length videos auto-regressively.
- Experiments demonstrate that our framework is capable of generating higher-quality, more consistent videos using fewer training resources.

2 Related Work

2.1 Diffusion Model For Text-to-Image Generation

Denoising Diffusion Probabilistic Model [9, 32] has shown impressive results in generating high-quality images, outperforming previous approach generative adversarial networks (GANs) [6]. Training with text guidance, users can generate images with text input. GLIDE [20] employs classifier-free guidance and train diffusion model in large-scale text-image pairs. DALLE-2 [25] further uses the latent space of CLIP [23] as a condition to improve performance. Imagen [29] employs a T5 [24] and cascaded diffusion models to generate high resolution images. Latent Diffusion Model (LDM) [25] proposes to forward the diffusion process in latent space, which is more efficient than other diffusion models. Although existing powerful T2I models allow image generation with free text, many works are proposed to enhance the control of image generation. Textual Inversion [5] and Dreambooth [28] propose to finetune the T2I model in a few images so that we can generate personalized content. Prompt2Prompt [7] proposes to use cross-attention maps to manipulate generated content and Null-Text Inversion [17] improves it to allow real image editing through optimization. Plug-and-Play [33] edit the image by manipulating different network layers. InstructPix2Pix [2] finetune the T2I model on generated instruct-image pairs, which allow editing the image by instruction. Most recently, ControlNet [39] and T2I-Adapter [18] add extra layers to the T2I models and train the model to generate images conditioned on specific control maps such as edge, pose, depth map, which is a milestone in controllable generation. We get inspiration from ControlNet and extend it to controllable video generation.

2.2 Video Editing and Text-to-Video Generation Models

GAN-based translation models such pix2pix [13] and vid2vid [35] are domain-specific and can not be generalized to other domains. Ebsynth [14] uses optical flow and structure from an input video to generate style transfer videos conditioned on a style image, while it is difficult to handle the error propagation. Text2Live [1] proposes editing the video on a flattened texture map, but optimization during inference is required. Diffusion-based models have also been used to generate videos with open text prompts. Some methods are looking to generate video based on T2I models. Tune-A-Video [37] proposes one-shot video generation by finetuning latent diffusion on a single video. FateZero [22] and Text2Video-Zero [15] explore the latent space of DDIM and generate video directly with pre-trained T2I models. However, it is a challenge to generate consistent videos with image models and therefore many video pre-training models are proposed to generate consistent videos. For example, transformer-based architecture such as NUWA [36], CogVideo [12], Phenaki [34] is capable of generating open domain videos with text. Video Diffusion Models (VDM) [11] extend text-to-image diffusion models to video generation. Imagen Video [8] improves VDM and trains on large-scale image-text pairs with cascaded diffusion and v-prediction parameterization. Make-A-Video [31] and MagicVideo [40] also train the video diffusion model on large-scale datasets and are able to generate impressive videos. However, these video models cannot generate controllable content since only free text is not strict enough. Currently, Gen-1 [4] proposes a diffusion model that generates videos conditioned on depth maps and text, which is the most relevant work to us.

3 Method

In this section, we will first introduce the diffusion models for image generation, including Latent Diffusion Model [9] and ControlNet [39]. Next, we will introduce the model architecture of the Video-ControlNet, which is based on the image diffusion models. Subsequently, we present a novel residual-based noise initialization approach that incorporates prior motion information for consistent video generation. Lastly, we will elaborate on the training and inference procedure of the Video-ControlNet with the proposed first-frame conditioning mechanism.

3.1 Preliminary: Latent Diffusion Model and ControlNet

Given an input signal x_0, a diffusion forward process is defined as:

$$p_\theta(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_{t-1}}x_{t-1}, \beta_t I), \quad t = 1, ..., T \tag{1}$$

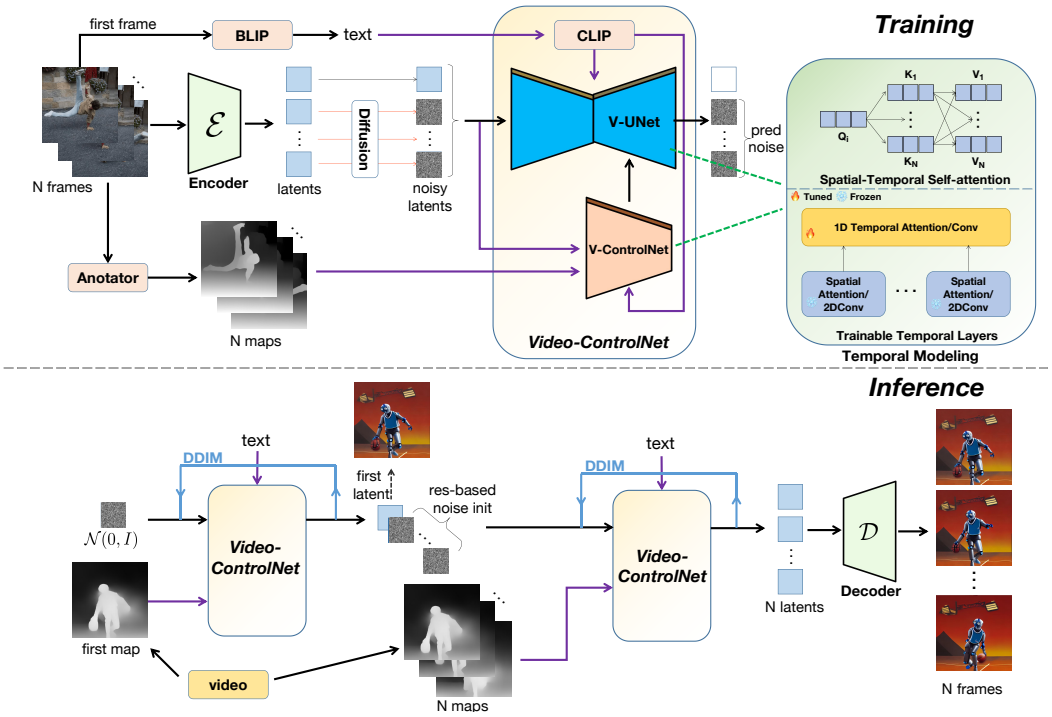

Figure 2: **Model Architecture:** We apply spatial-temporal self-attention and trainable temporal layers to both the image UNet and the image ControlNet, enabling the model to generate videos. **Training:** The model input includes video and its text prompt from BLIP captioning [16] and control maps from an annotator (e.g. depth estimation model). Each frame is passed to the encoder to get the latent code. We add residual-based motion noise to each latent except for the first frame and train the model to predict the subsequent noise conditioned on the first frame. We treat images as videos of one frame and we train on videos and images jointly. **Inference:** After training, our model is able to generate the first frame conditioned on its control map. The generated first frame is then used to generate subsequent frames with the motion prior. We can also auto-regressively generate longer videos conditioned on the last frame of the previous iteration.

where T is the total timestep of the diffusion process. A noise depending on variance β_t is gradually added to x_{t-1} to obtain x_t at the next timestep and finally reach $x_T \in \mathcal{N}(0, I)$. The goal of the diffusion model [9] is to learn to reverse the diffusion process (denoising). Given a random noise x_t, the model predicts the added noise at the next timestep x_{t-1} until the origin signal x_0.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad t = T, ..., 1 \tag{2}$$

We fix the variance $\Sigma_\theta(x_t, t)$ and utilize the diffusion model with parameter θ to predict the mean of the inverse process $\mu_\theta(x_t, t)$. The model can be simplified as denoising models $\epsilon_\theta(x_t, t)$, which are trained to predict the noise of x_t with a noise prediction loss:

$$min_\theta ||\epsilon - \epsilon_\theta(x_t, t, c_p)||_2^2 \tag{3}$$

where ϵ is the added noise to the input image x_0, the model learn to predict the noise of x_t conditioned on text prompt c_p at timestep t.

In latent diffusion [25] , the signal in the diffusion process is a compressed latent code z rather than the image signal x. Using an auto-encoder, the image x is encoded by an encoder $\mathcal{E}$ to obtain the latent code $z = \mathcal{E}(x)$, and the model learns to denoise in latent space. During inference, the reconstructed latent code z_0 can be reconstructed by a decoder $\mathcal{D}$, $x_0 = \mathcal{D}(z_0)$ to obtain the generated image.

ControlNet [39] is a neural network architecture that enhances pretrained image diffusion models with task-specific conditions by utilizing trainable layers copied from the original diffusion model.

These layers are then fine-tuned based on specific control maps such as edge, depth, and segmentation inputs. The loss with additional control can be formulated as:

$$min_\theta||\epsilon - \epsilon_\theta(x_t, t, c_p, c_f)||_2^2 \tag{4}$$

where the control map c_f is an additional control. Our research draws inspiration from ControlNet and expands its application into video synthesis. In the case of a video, the input signal x and control c_f is extended to a sequence of N frames.

3.2 Spatial-Temporal Modeling for Video Generation

T2I models are designed to generate images in the spatial domain. Our proposal is to restructure the controllable T2I model known as ControlNet, transforming it into a T2V model that boasts spatial-temporal structures for the generation of videos. Specifically, we make two kinds of architectural refactoring: (i) We introduce an additional temporal layer after each 2-dimensional (2D) layer, including 2D convolution and 2D attention, similar to previous works [31, 4]. (ii) We modify the 2D spatial self-attention layer to incorporate temporal information, resulting in spatial-temporal self-attention that facilitates fine-grained modeling.

Specifically, given an input tensor $h \in R^{F\times C\times H\times W}$, where F, C, H, W denote the number of frames, channels, image height and width, respectively. As illustrated on the right of Figure 2, the feature of each frame $h_i \in R^{C\times H\times W}$ are passed through a 2D convolution layer $Conv_{2D}()$, and then sent jointly to a trainable 1D convolutional layer $Conv_{1D}()$ that works in the temporal domain. The $Conv_{1D}()$ layer is initialized as the identity function, which results in independent processing of each frame when training is not taken. In a similar manner, following the 2D cross-modal attention layer, we introduce another attention layer for temporal coherence attention. Initially, we send each h_i through each cross-modal spatial attention layer and then collectively feed them into an additional temporal attention layer that are initialized as the identity function as well. As ControlNet's network structure is identical to the original UNet, we replicate the same process to accommodate a sequence of control maps as input.

Furthermore, we propose adjusting the spatial self-attention mechanism to incorporate spatial-temporal self-attention across frames for fine-grained modeling, which can be formulated as:

$$SelfAttn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V, \tag{5}$$

$$Q = W^Q\bar{v}_i, K = W^K[\bar{v}_0, ..., \bar{v}_{N-1}], V = W^V[\bar{v}_0, ..., \bar{v}_{N-1}] \tag{6}$$

where $\bar{v}_i$ denotes the token sequence of frame i, and $[\bar{v}_0, ..., \bar{v}_{N-1}]$ denotes the concatenation of the N frames.

We concatenate features K, V of N frames so that each position has a global perception of video noise.

After we get the new K, V matrix with the original parameters W^K and W^V as Eq. 6, we apply the same operation as the original self-attention (Eq. 5).

Algorithm 1: Residual-based Noise Initialization

We sample a noise x with N frames
$InputNoise : x = [x^n \leftarrow \mathcal{N}(0, I), n = 0, ..., (N-1)]$
$InputVideo : v = [v^n, n = 0, ..., (N-1)]$
R_{thres} : The threshold for residual change
$DownSample()$: Resize to align the size of noise
i=1 **for** *i<N* **do**
    $res = norm(v^i - v^{(i-1)})$
    $res_{mask} = res > R_{thres}$
    res_{mask} = $DownSample(res_{mask})$
    $x^i = [x^i - x^{(i-1)}] * res_{mask} + x^{(i-1)}$
end

3.3 Residual-Based Noise Initialization

A video comprises a sequence of images that feature numerous identical pixels across frames. In video compression, it is typical to store keyframes along with residuals/motion, instead of individual pixels for each frame. Inspired by this, we propose to introduce residual prior from the input video into the diffusion process.

Intuitively, our goal is to maintain the same noise in areas where the pixels of the input video remain unchanged between frames and use different noise in areas where the pixels of the input video change between frames. Previous research on the diffusion model [7] has supported the idea that noise in early denoising steps can determine the layout of the generated image.

Therefore, instead of randomly sampling different noise for each frame, we propose to sample noise based on the residual of the input video frames, which captures the motion change of each frame. Our algorithm 1 involves targeting the locations where RGB values exceed a predefined threshold R_{thres}. To generate residual-based noise, we first randomly sample Gaussian noise for each frame. Then, we compute the residual between two frames and apply it to reset the noise distribution, which ensures that static regions have the same noise between frames, while moving regions have different noise. Meanwhile, the threshold empowers us to regulate the video's smoothness. The higher the threshold, the more areas will maintain the same noise as their preceding frame. If the threshold is set to one, the noise will be the same for all frames, which may lead to artifacts. If the threshold is set to zero, the noise for each frame is sampled independently, which may lead to flickering.

3.4 Training and Inference with First-frame Conditioning

Training: A naive approach for temporal learning in a video diffusion model would be to train the model to predict the entire video sequence. However, this would require a large amount of video data to learn the diversity that has already presented in the image domain. To enhance the effectiveness of model training, we propose a first-frame conditioning method that generates video sequences conditioned on the first frame. This approach reduces the need for the model to memorize video content in the training set and instead focuses on learning to reconstruct motion, which makes it possible to achieve better results with a smaller video dataset. As shown in Figure 2, during training, we do not add noise to the first frame, so the model learns to generate subsequent frames based on the first frame v^1 of input video, text prompt c_p, and control maps c_f. Moreover, we introduce an extra condition, the latent code of the first frame $\mathcal{E}(v^1)$. Thus the loss function can be formulated as:

$$min_\theta ||\epsilon - \epsilon_\theta(x_t, t, c_p, c_f, \mathcal{E}(v^1))||_2^2 \tag{7}$$

By adopting this approach, our model can effectively utilize the motion information from the control maps and follow the content from the first frame. This simple yet effective strategy not only allows our model to generalize the domain from image to video, but also to auto-regressively generate longer videos. Additionally, we train image and video simultaneously, which enable the model to generate images with control as well.

Inference: We generate the initial frame, denoted as v^1, by providing the model with Gaussian noise in the form of a single frame x^1 along with conditioning factors including a text prompt c_p and a first frame control map c_f^1.

$$v^1 = VideoControlNet(x^1, c_p, c_f^1) \tag{8}$$

Once we obtain the first frame v^1 of the video, we generate the following frames conditioned on its latent code $\mathcal{E}(v^1)$:

$$v = VideoControlNet(x, c_p, c_f, \mathcal{E}(v^1)) \tag{9}$$

With our proposed method of first-frame conditioning, our model is capable of generating video sequences with greater diversity than what is present in the training data. Additionally, our model has a distinct advantage in creating longer videos by utilizing previously generated frames as the initial frame in the subsequent iteration. This allows us to use an auto-regressive approach to produce videos of any length, which sets us apart from other video diffusion models that are limited to generating videos only once.

Classifier-Free Guidance: Classifer-free guidance [10] with a guidance scale ω_t to sample the noise can be formulated as:

$$\hat{\epsilon}_\theta(x_t, t, c_p, c_f) = \epsilon_\theta(x_t, t, \emptyset, c_f) + \omega_t(\epsilon_\theta(x_t, t, c_p, c_f) - \epsilon_\theta(x_t, t, \emptyset, c_f)) \tag{10}$$

where $\emptyset$ denotes a null-text prompt, and $\epsilon_\theta(x_t, t, \emptyset, c_f)$ represents negative representation. Based on this, we incorporating a sampling strategy in [4] that treats noise prediction of video generated frame-by-frame as a negative representation needed to be avoided. Consequently, the final prediction of noise is calculated as:

$$\begin{aligned}\hat{\epsilon}_\theta(x_t, t, c_p, c_f) =& \epsilon_{\theta I}(x_t, t, \emptyset, c_f) + \omega_v(\epsilon_\theta(x_t, t, \emptyset, c_f) - \epsilon_{\theta I}(x_t, t, \emptyset, c_f)) + \\ & \omega_t(\epsilon_\theta(x_t, t, c_p, c_f) - \epsilon_\theta(x_t, t, \emptyset, c_f))\end{aligned} \tag{11}$$

Here, ω_v denotes the scale of video guidance, and $\epsilon_{\theta I}(x_t, t, \emptyset, c_f)$ denotes the prediction that each video frame is independently predicted. Just as a larger w_t can enhance text guidance, a larger w_v will result in a smoother overall effect.

4 Experiments

4.1 Implementation Details

DataSet Settings: 100k video clips sourced from the Internet and 100k image-text pairs obtained from Laion [30].

Training Settings: Our model is initialized with pre-trained weights from Stable Diffusion v1.5 [25] and ControlNet [39]. We only train the temporal layers with an 8:2 ratio of video to image rate. The resolution is set to 512×512, the batch size to 16, the learning rate to 10^{-5}, and the total number of training steps to $10k$. The model is evaluated using three control types: canny edge maps [3], hed edge maps [38], and depth maps.[26].

Inference Settings: The noise initialization threshold is set to 0.1, the scale for text guidance is 10.0, the scale for video guidance is 1.5, and DDIM uses 20 sampling steps.

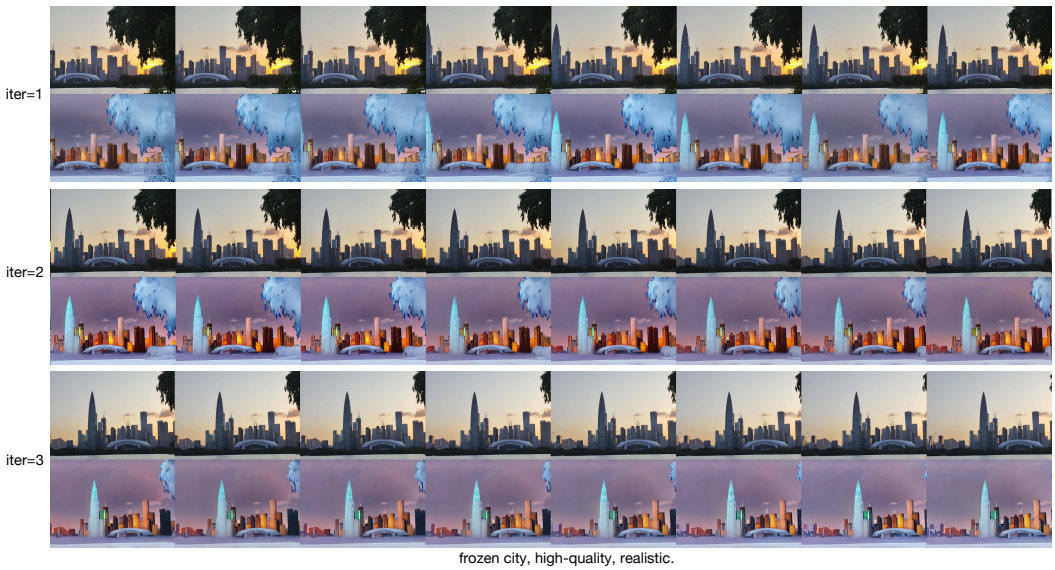

Figure 3: **Auto-Regressive Generation**: Our model is able to generate long videos auto-regressively. The first row is the first iteration, the second row is generating conditioned the last frame of the first iteration, and so on the third iteration.

4.2 Qualitative Results

Generating Videos with Different Types of Controls: We showcase three types of controls extracted from video to demonstrate our system's capacity to generate videos conditioned on various control types, as shown in Figure 1 and more results in supplementary material. Through experimentation,

we found that depth maps provide less structural information than edge maps, resulting in more diverse video outputs. Edge maps, on the other hand, produce videos with enhanced details but a lesser degree of variability. For instance, in the first row of Figure 1, we transform a human into a cartoon robot using depth map control. In contrast, using edge maps in the third row still results in a human-robot transformation, but with more intricate details retained, such as the clothing pattern.

Auto-Regressive Long Video Generation: As illustrated in Section 3.4, we present the outcomes of auto-regressively producing videos in Figure 3. The source video consists of 24 frames, and our objective is to convert a city at sunset into a frozen city. We generate the edited video through three iterations, each comprising eight frames. The generated video exhibits consistency across various iterations, which attests to the efficacy of our proposed approach in producing long videos through auto-regression based on the last frame of the previous iteration.

4.3 Comparison with Previous Methods

We compare our results with Gen-1 [4] and Text2Video-Zero [15] that have the same settings that generating videos with additional controls. Gen-1's results are obtained by running its product and Text2Video-Zero from their open-source code.

Qualitative Comparison: To demonstrate the superior performance of our model compared to others, we chose a video with significant changes in motion. As shown in Figure 4, the first row depicts the original video, from which we extracted depth maps and used a prompt 'a dog running through a field of poles, cyberpunk style' to create a style transfer video. In the second row, we present the most consistent and text-aligned video generated by our model. The result of Gen-1, shown in the third row, are less consistent and lack details of the dog. The last row exhibits the results of Text2Video-Zero, which fails to maintain consistency, as pointed out by the red circles. More Comparison will be shown in supplementary material.

Figure 4: **Qualitative Comparison:** We choose a hard case of a fast moving dog. Gen-1's result falls short of generating a detailed dog, whereas the result produced by Text2Video-Zero yields inconsistency of a dog. Ours exhibits a superior caliber compared to the aforementioned alternatives.

Quantitative Comparison: We adopt depth maps from 20 video clips from Davis [21] and in-the-wild videos, which are used to generate videos based on a given text prompt. To assess consistency, we measured **the depth map errors** between input and output videos. To evaluate text alignment, we calculated the cosine similarity between the X-CLIP [19] video embeddings for the output videos and text embeddings for the given prompts. As presented in Table 1, all methods displayed strong text alignment. Our method yielded the least depth map errors compared to the other two.

User Study: 18 participants were surveyed to evaluate the textual alignment and consistency of the generated videos by utilizing a rating scale ranging from 1 to 5. The data presented in Table 1

indicates that our model yields videos that demonstrate greater consistency aligned the quantity result of depth error findings, with the text alignment score exhibiting comparable results.

Table 1: Comparison of quantitative and user study with existing methods for T2V conditioned on depth maps in terms of text alignment and consistency.

Type	Text Align ↑	Dep Errors ↓	Text Align(User) ↑	Consis(User) ↑
Gen-1 [4]	0.252	0.112	4.25	3.89
Text2Video-Zero [15]	0.255	0.139	3.68	3.21
Ours	0.259	**0.091**	4.18	**4.18**

Figure 5: **Ablation Study of Different Thresholds:** We present three thresholds: 0.0, 0.1 and 1.0. Random noise(thres=0.0) allows for large changes but may result in flickering (e.g. the wave turning across frames). Same noise(thres=1.0) for each frame leads to smoothness and artifacts. Residual-based noise is a good trade-off for both changeable and consistent.

4.4 Ablation Study of Residual-based Noise Initialization

We analyze the effects of three types of noise: identical noise (threshold=1.0), distinct noise (threshold=0.0), and motion-enhanced noise (threshold=0.1). It is important to note that this threshold also regulates the smoothness of the resulting video. As shown in Figure 5, when there is no residual control, the background flickers, as seen in the second row. For example, the water changes color frequently. The third row displays the threshold we selected, which effectively balances consistency and smoothness to produce satisfactory results. The final row shows the consequence of using the same noise for each frame, a smooth video, but with severe artifacts. Overall, incorporating residual-based noise can introduce motion from the input video, resulting in videos with reduced flickering and better consistency.

5 Conclusion

In this paper, we propose a controllable T2V framework that is capable of generating videos conditioned on text prompts and control maps. With the proposed motion-enhanced noise initialization and first-frame conditioning strategy, our model can generate coherent, text-aligned, arbitrarily long videos after temporal learning in a small video dataset. While our method achieves impressive results, it still has some known limitations. For example, our T2V model relies on a T2I model and shares the same bad cases. If the input control maps extracted from another model is not consistent, it also brings errors to the generated videos. In the future, it's worth conducting research on the stability and controllability of the video generation models.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing, Apr 2022.

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[3] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 679–698, Jan 2009.

[4] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023.

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[8] 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.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models., Jan 2020.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance.

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[12] 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.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov 2017.

[14] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics*, 38(4):1–11, Jul 2019.

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

[17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

[18] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[19] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022.

[20] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

[21] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and LucVan Gool. The 2017 davis challenge on video object segmentation, Apr 2017.

[22] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[23] 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.

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and PeterJ. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, Oct 2019.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents.

[26] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1623–1637, Aug 2020.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Sep 2022.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, Aug 2022.

[29] 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.

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[31] 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.

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, Oct 2020.

[33] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

[34] 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 description. *arXiv preprint arXiv:2210.02399*, 2022.

[35] Tingchun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis, Oct 2019.

[36] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022.

[37] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022.

[38] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Feb 2016.

[39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[40] 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.