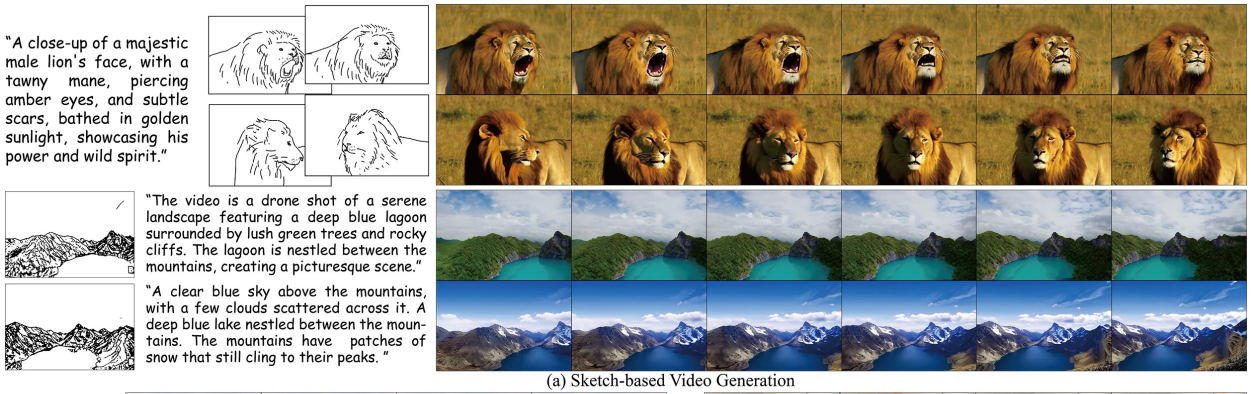
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

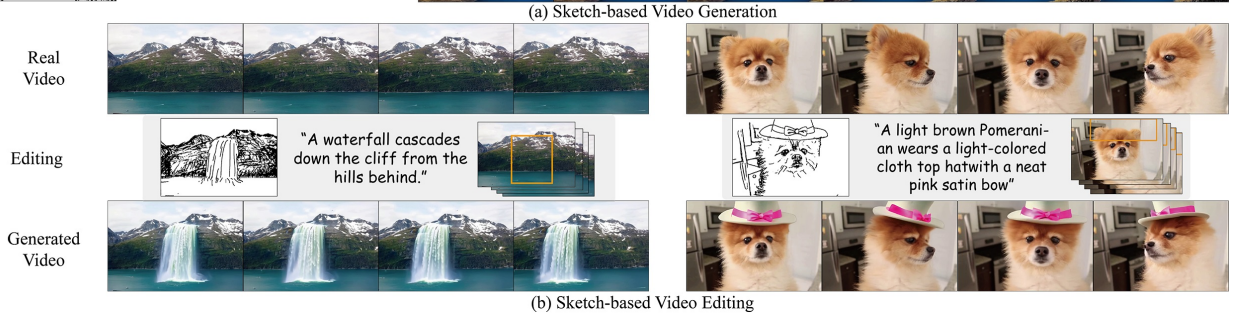
SketchVideo: Sketch-based Video Generation and Editing

Feng-Lin Liu[1,2] Hongbo Fu[3] Xintao Wang[4] Weicai Ye[4] Pengfei Wan[4] Di Zhang[4] Lin Gao[1,2*]

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Hong Kong University of Science and Technology [4]Kuaishou Technology



(a) Sketch-based Video Generation



(b) Sketch-based Video Editing

Figure 1. Our method enables high-quality video generation (a) and editing (b) based on sketch and text inputs. (a) Top: With the same text prompt, different keyframe sketches lead to results with similar semantics but diverse sketch-faithful geometry. (a) Bottom: With the same sketches, varied text prompts yield diverse appearances. (b) Users can also edit real videos by drawing on keyframe sketches, with edits automatically propagated even when edited objects in original videos have translation and rotation.

Abstract

Video generation and editing conditioned on text prompts or images have undergone significant advancements. However, challenges remain in accurately controlling global layout and geometry details solely by texts, and supporting motion control and local modification through images. In this paper, we aim to achieve sketch-based spatial and motion control for video generation and support fine-grained editing of real or synthetic videos. Based on the DiT video generation model, we propose a memory-efficient control structure with sketch control blocks that predict residual features of skipped DiT blocks. Sketches are drawn on one or two keyframes (at arbitrary time points) for easy interaction. To propagate such temporally sparse sketch conditions across all frames, we propose an inter-frame attention mechanism to analyze the relationship between the keyframes and each video frame. For sketch-based video editing, we design an additional video insertion module that maintains consistency between the newly edited content and the original video's spatial feature and dynamic motion. During inference, we use latent fusion for the accurate preservation of unedited regions. Extensive experiments demonstrate that our SketchVideo achieves superior performance in controllable video generation and editing.

1. Introduction

Diffusion-based text-to-image [12, 51, 53] and text-to-video [27, 35, 43, 73, 84] models advance significantly due to improvements in datasets [4, 42, 55, 61] and denoising network architectures [12, 51]. While text prompts effectively describe high-level semantics, they lack control of scene

*Corresponding Author: Lin Gao (gaolin@ict.ac.cn)

layouts and geometric details. To address this, existing video generation methods [2, 25, 73] utilize images as additional conditions but raise the questions of how to generate input images and achieve detailed editing. Sketching serves as a user-friendly interaction tool to capture spatial content and shape details accurately. One or two sketches are already sufficient to convey desired scene structures and motion information for short video clips (around 6 seconds), which are the target of our and most existing video generation methods. However, using such sparse keyframe sketches presents several challenges, including reasonably completing the missing frames, improving memory efficiency, and addressing the limited size of video datasets.

A naïve solution is to translate the input keyframe sketches into images and then utilize interpolation methods [16, 37, 70] for video generation. However, it is nontrivial to ensure consistency during keyframe sketch-to-image generation, which significantly affects the video quality. This approach also struggles to generate extrapolation frames when applying conditions in intermediate frames rather than beginning and ending time points. Another possible approach is to utilize white placeholders to fill missing condition frames and directly apply ControlNet [77] into video models, similar to SparseCtrl [22]. However, this requires the same network to process both sketches and white placeholders simultaneously, while the pretrained blocks handle tasks far from this sparse propagation. Additionally, for DiT-based video frameworks [27, 84], the traditional strategy [11] that copies half of the pretrained model as a condition network easily causes the out-of-memory issue.

To address these issues, we propose a novel sketch condition network specifically designed for the DiT-based video generation architecture (CogVideoX [73] in our work). Following ControlNet [11, 77], we employ a trainable copy of CogVideoX's DiT block to process only the sketch inputs and generate control features. No white placeholder is processed to align with the pretrained weights and reduce learning complexity. To propagate these keyframe features, we design an inter-frame attention mechanism that captures the relationship between the control keyframes and all video frames. Our approach computes query and key features from noisy latent while extracting value features from sketch conditions. This design leverages frame-to-frame similarity for control propagation. The above components consist of a single sketch control block. Instead of copying a half number of the pretrained blocks to construct the control network [11], we use only 5 sketch control blocks out of the 30 DiT blocks available in CogVideoX-2b. We design a novel uniformly distributed skip structure to add the control signals to different levels of features in discrete blocks (0, 6, 12, 18, and 24), achieving effective spatial control while improving memory efficiency. During training, an external image dataset is incorporated to solve the challenge of limited video data.

Beyond generation, interactive editing of real or synthetic videos further enhances creative flexibility. Existing methods [34, 40, 45, 71] achieve interesting text-based editing or effectively propagate single image editing into videos. Despite their effectiveness in appearance modification, they struggle with shape manipulation and object insertion, as they preserve the original temporal motion. Such information is missing for newly introduced content. Moreover, precisely identifying and preserving unedited regions for localized editing remains a challenge.

We propose a sketch-based editing method for detailed local modification. Rather than editing a single image and propagating changes, we directly construct an editing network based on our sketch control network. To analyze the relationship between edited regions and the original video, we incorporate a video insertion module that takes the original video with masked edited regions as inputs. The modified sketch control blocks generate residual features that capture temporally and spatially coherent contents in the edited areas. To accurately preserve unedited regions and achieve seamless fusion, the newly edited regions are blended with the original video in the latent space.

Extensive experiments demonstrate that our method outperforms existing approaches in video generation and editing. Our contributions are summarized as follows: 1) We propose SketchVideo, a novel sketch-based video generation and editing framework that enables detailed geometry control and manipulation using keyframe sketches, as shown in Fig. 1. 2) We design a sketch condition network that predicts skipped control features for the DiT framework, with an inter-frame attention mechanism to propagate one or two sketch conditions across the video. 3) We propose a video insertion module that analyzes the relationship between drawn sketches and original videos, utilizing a latent fusion strategy to preserve unedited regions accurately.

2. Related Work

Sketch-based Image Generation. GAN-based methods [21] have achieved great success in category-restricted sketch-to-image translation [13, 14, 50, 65, 75, 85]. Recently, diffusion-based text-to-image models [51, 53] handle general categories with conditional models like ControlNet [77], T2I-Adapter [41] and further advances [30, 46, 48, 76, 83]. Beyond U-Net, the DiT backbone [12] enables image generation, with PIXART-δ [11] utilizing the first half of the pretrained model's blocks as a trainable network to predict corresponding control residual features. Video generation, however, introduces additional challenges. For ease of interaction, we expect sketches specified only for a sparse set of keyframes, making it difficult to generate frames without sketch inputs. Additionally, video genera-

tion costs significantly higher memory resources, making methods [11, 77] that replicate half of the base model as a sketch encoder easily out of memory.

Diffusion-based Video Generation. VDM [26] pioneered diffusion-based video generation with a 3D U-Net denoising network. To improve quality, subsequent works [2, 3, 9, 10, 23, 25, 33, 64] integrated temporal modules into text-to-image models [51] to enable text- and image-conditioned video synthesis. Considering the efficiency, the DiT architecture is further used in Sora [43] and open-source projects [35, 60, 84]. Despite these advancements, subtle flickering artifacts remain in the results. CogVideoX [27, 73] further proposes a 3D full attention that merges the spatial and temporal attention, facilitating the generation of long-duration and high-resolution videos.

Building on existing video generation models, various conditions have been introduced to control the generation, such as camera movement [24, 66, 72], subject identity [31, 68], key-point trajectory [36, 56, 81] and example motions [67, 79]. However, they often lack control over spatial layouts and geometric details. Some methods [15, 18, 80] address this by extending image-based ControlNet [77] to videos but require all-frame conditions that are tedious for sketch interaction. SparseCtrl [22] tackles this by using white images for completions. However, due to the restriction of its base model [23] and a simple network design, its results suffer from temporal flickering. Similar ideas have been applied to cartoon interpolation [70] and colorization [28] with line art as input, but their outputs are limited to cartoon style. Our method utilizes sparse inputs, including hand-drawn sketches on one or two keyframes, to generate temporally stable and realistic videos. Additionally, our method further supports the sketch-based detailed editing of existing videos.

Deep Learning-based Video Editing. Pioneer works achieve video editing by style transfer [29, 52], GAN inversion [39, 62], and layered representations [32]. The advent of diffusion models further provides new editing paradigms. One category of such methods leverages image generation models to achieve compelling editing results, using temporal consistency techniques such as layered representations [6, 44], cross-frame attention [5, 17, 19, 57] and pixel warping [71]. A second category of methods employ video generation models. These works propagate the edits applied on the first frame into the other frames [34, 45], or utilize an inpainting strategy to achieve text-based editing [82] and motion modification [40]. To achieve text-based large-scale shape modification while maintaining motion features, a space-time feature loss [74] is designed to guide the inference process. Our method moves beyond traditional text- or image-based editing approaches, allowing users to draw one or two keyframe sketches at arbitrary time points for interactive video editing. Additionally, our method effectively handles sketch-based shape manipulation and dynamic object insertion, which are challenging for previous works.

Controllable Attention Mechanism. Attention across frames is initially used to ensure temporal consistency in AnimateDiff [23] and subsequent video generation models. Subsequent works, such as VideoBooth [31] and StillMoving [8], utilize it to learn identity-aware features for customized video generation. For video editing, existing works [17, 20, 57] utilize cross-frame attention to capture the temporal motion of input videos, enabling effective propagation of editing operations. Our method applies this idea to pixel-aligned sketch-based video generation and propagates the spatial geometric conditions instead of identity customization. We use a new feature derivation strategy for enhanced spatial control, differentiating our approach from traditional cross-attention mechanisms.

3. Methodology

This section introduces our sketch-based video generation and editing framework. In Sec.3.1, we provide an overview of CogVideoX-2b [73], a pretrained text-to-video generation method, which we will use for sketch-based video generation in our work. In Sec.3.2, we describe our sketch condition network specifically designed for the DiT architecture, which contains sketch control blocks to predict residual features. Within each control block, an inter-frame attention mechanism is designed to propagate the input sketches. In Sec.3.3, we detail the editing framework, which incorporates a video insertion module and latent fusion to preserve the features of the original video.

3.1. Preliminary

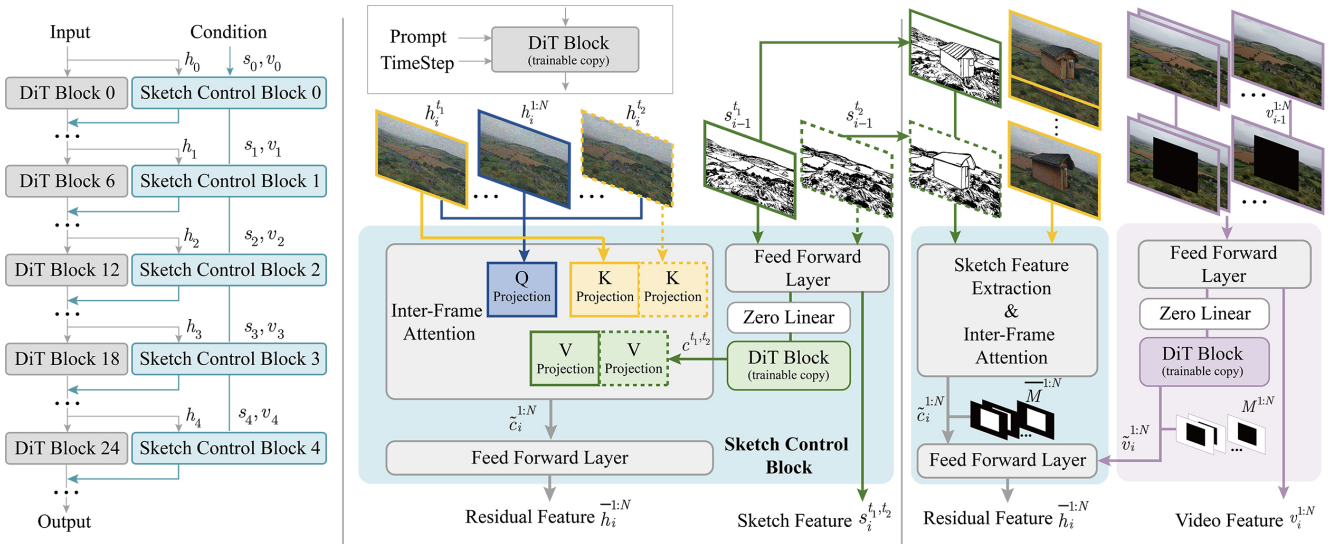
CogVideoX [73] is a text-to-video generation model that employs a 3D causal VAE and builds diffusion within the latent space. In CogVideoX-2b, the VAE model performs 8×8 spatial and 4× temporal downsampling, followed by a DiT architecture with 30 blocks for video generation. Within each block, a 3D full attention processes the concatenated text embeddings and patchified video latents, followed by a feed-forward layer to output the features. The 3D full attention merges the commonly used separate spatial and temporal attention [2, 23, 84] to improve the temporal coherence. The training objective of diffusion is:

$$L(\theta) := \mathrm{E}_{t,x_0^{1:N},y,\epsilon}\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0^{1:N} + \sqrt{1-\bar{\alpha}_t}\epsilon, t, y)\right\|^2, \quad (1)$$

where t is sampled between 1 and T (denosing steps), ϵ is the random noise, y is text prompts, and $x_0^{1:N}$ is the video data with N frames. The v-prediction [54] and zero SNR [38] are utilized for the diffusion setting.

3.2. Sketch-based Video Generation

Given text prompts and one or two keyframe sketches with corresponding time points t_1, t_2, our method generates



(a) Sketch Condition Network (b) Control Block for Generation (c) Control Block for Editing

Figure 2. Our framework for sketch-based video generation and editing. (a) Our sketch condition network for the DiT-based video generation architecture has a skip structure and five sketch control blocks that predict residual features. (b) For generation, features are extracted from temporally sparse input sketches and propagated through inter-frame attention. The input sketches are provided for one or two keyframes (the second sketch is shown as a dotted line). In the top left corner of (b), the prompt and timestep inputs are shown. (c) For editing, the same sketch control block (b) is utilized, with an additional video insertion module and video masks $M^{1:N}$ to analyze the relationship between edited and unedited regions. The 3D causal VAE is omitted to save space.

video clips that respect the input text prompts and sketches. As shown in Fig. 2, we design a sketch condition network for effective control. The input sketches are encoded into the latent space by the pretrained VAE, followed by patchifying and time-aware positional embedding to generate sketch latents $s_0^{t_1,t_2}$.

Skip Residual Structure. In sketch-based image generation, methods like ControlNet [77] and PIXART-δ [11] copy the half of the base model as their sketch encoder to fully utilize the pretrained text-to-image models. However, applying this to video generation, as done in SparseCtrl [22], is memory inefficient because it requires adding half of the base model's parameters. Unlike the U-Net architecture, the DiT network does not have an explicit encoder and decoder. Therefore, the assumption in PIXART-δ [11] that the first half of blocks serve as the encoder can be improved.

Instead of borrowing local consecutive blocks of the base model as the encoder and predicting their residual features, we recognize that blocks at different depths process distinct feature levels, which should be considered for sketch control. As illustrated in Fig. 2 (a), we propose a novel skip residual structure that reduces the number of blocks while enabling effective sketch control and high-quality generation. Our sketch condition network contains 5 sketch control blocks, uniformly distributed across the pretrained generation network to predict residual features for blocks 0, 6, 12, 18, and 24 of the original video generation model. This structure efficiently integrates sketch control information into multiple feature levels, enhancing the analysis of input conditions and original semantic features.

Sketch Control Block. In the i-th sketch control block, the input consists of hidden video features $h_i^{1:N}$ and sketch features $s_{i-1}^{t_1,t_2}$, and the output is the residual features $\overline{h}_i^{1:N}$ and updated sketch features $s_i^{t_1,t_2}$. As shown in Fig. 2 (b), the sketch feature $s_{i-1}^{t_1,t_2}$ is processed:

$$s_i^{t_1,t_2} = \text{FeedForward}(s_{i-1}^{t_1,t_2}), \tag{2}$$

where the output sketch feature $s_i^{t_1,t_2}$ is used for sketch control propagation and as the input to the next sketch control block.

To propagate the sketch inputs, a direct approach [22] replaces missing sketches with white images and employs a trainable copy of the pretrained DiT block to predict residual features. However, as shown in Fig. 5, this leads to fuzzy details in challenging cases, as the network randomly processes both sketches and white images. This mixed input is far from the pretrained model's task. To address the above issue, we employ a trainable copy of pretrained DiT block to process only the sketch inputs (no white placeholder), aligning with the pretrained weights and reducing learning difficulty. The resulting keyframe sketch features, denoted as $c_i^{t_1,t_2}$, are propagated to all frames through an inter-frame attention approach.

Inter-frame Attention. We utilize the input hidden features of all frames to calculate Q and the hidden features corresponding to the control frames to calculate K. During attention computation, this approach captures the internal

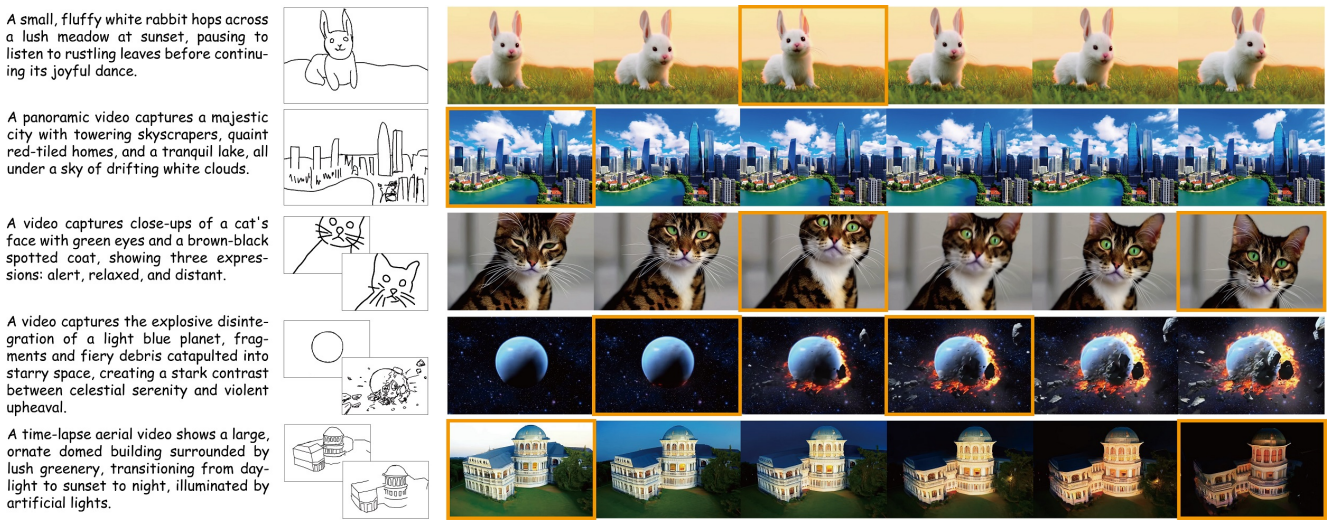


Figure 3. The sketch-based video generation results. Left: The input text prompts and sketches. Right: The video generation results. It can be seen that the generated results show high quality and good faithfulness with the input sketches. Our method can handle one/two keyframe sketch(es) at arbitrary user-specified time points (the frames corresponding to the input time points are highlighted by orange).

relationship between all frames and control keyframes, allowing propagation of V (derived from the keyframe sketch features). Our inter-frame attention is distinct from typical cross-frame attention, which uses both K and V from control conditions; instead, we leverage the inter-frame similarity within the input noisy hidden video features and progressively insert the sketches' spatial and temporal information. The output $\widetilde{c}_i^{1:N}$ is computed as: $\text{Attention}(Q,K,V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$, with

$$\text{Q} = W_q \cdot h_i^{1:N}, K = W_k \cdot h_i^{t_1,t_2}, V = W_v \cdot c_i^{t_1,t_2}, \quad (3)$$

where W_q, W_k, W_v are trainable linear projection weights. The output of the inter-frame attention is fed into a feed-forward layer to generate final residual features $\overline{h}_i^{1:N}$.

Training Strategy. To train the sketch condition network, we employ a hybrid training strategy in two stages. In the first stage, to accelerate convergence and address the issue of limited video data, the network is trained both for image generation at arbitrary time points and video generation with one or two keyframe sketches (randomly selected from corresponding sketch videos). In the second stage, video data alone is used to improve the temporal coherence.

3.3. Sketch-based Video Editing

For editing, given real or synthetic videos, users select one or two keyframes at arbitrary time points and modify the extracted sketches, with additional inputs of text prompts and masks $M^{1:N}$ that label regions to be edited for all the frames. Our method then generates realistic, local editing results. The input video is multiplied by the inverted masks to remove information from the edited regions and is subsequently encoded to generate a masked video latent representation $v_0^{1:N}$.

Video Insertion Module. For sketch-based editing, newly generated contents within the mask regions should be coherent with the original spatial and temporal features in the unedited regions. Thus, we design a video insertion module that analyzes the relationship between the input sketches and the original video. The video insertion module takes $v_{i-1}^{1:N}$ (input video latent or generated by a previous control block) as input and predicts the updated video features $v_i^{1:N}$, similar to the sketch generation process. Since video features are not temporally sparse, we use a trainable copy of CogVideoX-2b's DiT block to directly generate video insertion features $\widetilde{v}_i^{1:N}$. The sketch branch output $\widetilde{c}_i^{1:N}$ and video branch output $\widetilde{v}_i^{1:N}$ are multiplied by their respective masks and concatenated:

$$\text{Concat}(\widetilde{c}_i^{1:N} * M^{1:N}, \widetilde{v}_i^{1:N} * \overline{M}^{1:N}), \quad (4)$$

which serves as inputs to the feed-forward layers, producing final residual features incorporating the original video and sketch control information. This design ensures seamless integration of new contents with the original videos, effectively propagating edits across frames for dynamic motion.

Training Strategy. Directly training the video editing network would lead to low fidelity with the input sketches, possibly because of the challenging interaction between the input sketches and videos. So we finetune it from the pretrained sketch condition network for generation and add the new video insertion module. The pretrained model already has good sketch fidelity and only requires learning video information. The network is trained in a self-supervised inpainting manner, with randomly generated masks to imitate real-world editing.

Inference Latent Fusion. Although the original videos are encoded in the condition network, as shown in Fig. 8,

Figure 4. The sketch-based video editing results. For each example, the text prompts and sketches are shown on the left. On the right, the input real videos are shown at the top, while the edited results with the control keyframe highlighted in orange are shown at the bottom. The editing region masks are manually provided by users, highlighted as orange boxes. Our method generates realistic local editing results.

fine details might be lost during editing. To address this, we propose a latent fusion approach at inference. Specifically, we apply DDIM inversion [58] to generate noisy latent codes of input videos across the inference steps. At Steps 25 and 49 (out of 50 total steps), the latent codes in the unedited regions are replaced with these inversion latent codes, ensuring better preservation of the original video's details and improving the coherence of the edited regions.

4. Evaluation

4.1. Implementation Details

We implement SketchVideo based on CogVideoX-2b [73], trained on a subset of OpenVid [42] and LAION [55] datasets, with paired sketches from [7]. Training uses 8 NVIDIA H800 GPUs with a batch size of 8 and gradient accumulation of 4. For generation, one or two keyframe sketches are randomly sampled from the video, with 10,000 steps for each training stage. For editing, randomly drawn masks are applied and trained for 20,000 steps. For ease of reading, the input text prompts are simplified in figures. Additional implementation details and full input texts are available in the supplementary material.

4.2. Results

Our method generates high-quality videos from one or two keyframe sketches and text prompts. As shown in Fig. 3, our method can accurately control the object shape and scene layout, such as the rabbit's pose (1st row) and the position of lakes and buildings (2nd row). Text-only inputs cannot achieve such detailed geometry control. Our method also achieves interesting dynamic motion interpolation and extrapolation, generating smooth and realistic transitions, as seen in the cat's head shaking (3rd row) and the building's movement (last row). This allows control over both spatial layout and dynamic motion.

Our method also supports sketch-based video editing.

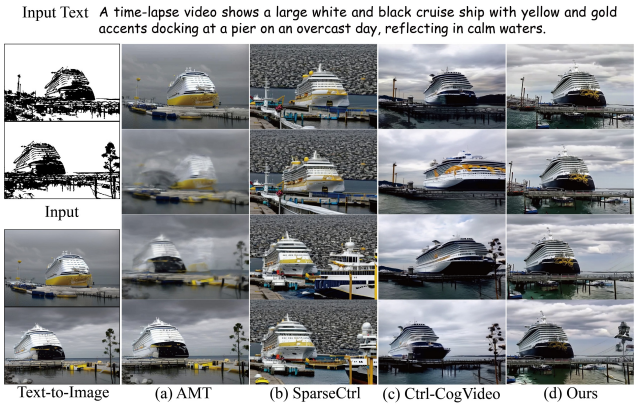


Figure 5. The comparison results of sketch-based video generation. Text prompts are shown on the top. On the left, we show the input sketches and sketch-based image generation results by ControlNet [77]. On the right, we show the results of the compared approaches, including AMT [37], SparseCtrl [22], Ctrl-CogVideo [77], and ours. Our method produces better results, especially for the intermediate frames.

Users specify bounding boxes and draw sketches in those regions, and our method generates photorealistic, seamlessly integrated content. As shown in Fig. 4, the new contents blend well with unedited regions and have dynamic motions like hat/hair rotation with original objects (1st row) or interesting fish swimming (2nd row). Our method can handle diverse editing cases, including object insertion, component replacement, and object removal. The unedited regions are well-preserved thanks to our latent fusion approach.

4.3. Comparison

For sketch-based generation, we compare our methods with three methods given the same keyframe sketches text prompts as inputs. For SparseCtrl [22], we use the official pretrained model and extract sketches by HED [69]. It is trained on videos in the WebVid-10M dataset [1] instead

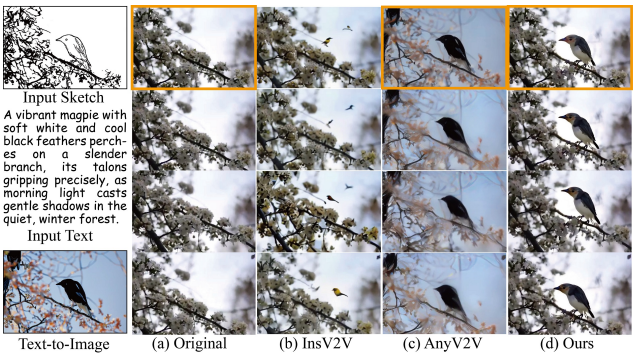


Figure 6. The comparison results of sketch-based video editing. On the left, we show the drawn editing sketches (for the frames highlighted in orange), text prompts, and sketch-based image generation results by ControlNet [77]. On the right, we show the original videos and editing results by the compared methods, including InsV2V [17], AnyV2V [34], and ours. Our method generates the most realistic results and preserves unedited regions well.

of the OpenVid-1M dataset [42] used in our method. We extend SparseCtrl to CogVideoX-2b [73], using PIXART-δ [11] as the sketch condition encoder (with 5 DiT blocks same as our method) and white images to complete the missing condition frames. We also compare with an interpolation baseline, which uses ControlNet to translate two-frame sketches into images and then interpolates them with AMT [37]. For the sketch-base editing task, we compare ours with a text-based video editing method InsV2V [17] and a first-frame editing method AnyV2V [34]. We compare with additional methods [16, 20, 45, 70], as shown in supplemental material.

As shown in Fig. 5, ControlNet [77] translates sketches into realistic images, which, however, lack temporal consistency and vary with shading and content. The interpolation results of AMT [37] exhibit fuzzy details and artifacts. SparseCtrl [22], based on AnimateDiff [23], exhibits temporal flickering, such as the suddenly appearing tree and distorted tower top (see the orange boxes in the 2nd row). Extending SparseCtrl to CogVideoX-2b [73] (Ctrl-CogVideo) still generates fuzzy details in intermediate frames, possibly due to the CogVideoX-2b's pretrained self-attention being designed for dense inputs instead of sparse sketches. Our method generates realistic videos with clear details and good temporal coherence, with even small details like the electric wires in the top-right corner accurately propagated.

For sketch-based editing (Fig. 6), InsV2V [17] generates interesting results with birds but lacks control over shape and geometry through text prompts alone. For image-based video editing method, we utilize ControlNet to edit the first frame and then propagate editing into the video by AnyV2V [34]. However, since the motion is borrowed from the original video, AnyV2V struggles with new content, leading to fuzzy details and distortion in the bird, as well as changes in unedited regions due to ControlNet's inability to retain the original features. In contrast, our method produces more realistic results with faithful representations of the sketch. More generation and editing results are available in the supplemental material and video.

Methods	LPIPS ↓	CLIP ↑	Fidelity	Consistency	Realism
AMT	29.17	96.12	3.13	3.51	3.57
SparseCtrl	44.85	96.48	2.79	2.94	2.83
Ctrl-CogVideo	32.23	98.04	2.86	2.47	2.50
Ours	27.56	98.31	1.21	1.08	1.11

Table 1. The quantitative results of sketch-based video generation comparison. The LPIPS and CLIP numbers are scaled up 100×, with each cell colored to indicate the best.

Methods	LPIPS ↓	CLIP ↑	PSNR↑	Fidelity	Preservation	Realism
InsV2V	13.61	95.39	16.84	2.58	2.26	2.61
AnyV2V	11.92	93.47	13.68	2.35	2.69	2.34
Ours	9.74	98.34	36.48	1.07	1.05	1.04

Table 2. The quantitative results of sketch-based video editing comparison. The LPIPS and CLIP numbers are scaled up 100×, with each cell colored to indicate the best.

We follow SparseCtrl [22] and use the LPIPS [78] metric to measure sketch faithfulness between the input sketches and those extracted from the corresponding frames of the generated videos. We use the CLIP [49] similarity to assess temporal coherence. For sketch-based generation, we test 200 random examples from OpenVid [42], using the first and last frame sketches and the corresponding text prompts in the dataset. As shown in Table 1, our method achieves the lowest LPIPS and highest CLIP scores, indicating its superior performance. For sketch-based editing, we use additional an MSE metric to measure unedited region preservation. We utilize 10 examples with hand-drawn sketches as input. In Table 2, our method outperforms existing methods across all metrics, demonstrating its superiority.

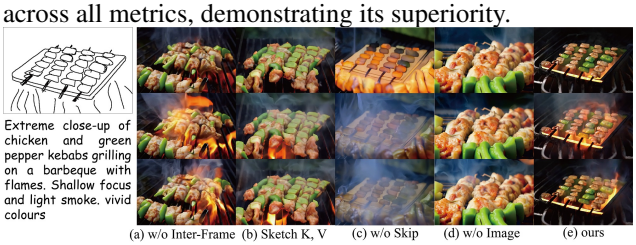
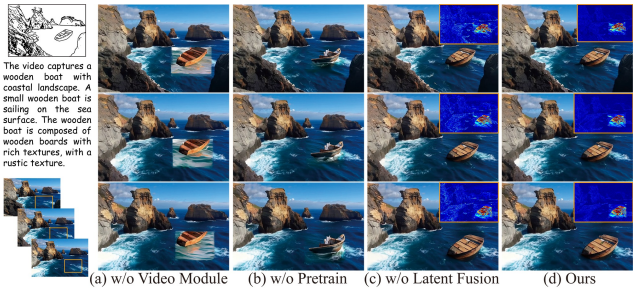


Figure 7. The ablation study results of sketch-based video generation. Our method generates more realistic and sketch-faithful results than the baselines.

4.4. Ablation Study

We conduct an ablation study to evaluate the effectiveness of each key component. Due to the restriction of computing resources, all the models are trained with a batch size of 2. For sketch-based generation, removing inter-frame attention and concatenating sketches with frame features at the temporal dimension results in strange control with low sketch faithfulness (Fig. 7 (a)). Similarly, if we replace the inter-frame attention with a typical cross-attention that uses



(a) w/o Video Module (b) w/o Pretrain (c) w/o Latent Fusion (d) Ours

Figure 8. The ablation study results of sketch-based video editing. Our method generates realistic results faithful to sketches. We utilize heat maps (top right) to visualize the difference between the edited and original frames.

Metric	w/o Inter-Frame Sketch K,V	w/o Skip	w/o Image	Ours
LPIPS ↓	36.33	32.59	31.91	34.58	30.79
CLIP ↑	98.10	98.19	97.60	98.24	98.48

Metric	w/o Video module	w/o Pretrain	w/o Latent Fusion	Ours
LPIPS ↓	9.31	12.60	9.77	9.74
CLIP ↑	97.97	98.12	98.44	98.34
PSNR ↑	33.61	36.05	31.69	36.48

Table 3. The qualitative results of ablation study for sketch-based generation and editing, with each cell colored to indicate the best and second best.

sketch features for both Key and Value (b), the sketch fidelity is also degraded due to the lack of effective control propagation. If we remove the skip structure and predict residual features for the first 5 blocks (c), the realism of the generated results is negatively influenced due to less control ability. Removing image data during training (d) causes significant geometry mismatch with input sketches. Compared to these alternatives, our method produces the most realistic and faithful results. As shown in Table 3, our method has the lowest LPIPS values (supporting the best sketch fidelity) while the CLIP metric has similar results (human eyes cannot recognize temporal coherence difference).

In sketch-based video editing (Fig. 8), removing the video insertion module (a) and solely utilizing the sketch condition encoder results in good sketch fidelity but obvious inconsistency with unedited regions. Training the sketch condition encoder from scratch without generation pretraining (b) reduces the faithfulness to the input sketches in the edited regions, which is also validated by the high LPIPS in Table 3. We use heat maps to reveal differences between edited and original frames. Removing the latent fusion strategy changes unedited region details (c), such as mountains and waves, as shown in the heat maps. The low PSNR value in unedited regions also supports this, as shown in Table 3. Our method maintains strong sketch fidelity, consistency, and unedited region preservation.

4.5. User Study

We conducted a user study to validate that our method outperforms existing approaches. For video generation, we randomly selected 10 examples from the test set in Sec. 4.3. In the questionnaire, participants were shown the results of different methods in random order. 20 participants ranked the methods results by three criteria: Sketch Fidelity, Temporal Consistency, and Video Realism. The rankings provided by the participants were used as scores. As shown in Table 1, our method outperforms existing methods in all criteria, demonstrating its superior performance. For sketch-based editing, we used the same 10 examples from the test set in Sec. 4.3. The same participants ranked the methods based on three criteria: Sketch Fidelity, Unedited Region Preservation, and Video Realism. As shown in Table 2, our method also outperforms existing approaches in editing.

5. Conclusion and Discussion

We presented SketchVideo, a unified method for sketch-based video generation and editing. For generation, we proposed a sketch condition network that predicts residual features for skipped DiT blocks of the base models to save the memory and achieve effective control. An inter-frame attention is further proposed to propagate the keyframe sketches, achieving interesting motion interpolation or extrapolation. We also introduce a hybrid image and video training strategy. For editing, we incorporate a video insertion module to ensure the newly generated content is spatially and temporally coherent with the original video. During inference, a latent fusion approach preserves unedited regions accurately.

Limitation and Future Work. While our method generates high-quality videos, its capabilities are still limited by the pretrained text-to-video model. Enhancing performance with more powerful pretrained models and generating long videos instead of video clips is a potential avenue. Additionally, similar to image generation, our method struggles with too complex scenarios, such as human hands and interaction between multiple objects. Incorporating 3D priors [59, 63, 86] like SMPL-X [47] helps address these issues in human scenarios. Moreover, while our method focuses on geometry control, exploring appearance customization through tools like color strokes presents an intriguing area for future research.

6. Acknowledgments

This work was sponsored by CCF-Kuaishou Large Model Explorer Fund, Beijing Municipal Science and Technology Commission (No. Z231100005923031), Innovation Funding of ICT, CAS (No. E461020) and National Natural Science Foundation of China (No. 62322210). The authors would like to acknowledge the Nanjing Institute of InforSuperBahn OneAiNexus for providing the training and evaluation platform.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, pages 1708–1718, 2021. 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 2, 3

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22563–22575, 2023. 3

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1

[5] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *Int. Conf. Comput. Vis.*, pages 23149–23160, 2023. 3

[6] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Int. Conf. Comput. Vis.*, pages 22983–22993, 2023. 3

[7] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7915–7925, 2022. 6

[8] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *ACM Trans. Graph.*, 43(6):244:1–244:11, 2024. 3

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *CoRR*, abs/2310.19512, 2023. 3

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7310–7320, 2024. 3

[11] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models. *CoRR*, abs/2401.05252, 2024. 2, 3, 4, 7

[12] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Int. Conf. Learn. Represent.*, 2024. 1, 2

[13] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: deep generation of face images from sketches. *ACM Trans. Graph.*, 39(4):72, 2020. 2

[14] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L. Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: deep face generation and editing with disentangled geometry and appearance control. *ACM Trans. Graph.*, 40(4):90:1–90:15, 2021. 2

[15] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023. 3

[16] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *Int. Conf. Learn. Represent.*, 2023. 2, 7

[17] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *Int. Conf. Learn. Represent.*, 2024. 3, 7

[18] Ernie Chu, Shuo-Yen Lin, and Jun-Cheng Chen. Video controlnet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models. *CoRR*, abs/2305.19193, 2023. 3

[19] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. *arXiv preprint arXiv:2405.12211*, 2024. 3

[20] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *Int. Conf. Learn. Represent.*, 2024. 3, 7

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. 2

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *Eur. Conf. Comput. Vis.*, pages 330–348, 2024. 2, 3, 4, 6, 7

[23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *Int. Conf. Learn. Represent.*, 2024. 3, 7

[24] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *CoRR*, abs/2404.02101, 2024. 3

[25] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *CoRR*, abs/2211.13221, 2022. 2, 3

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Adv. Neural Inform. Process. Syst.*, 35:8633–8646, 2022. 3

[27] 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. 1, 2, 3

[28] Zhitong Huang, Mohan Zhang, and Jing Liao. LVCD: reference-based lineart video colorization with diffusion models. *CoRR*, abs/2409.12960, 2024. 3

[29] Ondrej Jamriska, Sárka Sochorová, Ondrej Texler, Michal Lukác, Jakub Fiser, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Trans. Graph.*, 38 (4):107:1–107:11, 2019. 3

[30] Rui Jiang, Guang-Cong Zheng, Teng Li, Tian-Rui Yang, Jing-Dong Wang, and Xi Li. A survey of multimodal controllable diffusion models. *Journal of Computer Science and Technology*, 39(3):509–541, 2024. 2

[31] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6689–6700, 2024. 3

[32] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Trans. Graph.*, 40(6):210:1–210:12, 2021. 3

[33] 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. In *Int. Conf. Comput. Vis.*, pages 15908–15918, 2023. 3

[34] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 2, 3, 7

[35] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 1, 3

[36] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *CoRR*, abs/2406.15339, 2024. 3

[37] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 6, 7

[38] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, pages 5392–5399, 2024. 3

[39] Feng-Lin Liu, Shu-Yu Chen, Yu-Kun Lai, Chunpeng Li, Yue-Ren Jiang, Hongbo Fu, and Lin Gao. Deepfacevideoediting: sketch-based deep editing of face videos. *ACM Trans. Graph.*, 41(4):167:1–167:16, 2022. 3

[40] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *CoRR*, abs/2405.13865, 2024. 2, 3

[41] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 2

[42] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *CoRR*, abs/2407.02371, 2024. 1, 6, 7

[43] OpenAI. Sora overview:https://openai.com/index/sora/, 2024. 1, 3

[44] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Qifeng Chen. Codef: Content deformation fields for temporally consistent video processing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8089–8099, 2024. 3

[45] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *CoRR*, abs/2405.16537, 2024. 2, 3, 7

[46] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *CoRR*, abs/2403.12036, 2024. 2

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 8

[48] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *CoRR*, abs/2408.06070, 2024. 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 7

[50] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2287–2296, 2021. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10674–10685, 2022. 1, 2, 3

[52] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos and spherical images. *Int. J. Comput. Vis.*, 126(11):1199–1219, 2018. 3

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst.*, 2022. 1, 2

[54] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Int. Conf. Learn. Represent.*, 2022. 3

[55] 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. In *Adv. Neural Inform. Process. Syst.*, 2022. 1, 6

[56] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH*, page 111. ACM, 2024. 3

[57] Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. *CoRR*, abs/2403.09334, 2024. 3

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 6

[59] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. *Vis. Intell.*, 2(1), 2024. 8

[60] Vchitect Team and Shanghai Artificial Intelligence Laboratory. Vchitect-2.0: Parallel transformer for scaling up video diffusion models, 2024. 3

[61] Zheng Tianpeng, Chen Yanxiang, Wen Xinzhe, Li Yancheng, and Wang Zhiyuan. Research on diffusion model generated video datasets and detection benchmarks. *Journal of Image and Graphics*, pages 1–13, 2024. 1

[62] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia Conference Papers*, pages 29:1–29:9. ACM, 2022. 3

[63] Guangming Wang, Lei Pan, Songyou Peng, Shaohui Liu, Chenfeng Xu, Yanzi Miao, Wei Zhan, Masayoshi Tomizuka, Marc Pollefeys, and Hesheng Wang. Nerf in robotics: A survey. *CoRR*, abs/2405.01333, 2024. 8

[64] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *CoRR*, abs/2308.06571, 2023. 3

[65] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8798–8807, 2018. 2

[66] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, page 114, 2024. 3

[67] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *CoRR*, abs/2406.17758, 2024. 3

[68] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *CoRR*, abs/2408.13239, 2024. 3

[69] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Int. Conf. Comput. Vis.*, pages 1395–1403, 2015. 6

[70] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *CoRR*, abs/2405.17933, 2024. 2, 3, 7

[71] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender A video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia Conference Papers*, pages 95:1–95:11. ACM, 2023. 2, 3

[72] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH*, page 113, 2024. 3

[73] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 3, 6, 7

[74] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8466–8476, 2024. 3

[75] Chen Yuan, Zhao Yang, Zhang Xiaojuan, and Liu Xiaoping. Sketch colorization with finite color space prior. *Journal of Image and Graphics*, 29(4):978–988, 2024. 2

[76] Liang Yuan, Dingkun Yan, Suguru Saito, and Issei Fujishiro. Diffmat: Latent diffusion models for image-guided material generation. *Visual Informatics*, 8(1):6–14, 2024. 2

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3813–3824, 2023. 2, 3, 4, 6, 7

[78] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 7

[79] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Motioncrafter: One-shot motion customization of diffusion models. *CoRR*, abs/2312.05288, 2023. 3

[80] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *Int. Conf. Learn. Represent.*, 2024. 3

[81] Zhenghao Zhang, Junchao Liao, Menghao Li, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *CoRR*, abs/2407.21705, 2024. 3

[82] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris N. Metaxas, and Licheng Yu. AVID: any-length video inpainting with diffusion model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7162–7172, 2024. 3

[83] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong.

Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[84] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 2, 3

[85] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2242–2251, 2017. 2

[86] Siting Zhu, Guangming Wang, Dezhi Kong, and Hesheng Wang. 3d gaussian splatting in robotics: A survey. *CoRR*, abs/2410.12262, 2024. 8