CCEdit: Creative and Controllable Video Editing via Diffusion Models

Ruoyu Feng[1,2 *], Wenming Weng[1,2], Yanhui Wang[1,2],
Yuhui Yuan[2], Jianmin Bao[2], Chong Luo[2 †], Zhibo Chen[1 †], Baining Guo[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia
https://ruoyufeng.github.io/CCEdit.github.io/

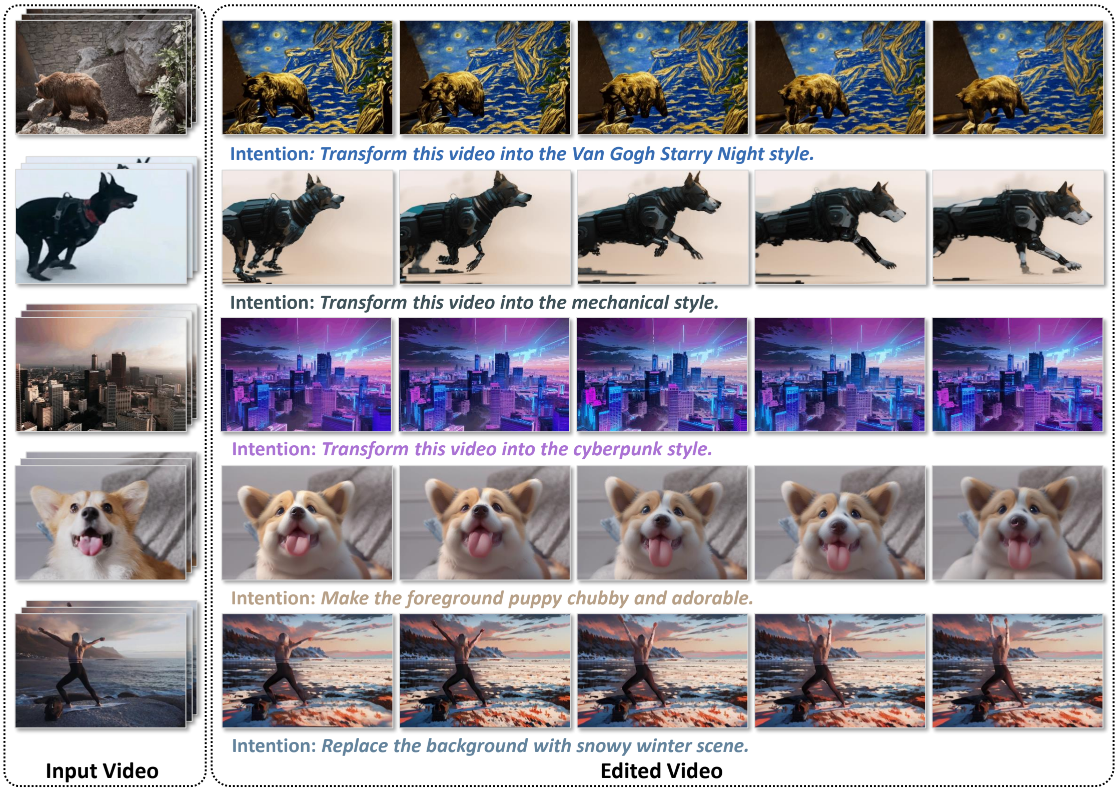
Intention: Transform this video into the Van Gogh Starry Night style.
Intention: Transform this video into the mechanical style.
Intention: Transform this video into the cyberpunk style.
Intention: Make the foreground puppy chubby and adorable.
Intention: Replace the background with snowy winter scene.
Input Video
Edited Video

Figure 1. Built upon diffusion models, CCEdit provides users with a powerful and flexible set of video editing capabilities, including style transfer (row 1-3), foreground modifications (row 4), and background replacement (row 5).

Abstract

In this paper, we present CCEdit, a versatile generative video editing framework based on diffusion models. Our approach employs a novel trident network structure that separates structure and appearance control, ensuring precise and creative editing capabilities. Utilizing the foundational ControlNet architecture, we maintain the structural integrity of the video during editing. The incorporation of an additional appearance branch enables users to exert fine-grained control over the edited key frame. These two side branches seamlessly integrate into the main branch, which is constructed upon existing text-to-image (T2I) generation models, through learnable temporal layers. The versatility of our framework is demonstrated through a diverse range of choices in both structure representations and personalized T2I models, as well as the option to provide the edited key frame. To facilitate comprehensive evaluation, we introduce the BalanceCC benchmark dataset, comprising 100 videos and 4 target prompts for each video. Our extensive user studies compare CCEdit with eight state-of-the-art video editing methods. The outcomes demonstrate CCEdit's substantial superiority over all other methods.

[*] This work is done when Ruoyu Feng is an intern with MSRA.
[†] Corresponding author.

1. Introduction

In recent years, the domain of visual content creation and editing has undergone a profound transformation, driven by the emergence of diffusion-based generative models [11, 20, 46]. A large body of prior research has demonstrated the exceptional capabilities of diffusion models in generating diverse and high-quality images [38, 40, 42] and videos [5, 21, 44], conditioned by text prompts. These advancements have naturally paved the way for innovations in generative video editing [7, 25, 34, 36, 50, 53, 54, 58].

Generative video editing, despite its rapid advancement, continues to face a series of significant challenges. These challenges include accommodating diverse editing requests, achieving fine-grained control over the editing process, and harnessing the creative potential of generative models. Diverse editing requirements include tasks such as stylistic alterations, foreground replacements, and background modifications. Generative models, while powerful and creative, may not always align perfectly with the editor's intentions or artistic vision, resulting in a lack of precise control. In response to these challenges, this paper introduces CCEdit, a versatile generative video editing framework meticulously designed to strike a harmonious balance between controllability and creativity while accommodating a wide range of editing requirements.

CCEdit achieves its goal by effectively decoupling structure and appearance control in a unified *trident network*. This network comprises three essential components: the main text-to-video generation branch and two accompanying side branches dedicated to structure and appearance manipulation. The *main branch* leverages a pre-trained text-to-image (T2I) diffusion model [40], which is transformed into a text-to-video (T2V) model through the insertion of temporal modules. The *structure branch*, implemented as ControlNet [55], is responsible for digesting the structural information extracted from each frame of the input video and seamlessly infusing it into the main branch. Simultaneously, the *appearance branch* introduces an innovative mechanism for precise appearance control, when an edited reference frame is available. The structure and appearance branches are effectively integrated into the central branch through learnable temporal layers. These layers serve not only as a cohesive link, aggregating information from side branches, but also as a crucial element ensuring temporal consistency across the generated video frames.

In highlighting the versatility of our framework, we provide a wide range of control choices for both structure and appearance manipulation. For structure control, users can choose from various types of structural information, including line drawings [8], PiDi boundaries [47], and depth maps [39], all of which can serve as input to the structure branch. On the appearance control front, the main branch already provides an inherent mechanism, allowing control through text prompts. Additionally, personalized T2I models from the Stable Diffusion community, such as DreamBooth and LoRA [22, 41], can be integrated as plugins into CCEdit, offering greater flexibility and creativity. More importantly, the appearance branch can accommodate the referenced key frame, facilitating fine-grained appearance control. Notably, all these control options are seamlessly integrated within the same framework, yielding editing outcomes that demonstrate both temporal coherence and precision. This not only underscores the versatility of our solution but also ensures ease of adoption, making it a compelling choice for AI-assisted video editing.

To address the challenges inherent in evaluating generative video editing methods, we introduce the *BalanceCC benchmark* dataset. Comprising 100 diverse videos and 4 target prompts for each video, this dataset includes detailed scene descriptions and attributes related to video category, scene complexity, motion, among others. These descriptions are generated with the assistance of the cutting-edge GPT-4V(ision) model [1, 31–33] and then refined by human annotators. Through extensive experimental evaluations on this dataset, we not only confirm the outstanding functionality and editing capabilities of CCEdit, but also underscore the comprehensiveness of the benchmark dataset. We firmly believe that BalanceCC stands as a robust and all-encompassing evaluation platform for the dynamic field of generative video editing.

2. Related Work

2.1. Diffusion-based Image and Video Generation

Diffusion models (DM) [11, 16, 20, 29, 30, 46] have demonstrated exceptional capabilities in the field of image synthesis. These models indeed help by learning to approximate a data distribution through the iterative denoising of a diffused input. What makes DMs truly practical is the incorporation of text prompt as condition to control the output image during the generative process [30, 37, 40, 42]. Apart from the proliferation of advanced techniques in the field of image synthesis, DMs have also excelled in video generation [5, 21, 30, 44]. This is achieved by integrating modulated spatial-temporal modules, enabling the synthesis of high-quality videos while maintaining temporal consistency.

2.2. Video Editing with Diffusion Models

Recent studies leverage the inherent generative priors of DMs for image editing [3, 10, 17, 27, 35, 48]. The same idea is also applied in the field of video editing. Unlike image editing, video editing involves not only the manipulation of appearance-based attributes but also requires the meticulous preservation of temporal coherence throughout frames. A lapse in maintaining this temporal coherence can

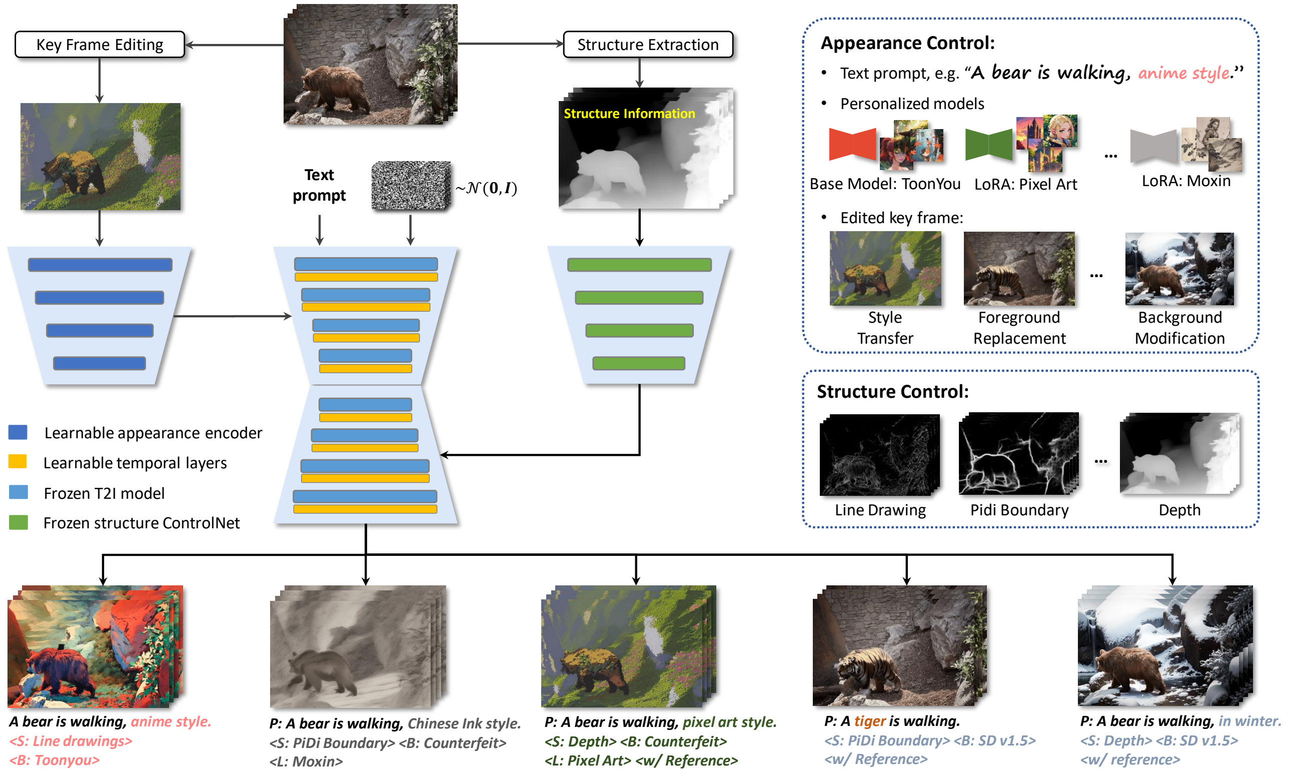
Key Frame Editing
Structure Extraction
Structure Information
Text prompt
$\sim\mathcal{N}(\mathbf{0}, \boldsymbol{I})$
Appearance Control:
Text prompt, e.g. "A bear is walking, anime style."
Personalized models
Base Model: ToonYou
LoRA: Pixel Art
...
LoRA: Moxin
Edited key frame:
Style Transfer
Foreground Replacement
...
Background Modification
Structure Control:
Line Drawing
Pidi Boundary
...
Depth
Learnable appearance encoder
Learnable temporal layers
Frozen T2I model
Frozen structure ControlNet
A bear is walking, anime style.
<S: Line drawings>
<B: Toonyou>
P: A bear is walking, Chinese Ink style.
<S: PiDi Boundary> <B: Counterfeit>
<L: Moxin>
P: A bear is walking, pixel art style.
<S: Depth> <B: Counterfeit>
<L: Pixel Art> <w/ Reference>
P: A tiger is walking.
<S: PiDi Boundary> <B: SD v1.5>
<w/ Reference>
P: A bear is walking, in winter.
<S: Depth> <B: SD v1.5>
<w/ reference>

Figure 2. **Illustration of our overall framework.** Structure and appearance information in the target video are modulated independently, and seamlessly integrated into the main branch. Structure control is conducted via the pre-trained ControlNet [55]. Appearance control is achieved precisely by the edited key frame. Details regarding the autoencoder and iterative denoising process are omitted for simplicity. "**P**", "**S**", "**B**", "**L**" indicate prompt, structure, base model, and LoRA, respectively.

result in visual artifacts, such as flickering and degradation.

Some generative video editing methods [6, 14, 23, 36, 49, 54, 56] strive to achieve training-free temporal consistency. They accomplish this by transitioning from spatial self-attention mechanisms within T2I diffusion models to temporal-aware cross-frame attention techniques. Some other methods [26, 43, 51, 58] perform per-video fine-tuning. They focus on optimizing the parameters of pre-trained T2I models according to the input video, aiming to achieve temporal coherence within the target video. However, this optimization for each input video can be time-consuming and inadequate tuning of the temporal modules might lead to suboptimal temporal coherence. Recent studies [15, 24, 53] have introduced trainable temporal layers to construct T2V generative models. These models are trained on extensive text-video paired datasets, and they are used in both video generation and editing tasks [12, 28].

Unlike previous work, this study does not seek a simple fix to existing T2I models for video editing, nor does it attempt to train a full-fledged T2V model. Instead, we introduce a unique network architecture tailored for video editing. Our approach involves dataset-level fine-tuning, circumvents the expenses associated with per-video tuning during inference time, and prioritizing the effective training of temporal layers to achieve robust model performance.

3. Approach

3.1. Preliminary

Diffusion models [20] are probabilistic generative models that approximate a data distribution $p(\mathbf{x})$ by gradually denoising a normally distributed variable. Specifically, DMs aim to learn the reverse dynamics of a predetermined Markov chain with a fixed length of T. The forward Markov chain can be conceptualized as a procedure of injecting noise into a pristine image. Empirically, DMs can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(\mathbf{x}_t, t)$ where $t = 1, ..., T$. These autoencoders are trained to predict a denoised variant of the noisy input $\mathbf{x}_t$. The corresponding objective can be simplified to

$$\mathbb{E}_{\mathbf{x}_0, t, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2]. \quad (1)$$

Latent diffusion models (LDMs) are trained in the learned latent representation space. The bridge between this latent space and the original pixel-level domain is established via a perceptual compression model. The perceptual compression model is composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, where $\mathbf{z} = \mathcal{E}(\mathbf{x})$ and $\mathbf{x} \approx \mathcal{D}(\mathcal{E}(\mathbf{x}))$. Then the optimization objective in Eq. (1) is modified as

$$\mathbb{E}_{\mathbf{z}_0,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\epsilon-\epsilon_\theta(\mathbf{z}_t,t)\|_2^2]. \quad (2)$$

3.2. The CCEdit Framework

The primary objective of our work is to empower creative control in video editing. Although creativity naturally emerges in generative models, achieving controllability is a more complex endeavor. To address this challenge, CCEdit strategically decouples the management of structure and appearance within a unified trident network. In Fig. 2, we provide an illustrative overview of the framework's architecture, which comprises three vital components.

The main branch. The main branch of our model fundamentally operates as a text-to-video generation network. It is built upon the well-established text-to-image model, Stable Diffusion [40]. We transform this model into a text-to-video variant by incorporating temporal layers into spatial layers of both the encoder and decoder. This entails the addition of a one-dimensional *temporal layer* with the same type as its previous *spatial layer*, *i.e.*, convolution blocks and attention blocks. Besides, we also use the skip connection and zero-initialized *projection out layer* of each newly added temporal layer for stable and progressive updating, which has been proven to be effective [15, 44, 55]. The zero-initialized projection out layer is instantiated as a linear layer. Formally, let $\mathcal{F}(\cdot;\Theta_s)$ be the 2D spatial block, $\mathcal{F}(\cdot;\Theta_t)$ be the 1D temporal block, and $\mathcal{Z}(\cdot;\Theta_z)$ be the zero-initialized projection out layer, where Θ_s, Θ_t, and Θ_z represent corresponding network parameters. The complete process of one pseudo-3D block that maps the input feature $\mathbf{u}$ to the output feature $\mathbf{v}$ is written as

$$\mathbf{v}=\mathcal{F}(\mathbf{u};\Theta_s)+\mathcal{Z}(\mathcal{F}(\mathcal{F}(\mathbf{u};\Theta_s);\Theta_t);\Theta_z), \quad (3)$$

where $\mathbf{u}$ and $\mathbf{v}$ are both 3D feature maps, *i.e.*, $\mathbf{u}\in\mathbb{R}^{l\times h\times w\times c}$ with $\{l,h,w,c\}$ as the number of frames, height, width, and the number of channels, respectively.

Moreover, we draw inspiration from AnimateDiff [15] and VideoLDM [5], which advocates the shared utilization of temporal layers among personalized T2I models such as DreamBooth [41] and LoRA [22]. The key aspect of it is training the temporal layers while keeping the spatial weights frozen. We follow this schedule to inherit the T2I model's compatibility and visual generation capability.

The structure branch. The introduction of the structure branch is motivated by the common need in video editing tasks to preserve frame structure for non-edited or style-transferred segments. Striking a delicate balance between maintaining faithful frame structure and allowing the generative model ample creative freedom poses a significant challenge. The structure branch is implemented with the pre-trained ControlNet [55]. To accommodate varying levels of structure control, we use various types of structure representation, including line drawings [8], PiDi boundaries [47], and depth maps [39], ensuring adaptability to control structure at different degrees.

Specifically, the structure representation from all frames is extracted individually and injected into the main branch. Each frame undergoes preprocessing to derive a structure representation, and the weights of the ControlNet are held in a frozen state during training, emphasizing the preservation of learned structural features. Formally, let $\mathcal{F}(\cdot;\Phi_c)$ denote the ControlNet that maps structure information into features, and $\mathcal{Z}(\cdot;\Phi_{z1})$ and $\mathcal{Z}(\cdot;\Phi_{z2})$ denote the two instances of zero convolutions in [55]. Then the process of adding structure control to the 3D-aware feature $\mathbf{v}$ is

$$\mathbf{v}_s=\mathbf{v}+\mathcal{Z}(\mathcal{F}(\mathbf{z}_t+\mathcal{Z}(\mathbf{c}_s;\Phi_{z1});\Phi_c);\Phi_{z2}), \quad (4)$$

where $\mathbf{z}_t$ denotes the noisy input in latent space, $\mathbf{c}_s$ denotes the structure condition of the video sequence, and $\mathbf{v}_s$ denotes the feature aware of structure information.

The appearance branch. In addition to using text prompts and incorporating personalized models for appearance control, we introduce a novel design—the appearance branch. This architectural innovation introduces a pioneering approach for fine-grained appearance control, allowing for the integration of an edited frame as a detailed reference in the context of video editing. Since the editing of key frame can be accomplished through precise user edits or by using advanced off-the-shelf image editing algorithms, the introduction of appearance branch provides our framework with greater creativity and controllability. Specifically, a key frame is initially assigned to the latent variable by the encoder $\mathcal{E}$. Subsequently, a neural network with similar architecture to the main branch's encoder extracts multi-scale features. The extracted features are incorporated into the main branch. Through this design, the appearance information from the edited key frame propagates to all frames via the temporal modules, effectively achieving the desired creative control in the output video. Formally, suppose $\mathcal{F}(\cdot;\Psi)$ is the encoder that maps the pixel-wise appearance of the key frame into features, $\mathcal{Z}(\cdot;\Psi_z)$ denotes the zero convolution projection out layer, $\mathbf{v}^j$ indicates the feature of the j-*th* frame, and $\mathbf{c}_a^j$ is the key frame. Then the process of adding appearance control to the features is as follows

$$\mathbf{v}_a^j=\mathbf{v}^j+\mathcal{Z}(\mathcal{F}(\mathcal{E}(\mathbf{c}_a^j);\Psi);\Psi_z), \quad (5)$$

where $\mathbf{v}_a^j$ is the j-*th* feature, aware of the edited appearance.

Training. Before training, we initialize the spatial weights of the main branch with pre-trained T2I models. Temporal weights are randomly initialized while the projection out layers are zero-initialized. We instantiate the model in the structure branch by pre-trained ControlNets [55]. As for the appearance branch, we copy the encoder of pre-trained T2I model and remove text cross-attention layers. During training, given the latent variables $\mathbf{z}_0=\mathcal{E}(\mathbf{x}_0)$ of an input video clip $\mathbf{x}_0$, diffusion algorithms progressively add noise to it and produce the noisy input $\mathbf{z}_t$. Given conditions of time step t, text prompt $\mathbf{c}_t$, structure information $\mathbf{c}_s$, and appearance information $\mathbf{c}_a^j$ of the key frame, the overall optimization objective is

$$\mathbb{E}_{\mathbf{z}_0,t,\mathbf{c}_t,\mathbf{c}_s,\mathbf{c}_a^j,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\epsilon-\epsilon_\theta(\mathbf{z}_t,t,\mathbf{c}_t,\mathbf{c}_s,\mathbf{c}_a^j)\|_2^2], \quad (6)$$

where ϵ_θ indicates the whole network to predict the noise added to the noisy input $\mathbf{z}_t$. We freeze the spatial weights in the main branch and the weights in the structure branch. Concurrently, we update the parameters of the newly incorporated temporal layers in the main branch, as well as the weights in the appearance branch. By default, the appearance branch takes the center frame of the video clip as input.

Inference with anchor prior. We find that, in some challenging cases, the edited video may exhibit large areas of flickering. This is often caused by inconsistent structural representations extracted by image-level pre-processing modules. Therefore, we propose a simple yet efficient strategy to improve the stability and quality of the result by modifying the start noise. Specifically, consider the individual noise sequence $[\epsilon_{\text{ind}}^1, ..., \epsilon_{\text{ind}}^l]$ and the edited center frame $\mathbf{c}_a^j$, where l and j indicate the frame numbers and the index of the edited key frame, respectively. The start noise ϵ^i for each frame is modified as

$$\epsilon^i = \epsilon_{\text{ind}}^i + \alpha\mathcal{E}(\mathbf{c}_a^j), \quad (7)$$

where α is the hyperparameter that controls the strength of prior, and $\mathcal{E}(\mathbf{c}_a^j)$ is the latent of the edited key frame. We call this strategy *anchor prior*, which is tailored for our pipeline of editing videos with an reference key frame. We empirically found that $\alpha = 0.03$ works well in most cases. The intuition behind it lies in that the video frames are usually similar to each other. The operation of adding noise to diffusion models tends to rapidly destroy high-frequency information while slowly degrading low-frequency information. Therefore, the anchor prior can be seen as providing a bit of low-frequency information to all frames while ensuring that the distribution remains almost unchanged (achieved by small α), thus becoming better starting points.

3.3. Editing for Long Videos

Video editing tools face a challenge in maintaining a consistent look and feel across clips that span tens of seconds, equivalent to hundreds of frames. The inherent limitation of generative models, processing only a dozen frames per inference due to memory constraints, introduces variability in results, even with a fixed random seed. CCEdit addresses this challenge with its fine-grained appearance control, enabling the editing of long videos into a cohesive look and feel through extension and interpolation modes.

In essence, let $L + 1$ represent the frames CCEdit processes in one run. For videos exceeding $L + 1$ frames, we select one key frame for every L frames. In the initial run, the first $L+1$ key frames undergo editing. Subsequent runs, in extension mode, treat the last edited frame from the previous run as the first frame. The edited result serves as a reference for the appearance branch. This process iterates until all key frames are processed. Transitioning to the interpolation mode, two adjacent frames become the first and last frames of an inference run to edit the $L - 1$ intermediate frames, and both edited frames serve as references for the appearance branch. This continues until all frames are edited. This meticulous process ensures consistent editing results throughout the entire video.

4. BalanceCC Benchmark

4.1. Overview

While generative video editing has gained considerable attention as a growing research field, the absence of a standardized benchmark for assessing the efficacy of different approaches poses a potential hindrance to the field's technical progression. Despite the recent introduction of TGVE 2023 [52] as an evaluation benchmark, it is crucial to note that the videos within this benchmark present challenges such as severe camera shake, complex scenes, blur, and low frame rates. In response, we introduce *BalanceCC*, a benchmark containing 100 videos with varied attributes, designed to offer a comprehensive platform for evaluating video editing, focusing on both controllability and creativity.

4.2. Benchmark Establishment

We curated a collection of 100 open-license videos suitable for legal, non-stigmatizing modifications. These videos range from 2 to 20 seconds in duration, each with a frame rate of about 30 fps. Besides, we utilize GPT-4V(ision) [1, 31–33] as an assistant to establish this benchmark. For each video, GPT-4V(ision) provides a description and assigns a complexity score to the scene using the center frame as a reference, with ratings from 1 (Simple) to 3 (Complex). Additionally, we manually annotate each video for camera movement, object movement, and categorical content, with motion rated on a scale from 1 (Stationary) to 3 (Quick), and categories that include humans, animals, objects, and landscapes. Following this, GPT-4V(ision) is tasked to craft target prompts for video editing, encompassing style, object, and background alterations, along with compound changes. This process, while akin to TGVE 2023 [52], we additionally introduce a "Fantasy Level" to indicate the imaginative and creative degree of the target prompt. These measures are intended to assist researchers in appraising the applicability of various methods to source videos and in gauging their potential. See supplementary for details on the prompting pipeline, specific instructions, principles of labeling, and illustrative examples.

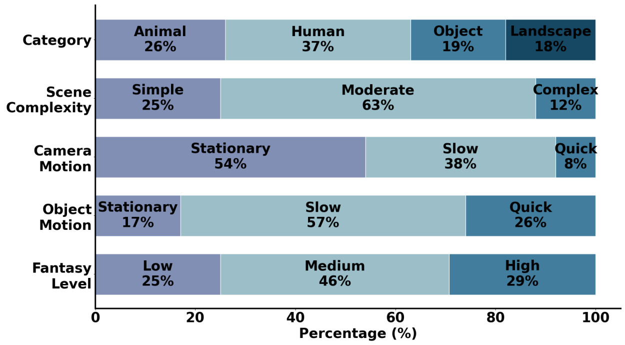

Figure 3. Illustration of the statistics on BalanceCC.

4.3. Statistics

The overall distribution of BalanceCC is illustrated in Fig. 3. For the data of original videos, the distribution across categories tends towards uniformity, yet the "Human" category is slightly more prevalent than others. This was a deliberate choice, as editing human subjects is more practically significant and, due to the complexity of human and facial structures, editing in the "Human" category presents more challenges. Regarding "Scene Complexity" and "Object Motion", videos with moderate and slow levels are slightly more common. In terms of "Camera Motion", videos of lower levels predominate (Stationary: 54%, Slow: 38%). Finally, regarding the "Fantasy Level" distribution in target prompts, there is a relatively balanced allocation, with a marginal inclination towards videos categorized at a moderate level.

We hope that the aforementioned categorization of the benchmark will better assist researchers and users in understanding the strengths and weaknesses of a method, thus enabling targeted improvements and fostering rapid development in the field.

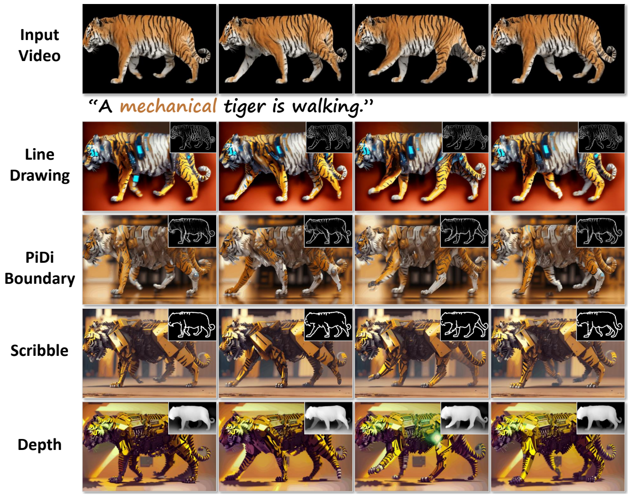

Figure 4. **Results under different structural guidance.**

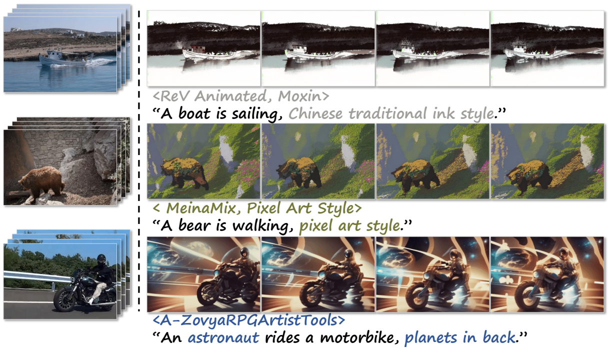
<ReV Animated, Moxin>
"A boat is sailing, Chinese traditional ink style."
< MeinaMix, Pixel Art Style>
"A bear is walking, pixel art style."
<A-ZovyaRPGArtistTools>
"An astronaut rides a motorbike, planets in back."

Figure 5. **Results of video style translation.** $\langle\cdot\rangle$ indicate the personalized T2I model we used.

5. Experiments

5.1. Implementation Details

Stable Diffusion-v1.5 is used as the base T2I model in the main branch. We use the pre-trained ControlNet [55] for the structure information guidance. The training dataset combines WebVid-10M [4] and a self-collected private dataset. We trained the temporal consistency modules and appearance ControlNet towards various types of structural information, including line drawings [8], PiDi boundaries [47], depth maps detected by Midas [39], and human scribbles. Depth maps are used by default. The control scales are set as 1. For the temporal interpolation model, we train it exclusively on depth maps, employing a smaller control scale of 0.5. This approach is adopted because its requirement for structural information is comparatively less than that of other models. During the training process, we first resize the shorter side to 384 pixels, followed by a random crop to obtain video clips with a size of 384×576. 17 frames at 4 fps are sampled from each video. The batch size is 32 and the learning rate is $3e-5$. We train each model for 100K iterations. During inference, we employ the DDIM [45] sampler with 30 steps, classifier-free guidance [19] of magnitude 9.

5.2. Applications

Controllable and creative style transfer. In CCEdit, the controllability and creativity of video style transfer are manifested in various dimensions. Two basic aspects include the diversity of structural information and the availability of off-the-shelf personalized models [9, 13]. The former enables users to customize the granularity and type of structural information retained from the original video, as depicted in Fig. 4. The latter allows users to edit the video into their desired domain, as shown in Fig. 5.

Video editing with precise appearance control. Sometimes, users require stronger control over the content they want to generate. For example, they may want to change

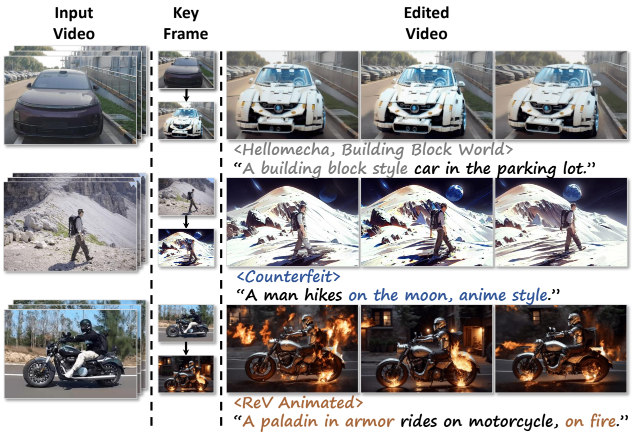
Input Video
Key Frame
Edited Video
<Hellomecha, Building Block World>
"A building block style car in the parking lot."
<Counterfeit>
"A man hikes on the moon, anime style."
<ReV Animated>
"A paladin in armor rides on motorcycle, on fire."

Figure 6. **Video editing results with customized center frame as reference.** The first row corresponds to customizing foreground, the second row corresponds to customizing background, and the third row is taking given reference image to affect the entire picture. $\langle\cdot\rangle$ indicate the personalized T2I model we used.

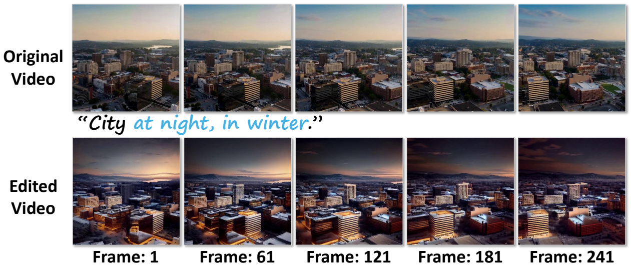

Figure 7. **Illustration of long video editing.** CCEdit achieves good consistency across over 240 frames. Zoom in for best view.

only the foreground, alter just the background, or edit the texture content of a video in a specific way. Therefore, CCEdit focuses more on precise appearance control by initially modifying the key frame with image editing techniques and then using it as a reference for the entire video. As depicted in Fig. 6, we first edit the center frames of the videos by Stable Diffusion Web UI [2], followed by utilizing these edited center frames as guides for the video editing process. Thanks to end-to-end network training, our method coherently propagates edits from the key frame throughout the entire video.

Long video editing. A seamless and visually appealing video typically necessitates a higher frame count and increased frame rate, elements that have been inadequately addressed by many contemporary video editing methodologies. CCEdit effectively resolves this through its hierarchical design for key frames editing, combined with iterative extension and a tailored temporal interpolation mechanism. This approach enables the editing of videos comprising up to hundreds of frames with 24 fps (frames per second). An example is shown in Fig. 7.

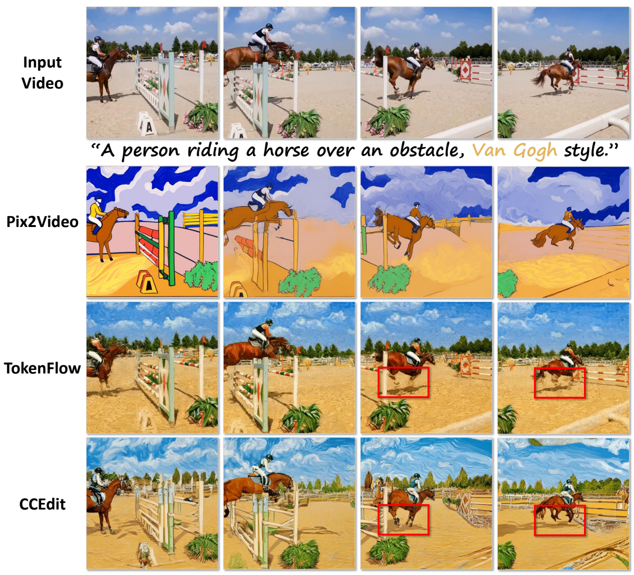

Figure 8. **Qualitative comparison results.** Red boxes reveals TokenFlow's inadequate local detail preservation, in contrast to our method's detailed, coherent output. Zoom in for best view.

5.3. State-of-the-Art Comparisons

Datasets. We employ a smaller segment of our proposed benchmark, designated as *mini-BalanceCC*. This subset encompasses 50 videos, each randomly selected from the original BalanceCC dataset, ensuring a representative distribution similar to that of the original collection.

Compared methods. To conduct an exhaustive comparison, we have selected eight representative video editing methodologies: Tune-A-Video [51], vid2vid-zero [49], Text2Video-zero [23], FateZero [36], Pix2Video [6], ControlVideo [56], Rerender A Video [54], and TokenFlow [14]. Method details are omitted for brevity, and can be found in supplementary. Regarding our approach, we employ depth maps as structure control. For the appearance control, we adopt the off-the-shelf method of PnP-Diffusion [48] with the same hyper-parameters to automatically edit the center frame of each video clip. To ensure fairness in comparison, Stable Diffusion-v1.5 is used as the base model for all methods.

Evaluation metrics. In our preliminary study, we observed that automatic metrics, such as CLIP-Score [18] to assess text alignment and frame consistency, do not fully align with human preferences [28, 52, 57]. We focused on collecting human preferences for a comprehensive user study, comparing our method against recent state-of-the-art techniques based on mean opinion score (MOS) and direct comparisons. We gathered 1,119 scoring results from 33 volunteers, each reflecting all indicators for an edited video. For automatic metric results, refer to the supplementary.

Method	Edit	Aes.	Tem.	Ove.	Win	Tie	Lose
Tune-A-Video [51]	3.24	3.01	2.72	2.77	16.4	6.9	76.7
vid2vid-zero [49]	3.00	2.38	2.11	2.35	10.6	4.6	84.8
Text2Video-Zero [23]	2.07	1.43	1.41	1.48	16.5	1.3	86.2
FateZero [36]	2.47	3.16	3.30	2.79	16.6	3.6	79.8
Pix2Video [6]	3.68	2.97	2.80	2.97	29.9	5.2	64.9
ControlVideo [56]	3.01	2.71	2.60	2.66	13.8	5.6	80.6
Rerender A Video [54]	2.40	2.69	2.82	2.50	11.1	0.0	88.9
TokenFlow [14]	3.78	3.61	**3.79**	3.58	32.4	14.7	52.9
CCEdit (Ours)	**4.06**	**4.00**	3.74	**3.87**	-	-	-

Table 1. **Left: Mean opinion scores (MOS) over different aspects of the generated video,** including editing accuracy (Edit), aesthetics (Aes.), temporal consistency (Tem.), and overall impression (Ove.). Scores range from 1 to 5. **Right: Win, Tie, and Lose percentage in side-by-side comparisons with CCEdit.**

Results. As illustrated in Tab. 1, CCEdit excels in both editing accuracy and aesthetic quality, and is just slightly inferior to TokenFlow in temporal smoothness. For overall impression, our approach achieved a MOS of 3.87 on a scale from 1 to 5. Among the eight reference methods, TokenFlow performed closest to ours, with an overall MOS of 3.58. The remaining seven methods scored between 1.5 to 3.0 on the MOS scale. As for direct comparisons, our method outperforms all eight reference schemes significantly. While TokenFlow remains the closest competitor, our CCEdit prevails in 52.9% of test cases against it, trails in 32.4%, and ties in 14.7% of cases.

Furthermore, Fig. 8 presents the qualitative results of the top three finalists (CCEdit, TokenFlow [14], and Pix2Video [6]). It shows that Pix2Video struggles to keep temporal coherence, while TokenFlow demonstrates noticeable blurring. In contrast, our method can accurately achieve the editing objective while maintaining the temporal coherence as well as the structure of the input video. Please see supplementary for more qualitative results.

5.4. Ablation Study

Appearance control. Fig. 9 illustrates the importance of taking the edited key frame as a reference in certain scenarios. Initially, translating video scenes into "cyberpunk" style (1st row) solely through prompt adjustments appears challenging, as this word is unfamiliar to the pre-trained T2I model weights and the temporal consistency modules. Providing a customized center frame allows the network to smoothly extend its appearance to adjacent frames, creating a cohesive video. Besides, we replicated the user study pipeline from Sec. 5.3 to evaluate the effectiveness of appearance control. The model without appearance control received a mean opinion score (MOS) of 2.88, significantly lower than the 3.87 scored by the process of editing one key frame first and then propagating to surrounding frames.

Anchor prior. Fig. 10 demonstrates the ablation study for our anchor prior. It reveals that the absence of the anchor prior may lead to regional flickering in the video sequence, while its presence effectively mitigates this issue.

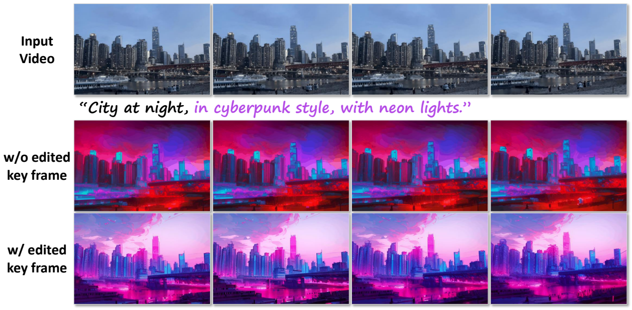

Figure 9. **Ablation study on appearance control.** In some challenging cases, appearance control is crucial to achieving the expected results.

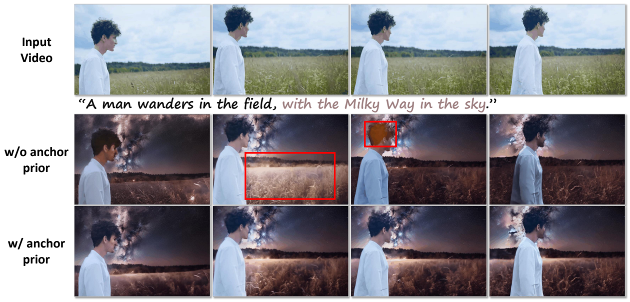
Input Video
"A man wanders in the field, with the Milky Way in the sky."
w/o anchor prior
w/ anchor prior

Figure 10. **Ablation study on anchor prior.** Our proposed anchor prior helps a lot in stabilizing the appearance across frames. The red boxes demonstrate the localized flickering in the frames.

6. Limitation and Future Works

In our approach, structural control is exerted by explicitly extracting the structural representation from the source video and sustaining it via the structure branch. However, it may encounter challenges when tasked with substantial structural alterations-exemplified by the conversion of a "cute rabbit" into a "majestic tiger." Addressing these complexities will be a primary objective of our future work.

7. Conclusion

This paper presents an innovative trident network architecture specifically designed for generative video editing. This unified framework enables precise and controllable video editing while broadening creative possibilities. To address the challenges in evaluating generative video editing approaches, we introduce the meticulously curated BalanceCC benchmark dataset. Our aim is to pave the way for researchers in the generative video editing domain and equip practitioners with indispensable tools for their creative workflows.

References

[1] Chatgpt can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, 2023. 2, 5

[2] AUTOMATIC1111. Stable Diffusion Web UI, 2022. 7

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 6

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 4

[6] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 3, 7, 8

[7] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[8] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022. 2, 4, 6

[9] Civitai. Civitai. https://civitai.com/, 2022. 6

[10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[13] Hugging Face. Hugging face. https://huggingface.co/, 2022. 6

[14] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3, 7, 8

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3, 4

[16] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451, 2023. 2

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 7

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[22] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2, 4

[23] 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. 3, 7, 8

[24] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

[25] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. *arXiv preprint arXiv:2310.10624*, 2023. 2

[26] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[28] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3, 7

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[30] Alexander Quinn 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. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[31] OpenAI. Gpt-4v(ision) system card. 2023. 2, 5

[32] OpenAI. Gpt-4v(ision) technical work and authors. https://cdn.openai.com/contributions/gpt-4v.pdf, 2023.

[33] OpenAI. Gpt-4 technical report, 2023. 2, 5

[34] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 2

[35] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[36] 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. 2, 3, 7, 8

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[39] René 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*, 44(3):1623–1637, 2020. 2, 4, 6

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[41] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 4

[42] 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. 2

[43] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 3

[44] 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. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 4

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 6

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 2

[47] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 2, 4, 6

[48] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 7

[49] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3, 7, 8

[50] Yuanzhi Wang, Yong Li, Xin Liu, Anbo Dai, Antoni Chan, and Zhen Cui. Edit temporal-consistent videos with image diffusion model. *arXiv preprint arXiv:2308.09091*, 2023. 2

[51] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3, 7, 8

[52] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 5, 7

[53] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 2, 3

[54] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2, 3, 7, 8

[55] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 4, 6

[56] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3, 7, 8

[57] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *arXiv preprint arXiv:2305.17431*, 2023. 7

[58] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 2, 3