CCEdit: Creative and Controllable Video Editing via Diffusion Models (Supplementary Material)

A. Details of the Trident Network

The detailed architecture of our proposed *trident network* is illustrated in Fig. 1. Specifically, in the appearance branch, the edited key frame $\mathbf{c}_a^j$ is encoded by the VAE encoder $\mathcal{E}$. Then it's fed into the encoder of *appearance branch*. Subsequently, the features extracted from each layer are fed into zero convolutions and the output are added to the corresponding features in the encoder side of *main branch*. On the right side, *i.e.*, the *structure branch*, structure information $\mathbf{c}_s$ of original video clip is encoded by the zero convolution and fed into structure branch encoder. Similar to the appearance branch, features extracted are fed into zero convolutions. Differently, the output are added to the corresponding features in the decoder side of main branch. The structure branch is instantiated by ControlNet [29]. Note that in the original paper of ControlNet, it consists a tiny network to encode the pixel-wise structure representation. Here we omit it for simplicity. Ultimately, the appearance information within the key frame is propagated to all frames through the temporal modules and the inherited structure information will ensure the structural fidelity, achieving the stable and controllable editing.

It is important to highlight that, we don't use a train-from-scratch tiny encoder to encode the condition as ControlNet [29] does in the appearance branch. Instead, we use the VAE encoder $\mathcal{E}$ to map the pixel-wise appearance into latent variable, which is in the same representation space as latent variable $\mathbf{z}_0$. The intuition behind is its inherent capacity to act as a natural bridge, mapping pixel-wise appearance into the latent space which is exactly the U-Net works in. Consequently, we are able to seamlessly copy the weights from the main branch encoder to initialize appearance branch, thereby accelerating and stabilizing the convergence process.

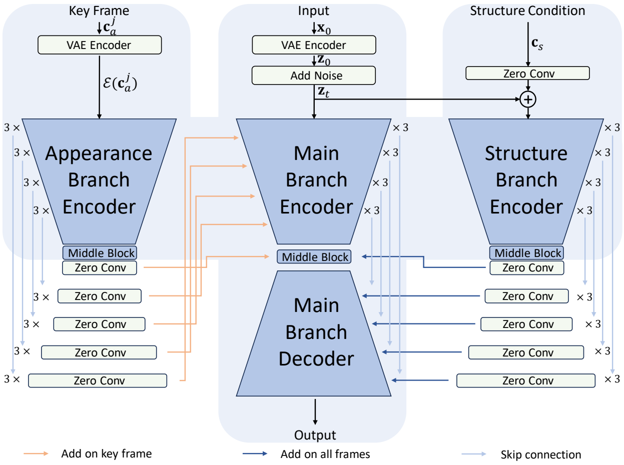


Figure 1. **Illustration of our proposed trident network. Left: Appearance branch. Middle: Main branch. Right: Structure branch.** Text prompts and time embedding are incorporated are omitted for simplicity.

B. BalanceCC Benchmark

Our objective is to develop a benchmark dataset specifically designed for tasks involving controllable and creative video editing. Therefore, we collected 100 open-license videos of different categories, including Animal, Human, Object, and Landscape. In addition, for each source video, we provided a text description and graded Camera Motion, Object Motion, and Scene Complexity on a scale from 1 to 3. For each video, there are four types of edits along with corresponding target prompts and Fantasy Levels (also ranging from 1 to 3), namely Style Change, Object Change, Background Change, and Compound Change. Our aim in doing so is to better compare the strengths and weaknesses of different methods and their areas of expertise, as well as to assist researchers in advancing their techniques. In this section, we provide details about how to prompt GPT-4V(ision) [1, 14–16] to assistant us to establish our proposed BalanceCC benchmark and some illustrative examples. The BlanceCC benchmark will be public soon.

B.1. Prompting Pipeline and Instructions

GPT-4V(ision) [1, 14–16] is a multi-modal model that possesses powerful capabilities in visual understanding, language comprehension, conversational skills, and a vast repository of knowledge. Consequently, we aim to leverage these dual capabilities to help us establish the BalanceCC benchmark. The process is akin to seeking advice from a wise person with extensive knowledge and excellent vision. Specifically, we first inform GPT-4V(ision) of our intention to create a benchmark dataset dedicated to video editing, explaining our requirements including scene complexity, original prompts, target prompts, editing types, and the corresponding fantasy levels. Then, we send the center frame of each video clip to GPT-4V(ision), allowing it to output the content we need in the specified format. In our initial attempts, we observed that GPT-4V(ision) still experienced some hallucinations, overly detailed descriptions and expansions, and instances of forgetfulness during interactions. Consequently, we made repeated and emphasized adjustments in our prompt. Additionally, we found that merely describing our needs was insufficient to achieve the desired results. Our solution was to provide corresponding examples as references, which significantly improved the quality of the content provided by GPT-4V(ision). The final prompt we used is as follows,

> Now I'm trying to build a benchmark for video editing. I need you to assist me in doing that. I will provide the center frame of each video for you. About the image, I hope you provide the following information to me:
>
> 1. Classify this video into one of "Human, Animal, Object, Landscape".
>
> 2. Describe this image, be brief, concise, and precise. Don't use too many adjectives.
>
> 3. Try to generate four text prompts of different types to edit this video. Be creative and imaginative. Offer me the corresponding "Editing Type", "Target Prompt", and "Fantasy Level" of each prompt. The "Editing Type" should be one of "Style Change, Object Change, Background Change, and Compound Change." About "Style Change", some examples are "old movies", "impressionist style", "Van Gogh style", "neon lights style", "cyberpunk style", "sepia-toned photo", "grayscale", "claymation style", "origami style", "oil painting style". About "Object Change", just change the object into other ones, like "dog to cat", "cat to tiger", "human to bear", "human to teddy bear", and even some specific identities like "Ironman". About "Background Change", just change the background, here are some examples, "in the Mars", "in the moon", "in the forest", "in the ocean", "in the castle". You can pick one of the examples I provided, and I hope you can also consider other ones that you think are interesting or suit this video. About "Compound Change", just combine what mentioned above. Please remember, be creative and imaginative, and don't be too outrageous. Besides, all targets including "Style Change, Object Change, Background Change, and Compound Change" should be provided for one video. The form of "Target Prompt" should be just like a description of an video, don't say something like "Transform the background into moon." Here is an example, the original prompt is "A majestic black swan gracefully floats on calm waters, with its reflection visible.", the "Target Prompt" can be "An elegant flamingo swan gracefully floats on calm waters, with its reflection visible, set against a backdrop of a mystical enchanted forest.". As for the "Fantasy Level" for each "Target Prompt", it indicates the degree of imagination. For example, if you change the cat to a tiger or change the background from autumn to winter, it can be seen as a relatively low degree of imagination. Transforming a cat into pixel tiger or tiger made of origami is relative high degree of imagination. Here is also 1-3 in total 3 levels. And similar to the description, be brief, concise, and precise.
>
> 4. Is the scene complex or not? Rank it from 1 to 3, corresponding to simple, moderate, and complex.

B.2. Human Refinement

Upon receiving initial outcomes from GPT-4V(ision), we engaged in a manual refinement and augmentation process. This primarily entailed the verification and rectification of existing annotations, along with the inclusion of additional details regarding the magnitude of camera and object motion within the video sequences. Specifically, our rule to define levels of different attributes is as follows:

Camera Motion: 1 corresponds to stationary, indicating minimal scene change and camera movement. 2 corresponds to slow movement, where the camera moves steadily and slowly. 3 corresponds to scenarios with intense camera shake and rapid movement.

Object Motion: 1 corresponds to stationary, where the target is almost motionless or has very minimal movement. 2 corresponds to slow movement, where the target follows a slow, simple, and regular trajectory (such as uniform linear motion). 3 corresponds to targets engaging in fast and complex movements (such as dancing and boxing).

Scene Complexity: 1 corresponds to scenes with a single target and a clean background. 2 corresponds to scenes with a few targets where both the targets and the background are not complex. 3 corresponds to scenes with multiple foreground targets, complex backgrounds, and intricate depth relationships.

Fantasy Level: 1 corresponds to simple target or background replacements and style transfers, such as transforming a dog into a cat or shifting to a Van Gogh painting style. 2 corresponds to more creative target and background replacements and style transfers, like replacing the background with a Martian landscape or turning an airplane into a dragon. 3 corresponds to complex and creative editing objectives combined together, with the Fantasy Level for Compound Change generally being 3.

B.3. Illustrative Examples

Four illustrative examples are shown in Fig. 5.

C. Experiments

C.1. Personalized T2I Models

As mentioned in the main text, our method can integrate off-the-shelf personalized models as plugins, enabling the generation of domain-specific results. In this section, we briefly introduce the principles and specific implementations of personalized models.

Stable Diffusion [19] is trained on a huge dataset that encompasses a broad spectrum of domains [22]. Although the Stable Diffusion model is highly versatile and capable of generating a wide array of images, it occasionally falls short in specific details, particularly when it comes to generating human faces and hands, where subtle variations can markedly influence the overall perception. Additionally, it often struggles to precisely meet users' expectations for specific content, styles, and attributes. Therefore, personalized T2I models are designed to address these challenges. Two respective methods are DreamBooth [21] and LoRA [8]. The former uses a unique string as an indicator to represent the corresponding domain or concept during training. Once trained, this indicator can be employed to transfer the expectations to the fine-tuned T2I model. DreamBooth faces challenges due to the extensive weight parameters, making communication less convenient. To use much less parameters and inherent the generalization of the base model, LoRA fine-tunes the model by preserving all original parameters and introducing the weight residuals ΔW to update the weights W. This process is formulated as $W' = W + \alpha \Delta W$, where α is the hyperparameter that controls the significance of the added ΔW. Typically, the parameters of ΔW are significantly fewer than those of W. Finally, two additional methods for creating robust personalized T2I base models are fine-tuning the entire model directly on the self-collected datasets and blending parameters from various models. Personalized T2I models play a crucial role in today's AI content generation. They empower both beginners and seasoned artists, as well as enthusiasts, to swiftly and autonomously produce stunning images and create new models. A significant objectives of our framework is to ensure compatibility with personalized T2I models, allowing creators to freely combine and perform highly creative edits on videos using models from the community.

In this paper, we collect several personalized T2I base models and LoRA weights from CivitAI [4] and explored different combinations, which are illustrated in Table 1. Similar to previous work [6], we employ the "trigger words" to activate these personalized models. α of all LoRA models is set as 0.9.

Model Name	Type
Counterfeit	T2I Base Model
ToonYou	T2I Base Model
rev Animated	T2I Base Model
HelloMecha	T2I Base Model
hellonijicute25d	T2I Base Model
A-Zovya Photoreal	LoRA
kMechAnimal	LoRA
Pixel Art Style	LoRA
fat animal	LoRA
Building Block World	LoRA
MoXin	LoRA
mechanical dog	LoRA

Table 1. Personalized models utilized in this paper, all sourced from CivitAI [4].

C.2. More Visualizations

Fig. 7 shows several visualized results of CCEdit. Please note that for optimal viewing of the video results, it is strongly recommended to access the website directly.

C.3. Comprehensive Comparison

C.3.1 Compared Methods

We compared our methods with eight state-of-the-art generative video editing methods: Tune-A-Video [26], vid2vid-zero [25], Text2Video-zero [9], FateZero [17], Pix2Video [3], ControlVideo [30], Rerender A Video [28], and TokenFlow [5]. The brief descriptions of these methods are as follows:

Tune-A-Video [26] propose the sparse attention mechanism to maintain the temporal coherence and optimize the network parameters through training on the source video. DDIM inversion [23] is utilized to preserve the structure of input video.

Vid2vid-zero [25] utilizes off-the-shelf image diffusion models and employs the null-text inversion module [11] for text-to-video alignment. Additionally, it incorporates a cross-frame modeling module to ensure temporal consistency and a spatial regularization module to maintain fidelity to the original video.

Text2Video-zero [9] introduces a method to enhance the latent codes of generated frames with motion dynamics, ensuring global scene and temporal consistency in the background. Additionally, it reprograms frame-level self-attention through cross-frame attention, focusing each frame on the first one to maintain the context, appearance, and identity of the foreground object.

FateZero [17] proposes to capture intermediate attention maps during inversion process, enhancing structural and motion information retention, and employs a novel spatial-temporal attention mechanism in the denoising UNet for improved frame consistency.

Pix2Video [3] involves two steps to conduct generative video editing: initially, using a structure-guided (e.g., depth) image diffusion model to edit an anchor frame based on text prompts, followed by a key step of progressively propagating these edits to subsequent frames. This is done via self-attention feature injection, adapting the core denoising phase of the diffusion model. Adjustments are then made to the latent code of each frame before continuing the process.

ControlVideo [30] leverages ControlNet [29] to ensure the structural consistency from input video clips. In addition, it introduces full cross-frame interaction in self-attention modules for appearance coherence, an interleaved-frame smoother to reduce flickering through frame interpolation.

Rerender A Video [28] propose to tackle the task of video editing by two parts: key frame translation and full video translation. Initially, it employs an adapted diffusion model to generate key frames, applying hierarchical cross-frame constraints to ensure coherence in shapes, textures, and colors. Subsequently, the framework extends these key frames to other frames using temporal-aware patch matching and frame blending techniques.

TokenFlow [5] propose the idea that the edited features convey the same inter-frame correspondences and redundancy as the original video features. Therefore, it propagates diffusion features based on inter-frame correspondences inherent in the model to ensure consistency in the diffusion feature space.

During the evaluation, all the videos consist of 17 frames at 6fps. We select depth maps as the structural representation. Additionally, to ensure fairness, the base model for all methods is Stable Diffusion v1.5.

C.3.2 Qualitative Results

The qualitative results for two videos are presented in Fig. 8 and Fig. 9. It can be observed that Tune-A-Video achieves effective editing that aligns well with the specified prompts, but falls short in maintaining temporal consistency and tends to produce overly contrasted images, possibly due to overfitting to the source video and excessively high default classifier-free guidance settings. Vid2vid-zero, Text2Video-Zero, and Pix2Video also struggle with insufficient temporal coherence. While FateZero exhibits better temporal coherence, its editing accuracy is not optimal. ControlVideo, despite its reasonable editing accuracy and temporal coherence, lacks a natural feel in its edited videos due to its global attention mechanism and interleaved-frame smoother technique. Rerender A Video demonstrates a limitation in executing precise edits, potentially due to an excessive dependence on detailed structural control mechanisms (line drawing and Canny edge of ControlNet). Such mechanisms restrict the method predominantly to minor stylistic alterations. TokenFlow achieves stable results in both temporal coherence and editing accuracy, yet it still encounters blurring issues in scenes with significant object motion or rapid camera movements (see the horse legs in Fig. 9). Finally, our approach demonstrates a notable capacity for maintaining temporal consistency, coupled with achieving exceptional accuracy in editing.

C.3.3 Quantitative Results

Evaluation Metrics. Our evaluation metrics include two aspects of both *automatic ones* and *user study* results. Automatic metrics are mainly conducted through the trained CLIP [7, 10, 18] model, similar to previous methods [3, 17, 26, 30]. Specifically, *"Tem-Con"* evaluates the temporal consistency of edited frames by calculating the similarity between successive frame pairs. Meanwhile, *"Tex-Ali"* quantifies frame-wise editing accuracy, represented as the cosine similarity between edited frames and target prompts. Additionally, the *PickScore* [10] is incorporated to predict the aesthetic quality and user preference of the edited videos. Regarding the user study, we designed an interface and invited 33 volunteers to score the videos and pickup the winners, receiving a total of 1119 ratings. Each rating corresponds to various aspects of a single video. Specifically, the aspects to be rated include: *"Editing Accuracy"*, representing whether the edited video accurately achieves the intended meaning of the target prompt; *"Aesthetics"*, denoting the visual appeal of the edited video; *"Temporal Consistency"*, indicating whether the video maintains coherence over time; and *"Overall Impression"*, which reflects the subjective overall rating of the video. The interface is illustrated in Fig. 6.

Results of Automatic Metrics. The results are illustrated in Tab. 2. Although our method ranked second in temporal consistency and first in text alignment in the table of user study presented in the main text, it did not particularly stand out in terms of corresponding objective metrics. This observation has been noted in many previous works [12, 27, 31], further emphasizing the significance of more advanced objective automatic metrics for the development of this field. Finally, our method achieved the best performance in the CLIP-based scoring function, PickScore, an indicator of human preference, demonstrating its superior alignment with human subjective perceptions.

More results of User Study. Beyond the general outcomes of the user study outlined in the main text, we have also conduct comprehensive statistics of different attributes at different levels, as shown in Tab. 4, Tab. 5, Tab. 6, Tab. 7, and Tab. 8, in terms of "Editing Type", "Scene Complexity", "Camera Motion", "Object Motion", and "Fantasy Level", respectively. The analysis and discussion are as follows:

"Editing Type": Results in Tab. 4 reveals that for most methods, "Style Change" got a higher score compared to others, which indicates that it is considered a relatively simple type of video editing task. This could be attributed to the fact that it does not necessitate semantic transformations of objects or backgrounds, but rather involves adjustments of lower-level representations. Surprisingly, the performance of various methods on "Compound Change" is somewhat better than on "Foreground Change" and "Background Change". This could be due to the increased complexity in editing necessitated by the requirement to preserve certain content from the original video in the latter cases.

"Scene Complexity": Tab. 5 shows the impact of scene complexity of the scores. The table shows that editing results in more complex scenes often receive lower scores, which is intuitively reasonable since increased complexity in the original videos typically heightens the difficulty of video editing. However, there are exceptions. FateZero and Pix2Video maintain stable or even improved scores as scene complexity increases, indicating that these two methods are relatively robust to varying levels of scene complexity.

"Camera Motion", and *"Object Motion"*: results in Tab. 6 and Tab. 7 illustrate that more intensive camera and object motion, consistently tends to result in lower scores. This is intuitive, as such scenarios invariably introduce greater challenges to the video editing task. Nevertheless, exceptions exist. FateZero, Pix2Video, and Rerender A Video exhibit insensitivity to "Camera Motion", showing stable performance regardless of it. Additionally, vid2vid-zero, FateZero, and ControlVideo demonstrate a relative insensitivity to "Object Motion", with their performance remaining stable or even improving in some cases.

"Fantasy Level": results in Tab. 8 indicate that, in addition to the attributes of the original video, the creativity of the editing objective also significantly impacts the difficulty of the edit. For most methods, as the "Fantasy Level" progresses from low to high, the corresponding scores decrease. However, there are exceptions. Tune-A-Video, Pix2Video, and ControlVideo show relative stability or even improvement in their performance across these varying levels of fantasy.

Method	Tem-Con ↑	Tex-Ali ↑	Pick ↑
Tune-A-Video [26]	0.937	0.284	0.206
vid2vid-zero [25]	0.933	0.284	0.209
Text2Video-Zero [9]	0.949	0.262	0.203
FateZero [17]	0.942	0.245	0.205
Pix2Video [3]	0.939	**0.285**	0.208
ControlVideo [30]	**0.950**	**0.285**	0.210
Rerender A Video [28]	0.928	0.247	0.201
TokenFlow [5]	0.949	0.270	0.210
CCEdit (Ours)	0.936	0.281	**0.213**

Table 2. **State-of-the-art comparison of automatic metrics.** "Tem-Con" represents temporal consistency, "Text-Ali" indicates textural alignment, and "Pick" represents to the PickScore [10].

Method	Pre-Processing	Inference	Total
Tune-A-Video [26]	545	22	567
vid2vid-zero [25]	148	230	378
Text2Video-Zero [9]	0	28	28
FateZero [17]	199	42	241
Pix2Video [3]	0	188	188
ControlVideo [30]	0	56	56
Rerender A Video [28]	76	96	172
TokenFlow [5]	182	27	209
CCEdit (Ours)	134	46	170

Table 3. **Runtime comparison (seconds).**

C.3.4 Runtime Analysis

Tab. 3 presents the runtime of various methods, detailing the time spent on pre-processing, inference, and the total duration, respectively. Pre-processing includes tasks of fine-tuning on the source video, performing inversion operations, caching attention maps, key frame editing, and others. The inference time represents the duration of the sampling process, along with all the associated operations. Overall, the time consumed by our method is not lengthy compared to other video editing techniques. It is worth noted that in our method, the time spent on key frame editing using PnP [24]) during pre-processing constitutes the majority of the total time, while the actual sampling time is relatively brief. It's attributed to the absence of any inversion and attention map operations. The only additional computational overhead arises from the extra network parameters introduced during the network forward process. In practical applications, one can opt for more advanced and lightweight image editing methods or manually make fine adjustments, thereby achieving the desired trade-off. This further demonstrates the practicality and flexibility of our approach.

C.4. Study on Control Scales

Structure Branch. Sometimes, the appearance of the edited key frame may structurally differ from the corresponding structure representation of the original video.

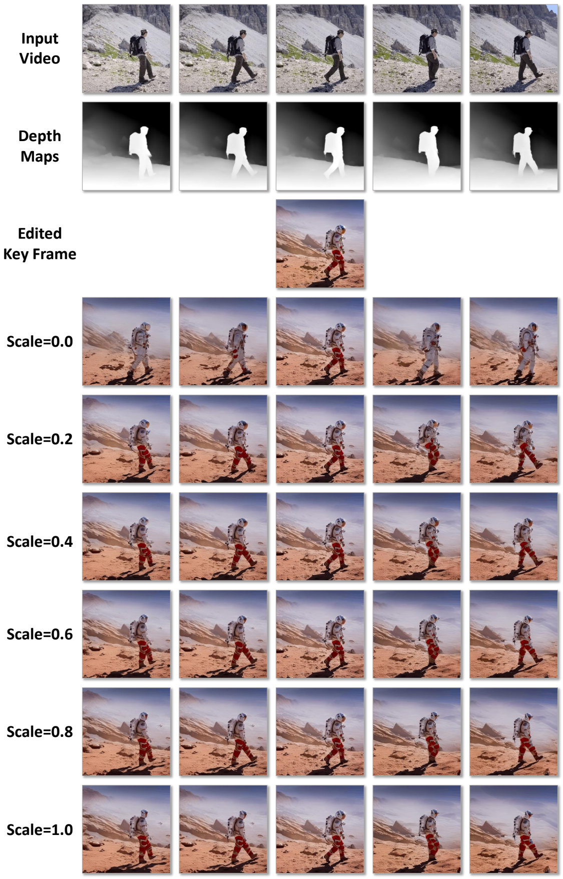


Figure 2. **Results at different scales of structure branch.** The target prompt is "An astronaut with a jetpack floats above a Martian landscape, with red rocky terrains and tall, alien-like mountains in the backdrop."

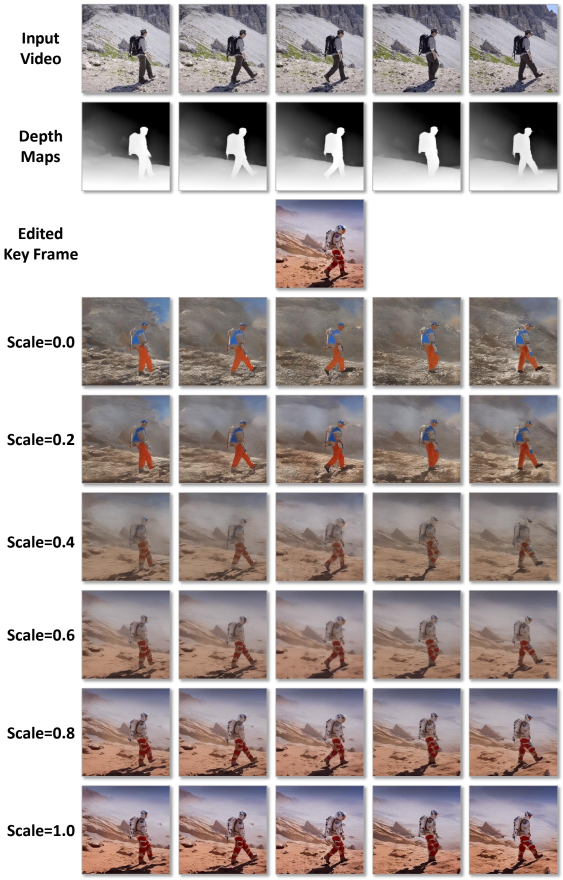


Figure 3. **Results at different scales of appearance branch.** The target prompt is "An astronaut with a jetpack floats above a Martian landscape, with red rocky terrains and tall, alien-like mountains in the backdrop."

Since the features of the structure branch are injected into the main branch through summation, the intensity of structure information infusion can be adjusted by modifying the coefficients (named control scale) applied to the features during this summation process. In such cases, reducing the control scale of the structure branch could help. This adjustment lessens its structural constraints on the results, allowing for a greater reliance on the information provided by the appearance branch and adherence to the coherence adjustments made by the temporal layers. The visualized results are shown in Fig. 2. It can be observed that in the edited key frame, the astronaut's silhouette appears markedly larger than that of the original person, a consequence of the voluminous spacesuit. When the structure control is relatively high (0.6∼1.0), the editing results show that the center frame remain consistent with the edited frame, while the structure of other frames is overly constrained by the structure representation. At a control scale of 0, the loss of structure information leads to the astronaut being unable to move correctly. However, with a moderate control scale (0.2∼0.4), a better trade-off is achieved in terms of appearance, structure, and motion. Note that in comparisons with other methods, to ensure fairness, our method consistently employed a control scale of 1.

Appearance Branch. Since the features of the Appearance Branch are also injected into the main branch through summation, the intensity of appearance information infusion can similarly be adjusted by tuning the summation coefficients of the appearance branch. The results are shown in Fig. 3. At a lower control scale (0∼0.2), the influence of appearance information is minimal, barely impacting the edited video. When the control scale is moderate (0.4∼0.6), appearance information begins to play a role. However, possibly due to conflicts with the priors of the main branch, this results in a somewhat dull and dark color tone in the visuals. Conversely, at a higher control scale (0.8∼1.0), appearance information exerts a decisive control over the overall appearance of the edited video.

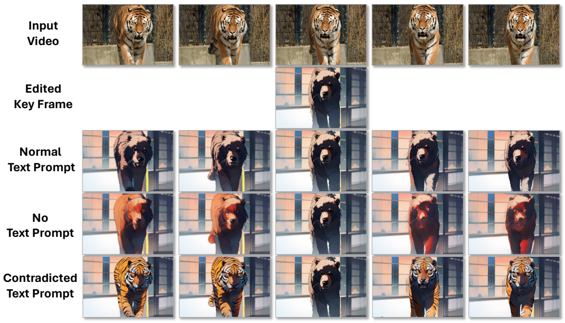


Figure 4. **Illustration of results with different text prompts**. The normal prompt is "A bear is walking". The contradicted text prompt is "A tiger is walking". The "ToonYou" personalized T2I model is used.

C.5. Study on Text Prompt

Another point worth exploring is whether text prompts are still necessary after introducing appearance control. To address this, we conducted a visual experiment. As shown in Fig. 4, providing a normal text prompt leads to correct results, whereas the absence of any text prompt results in significant distortions in the generated output. When given a text prompt that contradicts the appearance information, only the center frame retains the appearance information, while the other frames are controlled by the text prompt. Consequently, the conclusion is that text prompts are still necessary within this framework. We believe this may be due to the weights of the main branch and the structure branch being frozen during the training process. As a result, the entire editing process seems to involve the appearance branch exerting more detailed control over the image after the text prompt has already provided a coarse guide.

D. Limitation and Future Works

D.1. Structural Deviation

As described in the main text, a primary challenge that needs addressing in our video editing approach is the structural deviation (also the major issue mentioned in TokenFlow [5]) between the input and target videos. This deviation could stem from semantic changes inherent to the target or from alterations in the target's behavior. For instance, transitioning from a "cute rabbit" to a "fierce tiger" is challenging due to their fundamentally different physiological structures. Most existing methods struggle to overcome this hurdle and often only manage to modify their textural appearance. In our approach, adjusting the scale coefficient of structure branch and employing coarser-grained structure representations (like the skeleton) may alleviate this issue to some extent, but we believe it doesn't fundamentally solve the problem. Achieving changes in the target's behavior, such as transforming a "running bear" into a "dancing bear", is even more challenging. This complexity arises primarily because most contemporary generative video editing methods employ Text-to-Image (T2I) models at the image level. These models, devoid of prior knowledge concerning actions, encounter difficulties in editing motion.

We posit that a promising approach could be to integrate a pre-trained T2V (text-to-video) model, cleverly utilizing its priors to tackle these challenges.

D.2. Heavy Appearance and Structure Branch

In CCEdit, the appearance and structure branch utilize two heavy encoder to extract features, consisting significant amount of parameters. This may be unnecessary and could lead to issues such as increased GPU memory consumption and longer editing times during use. In the future, we plan to explore the adoption of more lightweight networks [13, 20] to address these concerns.

D.3. Flickering Problem.

We observed flickering in videos with higher frame rates or after frame interpolation, especially noticeable in high-frequency fine texture details. This is primarily attributed to our video editing operations being performed in the latent domain encoded by the 2D autoencoder. Introducing additional temporal layers in the autoencoder [2] is an promising way to solve this problem.

	Style Change				Object Change				Background Change				Compound Change			
Method	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.
Tune-A-Video [26]	3.13	2.91	2.96	2.87	3.32	3.00	2.74	2.63	3.06	2.94	2.62	2.75	3.48	3.13	2.43	2.83
vid2vid-zero [25]	2.62	2.50	2.29	2.42	3.16	2.21	1.79	2.16	2.71	2.57	2.29	2.29	3.38	2.38	2.14	2.52
Text2Video-Zero [9]	2.47	1.46	1.25	1.55	1.23	1.11	1.22	1.14	1.15	1.33	1.29	1.24	2.19	1.97	2.02	2.33
FateZero [17]	2.85	3.80	3.45	3.04	2.62	3.10	3.28	2.97	2.18	2.94	3.29	2.65	2.35	2.96	3.17	2.61
Pix2Video [3]	3.89	3.53	3.26	3.42	3.72	2.64	2.64	2.76	3.43	2.79	2.57	2.93	3.62	2.90	2.76	3.00
ControlVideo [30]	3.41	3.14	2.77	3.02	2.53	2.35	2.12	2.29	2.80	2.60	2.50	2.50	3.19	2.70	2.89	2.74
Rerender A Video [28]	2.71	3.10	3.00	2.90	2.44	2.29	2.47	2.12	2.78	2.89	3.11	2.67	2.24	2.69	2.93	2.48
TokenFlow [5]	4.06	3.94	**4.07**	3.89	3.73	3.27	**3.64**	3.36	3.38	**3.85**	3.53	3.54	4.06	3.56	**3.78**	3.72
CCEdit (Ours)	**4.19**	**4.27**	4.00	**4.04**	**4.00**	**3.81**	3.58	**3.77**	**3.65**	3.82	**3.54**	**3.60**	**4.26**	**4.04**	3.71	**3.91**

Table 4. **Quantitative comparison in terms of "Editing Type".**

	Simple				Moderate				Complex			
Method	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.
Tune-A-Video [26]	3.14	3.27	2.95	2.85	3.20	2.75	2.53	2.71	2.63	2.88	2.38	2.47
vid2vid-zero [25]	2.81	2.57	2.10	2.38	2.92	2.28	2.15	2.36	3.64	2.45	2.04	2.36
Text2Video-Zero [9]	2.38	1.86	1.71	1.98	1.93	1.43	1.42	1.32	1.83	1.21	1.12	1.11
FateZero [17]	2.32	2.89	3.07	2.64	2.57	3.38	3.42	2.91	2.88	3.00	3.25	2.88
Pix2Video [3]	3.38	2.62	2.48	2.66	3.82	3.03	2.85	3.13	3.04	3.64	3.55	3.55
ControlVideo [30]	3.08	3.33	3.03	3.12	2.97	2.41	2.46	2.43	3.33	2.83	2.52	2.83
Rerender A Video [28]	1.93	2.57	3.07	2.21	2.69	3.04	3.04	2.80	1.92	1.75	2.01	1.83
TokenFlow [5]	4.15	3.74	**4.02**	3.81	3.84	3.67	3.75	3.65	3.36	3.18	**3.35**	3.18
CCEdit (Ours)	**4.23**	**4.11**	3.84	**4.10**	**4.05**	**4.02**	**3.76**	**3.81**	**3.73**	**3.81**	3.33	**3.56**

Table 5. **Quantitative comparison in terms of "Scene Complexity".**

	Stationary				Slow				Quick			
Method	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.
Tune-A-Video [26]	3.17	2.94	2.81	2.91	3.42	3.18	2.58	2.97	2.86	2.29	2.71	2.57
vid2vid-zero [25]	3.03	2.41	2.11	2.38	2.94	2.31	2.03	2.28	2.50	2.17	2.35	2.07
Text2Video-Zero [9]	2.33	1.67	1.63	1.69	2.08	1.52	1.37	1.45	1.72	1.14	1.08	1.13
FateZero [17]	2.56	3.00	3.18	2.75	2.39	3.31	3.42	2.86	3.33	2.98	**3.17**	3.02
Pix2Video [3]	3.50	2.70	2.70	2.83	3.94	3.35	3.08	3.26	3.06	2.53	2.58	3.12
ControlVideo [30]	3.23	2.98	2.83	2.98	2.97	2.59	2.53	2.50	2.08	1.49	1.52	1.28
Rerender A Video [28]	2.25	2.81	2.97	2.47	2.53	2.74	2.89	2.61	3.02	2.45	2.04	2.41
TokenFlow [5]	4.02	3.70	**3.86**	3.72	3.52	3.52	3.80	3.48	3.11	3.23	3.12	2.98
CCEdit (Ours)	**4.06**	**3.96**	3.68	**3.85**	**4.10**	**4.14**	**3.90**	**3.95**	**3.79**	**3.42**	3.14	**3.21**

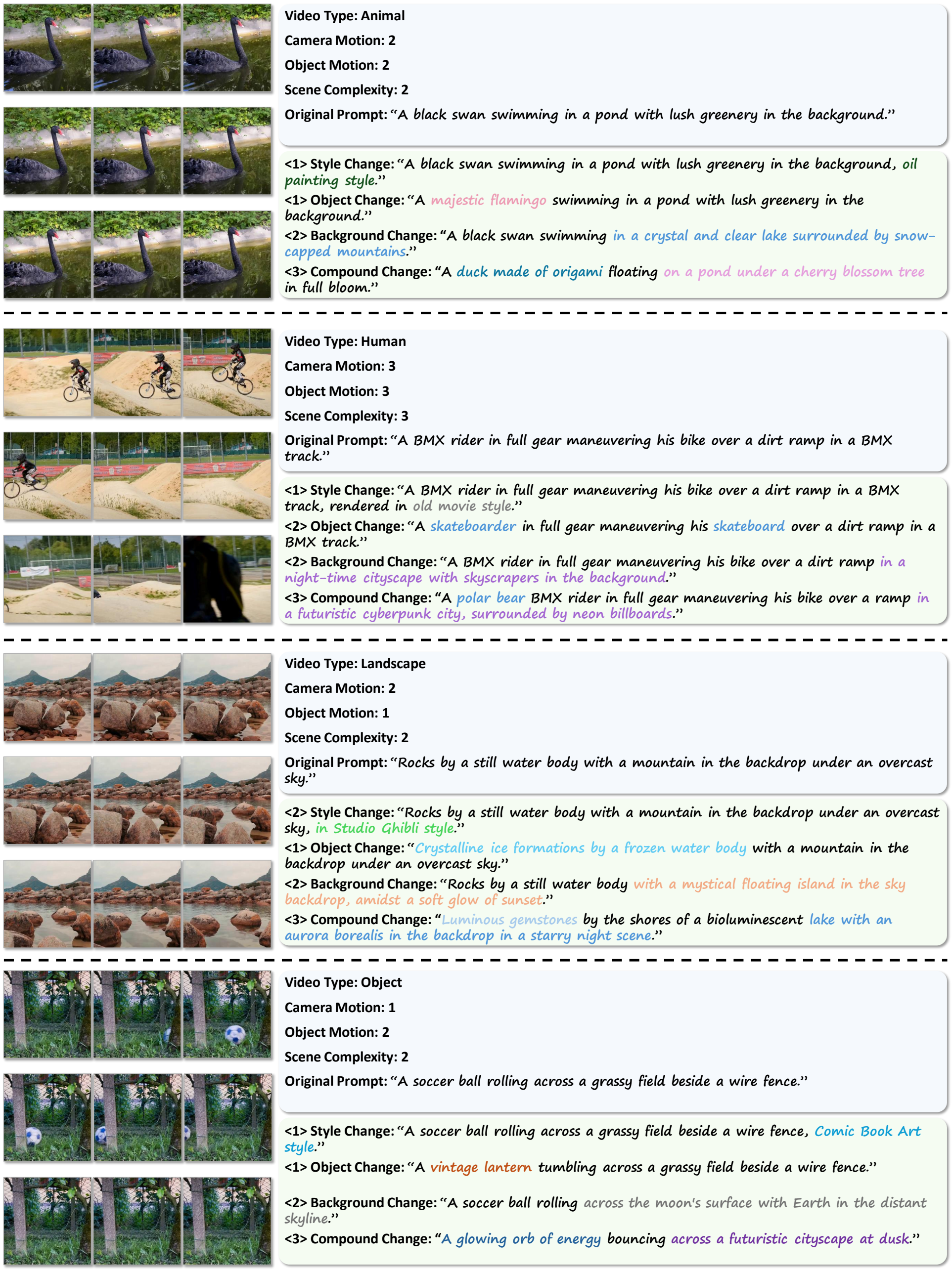
Table 6. **Quantitative comparison in terms of "Camera Motion".**

	Stationary				Slow				Quick			
Method	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.
Tune-A-Video [26]	3.27	3.01	2.55	2.82	3.32	3.15	2.77	2.81	3.13	2.72	2.67	2.73
vid2vid-zero [25]	2.67	2.07	1.53	2.07	2.95	2.37	2.21	2.37	3.47	2.80	2.47	2.67
Text2Video-Zero [9]	2.31	1.74	1.78	1.83	2.20	1.32	1.27	1.38	1.23	1.17	1.12	1.22
FateZero [17]	2.58	3.02	3.33	2.83	2.43	3.18	3.34	2.80	2.59	3.30	3.19	2.78
Pix2Video [3]	3.36	3.50	3.29	3.48	3.87	3.33	3.04	3.36	3.50	2.40	2.65	2.74
ControlVideo [30]	2.91	2.84	2.78	2.63	3.02	2.62	2.53	2.66	3.25	3.06	2.62	2.81
Rerender A Video [28]	1.90	2.70	2.80	3.23	2.52	2.90	3.02	2.67	2.47	2.35	2.47	2.35
TokenFlow [5]	3.94	3.47	4.01	3.73	3.89	3.81	3.81	3.66	3.59	3.35	**3.53**	3.41
CCEdit (Ours)	**4.10**	**4.27**	**4.22**	**4.00**	**4.18**	**4.10**	**3.83**	**3.97**	**3.78**	**3.63**	3.49	**3.53**

Table 7. **Quantitative comparison in terms of "Object Motion".**

	Low				Medium				High			
Method	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.	Edit	Aes.	Tem.	Ove.
Tune-A-Video [26]	3.11	3.11	2.68	2.68	3.18	2.79	2.82	2.82	3.46	3.18	2.54	2.79
vid2vid-zero [25]	3.05	2.47	2.11	2.47	2.52	2.30	2.11	2.15	3.48	2.44	2.12	2.52
Text2Video-Zero [9]	2.43	1.97	1.55	1.54	1.86	1.37	1.37	1.57	2.11	1.22	1.20	1.38
FateZero [17]	2.78	3.30	3.39	3.09	2.49	3.32	3.32	2.80	2.32	2.88	3.16	2.60
Pix2Video [3]	3.87	2.74	2.57	2.83	3.66	3.12	3.03	3.12	3.54	2.92	2.75	3.04
ControlVideo [30]	2.78	2.61	2.50	2.50	3.17	2.93	2.52	2.83	3.10	2.62	2.83	2.66
Rerender A Video [28]	2.69	2.85	2.62	2.74	2.34	2.71	2.93	2.45	2.39	2.67	2.91	2.52
TokenFlow [5]	4.19	3.88	3.88	3.81	3.90	3.69	**3.90**	3.72	3.54	3.38	**3.62**	3.38
CCEdit (Ours)	**4.23**	**4.01**	**3.69**	**4.00**	**4.01**	**4.05**	3.83	**3.86**	**4.02**	**3.95**	3.61	**3.76**

Table 8. **Quantitative comparison in terms of "Fantasy Level".**



Video Type: Animal
Camera Motion: 2
Object Motion: 2
Scene Complexity: 2
Original Prompt: "A black swan swimming in a pond with lush greenery in the background."

<1> Style Change: "A black swan swimming in a pond with lush greenery in the background, oil painting style."
<1> Object Change: "A majestic flamingo swimming in a pond with lush greenery in the background."
<2> Background Change: "A black swan swimming in a crystal and clear lake surrounded by snow-capped mountains."
<3> Compound Change: "A duck made of origami floating on a pond under a cherry blossom tree in full bloom."

Video Type: Human
Camera Motion: 3
Object Motion: 3
Scene Complexity: 3
Original Prompt: "A BMX rider in full gear maneuvering his bike over a dirt ramp in a BMX track."

<1> Style Change: "A BMX rider in full gear maneuvering his bike over a dirt ramp in a BMX track, rendered in old movie style."
<2> Object Change: "A skateboarder in full gear maneuvering his skateboard over a dirt ramp in a BMX track."
<2> Background Change: "A BMX rider in full gear maneuvering his bike over a dirt ramp in a night-time cityscape with skyscrapers in the background."
<3> Compound Change: "A polar bear BMX rider in full gear maneuvering his bike over a ramp in a futuristic cyberpunk city, surrounded by neon billboards."

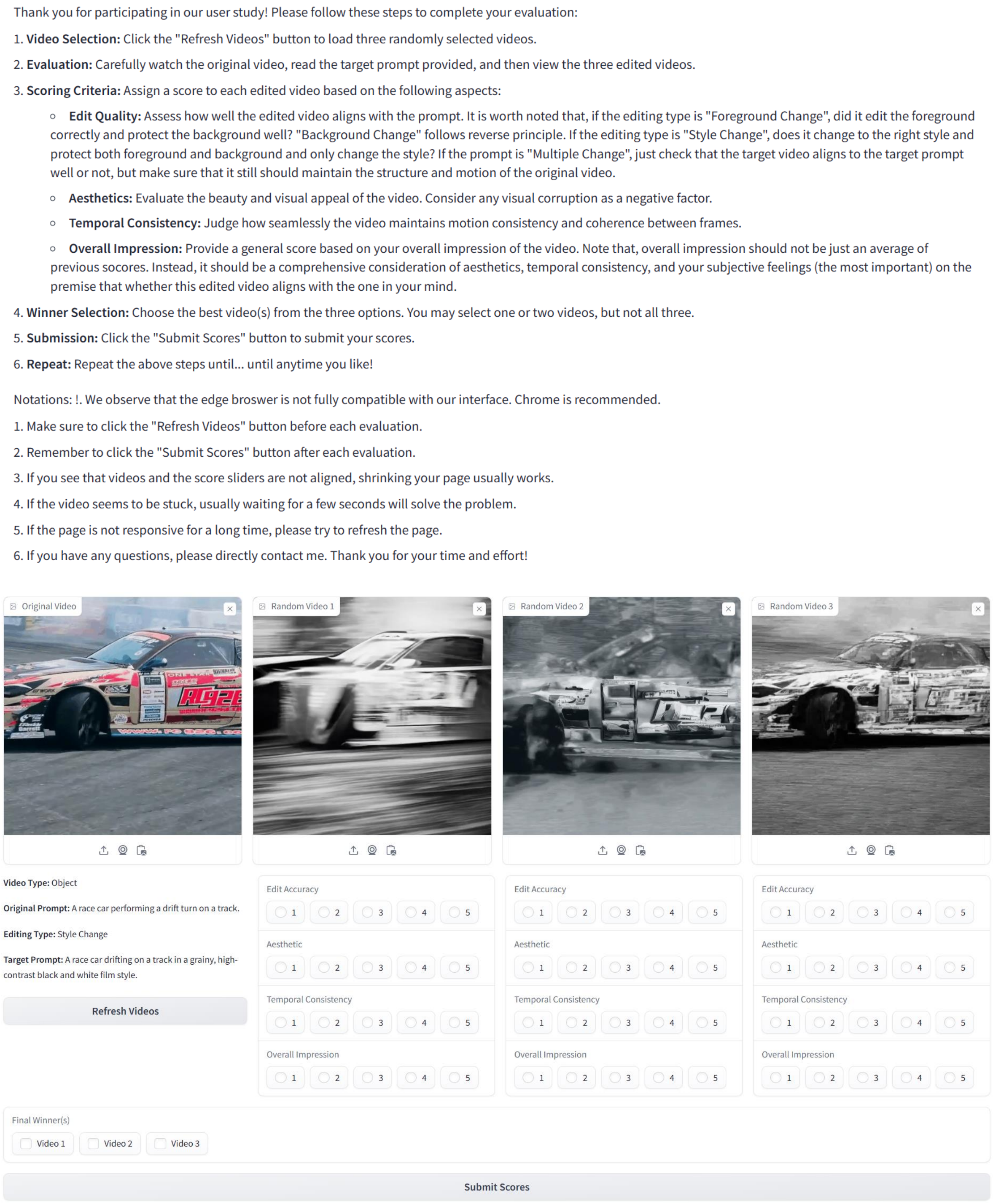
Video Type: Landscape
Camera Motion: 2
Object Motion: 1
Scene Complexity: 2
Original Prompt: "Rocks by a still water body with a mountain in the backdrop under an overcast sky."

<2> Style Change: "Rocks by a still water body with a mountain in the backdrop under an overcast sky, in Studio Ghibli style."
<1> Object Change: "Crystalline ice formations by a frozen water body with a mountain in the backdrop under an overcast sky."
<2> Background Change: "Rocks by a still water body with a mystical floating island in the sky backdrop, amidst a soft glow of sunset."
<3> Compound Change: "Luminous gemstones by the shores of a bioluminescent lake with an aurora borealis in the backdrop in a starry night scene."

Video Type: Object
Camera Motion: 1
Object Motion: 2
Scene Complexity: 2
Original Prompt: "A soccer ball rolling across a grassy field beside a wire fence."

<1> Style Change: "A soccer ball rolling across a grassy field beside a wire fence, Comic Book Art style."
<1> Object Change: "A vintage lantern tumbling across a grassy field beside a wire fence."
<2> Background Change: "A soccer ball rolling across the moon's surface with Earth in the distant skyline."
<3> Compound Change: "A glowing orb of energy bouncing across a futuristic cityscape at dusk."

Figure 5. **Illustrative examples of BalanceCC benchmark dataset.**



Thank you for participating in our user study! Please follow these steps to complete your evaluation:

1. **Video Selection:** Click the "Refresh Videos" button to load three randomly selected videos.
2. **Evaluation:** Carefully watch the original video, read the target prompt provided, and then view the three edited videos.
3. **Scoring Criteria:** Assign a score to each edited video based on the following aspects:
   - **Edit Quality:** Assess how well the edited video aligns with the prompt. It is worth noted that, if the editing type is "Foreground Change", did it edit the foreground correctly and protect the background well? "Background Change" follows reverse principle. If the editing type is "Style Change", does it change to the right style and protect both foreground and background and only change the style? If the prompt is "Multiple Change", just check that the target video aligns to the target prompt well or not, but make sure that it still should maintain the structure and motion of the original video.
   - **Aesthetics:** Evaluate the beauty and visual appeal of the video. Consider any visual corruption as a negative factor.
   - **Temporal Consistency:** Judge how seamlessly the video maintains motion consistency and coherence between frames.
   - **Overall Impression:** Provide a general score based on your overall impression of the video. Note that, overall impression should not be just an average of previous socores. Instead, it should be a comprehensive consideration of aesthetics, temporal consistency, and your subjective feelings (the most important) on the premise that whether this edited video aligns with the one in your mind.
4. **Winner Selection:** Choose the best video(s) from the three options. You may select one or two videos, but not all three.
5. **Submission:** Click the "Submit Scores" button to submit your scores.
6. **Repeat:** Repeat the above steps until... until anytime you like!

Notations: !. We observe that the edge broswer is not fully compatible with our interface. Chrome is recommended.

1. Make sure to click the "Refresh Videos" button before each evaluation.
2. Remember to click the "Submit Scores" button after each evaluation.
3. If you see that videos and the score sliders are not aligned, shrinking your page usually works.
4. If the video seems to be stuck, usually waiting for a few seconds will solve the problem.
5. If the page is not responsive for a long time, please try to refresh the page.
6. If you have any questions, please directly contact me. Thank you for your time and effort!

Video Type: Object

Original Prompt: A race car performing a drift turn on a track.

Editing Type: Style Change

Target Prompt: A race car drifting on a track in a grainy, high-contrast black and white film style.

Refresh Videos

Random Video 1	Random Video 2	Random Video 3
Edit Accuracy ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Edit Accuracy ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Edit Accuracy ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Aesthetic ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Aesthetic ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Aesthetic ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Temporal Consistency ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Temporal Consistency ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Temporal Consistency ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Overall Impression ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Overall Impression ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	Overall Impression ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Final Winner(s)

□ Video 1 □ Video 2 □ Video 3

Submit Scores

Figure 6. **Illustration of the interface to conduct user study.** Initially, we provided a description of the evaluation criteria at the top of the page, along with corresponding notes for consideration. Additionally, to reduce user burden and avoid the confusion of displaying multiple videos simultaneously, for each video's editing result, we randomly selected three from all nine options (eight comparative methods and our method) for users to rate on various criteria (Edit Accuracy, Aesthetic, Temporal Consistency, and Overall Impression). Finally, users were asked to choose the winner(s) among the three videos. Selecting multiple winners was allowed (up to two), but choosing none or all three was not permitted. Zoom in to see details.

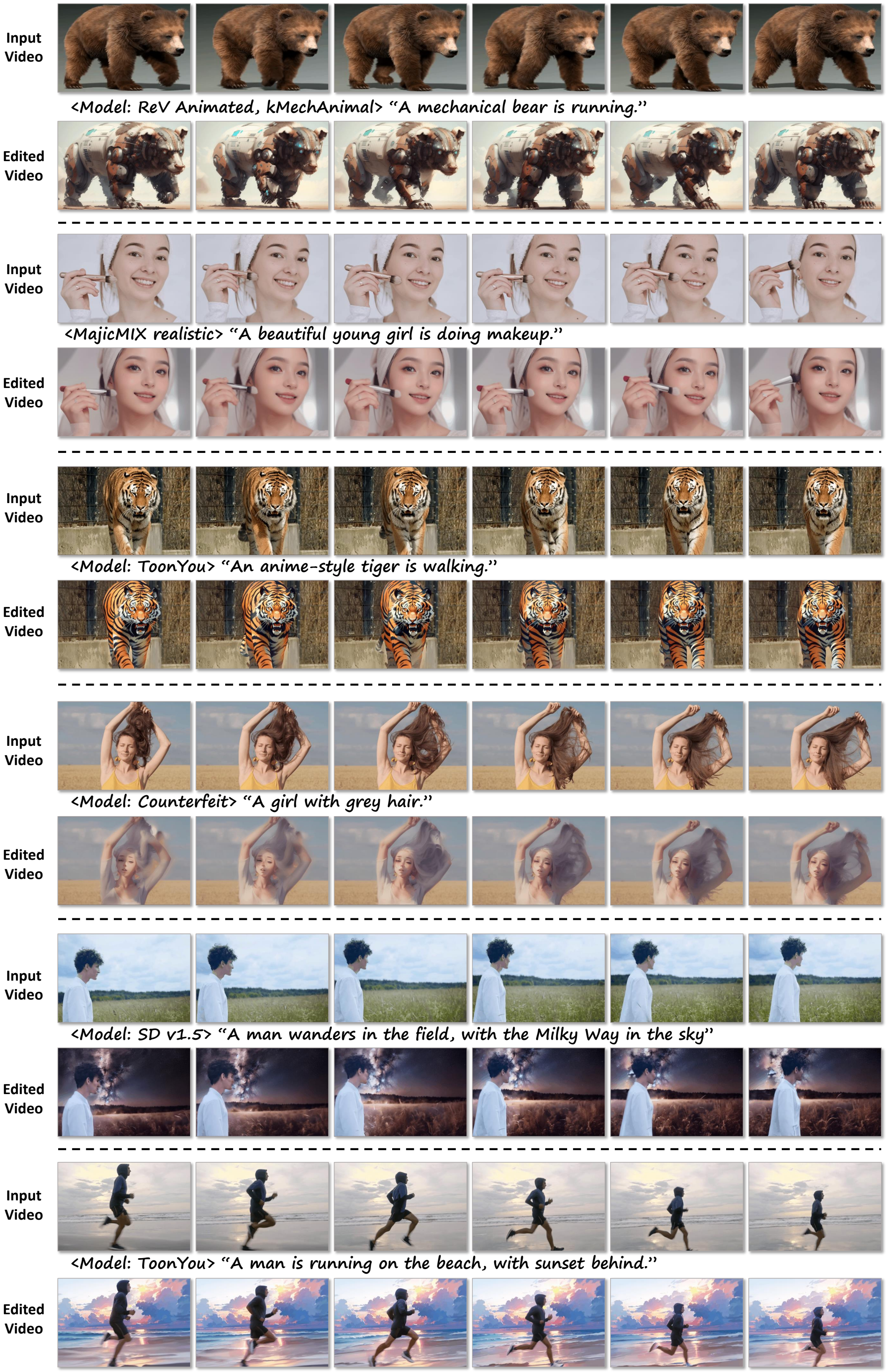

Input Video
<Model: ReV Animated, kMechAnimal> “A mechanical bear is running.”
Edited Video
Input Video
<MajicMIX realistic> “A beautiful young girl is doing makeup.”
Edited Video
Input Video
<Model: ToonYou> “An anime-style tiger is walking.”
Edited Video
Input Video
<Model: Counterfeit> “A girl with grey hair.”
Edited Video
Input Video
<Model: SD v1.5> “A man wanders in the field, with the Milky Way in the sky”
Edited Video
Input Video
<Model: ToonYou> “A man is running on the beach, with sunset behind.”
Edited Video

Figure 7. **Visualized results of CCEdit.** $\langle\cdot\rangle$ indicates the personalized T2I model we used.

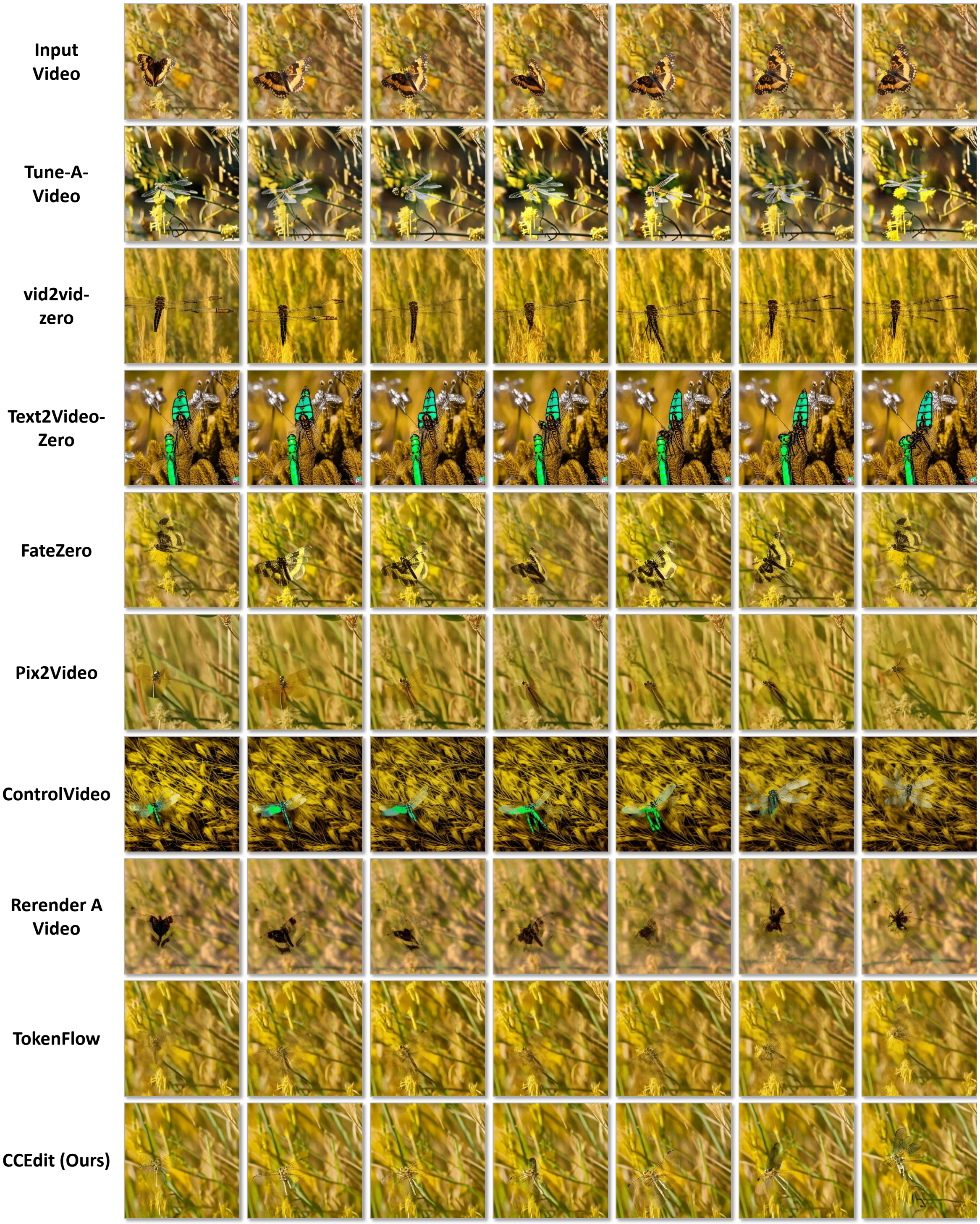


Figure 8. **Qualitative comparison of different methods.** The target prompt is “A dragonfly with shimmering wings perches on a plant amidst a field of golden grass.”

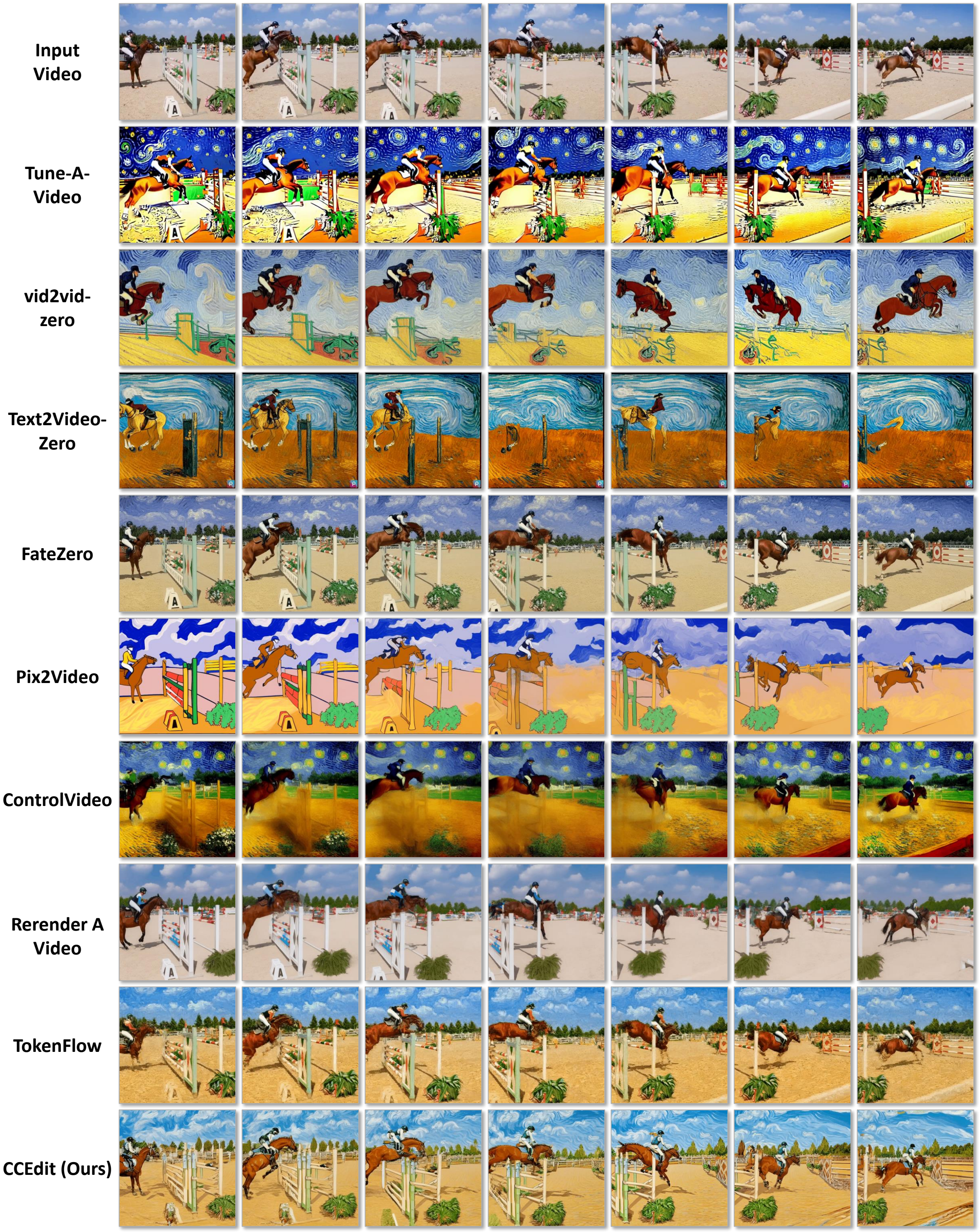


Figure 9. **Qualitative comparison of different methods.** The target prompt is “A rider on a horse jumping over an obstacle in an equestrian competition, rendered in Van Gogh style with swirling skies and vibrant colors.”

References

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

[2] 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. 7

[3] 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, 4, 5, 8, 9

[4] Civitai. Civitai. https://civitai.com/, 2022. 3

[5] 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, 4, 5, 7, 8, 9

[6] 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

[7] 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. 4

[8] 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. 3

[9] 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, 5, 8, 9

[10] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 4, 5

[11] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[12] 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. 4

[13] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 7

[14] OpenAI. Gpt-4v(ision) system card. 2023. 1

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

[16] OpenAI. Gpt-4 technical report, 2023. 1

[17] 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. 3, 4, 5, 8, 9

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[19] 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. 3

[20] Denis Zavadski Carsten Rother. Controlnet-xs. 2023. 7

[21] 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. 3

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3

[24] 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. 5

[25] 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, 5, 8, 9

[26] 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, 4, 5, 8, 9

[27] 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. 4

[28] 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. 3, 4, 5, 8, 9

[29] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 4

[30] 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, 4, 5, 8, 9

[31] 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. 4