ControlVideo: Adding Conditional Control for One Shot Text-to-Video Editing

Min Zhao[1,3], **Rongzhen Wang**[2], **Fan Bao**[1,3], **Chongxuan Li**[2*], **Jun Zhu**[1,3,4*]
[1]Dept. of Comp. Sci. & Tech., BNRist Center, THU-Bosch ML Center, THBI Lab, Tsinghua University
[2] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods , Beijing, China
[3]ShengShu, Beijing, China; [4]Pazhou Laboratory (Huangpu), Guangzhou, China
gracezhao1997@gmail.com; wangrz@ruc.edu.cn; bf19@mails.tsinghua.edu.cn;
chongxuanli@ruc.edu.cn; dcszj@tsinghua.edu.cn

Abstract

In this paper, we present *ControlVideo*, a novel method for text-driven video editing. Leveraging the capabilities of text-to-image diffusion models and ControlNet, ControlVideo aims to enhance the fidelity and temporal consistency of videos that align with a given text while preserving the structure of the source video. This is achieved by incorporating additional conditions such as edge maps, fine-tuning the key-frame and temporal attention on the source video-text pair with carefully designed strategies. An in-depth exploration of ControlVideo's design is conducted to inform future research on one-shot tuning video diffusion models. Quantitatively, ControlVideo outperforms a range of competitive baselines in terms of faithfulness and consistency while still aligning with the textual prompt. Additionally, it delivers videos with high visual realism and fidelity w.r.t. the source content, demonstrating flexibility in utilizing controls containing varying degrees of source video information, and the potential for multiple control combinations. The project page is available at https://ml.cs.tsinghua.edu.cn/controlvideo/.

1 Introduction

The endeavor of text-driven video editing is to seamlessly generate novel videos derived from textual prompts and existing video footage, thereby reducing manual labor. This technology stands to significantly influence an array of fields such as advertising, marketing, and social media content. Within this process, it is critical that the edited videos should *faithfully* preserve the content of the source video, maintain *temporal consistency* between generated frames and *align* with the target prompts. However, fulfilling all these requirements simultaneously presents considerable challenges.

Training a text-to-video model [1, 2] directly on extensive text-video data necessitates considerable computational resources. Recent advancements in large-scale text-to-image diffusion models [3–5] and controllable image editing [6–8] have been leveraged in zero-shot [9, 10] and one-shot [11, 12] methodologies for text-driven video editing. These developments have shown promising capabilities to edit videos in response to a variety of textual prompts, without requiring additional video data. However, despite the significant strides made in aligning output with text prompts, empirical evidence (see Figure 6 and Table 1) suggests that existing approaches still struggle to faithfully and adequately control the output, while also preserving temporal consistency.

To this end, we present *ControlVideo*, a novel approach for faithful and consistent text-driven video editing, based on a pretrained text-to-image diffusion model. Drawing inspiration from

[*] The Corresponding authors.

ControlNet [13], ControlVideo incorporates visual conditions such as Canny edge maps, HED boundaries and depth maps for all frames as additional inputs, thereby amplifying the source video's guidance. These visual conditions are processed by a ControlNet that has been pretrained on the diffusion model. Notably, such conditions provide a more accurate and flexible way of video control compared to text and attention-based strategies [6–8] currently used in text-driven video editing methods [9–12].

Furthermore, we have meticulously designed and fine-tuned the attention modules in both the diffusion model and ControlNet to enhance faithfulness and temporal consistency while preventing overfitting (see Sec. 3.2). Specifically, we transform the original spatial self-attention in both models into key-frame attention, aligning all frames with a selected one. Additionally, we incorporate temporal attention modules as extra branches in the diffusion model, succeeded by a zero convolutional layer [13] to preserve the output before fine-tuning. Given the observation that different attention mechanisms model the relationships between various positions, but consistently model the relationships between image features, we employ the original spatial self-attention weights as initialization for both key-frame attention and temporal attention in the corresponding network.

We conduct a systematic empirical study of the key components of ControlVideo (see Sec. 3.3), with the aim of informing future research into video diffusion model backbones for one-shot tuning. This study encompasses an analysis of the design of key and value and the parameters for finetuning in self-attention as well as the way to initialization and the incorporation of local and global positions for introducing temporal attention. Our results show the optimal performance is achieved by selecting a key frame as both key and value with finetuning W^O (see Figure 3) and incorporating temporal attention along with self-attention (key-frame attention in this work), in the main UNet except middle block, using pretraining weights (see Figure 4). We also thoroughly analyze the individual contributions of each component, as well as their combined effect (see Figure 5).

We collect 40 video-text pair data including the Davis dataset following the work [9, 11, 12] and others from the internet for evaluation. We compare with frame-wise Stable Diffusion and SOTA text-driven video editing methods [9, 11, 12] under various metrics. In particular, following [9, 12] we use CLIP [14] to quantify the text-alignment and temporal consistency and the SSIM [15] score to measure the faithfulness. Moreover, we perform a user study between ControlVideo and all baselines. Extensive results demonstrate ControlVideo substantially outperforms all these baselines in terms of faithfulness and temporal consistency with comparable text alignment performance.

Notably, our empirical results demonstrate the appealing ability of ControlVideo to produce videos with extremely realistic visual quality and very faithfully preserve original source content while following the text guidance. For instance, ControlVideo is able to makeup a person with maintaining the unique facial characteristics while all existing methods fail (see Figure 6). Moreover, ControlVideo is flexible to leverage different control types which contain varying degrees of information from the source video and thus provide flexible trade-off of the faithfulness and the editability of the video (see Figure 1). For instance, HED boundary provides detailed boundary information of the source video and is suitable for refine control such as facial video editing. Pose only contains the motion information from the source video and thus provides more flexible to change the subject and background, which is suitable for motion transfer. We also demonstrate the possibility to combine multiple controls to utilize the advantage of different control types.

2 Background

Diffusion Models. Diffusion models [16–19] gradually perturb data $x_0 \sim q(x_0)$ by a forward diffusion process:

$$q(x_{1:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t \boldsymbol{I}),$$

where β_t is the noise schedule, $\alpha_t = 1 - \beta_t$ and is designed to satisfy $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The data can be generated starting from Gaussian noise through the reverse diffusion process, where the reverse transition kernel $q(x_{t-1}|x_t)$ is learned by a Gaussian model: $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \boldsymbol{\mu}_\theta(x_t), \sigma_t^2 \boldsymbol{I})$. [18] shows learning the mean $\boldsymbol{\mu}_\theta(x_t)$ can be derived to learn a noise prediction network $\epsilon_\theta(x_t, t)$ via

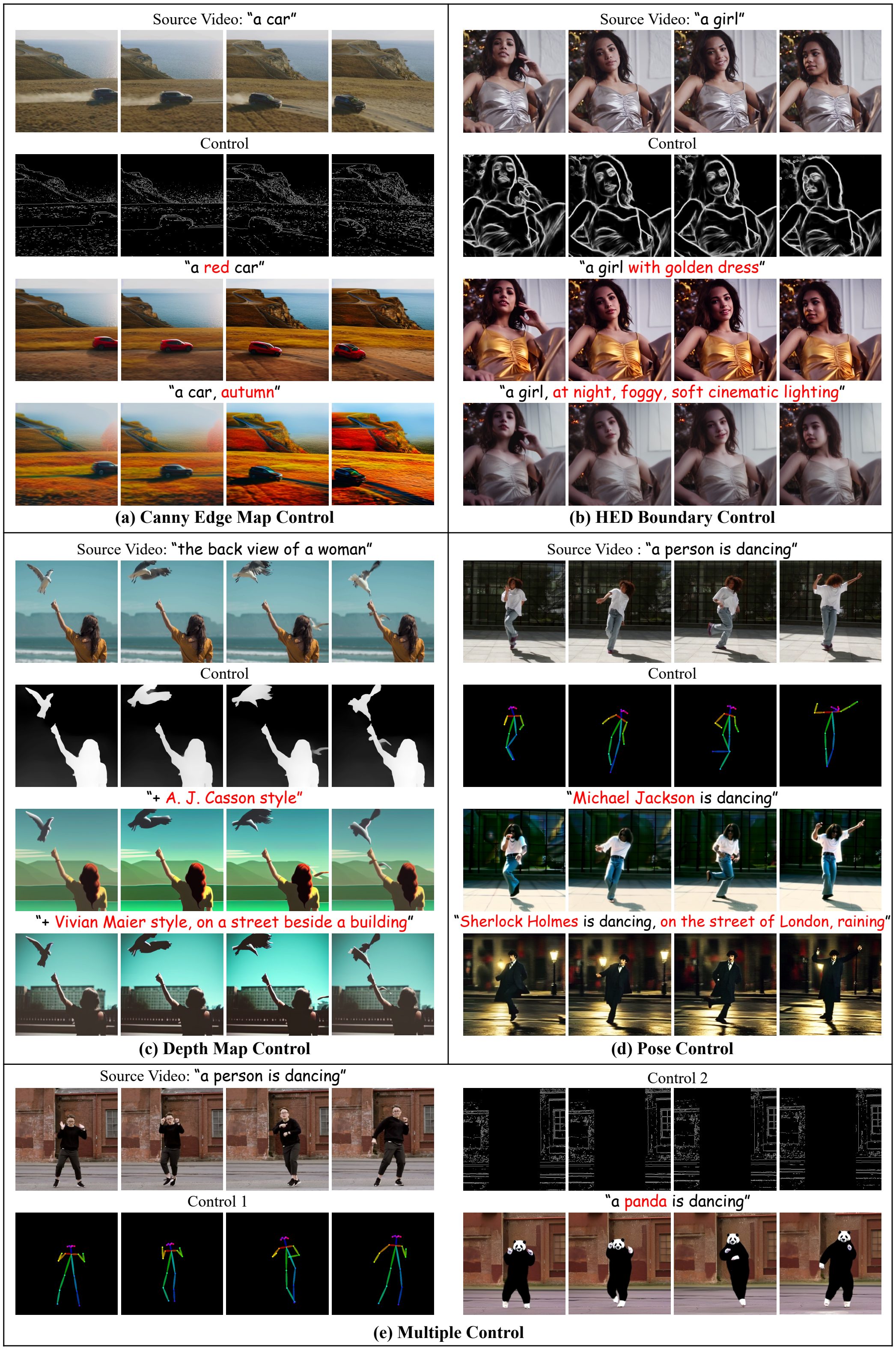

Source Video: "a car"
Control
"a red car"
"a car, autumn"
(a) Canny Edge Map Control
Source Video: "a girl"
Control
"a girl with golden dress"
"a girl, at night, foggy, soft cinematic lighting"
(b) HED Boundary Control
Source Video: "the back view of a woman"
Control
"+ A. J. Casson style"
"+ Vivian Maier style, on a street beside a building"
(c) Depth Map Control
Source Video : "a person is dancing"
Control
"Michael Jackson is dancing"
"Sherlock Holmes is dancing, on the street of London, raining"
(d) Pose Control
Source Video: "a person is dancing"
Control 1
Control 2
"a panda is dancing"
(e) Multiple Control

Figure 1: The main results of ControlVideo with different controls including (a) Canny edge maps, (b) HED boundary, (c) depth maps and (d) pose. ControlVideo can generate faithful and consistent videos in the task of *attributes, style, background editing* and *replacing subjects*. By choosing different control types, ControlVideo allows users to flexibly adjust the balance between faithfulness and editing capabilities. Multiple controls can be easily combined together for video editing.

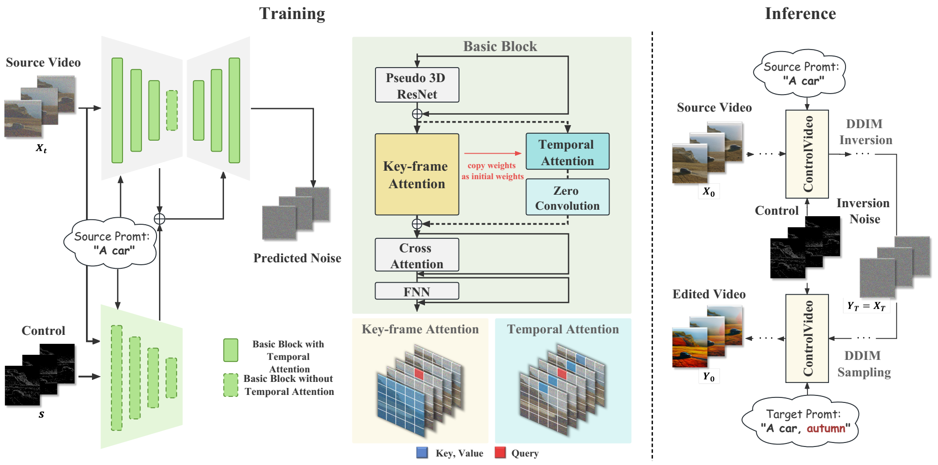


Figure 2: Flowchart of ControlVideo. ControlVideo incorporates visual conditions for all frames to amplify the source video's guidance, key-frame attention that aligns all frames with a selected one and temporal attention modules succeeded by a zero convolutional layer for temporal consistency and faithfulness. The three key components and corresponding fine-tuned parameters are designed by a systematic empirical study (3.3). Built upon the trained ControlVideo, during inference, we employ DDIM inversion to obtain the noise inversion X_T and then generate the edited video Y_0 using the target prompt starting from $Y_T = X_T$ via DDIM sampling.

a mean-squared error loss:

$$\min_{\theta} \mathbb{E}_{t,x_0,\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}||\epsilon - \epsilon_\theta(x_t, t)||^2. \tag{1}$$

DDIM inversion and DDIM sampling. Deterministic DDIM sampling [20] is one of ODE-based sampling methods [16, 21] to generate samples starting from $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ via the following iteration rule:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t). \tag{2}$$

DDIM inversion [20] can convert a real image x_0 to related inversion noise by reversing the above process, which can be reconstructed by DDIM sampling. Therefore, it is usually adopted in image editing [6, 7, 22].

Latent Diffusion Models. To reduce computational resources, latent diffusion model (LDM) [3] first leverages an encoder $\mathcal{E}$ to transform the image x_0 into lower-dimensional latent space $\boldsymbol{z}_0 = \mathcal{E}(x_0)$, which can be reconstructed by a decoder $x_0 \approx \mathcal{D}(\boldsymbol{z}_0)$, and then trains the noise prediction network $\epsilon_\theta(\boldsymbol{z}_t, t)$ in the latent space. For text-to-image generation, LDM learns a conditional noise prediction network $\epsilon_\theta(x_t, p, t)$ w.r.t. the mean-squared error:

$$\min_{\theta} \mathbb{E}_{t,x_0,p,\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}||\epsilon - \epsilon_\theta(x_t, p, t)||^2,$$

where p is the textual prompts.

ControlNet. To enable models to learn additional conditions c such as edge maps, ControlNet [13] constructs the noise prediction network $\epsilon_\theta(x_t, p, c, t)$ by adding a trainable copy to learn task-specific conditions on the locked large-scale T2I diffusion models and connecting the two components via the convolution layer with zero initialization.

3 ControlVideo

In this section, we present *ControlVideo*, a framework designed to enhance faithfulness and temporal consistency in text-driven video editing, built upon Stable Diffusion [3] and ControNet [13]. Specifically, in Section 3.1, we detail the training and sampling framework of ControlVideo. The architectural design of ControlVideo is explained in Section 3.2. Further, we conduct a comprehensive empirical examination of ControlVideo's key components in Section 3.3, analyzing the impact of each component individually and their collective influence when combined.

3.1 Training and Sampling Framework

Formally, given the source video $X_0 = \{x_0^i\}_{i=1}^N$[2] with N frames, source prompt p_s and the target prompt p_t, the goal of text-driven video editing is to generate a video $Y_0 = \{y_0^i\}_{i=1}^N$ that aligns with the target prompt p_t, faithfully preserves the content of the source video X_0, and maintains temporal consistency between the generated frames. Let $\epsilon_\phi(X_t, p, c, t)$ denote the text-to-video model in ControlVideo built upon Stable Diffusion [3] and ControlNet [13]. Let c denote the additional conditions (e.g., Canny edge maps, HED boundaries, and depth maps for all frames) similarly to ControlNet [13] and ϕ denote the parameters in the key-frame attention and temporal attention (see details in Sec. 3.2). We freeze all the other parameters in Stable Diffusion and ControlNet and therefore omit them in notation. Similarly to Eq. 1, we finetune $\epsilon_\phi(X_t, p, c, t)$ on the source video-text pair (X_0, p_s) via the mean-squared error loss as follows:

$$\min_\phi \mathbb{E}_{t,\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})} ||\epsilon - \epsilon_\phi(X_t, p_s, c, t)||^2.$$

Built upon $\epsilon_{\phi^*}(X_t, p_s, c, t)$, during inference, we employ DDIM inversion (see Eq. (2)) to obtain the noise inversion X_T, which encodes the information of the source video and then generate the edited video Y_0 using the target prompt p_t starting from $Y_T = X_T$ via DDIM sampling [20].

3.2 Key Components

In line with prior studies [11, 23], we first adopt pseudo 3D convolution layers by inflating the 2D convolution layers in ResNet to handle video inputs. Specifically, we replace the 3×3 kernels with $1 \times 3 \times 3$ kernels. The three key components of ControlVideo are explained as follows.

Adding Controls. Inspired by the recent advancements in image editing [6, 7], a natural approach is to generate the edited videos starting from the noise inversion of source video X_0 via DDIM inversion. However, as depicted in Figure 5 (row 3), such method leads to less faithful edited videos. To address this issue, we introduce additional visual conditions such as Canny edge maps, HED boundaries, and depth maps for all frames to amplify the source video's guidance to enhance faithfulness. Specifically, we leverage ControlNet [13] to process visual conditions, which has been pretrained on the Stable Diffusion. Formally, let $h_u, h_c \in R^{N\times H\times W\times C}$ denote the embeddings of encoder of Stable Diffusion and ControlNet with the same layer respectively. These features are combined by summing them as $h = h_u + \lambda h_c$, where λ is the control scale, and are fed into the decoder of the Stable Diffusion via skip connection. As shown in Figure 5 (row 4), such guidance from the source video significantly improves faithfulness (e.g. preserving the trees in the background of the source video in the "a car" → "a red car" case).

Since different control types encompass varying degrees of information derived from the source video, ControlVideo allows users to flexibly adjust the balance between faithfully maintaining the content of the source video and enabling more extensive editing capabilities. For instance, HED boundary provides detailed boundary information of the source video and is suitable for refined control such as facial video editing, which needs extremely refine control to preserve the identity and emotional accuracy by maintaining the individual's unique facial characteristics (see Figure 6). On the other hand, pose control provides greater flexibility in modifying the subject and background, making it suitable for applications like motion transfer. Moreover, we can flexibly combine multiple controls by weighted summing different control features to utilize the advantage of different control types: $h = h_u + \sum_i \lambda_i h_c$, where λ_i is the control scale of i-th control (see Figure 1).

Key-frame Attention. The spatial self-attention mechanism utilized in T2I diffusion models updates each frame individually, leading to temporal inconsistent video outputs. To address this issue, drawing inspiration from previous works [24, 25] that utilizes key frames to propagate edits throughout videos as well as recent advancements in video editing [11], we transform the original spatial self-attention in both Stable Diffusion and ControlNet into key-frame attention, which aligns

[2] We assume that raw videos have already been mapped to latent space throughout the paper.

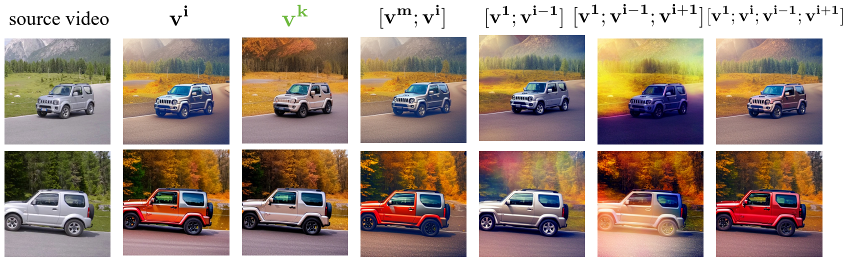

source video
$\mathbf{v^i}$
$\mathbf{v^k}$
$[\mathbf{v^m};\mathbf{v^i}]$
$[\mathbf{v^1};\mathbf{v^{i-1}}]$
$[\mathbf{v^1};\mathbf{v^{i-1}};\mathbf{v^{i+1}}]$
$[\mathbf{v^1};\mathbf{v^i};\mathbf{v^{i-1}};\mathbf{v^{i+1}}]$

Figure 3: Comparisons with different designs of key and value in self-attention. The green color marked our choice. See detail analysis in 3.3

all frames with a selected one. Formally, let v^i denote the embeddings of the i-th frame and let $k \in [1, N]$ represent the selected key frame. The key-frame attention is defined as follows:

$$Q = W^Q v^i, K = W^K v^k, V = W^V v^k,$$

where W^Q, W^K, W^V are the projection layer. The results show that there is no significant difference in using different key frame selections (see Appendix B.1). Therefore, we use the first frame as the key frame in this work. We employ the original spatial self-attention weights as initialization and finetune W^O by a systematic empirical study (see Sec. 3.3).

Temporal Attention. To improve faithfulness and temporal consistency of the edited video, we incorporate temporal attention modules as extra branches in the diffusion model. Based on the observation that different attention mechanisms consistently model the relationships between image features, we employ the original spatial self-attention weights as initialization. We add a zero convolutional layer [13] after each temporal attention module to preserve the output before fine-tuning. We incorporate temporal attention along with key-frame attention in the main UNet except middle block via a systematic empirical study(see Sec. 3.3).

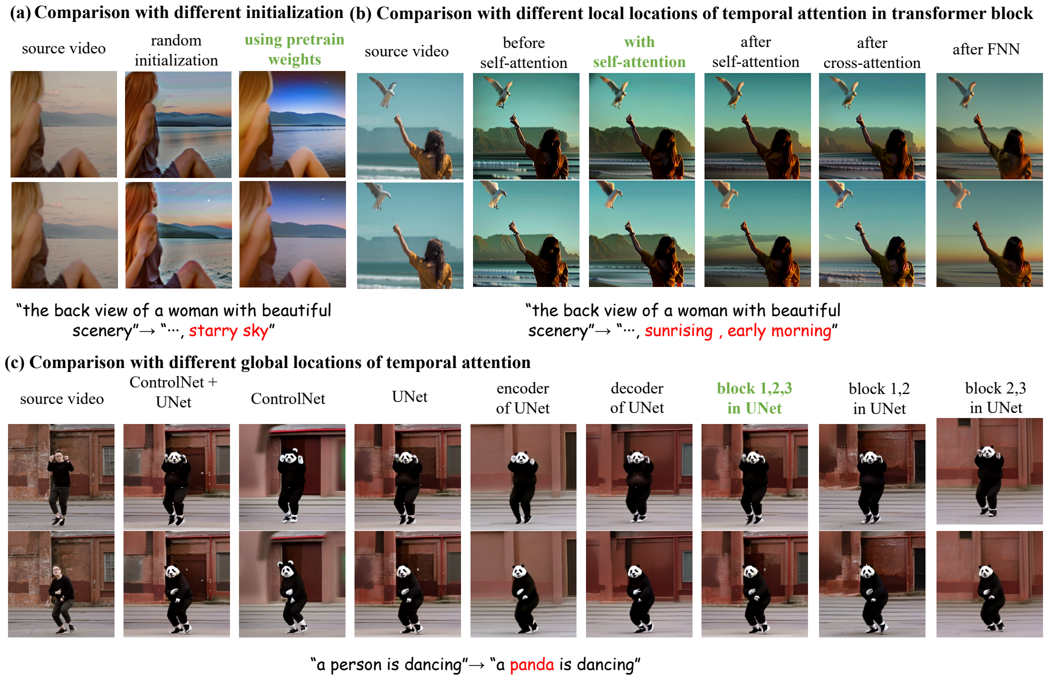
3.3 Analysis

In this section, we conduct a systematical empirical study by analyzing results on 20 video-text pair data and evaluate CLIP-temp, CLIP-text and SSIM (see Sec. 5) in Appendix B.

The Design of Key and Value in Self-Attention and Fine-tuned Parameters. Let $[;]$ denote the concat operation. We consider using these embeddings as key and value: (1) v^i: original spatial self-attention in T2I models. (2) v^k, which is our key-frame attention. We select four different key frames. (3) $[v^m; v^i]$ [9], where $m = Round(\frac{N}{2})$. (4) $[v^1; v^{i-1}]$ [11, 26]. (5) $[v^1; v^{i-1}; v^{i+1}]$, which includes bi-directional information. (6) $[v^1; v^i; v^{i-1}; v^{i+1}]$. As shown in Figure 3, key-frame attention shows the highest temporal consistency, implying that utilizing a key frame to propagate throughout videos is useful. There is no significant difference in different key frame selections (see Appendix B.1). In addition, adding the current frame features v^i shows less temporal inconsistency because the v^i contains different information between frames. For example, the color of the car turned red in $[v^m; v^i]$, $[v^1; v^i; v^{i-1}; v^{i+1}]$ following v^i (column 2). Further, we conduct following experiments to investigate finetune which parameters is more useful (see Appendix B.2): (1) W^Q. (2) W^O. (3) W^K, W^V. (4)W^Q, W^K, W^V. (5) W^Q, W^K, W^V, W^O. (6) add Lora [27] on W^Q, W^K, W^V, W^O. We find finetuning W^O shows good performance with less parameters.

The Way to Initialization and The Incorporation of Local and Global Positions for Introducing Temporal Attention. As shown in Figure 4(a), using pretrain spatial self-attention weights as initialization achieves better performance. Next, we explore following potential locations to incorporate temporal attention in transformer blocks: (1) before self-attention. (2) with self-attention. (3) after self-attention. (4)after cross-attention. (5) after FNN. As shown in Figure 4(b), before self-attention and with self-attention result in the best temporal consistency. This is because the input of these two locations is the same as spatial self-attention, which serves as the initial weight of temporal attention. Notably, with self-attention shows higher text alignment, making it our final choice. Moreover, we find the after FNN location yields the worst temporal consistency and should be avoided.

To investigate the optimal global location for adding temporal attention, we first conduct the following experiments: (1) ControlNet+UNet. (2) ControlNet. (3) UNet. (4) Encoder of UNet. (5) Decoder


(a) Comparison with different initialization

(b) Comparison with different local locations of temporal attention in transformer block

(c) Comparison with different global locations of temporal attention

Figure 4: Ablation studies of (a) the way to initialize and the incorporation of (b) local positions and (c) global positions for introducing temporal attention. The green color marked our choice. See detail analysis in 3.3.

of UNet. As shown in Figure 4(c), incorporating temporal attention only to the ControlNet fails to preserve the background and removing it does not decrease performance (all vs UNet). This suggests that ControlNet only extracts condition-related features (e.g. pose) and discards the other features (e.g. background), while U-Net, which is used for generation task, preserves all image information. As such, we ultimately choose to add temporal attention to UNet. Additionally, the decoder location achieves better performance than the encoder. This may be because, in U-Net, the decoder contains more information than the encoder by using skip connections to incorporate features from the encoder. Next, we investigate the location in UNet by following experiments: (1) all; (2) Block 1,2; (3) Block 1,3; (4) Block 2,3; (5) Block 1,2,3, which is UNet except middle block. As shown in Figure 4(c), the Block 1,2,3 shows similar performance with all while with less parameters, which is chosen as the final design.

Analyze the Role of Each Module and their Combination. As shown in Figure 5, adding controls provides additional guidance from the source video, thus improving faithfulness a lot. The key-frame attention improves temporal consistency a lot. The temporal attention improves faithfulness and temporal consistency. Combining all the modules achieves the best performance. Also, we can observe when controls contains more detail information of source video (e.g. HED boundary), adding control and key-frame attention can achieve relatively good results. When the controls contains less information (e.g. pose) or even non-existent, the temporal attention can greatly improve faithfulness.

4 Related Work

Diffusion Models for Text-driven Image Editing. Building upon the remarkable advances of text-to-image diffusion models [3, 4], numerous excellent methods have shown promising results in text-driven real image editing creation. In particular, several works such as Prompt-to-Prompt [6], Plug-and-Play [7] and Pix2pix-Zero [8] explore the attention control over the generated content and achieve SOTA results. Such methods usually start from the noise inversion and replace attention maps in the generation process with the attention maps from source prompt, which retrain the spatial layout of the source image. Despite significant advances, directly applying these image editing methods to video frames leads to temporal flickering.

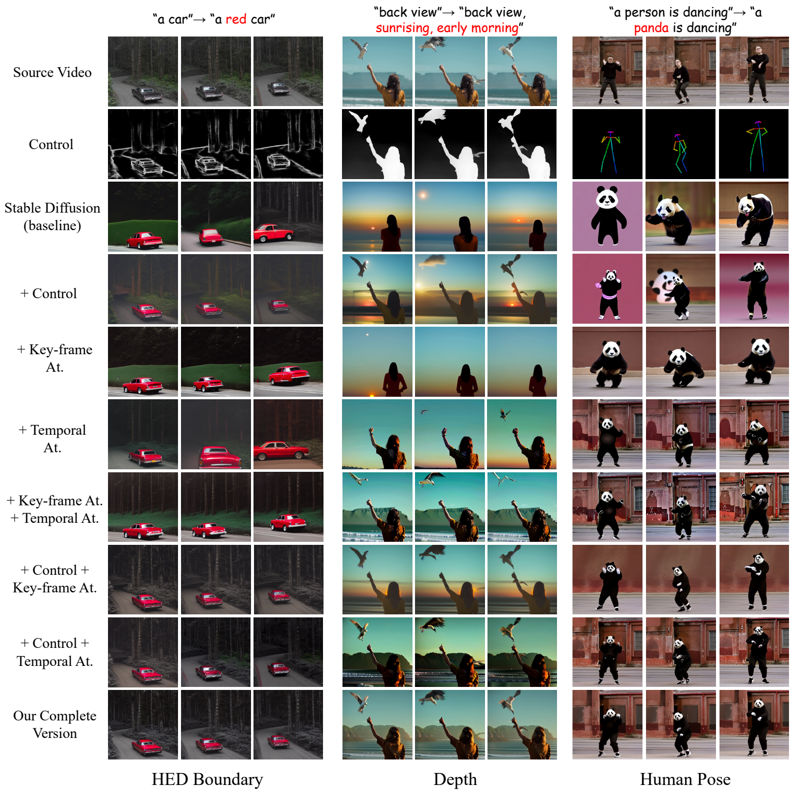

"a car"→ "a red car"
"back view"→ "back view, sunrising, early morning"
"a person is dancing"→ "a panda is dancing"
Source Video
Control
Stable Diffusion (baseline)
+ Control
+ Key-frame At.
+ Temporal At.
+ Key-frame At. + Temporal At.
+ Control + Key-frame At.
+ Control + Temporal At.
Our Complete Version
HED Boundary
Depth
Human Pose

Figure 5: Ablation studies of each module and their combination. At. represents attention. See detail analysis in 3.3.

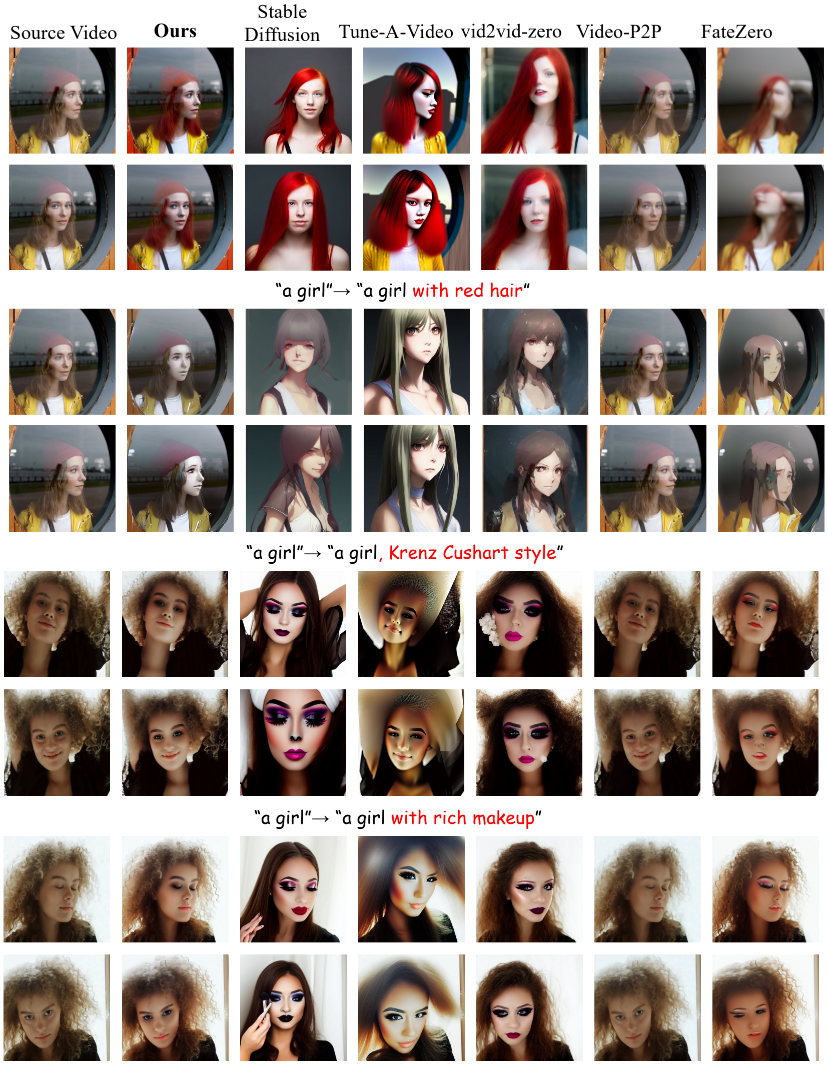
Diffusion Models for Text-driven Video Editing. Gen-1 [2] trains a video diffusion model on large-scale datasets and Dreamix [1] finetunes the pretrained Imagen Video [4] for text-driven video editing, achieving impressive performance. However, they require expensive computational resources. To overcome this, recent works build upon large-scale text-to-image diffusion models [3–5] and controllable image editing for this task on single text-video pair. In particular, Tune-A-Video [11] inflates the T2I diffusion model to text-to-video diffusion model and finetunes it on the source video and source prompt. Inspired by this, several works [10, 12, 26] propose to first optimize the null-text embedding for accurate video inversion and adopt attention map injection in the generation process, achieving superior performance. Despite the significant advances, empirical evidence suggests that they still struggle to faithfully and adequately control the output, while also preserving temporal consistency.

5 Experiments

Implementation Details. We collect 40 video-text pair data including DAVIS dataset [28] and other videos in the wild from website[3]. For fair comparisons, following previous works [9–12], we sample 8 frames with 512 × 512 resolution from each video. The Stable Diffusion 1.5 [3] and ControlNet 1.0 [13] are adopted in this work. We train the ControlVideo for 80, 300, 500 and 1500 iterations for canny edge maps, HED boundary, depth maps and pose respectively with learning rate 3×10^{-5}. The DDIM sampler [20] with 50 steps and 12 classifier-free guidance is used for inference. The control scale λ is set 1.

Evaluation Metrics. We evaluate the edited video from three aspects: temporal consistency, faithfulness and text alignment. Following previous works [9, 12], we report CLIP-temp for temporal

[3] https://www.pexels.com



"a girl"→ "a girl with red hair"

"a girl"→ "a girl, Krenz Cushart style"

"a girl"→ "a girl with rich makeup"

"a girl"→ "a girl with rich makeup"

Figure 6: Comparison with baselines. ControlVideo successfully preserves the identity and emotional accuracy by maintaining the individual's unique facial characteristics while others fail.

consistency and CLIP-text for text alignment. We also report SSIM [15] between each input-output pair for faithfulness. We perform user study to quantify text alignment, temporal consistency, faithfulness and the overall all aspects by pairwise comparisons between the baselines and our method. We provide more details about evaluation in Appendix A.

5.1 Text-driven Video Editing

Applications. We show the diverse video editing applications of ControlVideo including *attributes, style, background editing* and *replacing subjects* in Figure 1. For example, in the Figure 1 (a), ControlVideo change the color of car into red with others unchanged, suggesting the ability of local attributes editing. In Figure 1 (d), we show the dancing man has been changed into Michel Jacksons successfully. In the Figure 1 (e), with the guidance of canny edge map and pose, the dancing person is changed into a dancing panda with the highly faithful background. More qualitative results are available in Appendix D.

Comparisons. We compare ControlVideo with Stable Diffusion and the following state-of-the-art text-driven video editing methods: Tune-A-Video [11], vid2vid-zero [8] and FateZero [9], which are reproduced using public code. During inference, all baselines and ControlVideo use DDIM inversion

Table 1: Quantitative results. Text. and Temp. represent CLIP-text and CLIP-temp respectively. User study shows the preference rate of ControlVideo against baselines via human evaluation. A., T., F. and O. represent text alignment, temporal consistency, faithfulness and overall aspects.

	Metric			user study			
Method	Text.↑	Temp.↑	SSIM ↑	A. (%)↑	T.(%)↑	F.(%)↑	O.(%)↑
Stable Diffusion [3]	**0.278**	0.895	0.487	44	75	96	88
Tune-A-Video [11]	0.260	0.934	0.513	49	75	92	83
vid2vid-zero [8]	0.271	0.951	0.527	55	82	88	76
FateZero [9]	0.252	0.954	0.613	63	73	82	71
Ours	0.258	**0.961**	**0.647**	-	-	-	-

and DDIM sampling with the same setting. The quantitative and qualitative comparisons are reported in Table 1 and Figure 6. We also compare with Video-P2P [12]. We find it is more easy to reconstruct the video (see Figure 6) compared with other baselines and thus show quantitative results in the Appendix.

Extensive results demonstrate that ControlVideo substantially outperforms all these baselines in terms of faithfulness and temporal consistency with comparable text alignment performance. Notably, in the faithfulness evaluation in user study, we outperform the baseline by a significant margin (over 80%). As depicted in Figure 6, our method successfully preserves the identity and emotional accuracy by maintaining the individual's unique facial characteristics while others fail.

6 Conclusion

In this paper, we present ControlVideo to utilize T2I diffusion models for one-shot text-to-video editing, which introduces additional controls to preserve structure of source video, key-frame attention that aligns all frames with a key frame, and temporal attention using pre-trained spatial self-attention weights to improve faithfulness and temporal consistency. We demonstrate the effectiveness of ControlVideo by outperforming state-of-art text-driven video editing methods.

Table 2: Quantitative results about different choices of key and value in self-attention.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
v^i	0.263	0.905	0.635
$[v^m; v^i]$ [9]	0.260	0.939	0.642
$[v^1; v^{i-1}]$ [11, 26]	0.264	0.953	0.639
$[v^1; v^{i-1}; v^{i+1}]$	0.261	0.941	0.637
$[v^1; v^i; v^{i-1}; v^{i+1}]$	0.261	0.955	0.648
$v^k, k = 1$	0.263	0.954	0.655
$v^k, k = 3$	0.263	**0.961**	0.654
$v^k, k = 5$	0.261	0.958	**0.657**
$v^k, k = 7$	0.260	0.958	0.650

Table 3: Quantitative results about fine-tuned parameters of key-frame attention.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
W^Q	0.253	0.951	0.634
W^K, W^V	0.241	0.957	0.635
W^Q, W^K, W^V	0.241	0.958	0.635
W^Q, W^K, W^V, W^O	0.244	**0.961**	0.641
add Lora [27] on W^Q, W^K, W^V, W^O	0.237	0.957	0.630
W^O	0.246	0.960	**0.643**

A Implementation Details

We evaluate the human preference from text alignment, faithfulness, temporal consistency, and all three aspects combined. A total of 10 subjects participated in this section. Taking faithfulness as an example, given a source video, the participants are instructed to select which edited video is more faithful to the source video in the pairwise comparisons between the baselines and ControlVideo.

B Ablation Studies

In this section, we present the quantitative results of our ablation studies. Recognizing that the quantitative results may diverge from human evaluation, we ultimately prioritize human evaluation as our primary measure, while utilizing the quantitative results as supplementary references.

B.1 The Choices of Key and Value

As shown in Table 2, selecting a key-frame as key and value achieves high temporal consistency performance and there is no significant difference in using different key-frame selections.

B.2 The Fine-tuned Parameters.

) Based on the results presented in Table 3, we observe that fine-tuning W^O yields superior performance while utilizing fewer parameters, making it our ultimate selection.

B.3 The Incorporation of Local Positions for Introducing Temporal Attention

The quantitative results are shown in Table 4. We find before self-attention and with self-attention result in the best temporal consistency, and with self-attentinon has a slight higher text alignment score. In addition, we find after FNN location yields the worst temporal consistency and should be avoided. Moreover, using pretraining weights achieves better performance than random initialization.

Table 4: Quantitative results about different local locations for introducing temporal attention.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
before self-attention	0.225	0.917	0.589
after self-attention	0.241	0.909	**0.629**
after cross-attention	0.241	0.902	0.616
after FNN	0.251	0.908	0.630
with self-attention (random initialization)	0.230	0.909	0.585
with self-attention	0.238	**0.920**	0.621

Table 5: Quantitative results about different global locations for introducing temporal attention. These quantitative results diverge from human evaluation and we ultimately prioritize human evaluation as our primary measure and list them as references. See details in B.4.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
all	0.235	0.955	0.621
controlnet	0.243	0.959	0.643
unet	0.237	0.955	0.629
encoder	0.241	0.956	0.669
decoder	0.239	0.955	0.638
Block 1,2	0.239	0.957	0.632
Block 1,3	0.237	0.954	0.640
Block 2,3	0.242	0.956	0.656
Block 1,2,3	0.236	0.958	0.630

B.4 The Incorporation of Global Positions for Introducing Temporal Attention

The quantitative results are shown in Table 5, which diverges from human evaluation. We ultimately prioritize human evaluation as our primary measure and list them as references. From human evaluation aspects (see Figure 4 in the main text), we find adding temporal attention on Block 1,2,3 in UNet achieve good performance.

C Comparison with Baselines

In this section, we show the quantitative results of Video-P2P in Table 6.

D More Qualitative Results

In this section, we show more qualitative results with different control types in Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11. We can observe that ControlVideo can generate temporal consistent and faithful videos.

Table 6: Quantitative results. Text. and Temp. represent CLIP-text and CLIP-temp respectively. User study shows the preference rate of ControlVideo against baselines via human evaluation. A., T., F. and O. represent text alignment, temporal consistency, faithfulness and overall aspects.

	Metric			user study			
Method	Text.↑	Temp.↑	SSIM ↑	A. (%)↑	T.(%)↑	F.(%)↑	O.(%)↑
Stable Diffusion [3]	**0.278**	0.895	0.487	44	75	96	88
Video-P2P [12]	0.189	0.958	0.809	92	62	68	75
Ours	0.258	**0.961**	**0.647**	-	-	-	-

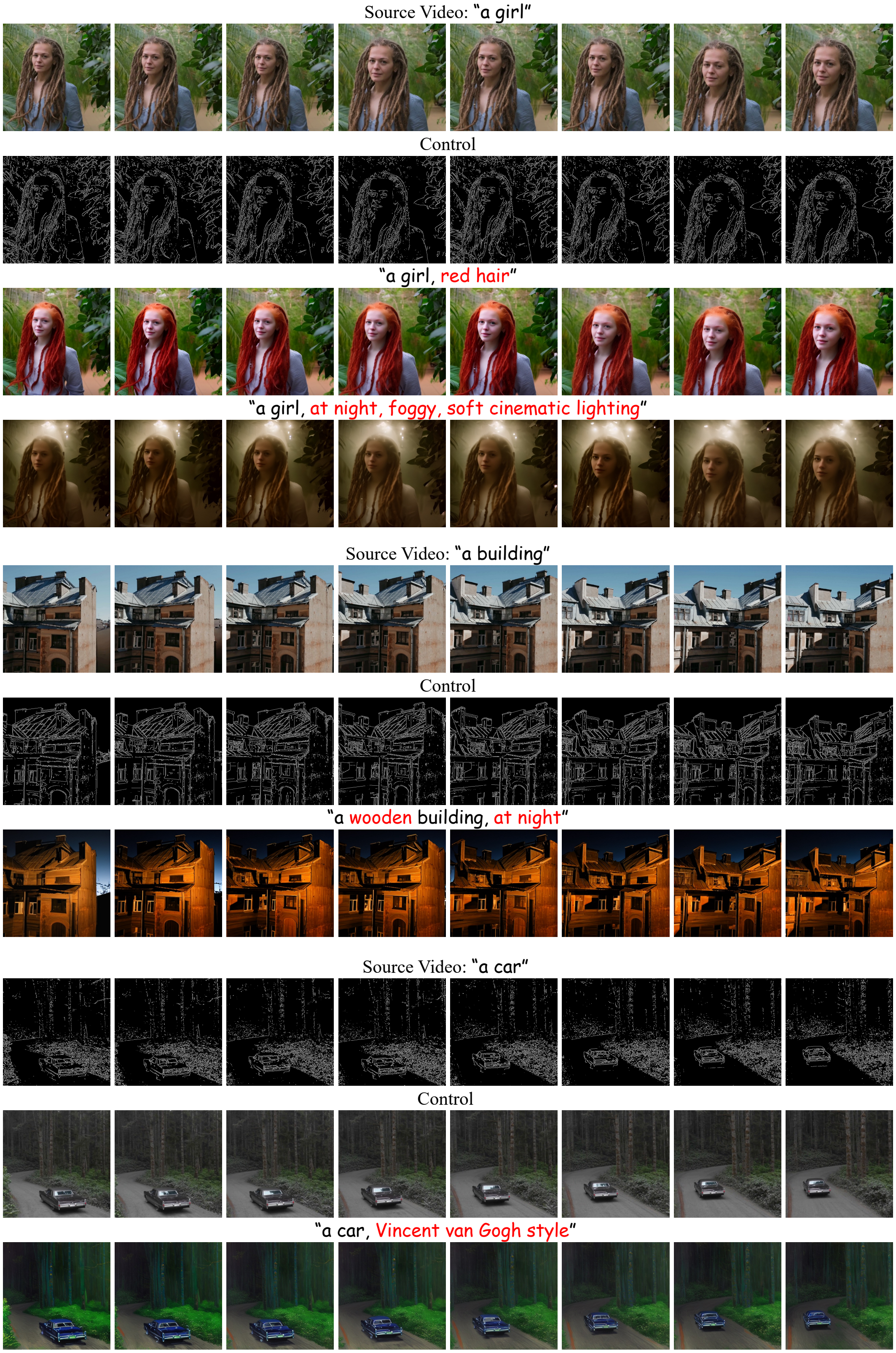

Source Video: "a girl"
Control
"a girl, red hair"
"a girl, at night, foggy, soft cinematic lighting"
Source Video: "a building"
Control
"a wooden building, at night"
Source Video: "a car"
Control
"a car, Vincent van Gogh style"

Figure 7: More qualitative results of ControlVideo with Canny edge maps.

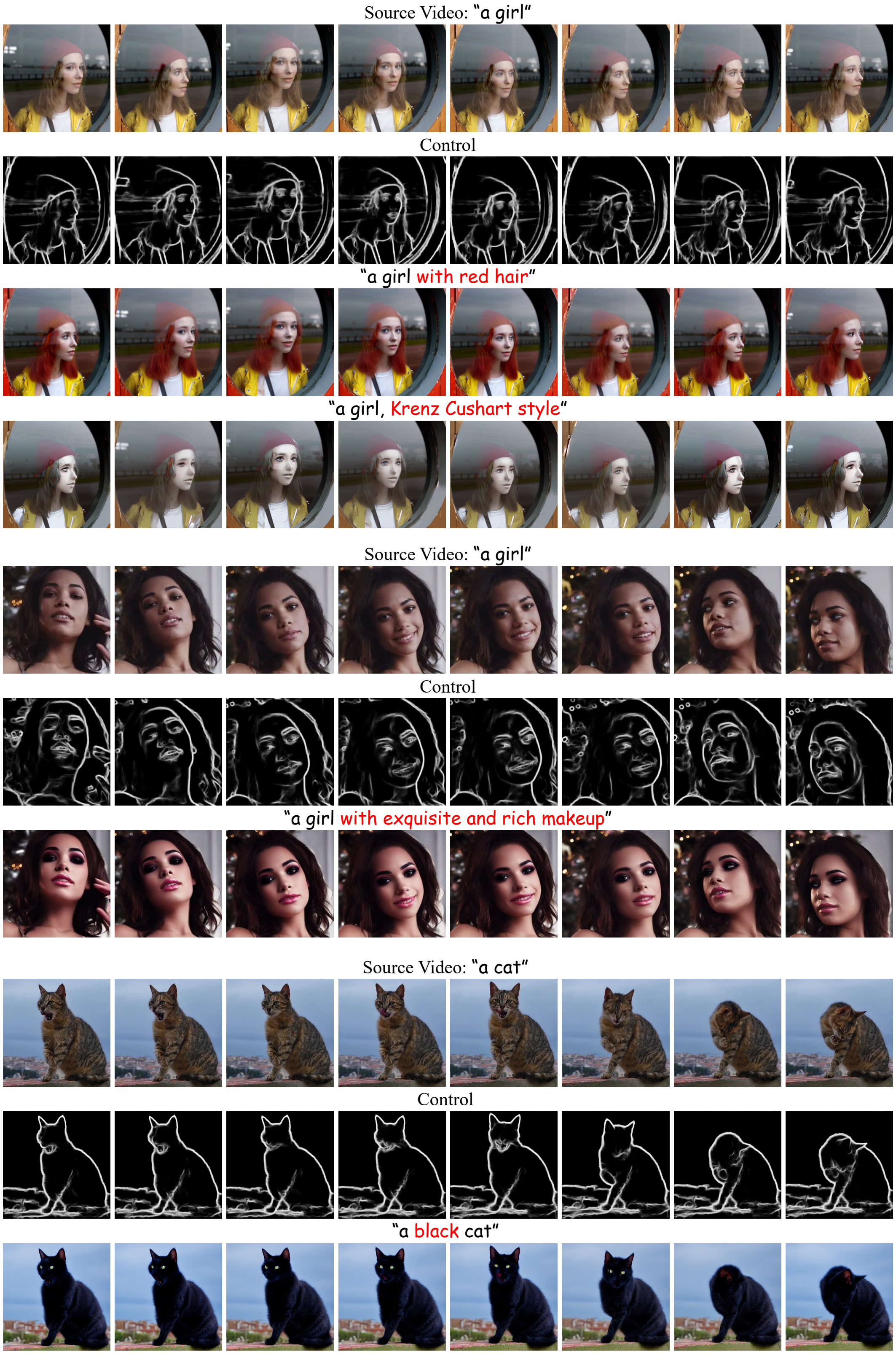


Figure 8: More qualitative results of ControlVideo with HED boundary.

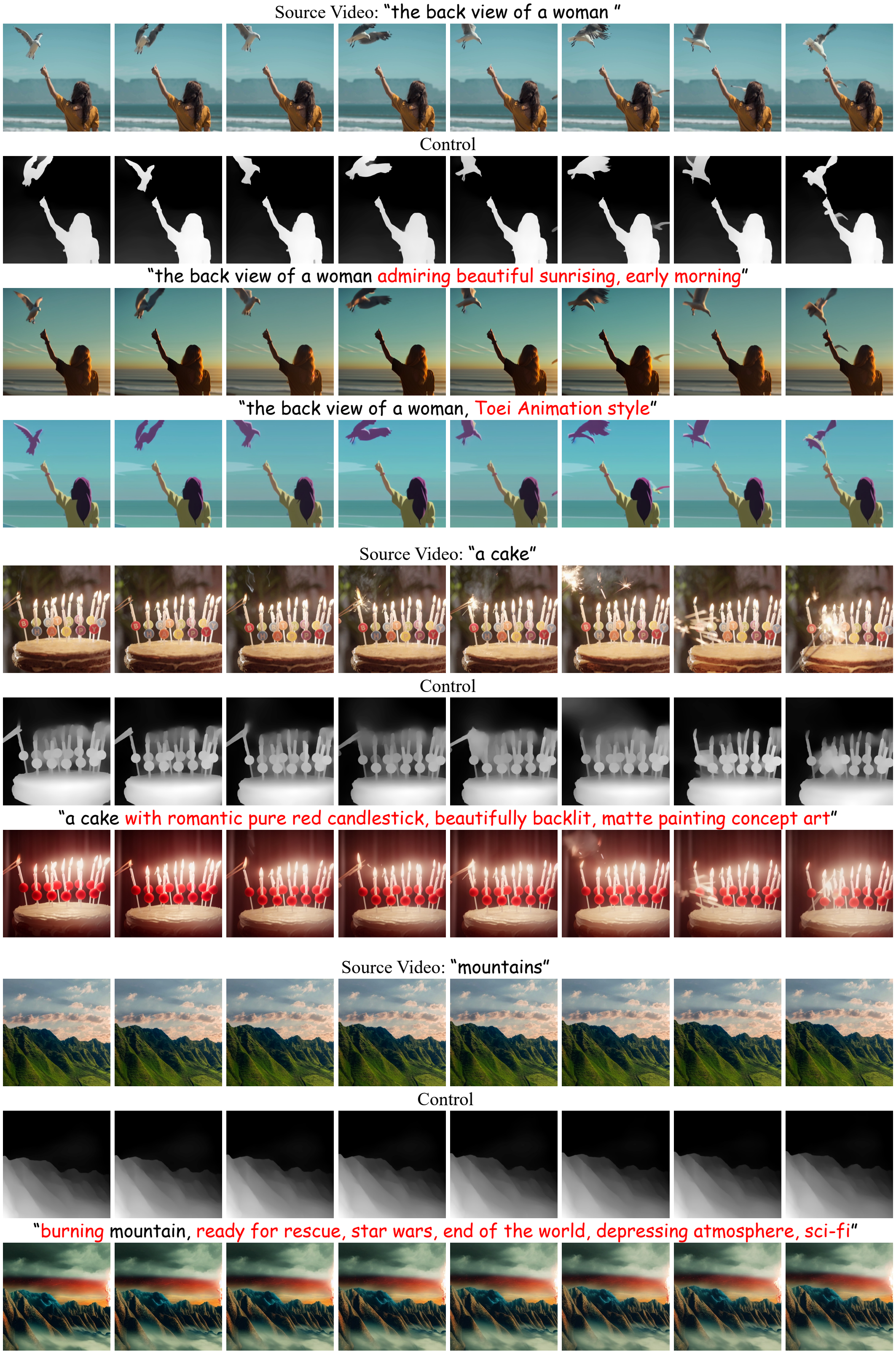

Source Video: "the back view of a woman "
Control
"the back view of a woman admiring beautiful sunrising, early morning"
"the back view of a woman, Toei Animation style"
Source Video: "a cake"
Control
"a cake with romantic pure red candlestick, beautifully backlit, matte painting concept art"
Source Video: "mountains"
Control
"burning mountain, ready for rescue, star wars, end of the world, depressing atmosphere, sci-fi"

Figure 9: More qualitative results of ControlVideo with depth maps.

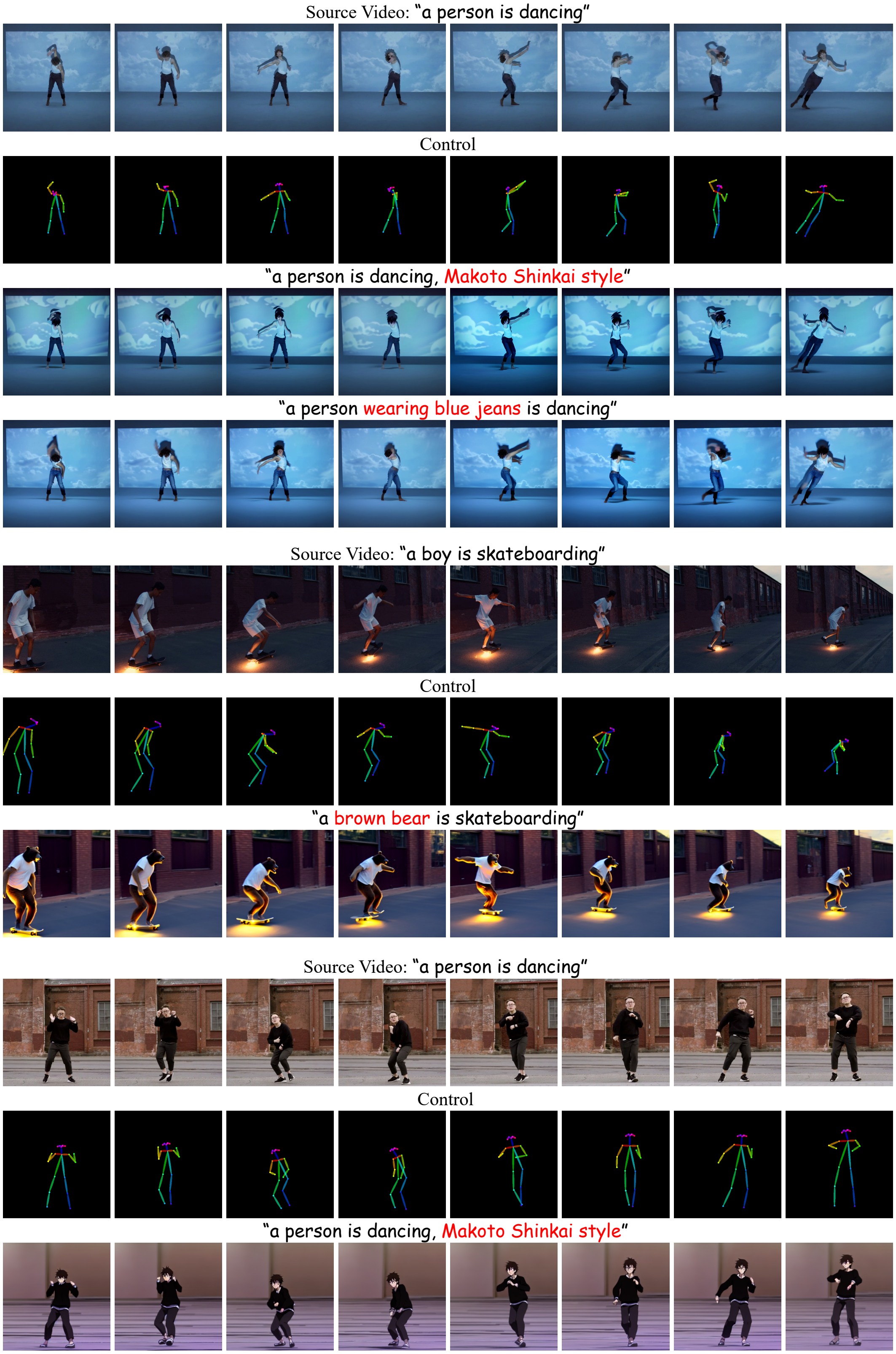

Source Video: "a person is dancing"
Control
"a person is dancing, Makoto Shinkai style"
"a person wearing blue jeans is dancing"
Source Video: "a boy is skateboarding"
Control
"a brown bear is skateboarding"
Source Video: "a person is dancing"
Control
"a person is dancing, Makoto Shinkai style"

Figure 10: More qualitative results of ControlVideo with pose.

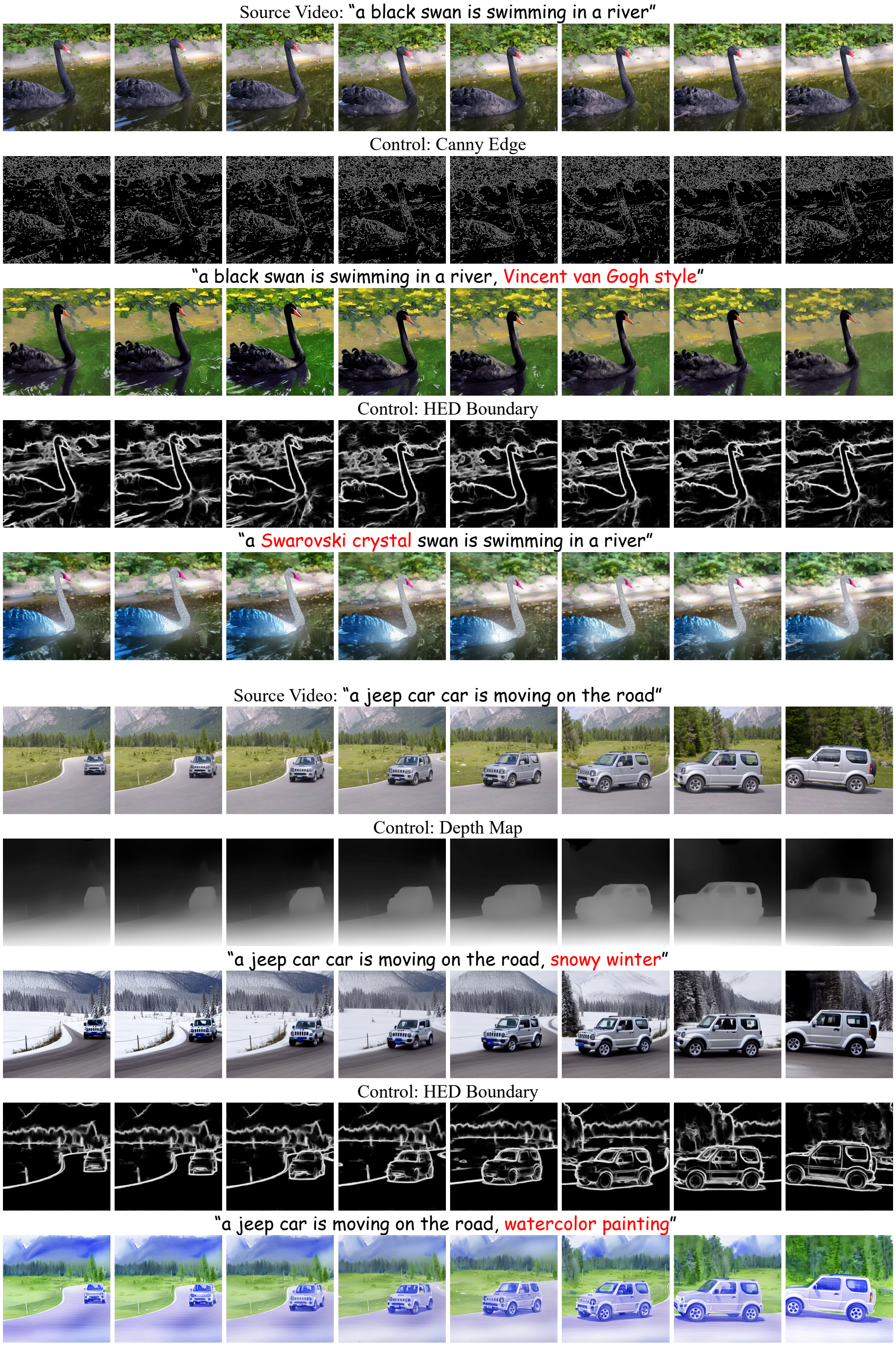

Source Video: "a black swan is swimming in a river"
Control: Canny Edge
"a black swan is swimming in a river, Vincent van Gogh style"
Control: HED Boundary
"a Swarovski crystal swan is swimming in a river"
Source Video: "a jeep car car is moving on the road"
Control: Depth Map
"a jeep car car is moving on the road, snowy winter"
Control: HED Boundary
"a jeep car is moving on the road, watercolor painting"

Figure 11: More qualitative results of ControlVideo on DAVIS dataset.

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*, 2023.

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

[8] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[9] 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.

[10] 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.

[11] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[12] 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.

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

[14] 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.

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[16] 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.

[17] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2021.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[21] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

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

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

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

[25] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020.

[26] 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.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[28] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.