Diffutoon: High-Resolution Editable Toon Shading via Diffusion Models

Zhongjie Duan[1] , **Chengyu Wang**[2] , **Cen Chen**[1*] , **Weining Qian**[1] and **Jun Huang**[2]
[1]East China Normal University, Shanghai, China
[2]Alibaba Group, Hangzhou, China
zjduan@stu.ecnu.edu.cn, {chengyu.wcy, huangjun.hj}@alibaba-inc.com,
{cenchen, wnqian}@dase.ecnu.edu.cn

Abstract

Toon shading is a type of non-photorealistic rendering task in animation. Its primary purpose is to render objects with a flat and stylized appearance. As diffusion models have ascended to the forefront of image synthesis, this paper delves into an innovative form of toon shading based on diffusion models, aiming to directly render photorealistic videos into anime styles. In video stylization, existing methods encounter persistent challenges, notably in maintaining consistency and achieving high visual quality. In this paper, we model the toon shading problem as four subproblems, i.e., stylization, consistency enhancement, structure guidance, and colorization. To address the challenges in video stylization, we propose an effective toon shading approach called Diffutoon. Diffutoon is capable of rendering remarkably detailed, high-resolution, and extended-duration videos in anime style. It can also edit the video content according to input prompts via an additional branch. The efficacy of Diffutoon is evaluated through quantitive metrics and human evaluation. Notably, Diffutoon surpasses both open-source and closed-source baseline approaches in our experiments. Our work is accompanied by the release of both the source code and example videos on Github.

1 Introduction

Toon shading [Barla *et al.*, 2006] plays a crucial role in the animation industry, aiming to render 3D computer-generated graphics in a stylized, flat manner. These techniques are extensively applied across diverse domains, including video game development and animation production [Hudon *et al.*, 2018]. As diffusion models [Sohl-Dickstein *et al.*, 2015] achieve impressive performance in image synthesis, we recognize their potential in video stylization. In this paper, we explore a novel approach to toon shading, aimed at directly transforming photorealistic videos into an animated style.

In recent years, Stable Diffusion [Rombach *et al.*, 2022], a diffusion model pre-trained on text-image datasets [Schuhmann *et al.*, 2022], has emerged as a powerful foundation for text-to-image synthesis. Within open-source communities, a wealth of fine-tuned models based on Stable Diffusion are available. Nevertheless, extending diffusion models to videos presents multiple challenges [Xing *et al.*, 2023]. Firstly, there is a significant issue with controllability. Applying diffusion models to videos often makes it challenging to preserve essential information from the original video, such as structure and lighting. Secondly, achieving consistency across frames is crucial, as processing frames independently can lead to undesirable flickering. Lastly, maintaining high visual quality is a concern. Although video platforms commonly support resolutions up to 1080P and even 4K, most diffusion models struggle to process videos at these high resolutions.

Prior studies have attempted to address these challenges. In controllable image synthesis, adapter-type control modules [Zhang *et al.*, 2023; Mou *et al.*, 2023] have demonstrated their capability for precise control. However, their application is confined to processing individual images, rendering them ineffective for video processing. To enhance video consistency, existing approaches are generally divided into two categories: training-free and training-based methods. Training-free methods [Yang *et al.*, 2023; Ceylan *et al.*, 2023] align content between frames through specific mechanisms, eliminating the need for training. Despite this advantage, their effectiveness is somewhat constrained. Conversely, training-based methods [Esser *et al.*, 2023; Guo *et al.*, 2023] tend to yield superior outcomes. Yet, the significant computational resources required pose a formidable challenge for training diffusion models on extensive video datasets. As a result, most video diffusion models are limited to managing a maximum of 32 frames. In the pursuit of improved visual quality, super-resolution [Wang *et al.*, 2021] holds the promise of enhancing video resolution. Nevertheless, these techniques might lead to additional complications, such as the loss of detailed information due to oversmoothing [Li *et al.*, 2022].

In this paper, we propose Diffutoon, a novel video processing method specifically designed for toon shading. We divide the toon shading problem into four subproblems: stylization, consistency enhancement, structure guidance, and colorization. For each subproblem, we provide a specific solution. Our proposed framework consists of a main toon shading pipeline and an editing branch. In the main toon shading pipeline, we construct a multi-module denoising model based

[*] Corresponding author.

on an anime-style diffusion model. ControlNet [Zhang *et al.*, 2023] and AnimateDiff [Guo *et al.*, 2023] are utilized in the denoising model to address controllability and consistency issues. To enable the generation of ultra-high-resolution content in long videos, we depart from the conventional frame-by-frame generation paradigm. Instead, we adopt a sliding window approach to iteratively update the latent embedding of each frame. Additionally, our method offers the capability to edit videos through the editing branch, which provides editing signals for the main toon shading pipeline. To improve the efficiency, we incorporate flash attention [Dao *et al.*, 2022] into the attention mechanisms, effectively mitigating excessive GPU memory usage. Remarkably, our approach can directly handle resolutions of up to 1536×1536. In our experiments, we first evaluate Diffutoon in the toon shading task, and then we evaluate the capability of editing some content according to given prompts. Comparative analyses are conducted with both open-source and closed-source approaches. Quantitative assessments and human evaluations consistently demonstrate the advantages of our approach over other methods. We have released the source code[1] and example videos[2]. The contribution of this paper includes:

- We introduce an innovative form of toon shading, aiming to release the potential of generative diffusion models in the field of non-photorealistic rendering.
- We propose an effective toon shading approach based on diffusion models, making it possible to transform photorealistic videos into an anime style and edit the content according to given prompts if required.
- Our implementation presents a robust framework for deploying diffusion models in video processing. This framework can achieve very high resolution and is capable of processing long videos.

2 Related Work

2.1 Stable Diffusion

Stable Diffusion [Rombach *et al.*, 2022] has become a widely adopted foundational model within both academic and open-source communities. Its architecture integrates a text encoder [Radford *et al.*, 2021], a UNet [Ronneberger *et al.*, 2015], and a VAE [Kingma and Welling, 2013]. To effectively adapt Stable Diffusion models for toon shading applications, it is crucial to fine-tune an anime-style image generation model specifically for image-to-image transformation tasks. By utilizing advanced training approaches such as LoRA [Hu *et al.*, 2021], Textual Inversion [Gal *et al.*, 2022], DreamBooth [Ruiz *et al.*, 2023], among others, we can easily customize a model to our needs. Furthermore, employing prompt engineering techniques [Cao *et al.*, 2023; Wang *et al.*, 2024] facilitates the refinement of text prompts, enhancing the ability to produce images of high aesthetics.

2.2 Fast Sampling of Diffusion Models

Diffusion models typically necessitate multiple iterative steps to generate clear images, resulting in slower generation speed than GANs [Goodfellow *et al.*, 2014]. The computational demands of video processing are amplified, as each frame requires processing. To confront this challenge, certain studies [Song *et al.*, 2020; Lu *et al.*, 2022; Duan *et al.*, 2023a] have introduced schedulers that manage the generation process and allow for the production of clear images with fewer iterations. While advancements in high-resolution image generation have been made, allowing for the scaling of low-resolution models to high-resolution applications [Jin *et al.*, 2023; He *et al.*, 2023], the computational burden associated with attention layers continues to be a substantial hurdle. Efficient attention mechanisms like flash attention [Dao *et al.*, 2022] have been pivotal in mitigating these concerns by reducing the memory and time requirements.

2.3 Controllable Image Synthesis

To augment the controllability of the outcomes, recent research such as ControlNet [Zhang *et al.*, 2023] and T2I-Adapter [Mou *et al.*, 2023] focuses on incorporating control signals into the generative pipeline. By attaching control modules, configured as adapters, to the UNet, a robust image-to-image conversion pathway can be established, enabling maintained or selective preservation of information from the source image. These innovations in controllable image-to-image generation methodologies have catalyzed research endeavors in the video-to-video domain. An illustrative example is Gen-1 [Esser *et al.*, 2023], which segregates video data into structural and content elements, utilizing depth estimation [Ranftl *et al.*, 2020] for structural integrity in the stylization process. In our work, we draw on these controlling techniques and integrate them within our proposed framework.

2.4 Temporal Diffusion Models

The foremost challenge when adapting diffusion models for video processing lies in achieving consistency across frames. Traditional methodologies that treat each frame independently often result in undesirable flickering. Some research efforts [Khachatryan *et al.*, 2023; Yang *et al.*, 2023] have introduced mechanisms such as cross-frame attention. This technique ensures alignment of content across adjacent frames without necessitating additional training. Furthermore, other studies [Blattmann *et al.*, 2023; Esser *et al.*, 2023; Guo *et al.*, 2023] have addressed the challenge of consistency by developing trainable modules specifically designed for video datasets. Notably, AnimateDiff [Guo *et al.*, 2023] has emerged as a particularly popular solution within open-source communities. In our approach, we incorporate motion modules to significantly enhance the coherence of the video.

2.5 Post-Processing Methods

The task of training diffusion models on lengthy videos is notably demanding. To mitigate issues related to long-term video consistency, several video post-processing techniques can be utilized. Among these, CoDeF [Ouyang *et al.*, 2023], FastBlend [Duan *et al.*, 2023b], and various video deflickering algorithms [Lei *et al.*, 2023] have shown promise in addressing the challenges of processing extended sequences. However, these methods sometimes encounter difficulties in

[1] https://github.com/Artiprocher/DiffSynth-Studio

[2] https://ecnu-cilab.github.io/DiffutoonProjectPage/

scenes characterized by rapid movement or substantial motion. The strategy proposed in this paper draws inspiration from such techniques, aiming to enhance the consistency of videos over longer duration.

3 Methodology

The architecture of Diffutoon is depicted in Figure 1. This setup encompasses a main toon shading pipeline alongside an editing branch. The pipeline is tasked with rendering the input video in an anime style, while the editing branch is specifically developed to produce an edited color video. This video is then seamlessly integrated into the main toon shading pipeline to facilitate anime video editing.

3.1 Toon Shading

We decompose the toon shading task into four subtasks: stylization, consistency enhancement, structure guidance, and colorization. Four models are employed to address each subtask, respectively:

- **Stylization**: For anime stylization, we utilize a personalized Stable Diffusion model [Rombach *et al.*, 2022][3]. In theory, our method is compatible with any open-source diffusion model that shares this architectural framework.
- **Consistency enhancement**: Several motion modules, based on AnimateDiff [Guo *et al.*, 2023], are integrated into the UNet to bolster temporal consistency in content.
- **Structure guidance**: Outline information is extracted from the input video to maintain structural integrity during generation using a ControlNet model [Zhang *et al.*, 2023]. This contrasts with methods like those of Esser et al. [Esser *et al.*, 2023], which rely on depth estimation; outlines are more congruent with flat-style animation rendering.
- **Colorization**: A second ControlNet model trained for super-resolution is used for colorization, improving quality even from low-resolution input videos. This model operates directly on the main toon shading pipeline and uses the edited color video as input when the editing branch is active.

As illustrated in the top part of Figure 1, the main toon shading pipeline involves several key steps. Given an input video containing N frames $\{\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^N\}$, we first generate an outline video and a color video. The outline video $\{\boldsymbol{o}^1, \boldsymbol{o}^2, \cdots, \boldsymbol{o}^N\}$ contains the structural information extracted from the input video, and the color video $\{\boldsymbol{c}^1, \boldsymbol{c}^2, \cdots, \boldsymbol{c}^N\}$ is the input video when the editing branch is disabled. Subsequently, the two videos serve as inputs to their respective ControlNet models, which in turn produce conditioning signals for the UNet. Simultaneously, the motion modules generate temporal signals. These four models constitute a large denoising model $\mathcal{E}$, employed iteratively to synthesize a visually consistent video.

In the denoising process, initially, the latent embedding of each frame is sampled from a Gaussian distribution:

$$\boldsymbol{x}_T = \{\boldsymbol{x}_T^i\}_{i=1}^N \sim \mathcal{N}(\boldsymbol{O}, \boldsymbol{I}), \tag{1}$$

where T is the number of iterative steps and each embedding is independent identically distributed. At each denoising step, we use classifier-free guidance [Ho and Salimans, 2021] to build a textual guidance mechanism, which consists of a positive side and a negative side. On the positive side, we use some empirical keywords (e.g., "best quality", "perfect anime illustration") as prompt τ for better aesthetics. Note that the motion modules are trained within 32 consecutive frames, we can only use the denoising model $\mathcal{E}$ in a sliding window with a size no larger than 32. The sliding windows with size d and stride s are:

$$\mathcal{W}(d, s) = \{[i, i + d - 1] : 1 \le i \le N, i \equiv 1 (\text{mod } s)\}, \tag{2}$$

where $s < d$ for a smooth transition between different sliding windows. In a sliding window $[l, r]$, the model output on the positive side is:

$$\{\boldsymbol{e}_{t,+}(l, i, r)\}_{i=l}^r = \mathcal{E}\Big(\{\boldsymbol{x}_t^i\}_{i=l}^r, \{\boldsymbol{o}_t^i\}_{i=l}^r, \{\boldsymbol{c}_t^i\}_{i=l}^r, t, \tau\Big). \tag{3}$$

The latent embeddings are initially stored in RAM and will be moved to GPU memory when the sliding window covers them. We adopt a linear combination of overlapping segments from different sliding windows:

$$\overline{\boldsymbol{e}}_{t,+}(i) = \sum_{(l,r)\in\mathcal{W}(d,s)} \frac{w(l, i, r)}{\sum_{(l',r')\in\mathcal{W}(d,s)} w(l', i, r')} \boldsymbol{e}_{t,+}(l, i, r). \tag{4}$$

The weight $w(l, i, r)$ is formulated as follows:

$$w(l, i, r) = \begin{cases} 1 + \epsilon - \left|i - \frac{l+r}{2}\right| / \frac{r-l}{2}, & \text{if } l \le i \le r, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\epsilon = 10^{-2}$ is the minimum weight of tail frames. This allows the information from each frame to be shared with other frames throughout the generation process. This mechanism implicitly implements a large size of sliding window, enhancing the long-term consistency of generated content. To avoid disintegrated parts on faces and hands, we employ a textual inversion τ' [Gal *et al.*, 2022] on the negative side[4], which involves 10 token embeddings to be processed by the text encoder. By replacing τ with τ' in (3) and (4), we can obtain the estimated noise on the negative side $\overline{\boldsymbol{e}}_{t,-}(i)$. Then, the guided estimated noise is:

$$\overline{\boldsymbol{e}}_t(i) = g \cdot \overline{\boldsymbol{e}}_{t,+}(i) + (1 - g) \cdot \overline{\boldsymbol{e}}_{t,-}(i). \tag{6}$$

The classifier-free guidance scale g is set to 7 by default. Based on empirical evidence, we skip the final attention layer of the text encoder, which can improve the visual quality slightly. The overall estimated noise of the whole video is:

$$\overline{\boldsymbol{e}}_t = \big(\overline{\boldsymbol{e}}_t(0), \overline{\boldsymbol{e}}_t(1), \cdots, \overline{\boldsymbol{e}}_t(n)\big) \in \mathbb{R}^{N\times H\times W\times C}. \tag{7}$$

After that, we utilize the DDIM [Song *et al.*, 2020] scheduler to control the generation process:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\boldsymbol{x}_t - \sqrt{1 - \alpha_t}\overline{\boldsymbol{e}}_t}{\sqrt{\alpha_t}}\right) + \sqrt{1 - \alpha_{t-1}}\overline{\boldsymbol{e}}_t, \tag{8}$$

[3] https://civitai.com/models/34553/aingdiffusion

[4] https://civitai.com/models/11772

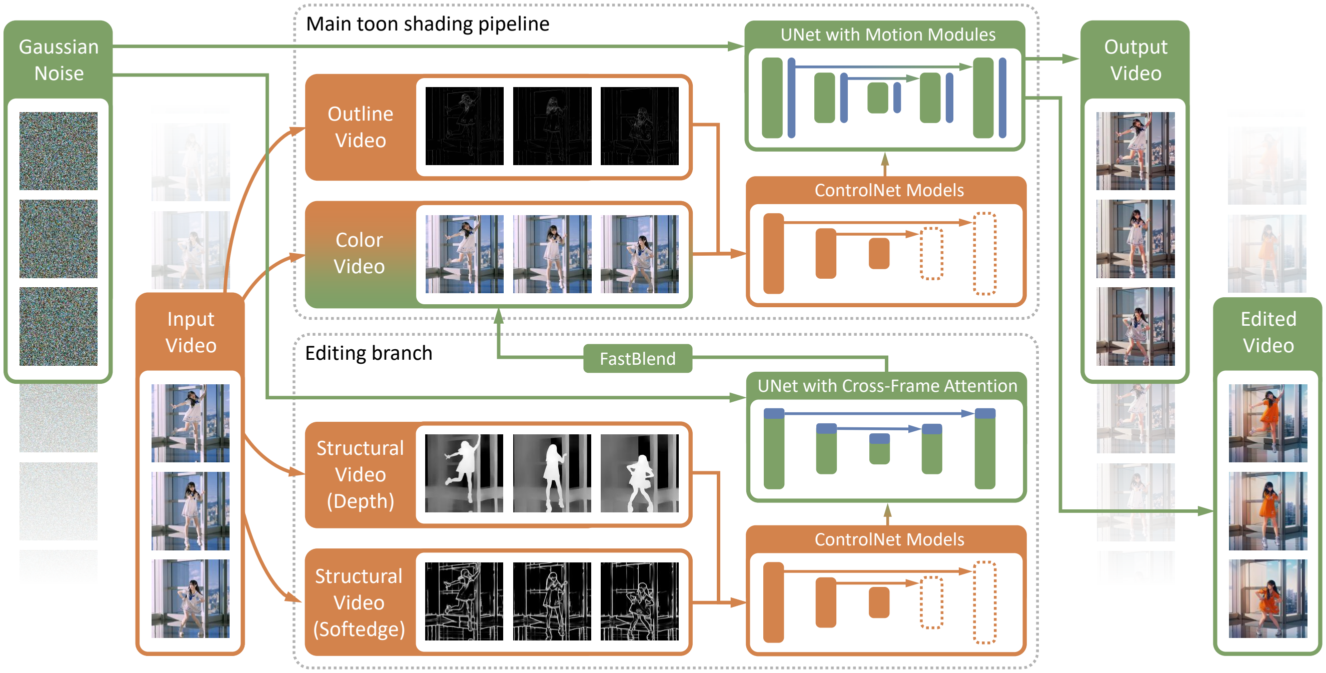

Figure 1: The overall architecture of Diffutoon, illustrated with the main toon shading pipeline at the top and the editing branch at the bottom. The editing branch is designed to generate editing signals in the format of a color video for integration into the main toon shading pipeline.

where α_t is the hyper-parameter that determines how much noise it contains in step t. We follow the implementation of DDIM in AnimateDiff [Guo *et al.*, 2023]. Despite the findings from recent studies suggesting that alternative schedulers, such as DPM-Solver [Lu *et al.*, 2022] and OLSS [Duan *et al.*, 2023a], can achieve superior visual quality within a specified number of steps, we decide to employ such a straightforward scheduler due to memory constraints. This decision is driven by the fact that these alternative schedulers need to store all latent tensors throughout the generation process, posing challenges for processing long videos. Additionally, we set the number of denoising steps T to only 10 for faster generation without compromising the resulting quality.

3.2 Adding Editing Signals to Toon Shading

In the main toon shading pipeline, the information in the input video is decomposed into outlines and colors. Practically, content can be edited by altering the outline or color video. However, given the scarcity of reliable video editing methods for structural information, our efforts primarily target color information editing. The ControlNet model employed for color video processing aids the UNet in generating high-quality videos, demonstrating notable fault tolerance to imprecision in video editing. Inspired by this insight, we have devised a specialized branch for video editing. An editing branch is incorporated to create text-guided editing signals, formatted as a color video. Its architecture is depicted in the lower section of Figure 1.

Analogous to the main toon shading pipeline, the synthesis of the editing signal is segmented into four subtasks:

- **Stylization**: The same Stable Diffusion model from the main pipeline is employed.
- **Consistency enhancement**: To boost consistency, we utilize cross-frame attention and FastBlend [Duan *et al.*, 2023b]. We eschew the motion modules [Guo *et al.*, 2023] that can degrade visual quality through their reliance on a modified DDIM scheduler. By employing the DDIM scheduler consistent with its training and supplementing it with cross-frame attention and FastBlend, we achieve reliable consistency. Notably, cross-frame attention is a proven effective technique [Yang *et al.*, 2023; Ceylan *et al.*, 2023], and FastBlend offers a model-free deflickering solution for post-processing.
- **Structure guidance**: Depth estimation [Ranftl *et al.*, 2020] and softedge [Xie and Tu, 2015] techniques represent structural information, guided by two ControlNet models. These configurations have been empirically validated by prior work [Esser *et al.*, 2023; Duan *et al.*, 2023c] to preserve structural integrity, particularly in extensive video editing scenarios.
- **Colorization**: Color is influenced by the input prompts. Occasionally, if classifier-free guidance falters, resulting in incorrect colors across frames, FastBlend corrects these inconsistencies by capitalizing on neighboring frame information.

The remaining elements of the editing branch align with those in the main toon shading pipeline. Although synthesized color videos from this branch may present some blurring, they preserve substantial coherence, effectively guiding the main pipeline to render a high-quality video.

3.3 Synthesizing High-Resolution Long Videos

We have developed Diffutoon within the DiffSynth framework [Duan *et al.*, 2023c], enabling the processing of long videos in latent space. To minimize GPU memory requirements and enhance computational efficiency, we incorporate flash attention [Dao *et al.*, 2022] across all attention-reliant components, including the text encoder, UNet, VAE, ControlNet models, and motion modules. This adaptation of memory-efficient attention facilitates the direct synthesis of exceptionally high-resolution videos. Furthermore, we employ a sliding window mechanism, empowering our pipeline to generate videos that are not only significantly detailed and high-resolution but also of extended duration.

4 Experiments

Our primary focus centers on high-resolution videos with rapid and substantial motion. We evaluate the efficacy of Diffutoon using 10 videos sourced from a video platform[5]. The total number of rendered frames reaches 5600, surpassing that of prior studies [Yang *et al.*, 2023; Jamriška *et al.*, 2019]. In our experiments, we achieve a video resolution of up to 1536×1536, resulting in visually impressive frames. The detailed settings of parameters are presented in Table 1.

Components	Parameter	Value
Main toon shading pipeline	frame height	1536
	frame width	1536
	classifier-free guidance scale	7
	inference steps	10
	sliding window size	16
	sliding window stride	8
	conditioning scale (outline)	0.5
	conditioning scale (color)	0.5
Editing branch	frame height	512
	frame width	512
	classifier-free guidance scale	7
	inference steps	20
	sliding window size	8
	sliding window stride	4
	conditioning scale (depth)	0.5
	conditioning scale (softedge)	0.5

Table 1: Parameter settings in the experiments.

4.1 Comparison with Baseline Methods

The evaluation encompasses two distinct tasks: toon shading, where we solely utilize the main toon shading pipeline to convert input videos into an anime style, and toon shading with editing signals, where manually crafted editing prompts are employed to modify the content during the rendering process. For both tasks, we engage in comparative evaluations with other state-of-the-art methods. This includes Rerender-a-video [Yang *et al.*, 2023], an open-source method that employs a specialized pipeline for video synthesis. To ensure fairness in comparison, we adapt Rerender-a-video by substituting its default model with the diffusion model from our approach. Moreover, we include several well-regarded closed-source models that have shown to be competitive relative to existing techniques. Among these, Gen-1 [Esser *et al.*, 2023]—though not explicitly designed for toon shading—is evaluated particularly in the second task. Additionally, DomoAI [DOMO.AI, 2024] provides a variety of models via Discord[6], from which we select the "Anime v2 - Japanese anime style" for our tests. Given DomoAI's length constraint, our experiments are limited to 10-second clips from each video. This thorough comparative analysis is executed to ascertain the efficacy of our methodology relative to both open-source and closed-source state-of-the-art methods across varied tasks.

Currently, accurately assessing video quality poses considerable challenges, and recent years have seen some controversy regarding the choice of evaluation metrics [Brooks *et al.*, 2022; Blattmann *et al.*, 2023]. In our experiments, we evaluate the quality of videos generated by each method across three dimensions: 1) **Aesthetics**: The visual appeal is quantified using an aesthetic score [Schuhmann *et al.*, 2022], which provides an indicator of the overall visual quality of the generated videos. 2) **Text-video similarity**: To assess the relevance of the generated videos to the provided text in the toon shading with editing signals task, we employ cosine similarity calculated by the CLIP model [Radford *et al.*, 2021]. 3) **Video consistency**: In line with the methodologies of Rerender-a-video [Yang *et al.*, 2023] and Pix2Video [Ceylan *et al.*, 2023], we use pixel-MSE as a metric to gauge video consistency. It is noteworthy that the services offered by DomoAI and Gen-1 are limited to 24 FPS (Frames Per Second), which does not match the original video frame rate. Consequently, calculating pixel-MSE for these methods is impractical. The quantitative outcomes are detailed in Table 2. Our approach substantially outperforms the baseline models in both tasks, underlining the efficacy of Diffutoon.

Task	Method	Metric		
		Aesthetic score ↑	CLIP score ↑	Pixel MSE ↓
Toon shading	Rerender-a-video	5.35	-	200.46
	DomoAI	6.26	-	-
	Diffutoon	**6.47**	-	**188.87**
Toon shading with editing signals	Rerender-a-video	5.40	28.63	266.23
	DomoAI	6.25	29.01	-
	Gen-1	6.11	28.91	-
	Diffutoon	**6.37**	**30.69**	**143.51**

Table 2: Quantitative results of each approach.

Beyond these metrics, we also conducted a human evaluation with 10 participants. In each session, participants were shown two videos: one produced by our method and another by a randomly chosen baseline method. Participants were tasked with selecting the video they perceived to have better visual quality. The preferences of the participants are summarized in Table 3. These findings reveal a strong preference for our method, indicating its ability to generate videos with superior visual appeal. This feedback further underscores the

[5] https://www.bilibili.com/

[6] https://discord.com/

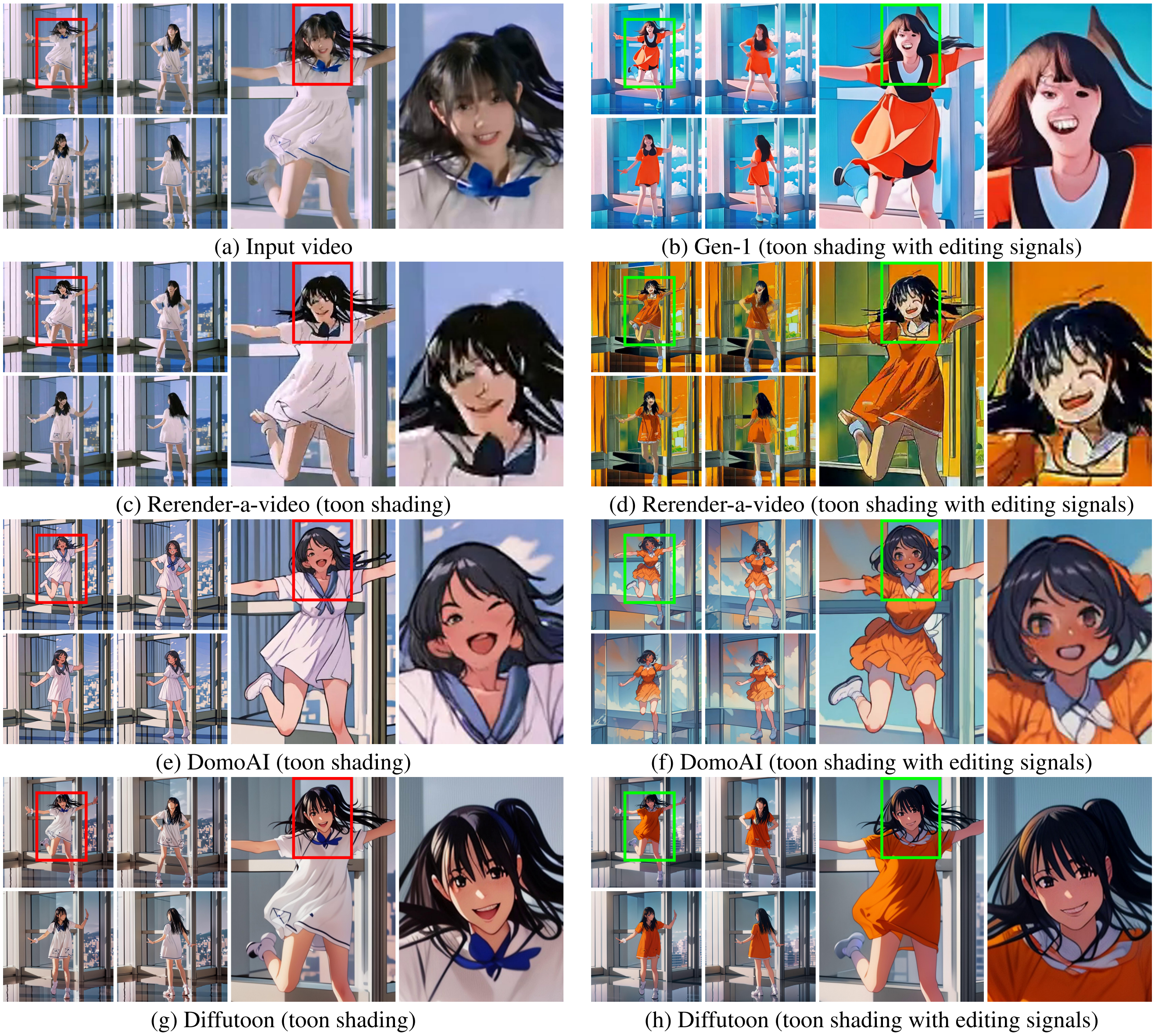
(a) Input video
(b) Gen-1 (toon shading with editing signals)
(c) Rerender-a-video (toon shading)
(d) Rerender-a-video (toon shading with editing signals)
(e) DomoAI (toon shading)
(f) DomoAI (toon shading with editing signals)
(g) Diffutoon (toon shading)
(h) Diffutoon (toon shading with editing signals)

Figure 2: Visual comparison with other methods. The prompt used for editing is "best quality, perfect anime illustration, a girl is dancing, smile, solo, orange dress, black hair, white shoes, blue sky ". Since the resolution of our generated video is extremely high, we enlarge some areas to view details. **We highly recommend readers to see the videos on our project page.**

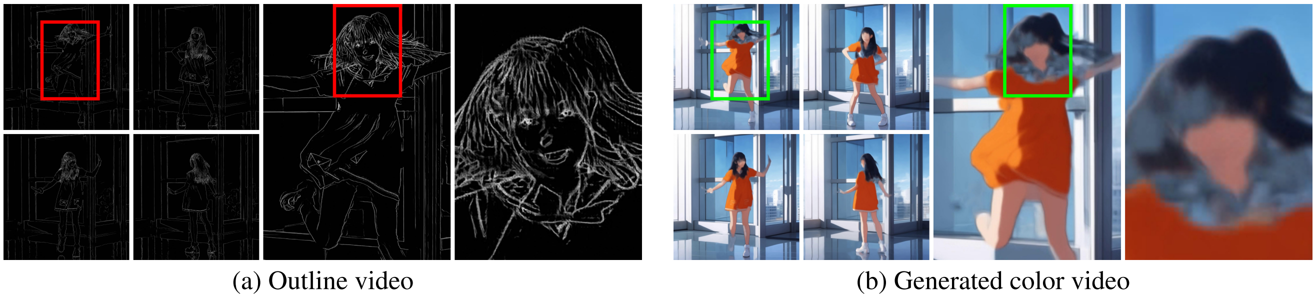

Figure 3: Intermediate results of Diffutoon. In the main toon shading pipeline, the video is synthesized according to the outline video and the color video. When the editing branch is enabled, the generated color video contains the editing signals.

Task	Baseline	Preference	
		Diffutoon	Other
Toon shading	Rerender-a-video	**98.21%**	1.79%
	DomoAI	**90.77%**	9.23%
Toon shading with editing signals	Rerender-a-video	**97.44%**	2.56%
	DomoAI	**82.35%**	17.65%
	Gen-1	**95.74%**	4.26%

Table 3: User preference in human evaluation.

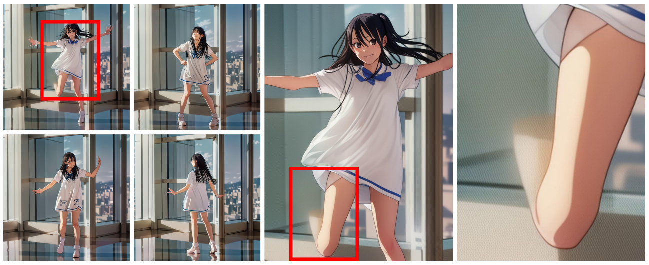

Figure 4: Video rendered without outline information.

advantage of our approach.

4.2 Case Study

Figure 2 displays video samples produced by various methods. In the original video (Figure 2a), a girl is depicted dancing with rapid movements, a scenario that poses a significant challenge for video processing algorithms. Gen-1 and Rerender-a-video exhibit difficulties in handling high-resolution videos, leading to facial distortions of the character. Moreover, in the videos processed by DomoAI (Figure 2e and Figure 2f), the third frame lacks content, and the character's movements in the fourth frame are misaligned with the original footage. This suggests DomoAI's limitations in accurately capturing and reproducing motion features and poses from the source video. In contrast, videos produced by Diffutoon (Figure 2g and Figure 2h) exemplify the method's capacity to preserve intricate details such as lighting, hair texture, and pose, all while adhering to an anime visual style. Remarkably, in the toon shading with editing signals task, Diffutoon demonstrates precise control over color manipulation based on provided textual cues. These outcomes vividly underscore the robustness and effectiveness of our method.

In Figure 3, we showcase the intermediate outcomes of employing Diffutoon, including the outline and color videos generated by the editing branch. The outline video effectively captures structural information necessary for rendering frames in an anime style, thus ensuring the integrity of the visual quality. However, the color video appears blurry, attributed to the rapid movements of the dancing girl, indicating that the editing branch, when functioning autonomously, struggles to maintain high video quality. These two videos play a crucial role in producing the high-resolution video depicted in Figure 2h, by providing foundational information for rendering. For additional video examples, please refer to our project page.

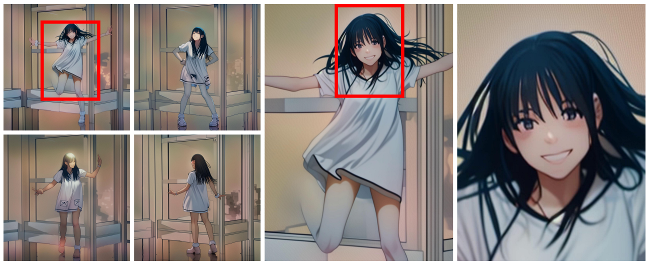

Figure 5: Video rendered without color information.

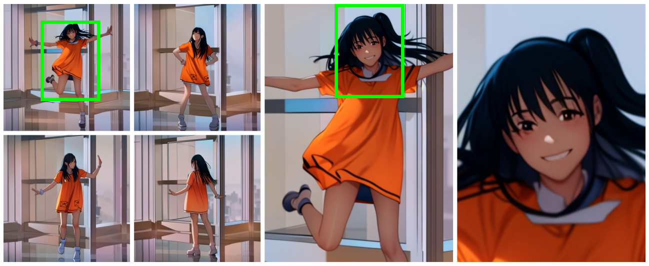

Figure 6: Video rendered by the editing branch with AnimateDiff.

4.3 Ablation Study

Given the extensive evaluations of motion modules by previous research [Xing *et al.*, 2023], we further investigated the impact of the two ControlNet models within our main toon shading pipeline. The videos rendered without each ControlNet model illustrate the deficits: without outline information, as seen in Figure 4, there is noticeable frame distortion, and without color guidance, as presented in Figure 5, evident flickering occurs. These outcomes confirm the critical nature of both outline and color components.

In the toon shading task that includes editing signals, we devised an alternative single-pipeline approach based on the editing branch, substituting FastBlend with AnimateDiff. The resulting video, shown in Figure 6, is unsatisfactory, exhibiting darkness and a lack of aesthetic appeal. As discussed in Section 3.2, the primary deficiency stems from AnimateDiff's reliance on a modified DDIM scheduler, which does not align well with the Stable Diffusion backbone, leading to subpar video quality. Nonetheless, this shortcoming minimally affects the main toon shading pipeline as the color consistency is secured by the ControlNet. This underscores the necessity for a distinct pipeline architecture.

5 Conclusion and Future Work

In this paper, we explore an innovative form of toon shading based on diffusion models, aiming to seamlessly transform photorealistic videos into anime styles. We have introduced an advanced toon shading methodology comprising a main toon shading pipeline and an editing branch. Our method is adept at processing high-resolution, lengthy videos and facilitates video editing through the editing branch. The comprehensive experimental results underscore the effectiveness of our approach. Looking ahead, our future work will concentrate on uncovering additional applications within the video processing domain.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant Number 62202170, Fundamental Research Funds for the Central Universities under grant Number YBNLTS2023-014, and Alibaba Group through the Alibaba Innovation Research Program.

References

[Barla *et al.*, 2006] Pascal Barla, Joëlle Thollot, and Lee Markosian. X-toon: An extended toon shader. In *Proceedings of the 4th international symposium on Non-photorealistic animation and rendering*, pages 127–132, 2006.

[Blattmann *et al.*, 2023] 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.

[Brooks *et al.*, 2022] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022.

[Cao *et al.*, 2023] Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–11, 2023.

[Ceylan *et al.*, 2023] 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.

[Dao *et al.*, 2022] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[DOMO.AI, 2024] Group DOMO.AI. Domo.ai, 2024. https://ai.domo.com/, Last accessed on 2024-01-18.

[Duan *et al.*, 2023a] Zhongjie Duan, Chengyu Wang, Cen Chen, Jun Huang, and Weining Qian. Optimal linear subspace search: Learning to construct fast and high-quality schedulers for diffusion models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 463–472, 2023.

[Duan *et al.*, 2023b] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, Jun Huang, and Mingyi Jin. Fastblend: a powerful model-free toolkit making video stylization easier. *arXiv preprint arXiv:2311.09265*, 2023.

[Duan *et al.*, 2023c] Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, Jun Huang, Fei Chao, and Rongrong Ji. Diffsynth: Latent in-iteration deflickering for realistic video synthesis. *arXiv preprint arXiv:2308.03463*, 2023.

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

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Guo *et al.*, 2023] 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.

[He *et al.*, 2023] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. *arXiv preprint arXiv:2310.07702*, 2023.

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

[Hu *et al.*, 2021] 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.

[Hudon *et al.*, 2018] Matis Hudon, Rafael Pagés, Mairéad Grogan, Jan Ondřej, and Aljoša Smolić. 2d shading for cel animation. In *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*, pages 1–12, 2018.

[Jamriška *et al.*, 2019] 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.

[Jin *et al.*, 2023] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *arXiv preprint arXiv:2306.08645*, 2023.

[Khachatryan *et al.*, 2023] 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.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lei *et al.*, 2023] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. Blind video deflickering by neural filtering with a flawed atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10439–10448, 2023.

[Li *et al.*, 2022] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[Lu *et al.*, 2022] 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. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[Mou *et al.*, 2023] 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.

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

[Radford *et al.*, 2021] 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.

[Ranftl *et al.*, 2020] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[Rombach *et al.*, 2022] 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.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Ruiz *et al.*, 2023] 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.

[Schuhmann *et al.*, 2022] 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.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

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

[Wang *et al.*, 2021] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[Wang *et al.*, 2024] Jiapeng Wang, Chengyu Wang, Tingfeng Cao, Jun Huang, and Lianwen Jin. Diffchat: Learning to chat with text-to-image synthesis models for interactive image creation. *arXiv preprint arXiv:2403.04997*, 2024.

[Xie and Tu, 2015] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[Xing *et al.*, 2023] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023.

[Yang *et al.*, 2023] 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.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.