Diffutoon: High-Resolution Editable Toon Shading via Diffusion Models

Zhongjie Duan¹, Chengyu Wang², Cen Chen¹, Weining Qian¹ and Jun Huang²
¹East China Normal University, Shanghai, China
²Alibaba 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 quantitative 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 [Hudson 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 adjust 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 $1536 \times 1536$. In our experiments, we first evaluate Diffutoon in the toon shading task, and then 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
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]. 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 \(\{v^1, v^2, \ldots, v^N\}\), we first generate an outline video and a color video. The outline video \(\{o^1, o^2, \ldots, o^N\}\) contains the structural information extracted from the input video, and the color video \(\{c^1, c^2, \ldots, c^N\}\) 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 \(E\), 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:

\[
x^i_T = \{x^i_T\}_{i=1}^N \sim N(O, I),
\]

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 \(\tau\) for better aesthetics. Note that the motion modules are trained within 32 consecutive frames, we can only use the denoising model \(E\) in a sliding window with a size no larger than 32. The sliding windows with size \(d\) and stride \(s\) are:

\[
W(d, s) = \{i, i + d - 1 : 1 \leq i \leq N, i \equiv 1(\text{mod} s)\},
\]

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:

\[
\{e_{t,+}(l, i, r)\}_{i=1}^r = E \left( \{x^i\}_{i=1}^r, \{o^i\}_{i=1}^r, \{c^i\}_{i=1}^r, t, \tau \right).
\]

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:

\[
\tau_{t,+}(i) = \sum_{(l, r) \in W(d, s)} \frac{w(l, i, r)}{\sum_{(l', r') \in W(d, s)} w(l', i, r')} e_{t,+}(l, i, r).
\]

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

\[
w(l, i, r) = \begin{cases} 1 + \frac{\epsilon}{\alpha} - \frac{|i - l| + |r - r'|}{2}, & \text{if } l \leq i \leq r, \\ 0, & \text{otherwise} \end{cases}
\]

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 \(\tau'\) [Gal et al., 2022] on the negative side, which involves 10 token embeddings to be processed by the text encoder. By replacing \(\tau\) with \(\tau'\) in (4) and (3), we can obtain the estimated noise on the negative side \(\tau_{t,-}(i)\). Then, the guided estimated noise is:

\[
\tau_{t}(i) = g \cdot \tau_{t,+}(i) + (1 - g) \cdot \tau_{t,-}(i).
\]

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:

\[
\tau_t = (\tau_t(0), \tau_t(1), \ldots, \tau_t(n)) \in \mathbb{R}^{N \times H \times W \times C}.
\]

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

\[
x_{t-1} = \sqrt{\alpha_t} x_t - \sqrt{1 - \alpha_t} \tau_t + \sqrt{1 - \alpha_t} \tau_{t,-},
\]

\(\tau_{t,-}\) can be moved to GPU memory when the sliding window covers them. We adopt a linear combination of overlapping segments from different sliding windows:

\[
\tau_{t,+}(i) = \sum_{(l, r) \in W(d, s)} \frac{w(l, i, r)}{\sum_{(l', r') \in W(d, s)} w(l', i, r')} e_{t,+}(l, i, r).
\]

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

\[
w(l, i, r) = \begin{cases} 1 + \frac{\epsilon}{\alpha} - \frac{|i - l| + |r - r'|}{2}, & \text{if } l \leq i \leq r, \\ 0, & \text{otherwise} \end{cases}
\]

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 \(\tau'\) [Gal et al., 2022] on the negative side, which involves 10 token embeddings to be processed by the text encoder. By replacing \(\tau\) with \(\tau'\) in (4) and (3), we can obtain the estimated noise on the negative side \(\tau_{t,-}(i)\). Then, the guided estimated noise is:

\[
\tau_{t}(i) = g \cdot \tau_{t,+}(i) + (1 - g) \cdot \tau_{t,-}(i).
\]

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:

\[
\tau_t = (\tau_t(0), \tau_t(1), \ldots, \tau_t(n)) \in \mathbb{R}^{N \times H \times W \times C}.
\]

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

\[
x_{t-1} = \sqrt{\alpha_t} x_t - \sqrt{1 - \alpha_t} \tau_t + \sqrt{1 - \alpha_t} \tau_{t,-},
\]
where $\alpha_t$ 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 platform3. 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 $1536 \times 1536$, resulting in visually impressive frames. The detailed settings of parameters are presented in Table 1.

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, Do-moAI [DOMO.AI, 2024] provides a variety of models via Discord6, 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.

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

<p>|</p>
<table>
<thead>
<tr>
<th>Components</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main toon shading pipeline</td>
<td>frame height</td>
<td>1536</td>
</tr>
<tr>
<td></td>
<td>frame width</td>
<td>1536</td>
</tr>
<tr>
<td></td>
<td>classifier-free guidance scale</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>inference steps</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>sliding window size</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>sliding window stride</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>conditioning scale (outline)</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>conditioning scale (color)</td>
<td>0.5</td>
</tr>
<tr>
<td>Editing branch</td>
<td>frame height</td>
<td>512</td>
</tr>
<tr>
<td></td>
<td>frame width</td>
<td>512</td>
</tr>
<tr>
<td></td>
<td>classifier-free guidance scale</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>inference steps</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>sliding window size</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>sliding window stride</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>conditioning scale (depth)</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>conditioning scale (softedge)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 1: Parameter settings in the experiments.

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>Metric</th>
<th>Aesthetic score ↑</th>
<th>CLIP score ↑</th>
<th>Pixel MSE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toon shading</td>
<td>Rerender-a-video</td>
<td>Diffutoon</td>
<td>5.35</td>
<td>6.47</td>
<td>200.46</td>
</tr>
<tr>
<td></td>
<td>DomoAI</td>
<td>6.26</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Toon shading</td>
<td>Rerender-a-video</td>
<td>Diffutoon</td>
<td>5.40</td>
<td>6.37</td>
<td>266.23</td>
</tr>
<tr>
<td>with editing signals</td>
<td>DomoAI</td>
<td>6.25</td>
<td>28.91</td>
<td>28.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gen-1</td>
<td>6.11</td>
<td>153.65</td>
<td>266.23</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Quantitative results of each approach.

---
3https://www.bilibili.com/
6https://discord.com/
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.
Table 3: User preference in human evaluation.

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Diffutoon</td>
</tr>
<tr>
<td>Toon shading</td>
<td>Rerender-a-video</td>
<td>98.21%</td>
</tr>
<tr>
<td></td>
<td>DomoAI</td>
<td>1.79%</td>
</tr>
<tr>
<td>Toon shading with editing signals</td>
<td>Rerender-a-video</td>
<td>97.44%</td>
</tr>
<tr>
<td></td>
<td>DomoAI</td>
<td>2.56%</td>
</tr>
<tr>
<td></td>
<td>Gen-1</td>
<td>82.35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17.65%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95.74%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.26%</td>
</tr>
</tbody>
</table>

Figure 4: Video rendered without outline information.

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.

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


