Research Article

AniSora: Exploring the Frontiers of Animation Video Generation in the Sora Era

Yudong Jiang[1], Baohan Xu[1], Siqian Yang[1], Mingyu Yin[1], Jing Liu[1]

[1]. Bilibili Inc.

Animation has gained significant interest in the recent film and TV industry. Despite the success of advanced video generation models like Sora, Kling, and CogVideoX in generating natural videos, they lack the same effectiveness in handling animation videos. Evaluating animation video generation is also a great challenge due to its unique artist styles, violating the laws of physics and exaggerated motions. In this paper, we present a comprehensive system, AniSora, designed for animation video generation, which includes a data processing pipeline, a controllable generation model, and an evaluation dataset. Supported by the data processing pipeline with over 10M high-quality data, the generation model incorporates a spatiotemporal mask module to facilitate key animation production functions such as image-to-video generation, frame interpolation, and localized image-guided animation. We also collect an evaluation benchmark of 948 various animation videos, the evaluation on VBench and human double-blind test demonstrates consistency in character and motion, achieving state-of-the-art results in animation video generation. Our evaluation benchmark will be publicly available at https://github.com/bilibili/Index-anisora.

Yudong Jiang, Baohan Xu, and Siqian Yang equally contributed to this work.

1. Introduction

The animation industry has seen significant growth in recent years, expanding its influence across entertainment, education, and even marketing. As demand for animation content rises, the need for efficient production processes is also growing quickly, particularly in animation workflows. Traditionally, creating high-quality animation has required extensive manual effort for tasks like

doi.org/10.32388/HIFC4X

creating storyboards, generating keyframes, and inbetweening, making the process labor-intensive and time-consuming. Previous efforts[1][2] to incorporate computer vision techniques have assisted animators in generating inbetween frames for animation. However, these methods often show effectiveness only within certain artistic styles, limiting their applicability to the varied demands of modern animations.

With recent advancements in video generation, there has been notable progress in generating high-quality videos across various domains. Inspired by Generative Adversarial Networks[3], Variational Autoencoders[4], and, more recently, transformer-based architectures[5][6], the field has seen remarkable improvements in both efficiency and output quality. However, most video generation methods are trained and evaluated on general-purpose datasets, typically featuring natural scenes or real-world objects[7][8]. The domain of animation video generation, which plays an important role ranging from entertainment to education, has received relatively little attention. Animation videos often rely on non-photorealistic elements, exaggerated expressions, and non-realistic motion, presenting unique challenges that current methods do not address.

In addition to the generation challenges, the evaluation of video generation is also inherently complex. Evaluating video generation quality requires assessing not only the visual fidelity of each frame but also temporal consistency, coherence, and smoothness across frames[9]. This challenge intensifies in animation, where unique artistic styles must remain consistent despite exaggerated motions and transformations. Progress in this field demands effective evaluation datasets tailored to animated video generation, enabling comprehensive testing of model adaptability to diverse styles, scene changes, and complex motions, thereby driving model optimization and innovation.

In this paper, as shown in Fig. 1, a full system **AniSora** is presented for animation video generation. First, our data processing pipeline offers over 10 million high-quality text-video pairs, forming the foundation of our work. Secondly, we develop a unified diffusion framework adapted for animation video generation. Our framework leverages spatiotemporal masking to support a range of tasks, including image-to-video generation, keyframe interpolation, and localized image-guided animation. By integrating these functions, our system bridges the gap between keyframes to create smooth transitions and enables dynamic control over specific regions, such as animating different characters speaking precisely. This enables a more efficient creative process for both professional and amateur animation creators. Fig. 2 demonstrates some examples generated by our model under image-to-video conditions.

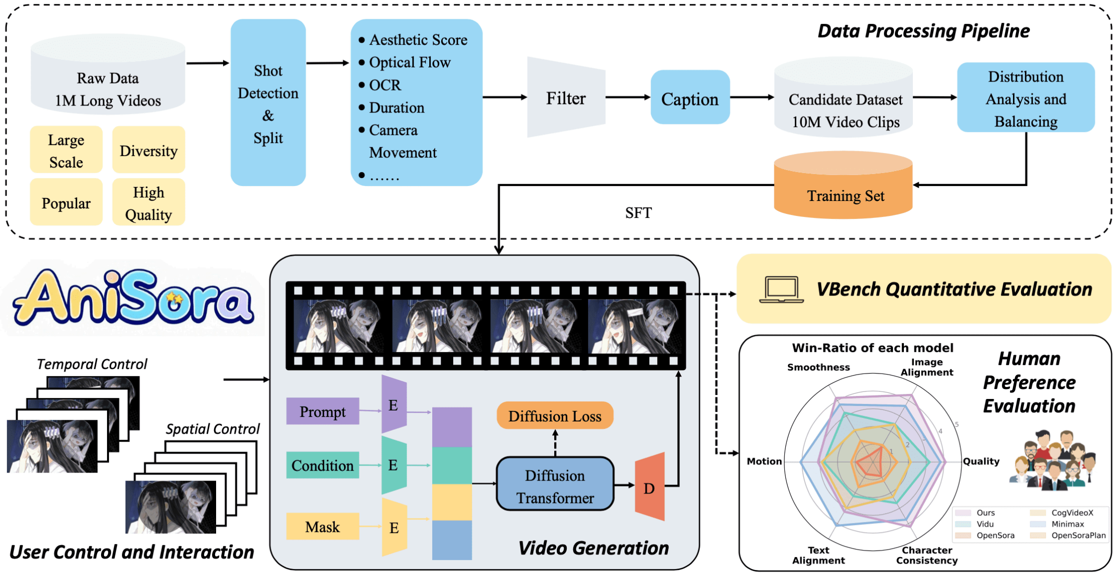


Figure 1. Overview. We propose **AniSora**, a comprehensive framework for animation video generation that integrates a high-quality animation dataset, a spatiotemporal conditional model, and a specialized animation video benchmark. The **Data Processing Pipeline** constructs a 10M video clip dataset derived from 1M diverse long animation videos. The **Video Generation** model employs a spatiotemporal conditional model, supporting various **User Control and Interaction** modes and enabling tasks such as frame interpolation, localized guidance, and so on. The benchmark set comprises 948 ground-truth videos spanning diverse styles, common motions, and both 2D and 3D animations. The prompt suite provides standardized prompts and guiding conditions, complemented by **Human Preference Evaluation** and a **Quantitative Evaluation** with eight objective metrics for assessing visual appearance and consistency. **AniSora** surpasses SOTA models, establishing a new benchmark for animation video generation.

doi.org/10.32388/HIFC4X

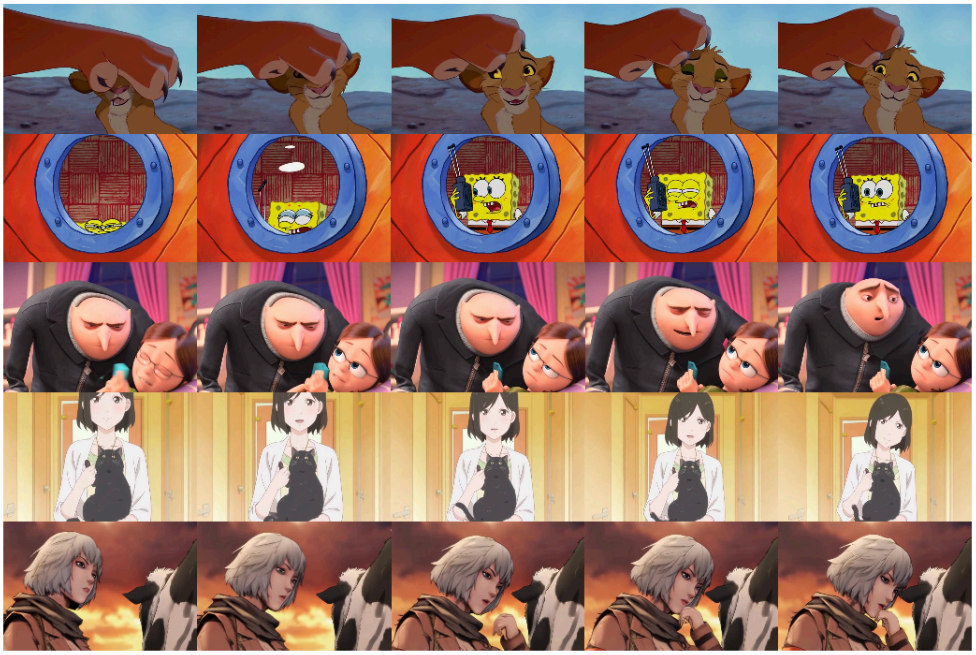


Figure 2. Our method can generate high quality and high consistency in various kinds of 2D/3D animation videos. These examples are generated under image-to-video settings conditioned on the leftmost frame. It is best viewed in color.

Additionally, we propose a benchmark dataset specifically designed for animation video evaluation. Unlike existing evaluation datasets, which primarily focus on natural landscapes or real-world human actions, our dataset addresses the unique requirements of animation video assessment. To achieve this, we collected 948 animation videos across various categories and manually refined the prompts associated with each video.

Our contributions can be summarized as follows:

- We develop a comprehensive video processing system that significantly enhances preprocessing for video generation.
- We propose a unified framework designed for animation video generation with a spatiotemporal mask module, enabling tasks such as image-to-video generation, frame interpolation, and localized image-guided animation.
- We release a benchmark dataset specifically for evaluating animation video generation.

2. Related Work

2.1. Video generation models

With the development of diffusion models, significant progress has been made in video generation over the past two years. Some research including[7][8][10][11] have demonstrated promising results in general video generation. Due to the limited available animation datasets, these models are not particularly effective for animation video generation.

2.2. Animation video datasets

Video data is one of the most critical elements for generation models, particularly for domain-specific data. However, obtaining high-quality animation video data is especially difficult compared to natural video datasets. Previous research has released some animation-related datasets, including ATD-12K[1], AVC[12]. While these datasets, collected from various animation movies, are helpful for some video interpolation and super-resolution tasks, they are limited by their small size. More recently, Sakuga-42M[13] has been proposed with 1.2M clips. It has improved compared to previous datasets that only contained a few hundred clips. Nevertheless, this remains insufficient for training video generation models, in contrast to general video datasets like Panda-70M[14] and InternVid-200M[15]. Additionally, 80% of its clips are low-resolution and less than 2 seconds, which hampers the generation of high-quality videos.

2.3. Evaluation of video generation models

Evaluating video generation models has remained a significant challenge in the past few years. Recently, Liu et al. have made great efforts to generate a diverse and comprehensive list of 700 prompts using LLM[16]. Besides, Huang et al. have proposed vbench for general video generation[9]. The authors have released 16 evaluation dimensions and prompt suites. Moreover, there is a notable absence of dedicated animation evaluation datasets, which limits the ability to benchmark models specifically designed for this genre. In[17], the authors primarily have focused on the performance of recent video generation models across various categories of datasets. Furthermore, they have also investigated some vertical-domain models like pose controllable human generation, and audio-driven animation.

While these works provide valuable insights into the capabilities of these models in generating diverse video content, they don't specifically address the unique requirements and challenges associated with animation video generation.

3. Dataset

We build our animation dataset according to the observation that *high quality text-video pairs are the cornerstone of video generation*, which is proved by recent researches[18]. In this section, we give a detailed description of the construction of our animation dataset and the evaluation benchmark.

Animation Dataset Construction

We build a pipeline to get high-quality text-video pairs among 1 million raw animation videos. First of all, we use scene detection[19] to divide raw animation videos into clips. Then, for each video clip, we construct a filter rule from four dimensions: text-cover region, optical flow score, aesthetic score, and number of frames. The filter rule is gradually built up through the observations in model training. In detail, the text-cover region score (obtained by[20]) can drop those clips with text overlay similar to end credits. Optical flow score[21] prevents those clips with still images or quick flashback scenes. Aesthetic score[22] is utilized to preserve clips with high artistic quality. Besides, we retain the video clips whose duration is among $2s$-$20s$ according to the number of the frames. After the four steps mentioned above, about 10% clips (more than 10 million clips) can be retained into training step. In addition, a few higher quality clips will be finally filtered from training set to further improve the model's performance. Specifically, during the training process, we adjust the proportions of specific training data (e.g., talking and motion amplitude) according to the observed performance.

Benchmark Dataset Construction

Moreover, to compare the generation videos between our model and other recent researches directly, we construct a benchmark dataset manually. 948 animation video clips are collected and labeled with different actions, e.g., talking, walking & running, eating, kissing, and so on. Among them, there are 857 2D animation clips and 91 3D clips. These action labels are summarized from more than 100 common actions with human annotation. Each label contains 10-30 video clips. The corresponding text prompt is generated by Qwen-VL2[23] at first, then is corrected manually to guarantee the text-video alignment.

doi.org/10.32388/HIFC4X

4. Method

In this section, we present an effective approach for animation video generation using a diffusion transformer architecture. Section 4.1 provides an overview of the foundational video diffusion transformer model. In section 4.2, we introduce a spatiotemporal mask module that extends the diffusion transformer model, enabling crucial animation production functions such as image-to-video generation, frame interpolation, and localized image-guided animation within a unified framework. These enhancements are essential for professional animation production. Finally, section 4.3 details the supervised fine-tuning strategy employed on the animation dataset.

4.1. Dit-based Video Generation Model

We adopt a DiT-based[6] text-to-video diffusion model as the foundation model. As shown in Fig. 3, the model leverages the three components to achieve coherent, high-resolution videos aligned with text prompts.

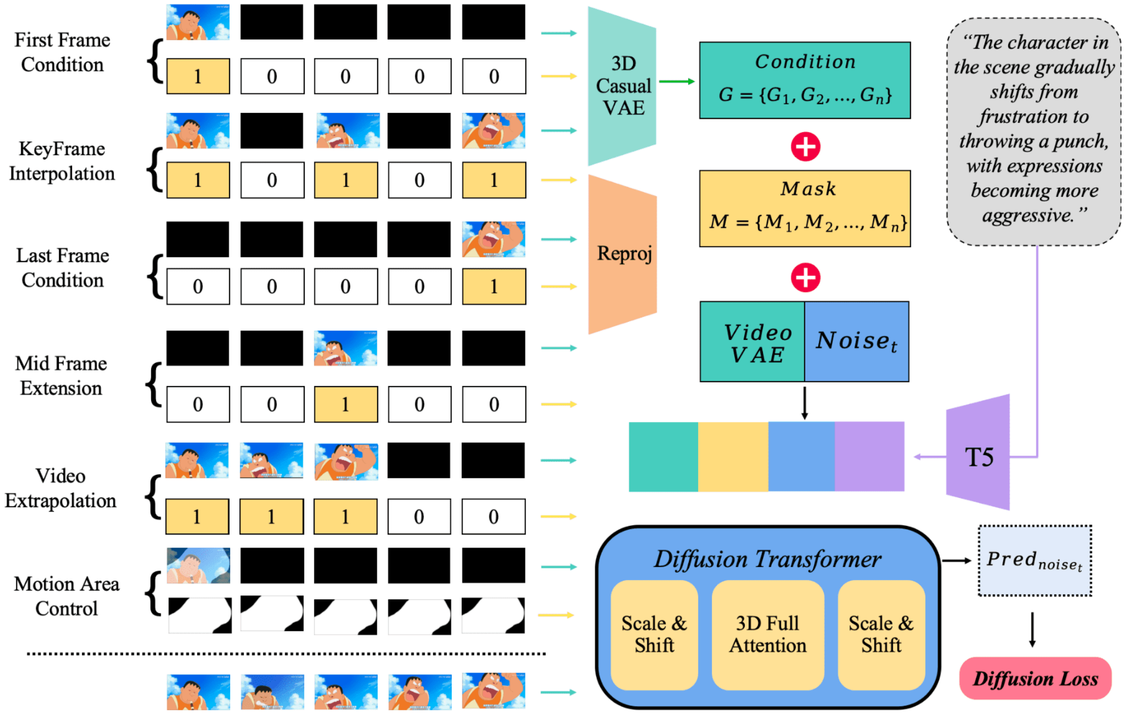


Figure 3. Method. This figure illustrates the Masked Diffusion Transformer framework for animation video generation, designed to support various spatiotemporal conditioning methods for precise and flexible animation control. A 3D Causal VAE compresses spatial-temporal features into a latent representation, generating the guide feature sequence G, while a reprojection network constructs the mask sequence M. These components, combined with noise and prompt's feature, serve as input to the Diffusion Transformer. The transformer employs techniques such as patchify, 3D-RoPE embeddings, and 3D full attention to effectively capture and model complex spatial-temporal dependencies. This framework enables seamless integration of features like keyframe interpolation, motion control, and mid-frame extension, streamlining animation production and enhancing creative possibilities.

3D Casual VAE used in video generation frameworks[24][25]serves as a specialized encoder-decoder architecture tailored for spatiotemporal data compression. This 3D VAE compresses videos across both spatial and temporal dimensions, significantly reducing the diffusion model computing. We follow the approach of Yang et al.[8] to extract latent features, transforming the original video with dimensions $(W, H, T, 3)$ into a latent representation of shape $(W/8, H/8, T/4, 16)$.

Patchify is a critical step for adapting vision tasks to transformer-based architectures[26]. Given an input video of size $T \times H \times W \times C$, it is split spatio into patches of size $P \times P$, and temporal into size Q resulting in $(T/Q) \times (H/P) \times (W/P) \times C$ patches. This method enables efficient high-dimensional data processing by reducing complexity while retaining local spatial information.

3D Full Attention is a module we propose for spatial and temporal modeling, inspired by the remarkable success of long-context training in large language models (LLMs)[27] and foundation video generation models[8][18].

Diffusion schedule applies Gaussian noise to an initial sample x_0 over T steps, generating noisy samples $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon$, where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The reverse process predicts ϵ by minimizing the mean squared error:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_0, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right].$$

To stabilize training, we use the v-prediction loss[28], where $v = \sqrt{1-\alpha_t} x_0 - \sqrt{\alpha_t}\epsilon$ and the loss becomes

$$\mathcal{L}_{v-\text{prediction}} = \mathbb{E}_{x_0, v, t}\left[\|v - v_\theta(x_t, t)\|_2^2\right].$$

This approach enhances stability and model performance.

4.2. Spatiotemporal Condition Model

Keyframe Interpolation creates smooth transitions between key-frames by generating intermediate frames, or ”in-between.” It is an essential stage in professional animation production and represents some of the most labor-intensive tasks for artists. We extend this concept to video generation conditioned on one or multiple arbitrary frames placed at any position within a video sequence.

Motion Control, as a technique within our framework, addresses the limitations of text-based control and enables precise control over motion regions. This approach enhances artists’ control over video content, allowing them to express their creativity while significantly reducing their workload.

4.2.1. Masked Diffusion Transformer Model

In the Masked Diffusion Transformer framework, we construct a guide feature sequence $G = \{G_1, G_2, \ldots, G_n\}$ by placing the VAE-encoded guide frame F_{p_i} at designated positions p_i, while setting $G_j = 0$ for all other positions $j \neq p_i$. A corresponding mask sequence $M = \{M_1, M_2, \ldots, M_n\}$ is generated, where $M_{p_i} = 1$ for guide frame positions and $M_j = 0$ otherwise. The mask is processed through a re-projection function, yielding an encoded representation

$Reproj(M)$. The final input to the Diffusion Transformer is the concatenation of noise, encoded mask, prompt's T5 feature, and guide sequence along the channel dimension:

$$X = Concat(Noise_t, Reproj(M), G, T5) \tag{1}$$

This setup integrates position-specific guidance and mask encoding, enhancing the model's conditioned generation capabilities.

4.2.2. Motion Area Condition

This framework can also support spatial motion area conditions inspired by Dai et.al[29]. Given the image condition F_{p_i}, and motion area condition is represented by mask M_F, the same shape with F_{p_i}. Motion area in M_F is labeled 1, other place is set to 0. As equation 1 in 4.2.1, for guide frame position p_i, set $M_{p_i} = M_F$. The data processing and training pipeline can be summarized as follows: **Constructing video-mask pairs**, we first construct paired training data consisting of videos and their corresponding masks. Using a foreground detector by Kim et.al[30], we detect the foreground region in the first frame of the video. This region is then tracked across subsequent frames to generate a foreground mask for each frame. **Union of foreground masks**, the per-frame foreground masks are combined to create a unified mask M_F, representing the union of all foreground regions across the video. **Video latent post-processing**, for the video latent representation z_0, non-moving regions are set to the latent features of the guide image, ensuring static areas adhere to the guide. **LoRA-based conditional training**, we train the conditional guidance model using Low-Rank Adaptation (LoRA) with a parameter size of 0.27B. This approach significantly reduces computational requirements while enabling efficient model training.

4.3. Supervised Fine-Tuning

We initialize our model with the pre-trained weights of CogVideoX, which was trained on 35 million diverse video clips. Subsequently, we perform full-parameter supervised fine-tuning (SFT) on a custom animation training dataset to adapt the model specifically for animation tasks.

Weak to Strong

Our video generation model adopts a weak-to-strong training strategy to progressively enhance its learning capabilities across varying resolutions and frame rates. Initially, the model is trained on 480P videos at 8fps for 3 epochs, allowing it to capture basic spatiotemporal dynamics at a lower frame rate.

Following this, the model undergoes training on 480P videos at 16fps for an additional 1.9 epochs, enabling it to refine its temporal consistency and adapt to higher frame rates. Finally, the model is fine-tuned on 720P videos at 16fps for 2.3 epochs, leveraging the previously learned features to generate high-resolution, temporally coherent video outputs. Additionally, we applied stricter filtering as in section3, producing a 1M ultra high-quality dataset for final-stage fine-tuning, significantly boosting high-resolution video quality.

Removing Generated Subtitles

The presence of a significant number of videos with subtitles and platform watermarks in our training data led to the model occasionally generating such artifacts in its outputs. To mitigate this issue, we performed supervised fine-tuning using a curated dataset of videos entirely free of subtitles and watermarks. This dataset, consisting of 790k video clips, was constructed through proportional cropping of videos containing subtitles and the selection of clean, subtitle-free videos. Full-parameter fine-tuning was then applied to the model, and after 5.5k iterations, we observed that the model effectively eliminated the generation of subtitles and watermarks without compromising its overall performance.

Temporal Multi-Resolution Training

Given the scarcity of high-quality animation data, we employ a mixed training strategy using video clips of varying durations to maximize data utilization. Specifically, a variable-length training approach is adopted, with training video durations ranging from 2 to 8 seconds. This strategy enables our model to generate 720p video clips with flexible lengths between 2 and 8 seconds.

Multi-Task Learning

Compared to the physically consistent motion patterns in the real world, animation styles, and motion dynamics can vary significantly across different works. This domain gap between datasets often leads to substantial quality differences in videos generated from guide frames with different artistic styles. We incorporate image generation into a multi-task training framework to improve the model’s generalization across diverse art styles. Experimental results demonstrate that this approach effectively reduces the quality gap in video generation caused by stylistic differences in guide frames.

Mask Strategy

During training, we unmask the first, last, and other frames obtained through uniform sampling with a 50% probability. This strategy equips the model with the ability to handle arbitrary guidance, enabling it to perform tasks such as in-betweening, first-frame continuation, and arbitrary frame guidance, as discussed in Section 4.2.1.

5. Experiment

5.1. Benchmark Evaluation

In this section, we give both objective and human evaluation results of our benchmark.

Automated Evaluation

To obtain the objective results, we choose several dimensions in VBench[9], e.g., motion smoothness, aesthetic quality, imaging quality, subject consistency, I2V subject consistency, I2V background consistency, and overall consistency. The former three metrics evaluate the visual quality, while the latter four reflect the degree of consistency. Especially, in VBench, overall consistency evaluates the text-video consistency, since they use ViCLIP[15] as the baseline model. In addition, we utilize a motion amplitude model, which is based on ActionCLIP[31] framework to evaluate the motion score of the generation clips. In detail, About 10 million animation video clips and their corresponding motion captions are collected into 6 degrees of movement amplitude (from stillness to significant motion) to finetune the action model. Finally, the motion score is obtained from the similarity score between the designed motion prompt and the participant video.

$$S_{motion} = \cos(MCLIP(V), MCLIP(T_m)), \tag{2}$$

where $MCLIP$ denotes the finetuning action model. V represents the generation video, and T_m denotes the designed motion prompt.

6 recent I2V investigations are involved into our evaluation: Open-sora-V1.2[10], Open-sora-plan-V1.3[11], Cogvideox-5B-V1[8], Vidu[32], Minimax[33] and AniSora(ours).

Tab. 1 gives the automated results from 8 metrics. We observe that our method performs well on subject consistency and motion smoothness, and closely on other 5 dimensions except motion score. These mainly because we conduct a thorough assessment of the balance between generation quality

doi.org/10.32388/HIFC4X

and motion magnitude, and find most generation clips with big motion results in distortion or unnatural segments. It is worth mentioning that the automated scores of AniSora are similar to those of GT, and the visual performance can refer to Fig. 2. Furthermore, the automated scores from VBench show the room for improvement across several dimensions, and we will provide our improved metrics soon.

Method	**Appearance**				**Consistency**			
	Motion	**Motion**	**Aesthetic**	**Imaging**	**I2V**	**I2V**	**Overall**	**Subject**
	Smoothness	**Score**	**Quality**	**Quality**	**Subject**	**Background**	**Consistency**	**Consistency**
Opensora-Plan(V1.3)	99.13	76.45	53.21	65.11	93.53	94.71	21.67	88.86
Opensora(V1.2)	98.78	73.62	54.30	68.44	93.15	91.09	**22.68**	87.71
Vidu	97.71	**77.51**	53.68	69.23	92.25	93.06	20.87	88.27
Cogvideo(5B-V1)	97.67	71.47	**54.87**	68.16	90.68	91.79	21.87	90.29
MiniMax	99.20	66.53	54.66	**71.67**	95.95	**95.42**	21.82	93.62
AniSora	**99.34**	45.59	54.31	70.58	**97.52**	95.04	21.15	**96.99**
AniSora-K	99.12	59.49	53.76	68.68	95.13	93.36	21.13	94.61
AniSora-I	99.31	54.96	54.67	68.98	94.16	92.38	20.47	95.75
GT	98.72	56.05	52.70	70.50	96.02	95.03	21.29	94.37

Table 1. Automated Performance Comparison of Different Methods. (Note that AniSora-K denotes the results with keyframe interpolation, and AniSora-I denotes the interpolated average results of AniSora)

Human Evaluation

To comprehensively evaluate our model, we introduce a brief and clear human blind testing for 6 dimensions: visual smoothness, visual motion, visual appeal, text-video, image-video, and character consistency. Correspondingly, visual smoothness, text-video, and character consistency are similar to motion smoothness, overall, and subject consistency, respectively. Moreover, image-video consistency is equal to I2V subject and I2V background consistency, visual appeal is equal to aesthetic quality and imaging quality, and visual motion is the same as the motion score mentioned in the automated evaluation. In detail, each participant labeled 6 dimensions (from 1 to 5, and 5 is the best) without prior knowledge of the generation methods. Tab. 2 shows the human evaluation results in a percentage format. We observe that Anisora outperforms the other methods across most dimensions; however, there is still substantial room for improvement, particularly in text-video consistency. We conducted a statistical analysis to evaluate the consistency of scores given by 12 raters across various dimensions. The results indicate that the Pearson correlation coefficients for individual dimensions range from 0.5 to 0.6, with an overall correlation coefficient of 0.56. This suggests that even human evaluators exhibit significant subjectivity and randomness when assessing the quality of generated videos across different dimensions. These findings highlight the importance of establishing consistent and objective evaluation criteria for assessing video generation quality.

Method	**Appearance**			**Consistency**		
	Visual Smooth	**Visual Motion**	**Visual Appeal**	**Text-Video Consistency**	**Image-Video Consistency**	**Character Consistency**
Opensora-Plan(V1.3)	38.1	38.92	47.88	55.82	43.52	34.72
Opensora(V1.2)	28.14	37.24	37.46	47.64	42.62	31.52
Vidu	58.78	47.9	65.48	60.8	56.5	54.26
CogVideoX(5B-V1)	46.64	49.3	56.06	68.82	56.36	48.88
MiniMax	65.98	**57.08**	71.56	**80.38**	67.88	65.82
AniSora(Ours)	**71.68**	51.06	**74.36**	71.56	**78.38**	**75.14**

Table 2. Human Evaluation Scores

doi.org/10.32388/HIFC4X

5.2. Spatiotemporal Mask module

Frame Interpolation

Tab. 1 presents the results of different interpolation settings on our benchmark dataset (AniSora-K and AniSora-I). Our evaluation process involved generating videos on our benchmark with various guidance conditions sampled at equal proportions, which can refer to Fig. 3. We then compute the average score of all samples, as well as specific statistical analysis for keyframe interpolation results. The performance indicates that single-frame guidance achieves competitive results whether the guiding frame is placed at the beginning, middle, or end of the frame sequence, which also consistently outperforms other methods. Adding more guiding frames further improves both character consistency and motion stability. We also observed from the motion score and smooth score that our baseline model achieves a balance between motion range and consistency, while keyframe guidance enables the model to produce animation videos with larger motion ranges and more realistic motion. More samples can be found in supplementary materials.

As shown in Fig. 4, our unified framework supports different interpolation settings, enabling these functions to meet the demands of professional animation production. We observe that more guiding frames contribute to a more stable character identity and more precise actions align with creators. Nonetheless, amateur creators can still obtain satisfactory results by using just the first or last frame.

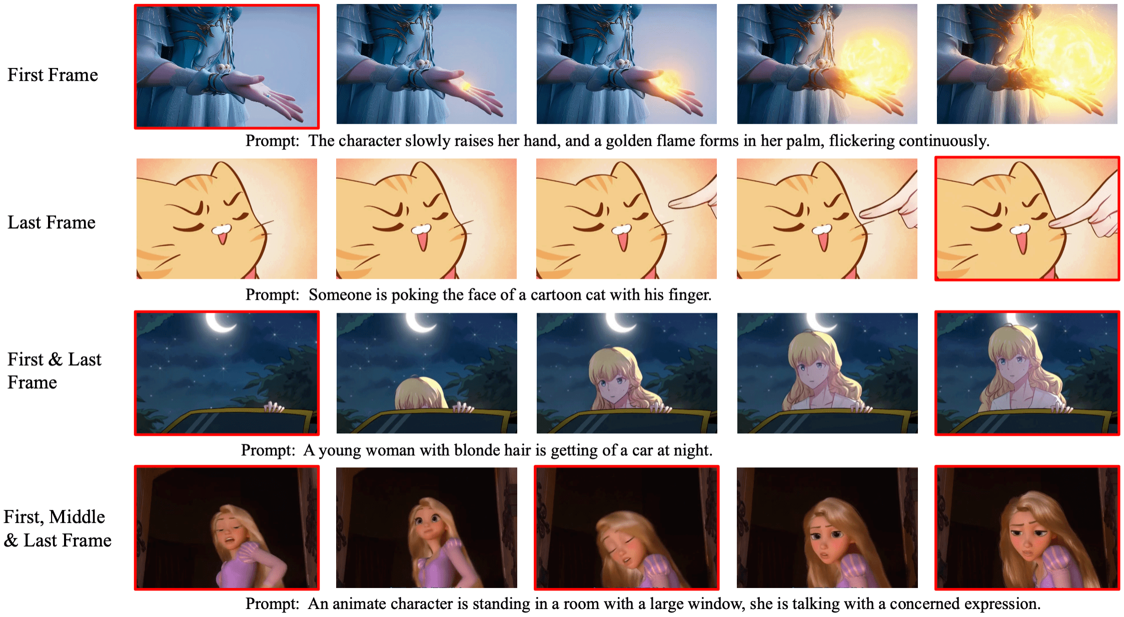

First Frame
Prompt: The character slowly raises her hand, and a golden flame forms in her palm, flickering continuously.
Last Frame
Prompt: Someone is poking the face of a cartoon cat with his finger.
First & Last Frame
Prompt: A young woman with blonde hair is getting of a car at night.
First, Middle & Last Frame
Prompt: An animate character is standing in a room with a large window, she is talking with a concerned expression.

Figure 4. Illustration of different interpolation strategies. The images highlighted in red indicate the provided reference images.

Motion Area Condition

The evaluation of motion area condition is constructed based on our benchmark dataset. For each initial frame, we performed saliency segmentation, followed by connected-component analysis to generate bounding boxes for each instance. Then we manually filtered the results to select high-quality motion area masks, resulting in 200 samples. Following the experiment settings in[29], we conducted the comparison of motion mask precision in Tab. 3. We also computed the score of AnimateAnything on our selected 200 samples. The lower score is primarily due to flickering and noise appearing outside the motion mask area. The results demonstrate the effectiveness of our spatial mask module in controlling movable regions. It is also noticeable that even without motion control, our generation model trained for animation video still shows a certain level of control. This may be due to the effective prompt-based guidance for the main subject, which aligns well with the defined motion mask. Fig. 5 also illustrates several motion mask guidance examples.

Method	Motion Mask Precision
AnimateAnything[29]	0.6141
Ours - No Control	0.4989
Ours - Motion Mask	**0.9604**

Table 3. Comparison of motion mask precision

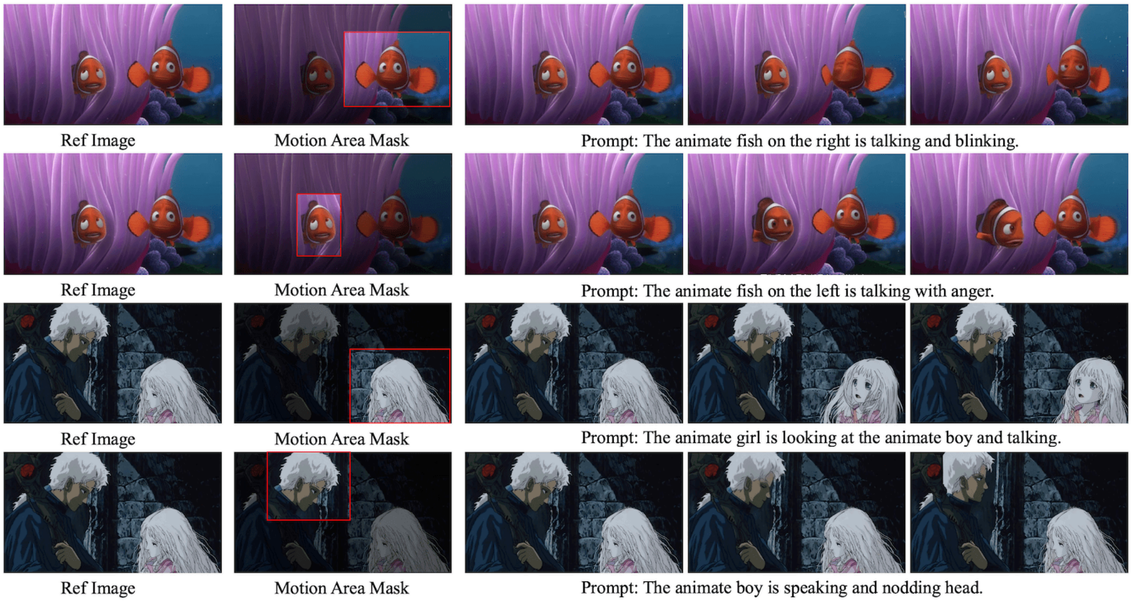

Ref Image
Motion Area Mask
Prompt: The animate fish on the right is talking and blinking.
Ref Image
Motion Area Mask
Prompt: The animate fish on the left is talking with anger.
Ref Image
Motion Area Mask
Prompt: The animate girl is looking at the animate boy and talking.
Ref Image
Motion Area Mask
Prompt: The animate boy is speaking and nodding head.

Figure 5. Examples of motion mask guidance. The first column shows the ref image, while the second column displays the mask. Animation creators can produce videos with fine-grained control over characters and backgrounds, ensuring alignment with various storylines.

5.3. Animation Video Training

2D and 3D Animation

Analysis using QWEN2[23] shows that 2D samples account for 85% of our dataset, yet 3D animation generation quality consistently surpasses that of 2D. Benchmark evaluations in Tab. 4 confirm 3D animations demonstrate superior visual appearance and consistency, a phenomenon unique to

doi.org/10.32388/HIFC4X

animation training. We attribute this gap to the pre-trained model's exposure to real-world video data. Unlike 2D animations with diverse motion patterns, 3D animations rendered by physics-based engines like Unreal Engine follow consistent physical laws, enabling better knowledge transfer during SFT. Consequently, improving generalization on 2D animation data remains more challenging than on 3D or real-world data.

Dims	**2D**	**3D**	**All**
Visual Smooth	70.23	**73.48**	71.68
Visual Motion	**51.14**	50.97	51.06
Visual Appeal	74.05	**74.74**	74.36
Text-Video Consistency	70.23	**73.21**	71.56
Image-Video Consistency	77.59	**79.37**	78.38
Character Consistency	**75.64**	74.52	75.14

Table 4. Human Evaluation Results between 2D and 3D Generation Clips

The Fig. 6 demonstrates some results of 2D and 3D animation generation. Artifacts are more prevalent in 2D generation results, such as exaggerated deformations, more diverse character appearances, and motions that break the physical rules. For instance, in the third row of 2D examples, tears appear to be floating in the air, making it more difficult for the model to capture the dynamic details accurately. In contrast, the motions rendered by the physics-based engines in 3D animations enable the model to achieve more reasonable results.

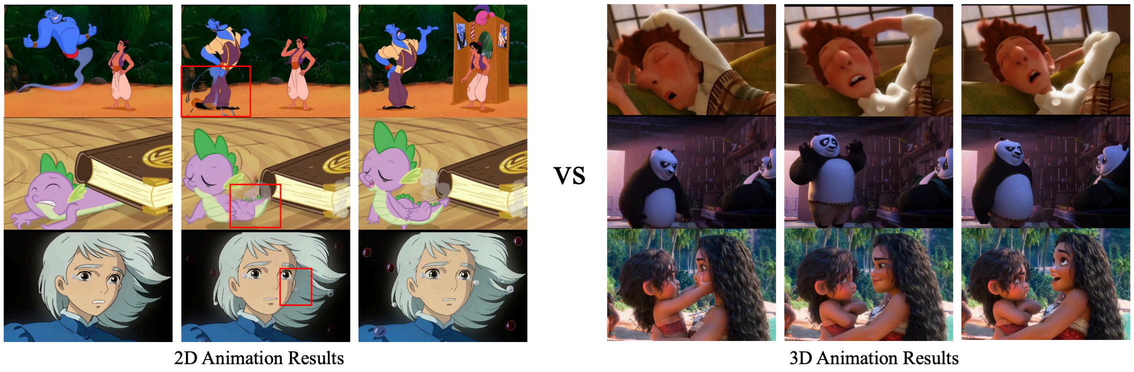


Figure 6. Comparison of 2D and 3D animation examples. The badcases in 2D animation are mainly due to exaggerated deformations, diverse appearances, and motions that violate physical laws.

Multi-Task Learning

The diversity of anime styles presents a challenge for video generation. Although our model performs well in most styles, unique styles may result in inconsistencies, particularly in character details. To address this, we applied multi-task learning, combining image and video training to enhance the model's adaptability to diverse styles.

We evaluated multi-task training using a manga with a unique artistic style. About 270 illustrations were used for the image generation task, while video training data remained the same as the baseline model. Additional illustrations served as first-frame conditions during video generation. After 5k training steps, as shown in Fig. 7, without incorporating images, the model struggles to fully understand such styles, resulting in flaws in character detail generation. While with the help of a small dataset of 270 images, the generated videos showed significantly greater stability and improved visual quality, particularly with highly distinctive guidance images. This approach effectively tailors animations to specific characters and mitigates domain gaps caused by variations in artistic styles, especially when high-quality animation data is limited.

doi.org/10.32388/HIFC4X

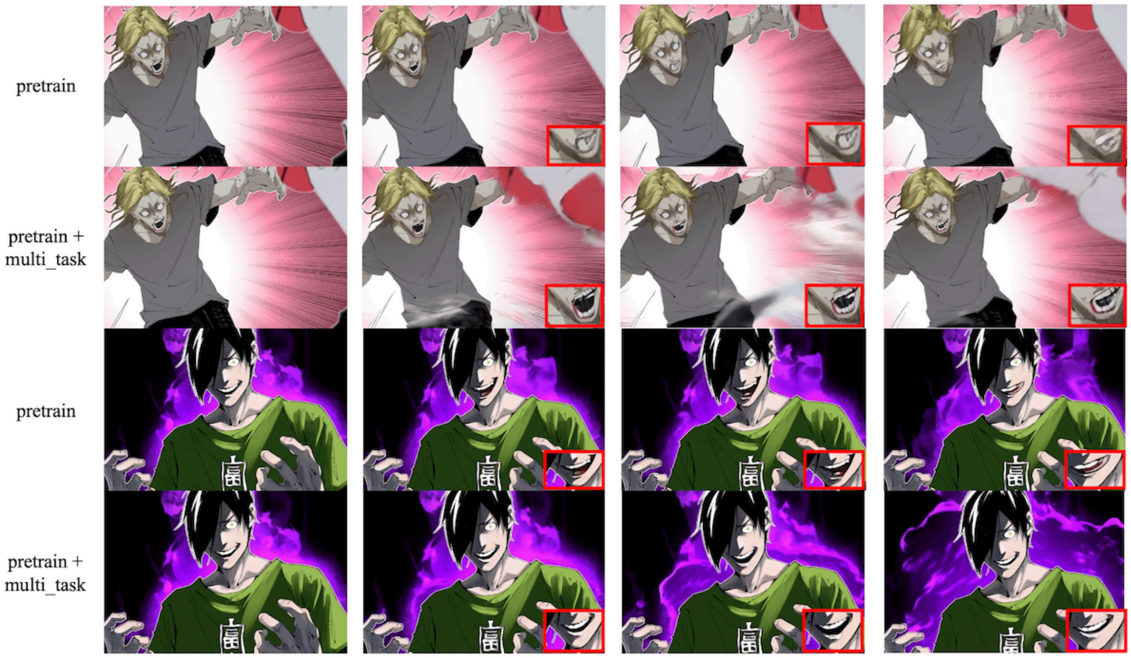


Figure 7. Comparison of results w/wo multi-task learning. The highlighted regions in red demonstrate significant improvements in stability and consistency after applying multi-task learning.

Low-resolution vs High-resolution

During the weak-to-strong training process, we observed that higher frame rates and resolutions enhance stability in visual details. As illustrated in Fig. 8, at 480P, facial features exhibit noticeable distortions, while at 720P, the model preserves both motion consistency and fine details. The higher resolution increases token representation for high-density areas, improving temporal consistency and overall content quality.

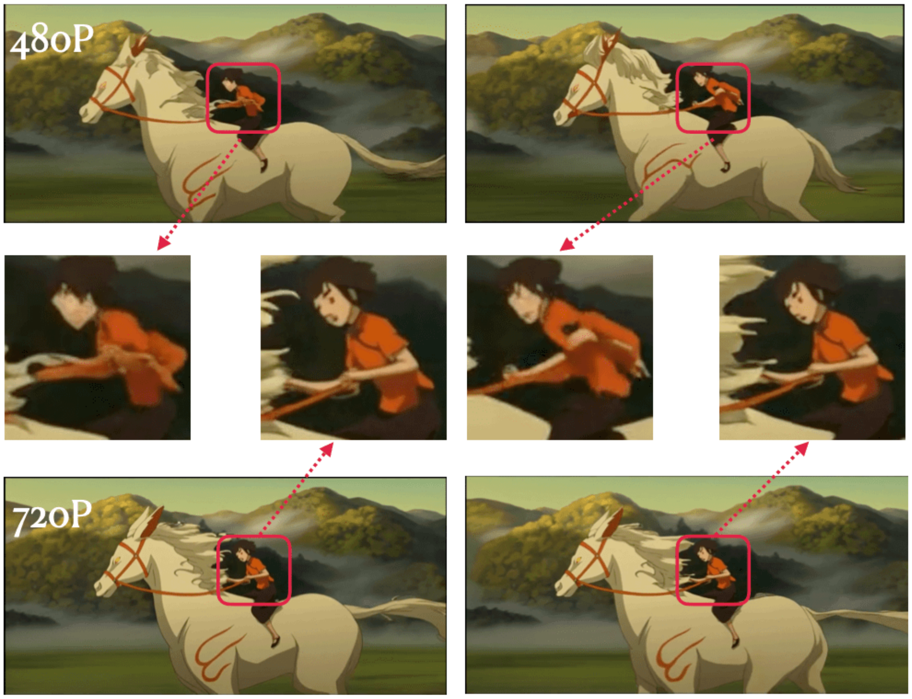
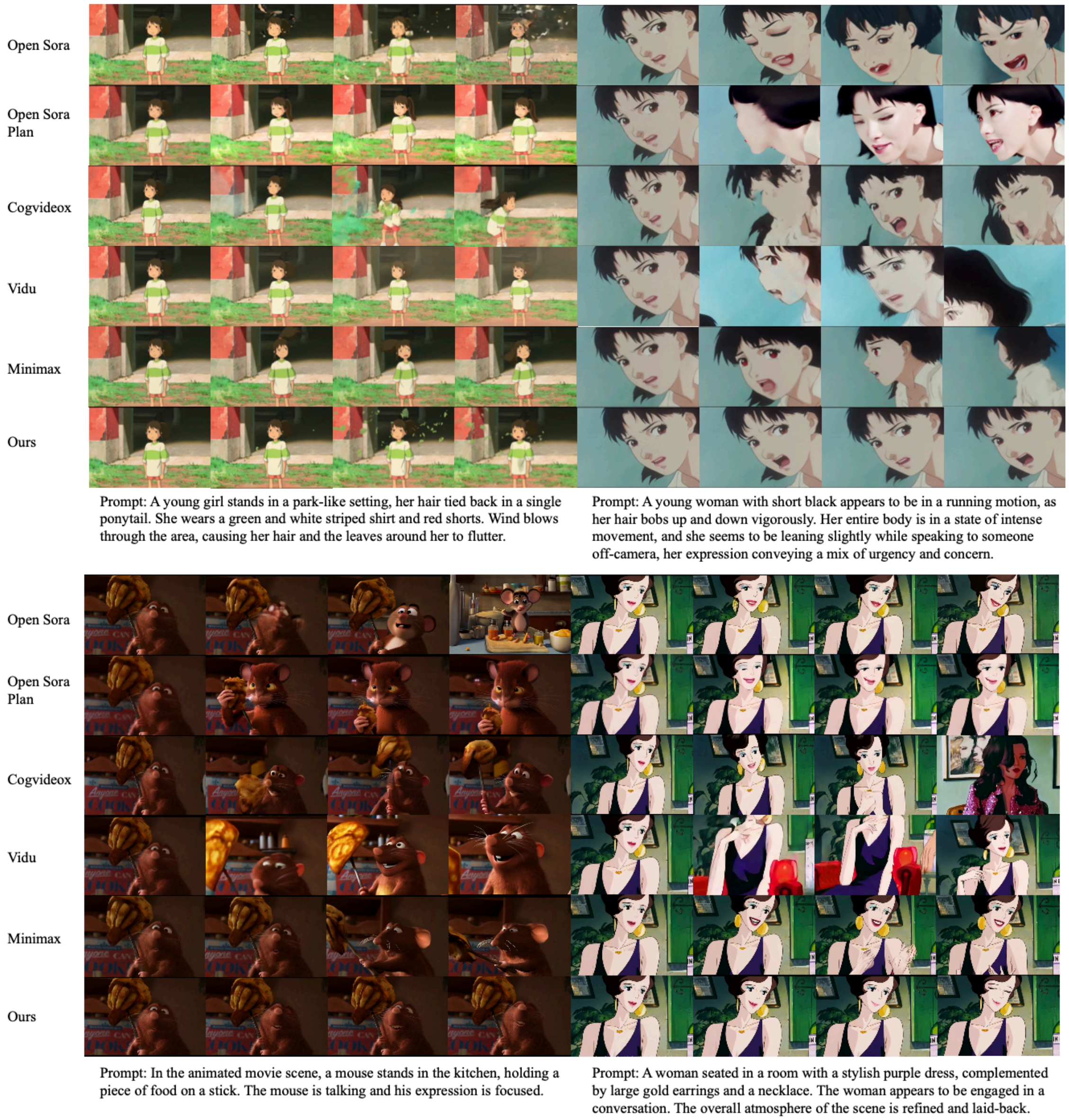


Figure 8. The figure compares the generation performance of 480P and 720P videos, highlighting that 720P achieves greater stability in the generation of details such as the character's facial features and hands.



Prompt: A young girl stands in a park-like setting, her hair tied back in a single ponytail. She wears a green and white striped shirt and red shorts. Wind blows through the area, causing her hair and the leaves around her to flutter.

Prompt: A young woman with short black appears to be in a running motion, as her hair bobs up and down vigorously. Her entire body is in a state of intense movement, and she seems to be leaning slightly while speaking to someone off-camera, her expression conveying a mix of urgency and concern.

Prompt: In the animated movie scene, a mouse stands in the kitchen, holding a piece of food on a stick. The mouse is talking and his expression is focused.

Prompt: A woman seated in a room with a stylish purple dress, complemented by large gold earrings and a necklace. The woman appears to be engaged in a conversation. The overall atmosphere of the scene is refined and laid-back.

Figure 9. Comparison of our method with others using the first frame in the leftmost column as the guiding condition. Existing methods often struggle with animation data, leading to issues such as character identity shifts, unnatural dynamics, and motion blur.

6. Conclusion

In this paper, our proposed AniSora, a unified framework provides a solution to overcoming the challenges in animation video generation. Our data processing pipeline generates over 10M high-quality training clips, providing a solid base for our model. Leveraging a spatiotemporal mask, the generation model can create videos based on diverse control conditions. Furthermore, our evaluation benchmark demonstrates the effectiveness of our method in terms of character consistency and motion smoothness. We hope that our research and evaluation dataset establish a new benchmark and inspire further work in the animation industry.

Despite the promising results, some artifacts and flickering issues are still present in our generated animation videos. In the future, we aim to develop a comprehensive automated scoring system specifically designed for animation video evaluation datasets, ensuring closer alignment with human subjective perceptions. Additionally, we plan to expand the current model architecture to incorporate guidance across multiple modalities, such as camera movements, trajectories, skeletal motions, and audio. To tackle the challenge posed by the limited availability of high-quality animation data, we will employ reinforcement learning techniques to further refine the model's performance.

References

1. [a, b]*Li S, Zhao S, Yu W, Sun W, Metaxas D, Loy CC, Liu Z (2021). "Deep Animation Video Interpolation in the Wild". In: CVPR.*
2. ^*Xing J, Liu H, Xia M, Zhang Y, Wang X, et al. ToonCrafter: Generative Cartoon Interpolation. arXiv preprint arXiv:2405.17933. 2024.*
3. ^*Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, et al. (2014). "Generative adversarial nets". NeurIPS. 27.*
4. ^*Kingma DP (2013). "Auto-encoding variational bayes". arXiv preprint arXiv:1312.6114. Available from: https://arxiv.org/abs/1312.6114.*
5. ^*Vaswani A (2017). "Attention is all you need". NeurIPS.*
6. [a, b]*Peebles W, Xie S (2023). "Scalable diffusion models with transformers". In: ICCV.*
7. [a, b]*Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127. 2023.*
8. [a, b, c, d, e]*Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072. 2024.*
9. [a, b, c]*Huang Z, He Y, Yu J, Zhang F, Si C, Jiang Y, Zhang Y, Wu T, Jin Q, et al. Vbench: Comprehensive benchmark suite for video generative models. In: CVPR; 2024.*

10. a, b Zheng Z, Peng X, Yang T, Shen C, Li S, Liu H, et al. Open-Sora: Democratizing Efficient Video Production for All [software]. 2024 Mar. Available from: https://github.com/hpcaitech/Open-Sora.

11. a, b PKU-Yuan Lab and Tuzhan AI etc. Open-Sora-Plan [software]. 2024 Apr. GitHub. doi:10.5281/zenodo.10948109.

12. ^ Wu Y, Wang X, Li G, Shan Y (2022). "AnimeSR: Learning Real-World Super-Resolution Models for Animation Videos". In: NeurIPS, 2022.

13. ^ Pan Z, Zhu Y, Mu Y. "Sakuga-42M Dataset: Scaling Up Cartoon Research". arXiv preprint arXiv:2405.07425. 2024.

14. ^ Chen TS, Siarohin A, Menapace W, Deyneka E, Chao HW, Jeon BE, Fang Y, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: CVPR; 2024.

15. a, b Wang Y, He Y, Li Y, Li K, Yu J, Ma X, Li X, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942. 2023.

16. ^ Liu Y, Cun X, Liu X, Wang X, Zhang Y, Chen H, Liu Y, et al. Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440. 2023.

17. ^ Zeng A, Yang Y, Chen W, Liu W (2024). "The Dawn of Video Generation: Preliminary Explorations with SORA-like Models". arXiv preprint arXiv:2410.05227.

18. a, b Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B, Ma C, et al. Movie Gen: A Cast of Media Foundation Models. arXiv preprint arXiv:2410.13720. 2024.

19. ^ Breakthrough (2024). PySceneDetect. Available from: https://github.com/Breakthrough/PySceneDetect.

20. ^ Baek Y, Lee B, Han D, Yun S, Lee H (2019). "Character Region Awareness for Text Detection". In: CVPR.

21. ^ princeton-vl (2020). RAFT. Available from: https://github.com/princeton-vl/RAFT.

22. ^ christophschuhmann (2022). "improved-aesthetic-predictor". Available from: https://github.com/christophschuhmann/improved-aesthetic-predictor.

23. a, b Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv preprint arXiv:2409.12191. 2024.

24. ^ Gupta A, Yu L, Sohn K, Gu X, Hahn M, Fei-Fei L, et al. (2023). "Photorealistic video generation with diffusion models". arXiv preprint arXiv:2312.06662. Available from: https://arxiv.org/abs/2312.06662.

25. ^ Yu L, Lezama J, Gundavarapu NB, Versari L, Sohn K, Minnen D, Cheng Y, et al. (2023). "Language model beats diffusion--tokenizer is key to visual generation". arXiv preprint arXiv:2310.05737. Available fro

m: https://arxiv.org/abs/2310.05737.

26. ^*Alexey D (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". arXiv preprint arXiv: 2010.11929. Available from: https://arxiv.org/abs/2010.11929.*
27. ^*Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024.*
28. ^*Salimans T, Ho J (2022). "Progressive distillation for fast sampling of diffusion models". arXiv preprint arXiv:2202.00512. Available from: https://arxiv.org/abs/2202.00512.*
29. a, b, c*Dai Z, Zhang Z, Yao Y, Qiu B, Zhu S, Qin L, Wang W (2023). "AnimateAnything: Fine-Grained Open Domain Image Animation with Motion Guidance". arXiv e-prints.*
30. ^*Kim T, Kim K, Lee J, Cha D, et al. Revisiting image pyramid structure for high resolution salient object detection. In: ACCV; 2022.*
31. ^*Wang M, Xing J, Liu Y (2021). "Actionclip: A new paradigm for video action recognition". arXiv preprin t arXiv:2109.08472. Available from: https://arxiv.org/abs/2109.08472.*
32. ^*Vidu. Available from: https://www.vidu.studio.*
33. ^*Minimax. "Available from: https://www.minimaxi.com."*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.