Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Exploring the Frontiers of Animation Video Generation in the Sora Era: Method, Dataset and Benchmark

Yudong Jiang [*†], **Baohan Xu** [*†], **Siqian Yang**[*†], **Mingyu Yin**[†], **Jing Liu**[†], **Chao Xu**, **Siqi Wang**, **Yidi Wu**, **Bingwen Zhu**, **Yue Zhang**, **Jinlong Hou**, **Huyang Sun**
Bilibili Inc.
{jiangyudong, xubaohan, yangsiqian, yinmingyu, liujing04, xuchao07, wangsiqi04, wuyidi, zhubingwen, zhangyue17, houjinlong, sunhuyang}@bilibili.com

Abstract

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 benchmark. 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, with specifically developed metrics for animation video generation. **Our entire project is publicly available on** *https://github.com/bilibili/Index-anisora/tree/main*

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 creating storyboards, generating keyframes, and inbetweening, making the process labor-intensive and time-consuming. Previous efforts[Siyao *et al.*, 2021; Xing *et al.*, 2024] 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[Goodfellow *et al.*, 2014], Variational Autoencoders[Kingma, 2013], and, more recently, transformer-based architectures[Vaswani, 2017; Peebles and Xie, 2023], 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[Blattmann *et al.*, 2023; Yang *et al.*, 2024]. 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 synthesis 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[Huang *et al.*, 2024]. For animation video generation, this challenge is amplified. Animation videos feature unique artist styles, including color and style that need to be consistent, even as characters undergo exaggerated motions and transformations. Traditional evaluation metrics are commonly used for real-world videos, which may not fully capture the consistency of the main characters and art style of this kind of video. Therefore, developing effective evaluation datasets and metrics customized for animation video generation is essential in this specialized field.

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 allows a more efficient creative process for both professional and amateur animation creators. Fig. 2 demonstrates some examples generated by our model under image-

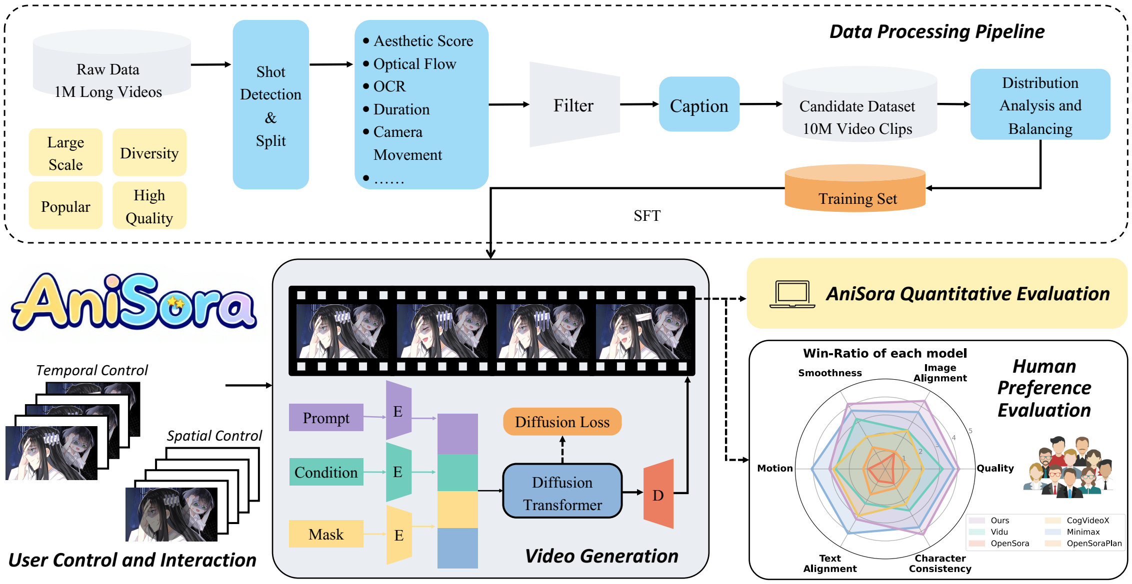

Figure 1: **Overview.** We propose **AniSora**, a comprehensive framework for animation video generation that integrates a high-quality animation dataset, a spatiotemporal conditional model, and a specialized animation video benchmark. The **Data Processing Pipeline** constructs a 10M video clip dataset 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 a **Quantitative Evaluation** with six objective metrics for assessing visual appearance and consistency. Additionally, **Human Preference Evaluation** confirms strong alignment with the proposed metrics. **AniSora** surpasses SOTA models, establishing a new benchmark for animation video generation.

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.

to-video conditions.

Additionally, we propose a benchmark dataset and evaluation metrics specifically designed for animation video evaluation. Currently, there is no existing evaluation dataset for this purpose, so we collected 948 animation videos across various categories and manually refined the prompts. Besides, existing evaluation standards struggle to effectively assess the quality of animation video generation. To address this gap, we have introduced innovative metrics that specifically assess animation video generation. These include character consistency, animation art style consistency, and distortion detection, which are crucial for maintaining the unique visual identity of animation videos.

Our contributions can be summarized as follows:

- We develop a comprehensive video processing system that significantly enhances preprocessing for animation 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.
- To the best of our knowledge, we for the first time released a benchmark dataset and evaluation metrics 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[Blattmann *et al.*, 2023; Yang *et al.*, 2024; Zheng *et al.*, 2024; Lab and etc., 2024] 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[Siyao *et al.*, 2021], AVC[Wu *et al.*, 2022]. While these datasets, collected from various animation movies, are helpful for interpolation and super-resolution tasks, they are limited by small size. More recently, Sakuga-42M[Zhenglin Pan, 2024] has been proposed with 1.2M clips. It has improved compared to previous datasets that only contained a few hundred clips. However, this remains insufficient for training video generation models, in contrast to general video data sets such as Panda-70M[Chen *et al.*, 2024] and InternVid-200M[Wang *et al.*, 2023]. 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[Liu *et al.*, 2023]. Besides, Huang et al. have proposed vbench for general video generation[Huang *et al.*, 2024]. The authors have released 16 evaluation dimensions and prompt suites. While these dimensions are still insufficient to comprehensively evaluate all aspects of animation video generation. Moreover, there is a notable absence of dedicated animation evaluation datasets, which limits the ability to benchmark models specifically designed for this genre. In [Zeng *et al.*, 2024], the authors have focused mainly on the performance of recent video generation models in various categories of datasets. Furthermore, they have also investigated some vertical-domain models like pose-controllable 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 [Polyak *et al.*, 2024]. In this section, we give a detailed description 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[Breakthrough, 2024] to divide raw 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[Baek *et al.*, 2019] can drop those clips with text overlay similar to end credits. Optical flow score [princeton vl, 2020] prevents those clips with still images or quick flashback scenes. Aesthetic score [christophschuhmann, 2022] 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 frames. Furthermore, we collected 0.5M high-quality animation videos along with their corresponding captions to create video-text pairs, which were used to fine-tune Qwen-VL2[Wang *et al.*, 2024]. After fine-tuning, the model provides more accurate descriptions of characters, scenes, and action details in animation content. After the steps mentioned above, about 10% clips (more than 10M clips) with captions can be retained in the training step.

In addition, since occupationally-generated animation videos typically have significantly higher production costs and quality compared to user-generated animation content, we fine-tuned Qwen-VL2 based on these data. This model is then utilized to filter higher-quality clips 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: Since there is currently no benchmark dataset specifically designed for animation content, we construct a benchmark dataset manually to compare the generation videos between our model and other recent researches. 948 animation clips are collected and labeled with different actions, e.g., talking, walking, running, eating, 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 fine-tuned Qwen-VL2 (mentioned above) at first, then is corrected manually to guarantee the text-video alignment. (More details in section 5 and 6)

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 diffusion transformer model. In section 4.2, we introduce a spatiotemporal mask module that extends the 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[Peebles and Xie, 2023] 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.

3D Casual VAE used in video generation frameworks[Gupta *et al.*, 2023; Yu *et al.*, 2023]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

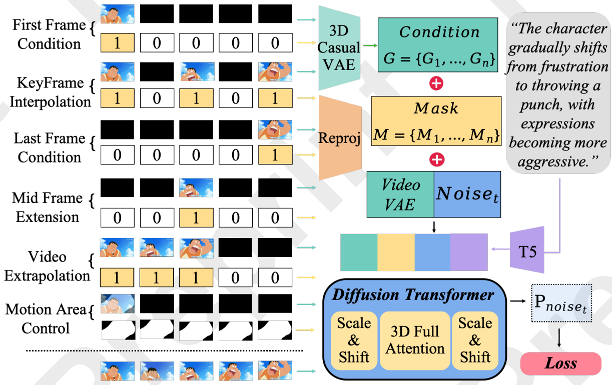

Figure 3: **Method.** This figure illustrates the Masked Diffusion Transformer framework for animation video generation, designed to support various spatiotemporal conditioning methods for precise and flexible animation control. A 3D Causal VAE compresses spatial-temporal features into a latent representation, 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 spatial-temporal dependencies. This framework seamlessly integrates keyframe interpolation, motion control, and mid-frame extension, simplifying animation production and enhancing creative possibilities.

al. [Yang *et al.*, 2024] 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 [Alexey, 2020]. 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 We propose 3D full attention module for spatial and temporal modeling considering the great success of long-context training in LLM[Dubey *et al.*, 2024] and foundation video generation model[Yang *et al.*, 2024; Polyak *et al.*, 2024].
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 $\epsilon \sim \mathcal{N}(0, I)$. 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 [Salimans and Ho, 2022], where $v = \sqrt{1-\alpha_t}\, x_0 - \sqrt{\alpha_t}\, \epsilon$ and the loss becomes

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

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 Our framework enables precise control over motion regions addressing the limitations of text-based control in these aspects. This approach enhances artists' control over video content, allowing them to express their creativity while significantly reducing their workload.

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:

$$X = Concat(Noise_t, Reproj(M), G, T5) \qquad (1)$$

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

Motion Area Condition

This framework can also support spatial motion area conditions inspired by Dai et.al[Dai *et al.*, 2023]. 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, 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 corresponding masks. Using a foreground detector by Kim et.al [Kim *et al.*, 2022], 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.

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 in the appendix 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.

In practice, we also employed several other effective training strategies, such as *weak to strong training*, *generated subtitle removal*, and *temporal multi-resolution training*. Detailed training procedures can be found in the appendix.

5 Benchmark

To evaluate the effects of the animation video generation models, we build a comprehensive benchmark dataset, as mentioned in section 3. To give a fair comparison of different methods, we define 6 basic dimensions to describe the quality of the generated videos. Then, we introduce the human annotation and evaluation metrics, which are partly based on the annotation results.

5.1 Evaluation Dimensions

In fact, the essential concepts for evaluating a animation video generation model are *visual appearance* and *visual consistency*. Visual appearance describes the basic quality, which is only concerned with the generation of videos themselves, including visual smoothness, visual motion, and visual appeal, while visual consistency considers the text-video, image-video, and character consistency, respectively.

Inspired by Vbench[Huang *et al.*, 2024], we first adopted a similar approach to evaluate visual appearance and consistency, such as calculating the clip score between frames to analyze distortions and using aesthetic scores to estimate aesthetic quality. However, we found that these scores often showed significant discrepancies from subjective human experiences on animation data and there was no clear distinction between different methods. As shown in Fig.4, the bottom video received a lower score due to its larger motion amplitude. However, the top video contains visible distortions that are easily noticeable to humans, yet these are not reflected in the scoring metrics.

Therefore, we adopted a regression-based approach for certain dimensions to learn human scoring standards. A detailed description of six evaluated metrics is given as follows. These evaluation criteria are specifically designed for animation data and align closely with human subjective experience.

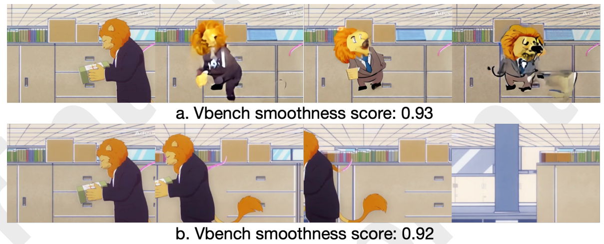

Figure 4: Video generated by Opensora-V1.2 (top) and Vidu (bottom). The top video received a higher score despite containing noticeable distortions.

5.2 Human Annotation

In the experiment, each case in the benchmark has 6 generation video clips with different participant models. 20 expert volunteers give their rating (1-5, 5 is the best) from above 6 dimensions, the detail scores are shown in Section 6.1.

5.3 Visual Appearance

We evaluate the basic quality of a video clip from three aspects: visual smoothness, visual motion, and visual appeal. We constructed an evaluation model training set to develop an evaluation model aligned with human scoring. We generated 5000 animated video clips using various models described in Section 6.1. These samples were manually annotated following the approach outlined in Section 5.2, serving as the ground truth for training the evaluation model.

Visual Smoothness Our goal was to learn the human standards to evaluate video smoothness and to be able to identify animated videos with distortions. Thus, we trained a model to regress human-generated scores to avoid the impact of motion on visual content. The training set consists of generated clips and annotations, mentioned in Section 5.3. To enhance the robustness of the model, we also incorporated hundreds of anime videos as the highest-scoring examples.

Visual Motion We employ a model based on the Action-CLIP [Wang *et al.*, 2021] framework to train a motion-scoring model that evaluates the magnitude of the primary motion in animation videos. About 2 thousand 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.

$$S_{motion} = Cos(MCLIP(V), MCLIP(T_m)), \quad (2)$$

where $MCLIP$ denotes the finetuning action model. V represents the generation video and T_m denotes the designed motion prompt. (Details is provided in supplementary)

Visual Appeal We define visual appeal score to reflect the basic effects of video generation. As discussed in Section 5.1, the aesthetic score models used in previous studies were trained on real-world datasets such as LAION-5B, simply calculating aesthetic scores on animation data does not provide sufficient differentiation, as the results from various methods are indistinguishable.

Using the key frame extraction method to collect the key frames in the video first, then we train a regress model on the evaluation model training set to learn human aesthetic standards. The formulation shows as follows:

$$S_{appeal} = Aes(SigLIP(I_{0,1,...,K})) \quad I_i \in KeyFrm(V), \tag{3}$$

where $KeyFrm$, $SigLIP$ and Aes denote the key frame extraction method, feature encoder method and aesthetic evaluation method, and K denotes the number of the keyframes.

5.4 Visual Consistency

Three factors are considered to evaluate the visual consistency of the generation video: text-video, image-video, and character consistency, respectively.

Text-video Consistency To evaluate the text-video consistency, we finetune the vision encoder and the text encoder modules with a regression head according to animation video-text pairs according to the training set in Section 5.3. The formulation is shown as follows:

$$S_{tvc} = Reg(E_v, E_t), \tag{4}$$

where Reg denotes the regression head, and E_v, E_t denote the vision and text encoder.

Image-video Consistency In the I2V situation, the participant image, as an input factor, should ensure that its style is consistent with the generated videos. Similar to text-video consistency, we combine a vision encoder with a regression head to evaluate the score. The model is also fine-tuned on the training set in Section 5.3. The formulation lists as follows:

$$S_{ivc} = Reg(E_v(V), E_v(I_p)), \tag{5}$$

where V and I_p denote the participant video clip and the input image.

Character Consistency Character consistency is a crucial factor in animation video generation. When the character generated by the protagonist in the animation changes, even if the quality of the video is great, it still has the risk of infringement. Hence, we design a set of procedures including detection, segmentation, and recognition. We apply GroudingDino [Ren *et al.*, 2024] and SAM [Ravi *et al.*, 2024] to achieve character mask extraction for each frame in the videos. Then, we finetune a BLIP-based model [Li *et al.*, 2022] to establish connections between each mask and the animation IP character. In detail, thousands of source video clips with their characters labeled are treated as training sets to obtain and store the reliable features from BLIP-based model. In the evaluation step, we get the score of character consistency by calculating the cosine similarity between the generated and stored character's features.

$$S_{IPc} = \frac{1}{S}\sum_{i}^{S} Cos(BLIP(M_i), fea_c), \tag{6}$$

where S denotes the number of sample frames, M_i denotes the mask obtained from GroudingDino and SAM methods, and fea_c denotes the stored character's features.

6 Experiment

6.1 Benchmark Evaluation

In this section, we give both objective and human evaluation results of our benchmark. Six recent i2v models are involved in our evaluation: Open-sora [Zheng *et al.*, 2024], Open-sora-plan [Lab and etc., 2024], Cogvideox [Yang *et al.*, 2024], Vidu-1.5 [Vidu,], Minimax-I2V01 [Minimax,] and Anisora(ours). Tab. 1 gives the detailed scores from 6 dimensions in the benchmark evaluation and the overall scores of the human evaluation. We observe that our model performs better than the other five methods on most dimensions, especially on visual smoothness and character consistency, except on the visual motion dimension. These mainly because we conduct a thorough assessment of the balance between generation quality and motion magnitude, and find most generation clips with big motion results in distortion or unnatural segments. In order to prove our benchmark is applicable to anime scenarios, we also evaluated our benchmark dataset using seven relevant dimensions in VBench benchmark. Due to the space limitation, we outline the results in Appendix. We observed that certain dimensions, including Motion Smoothness, Aesthetic Quality, I2V Background, and Overall Consistency, lacked sufficient discriminative power. In particular, some poorly generated results received higher scores than the ground truth, highlighting a discrepancy that fails to accurately capture human perception and experience.

Fig. 5 illustrates the detailed correlations among 6 dimensions between human evaluation and benchmark results. Obviously, they are highly correlated with each other.

6.2 Spatiotemporal Mask Module

Frame Interpolation Tab. 1 presents the results of different interpolation settings on benchmark dataset. 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 a 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 character consistency. We also observed from the motion 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 the appendix.

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 [Dai *et al.*, 2023], we conducted the comparison of motion mask precision in Tab. 2. 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

Models	Human Evaluation	Visual Smooth	Visual Motion	Visual Appeal	Text-Video Consistency	Image-Video Consistency	Character Consistency
Vidu-1.5	60.98	55.37	**78.95**	50.68	60.71	66.85	82.57
Opensora-V1.2	41.10	22.28	74.9	22.62	52.19	55.67	74.76
Opensora-Plan-V1.3	46.14	35.08	77.47	36.14	56.19	59.42	81.19
CogVideoX-5B-V1	53.29	39.91	73.07	39.59	67.98	65.49	83.07
MiniMax-I2V01	69.63	69.38	68.05	**70.34**	**76.14**	78.74	89.47
AniSora(Ours)	**70.13**	**71.88**	48.45	65.38	74.26	**82.66**	**94.88**
AniSora(Interpolated Average)	-	70.78	53.02	64.41	73.56	80.62	91.59
AniSora(KeyFrame Interpolation)	-	70.03	58.1	64.57	74.57	80.78	91.98

Table 1: Benchmark Evaluation Results.

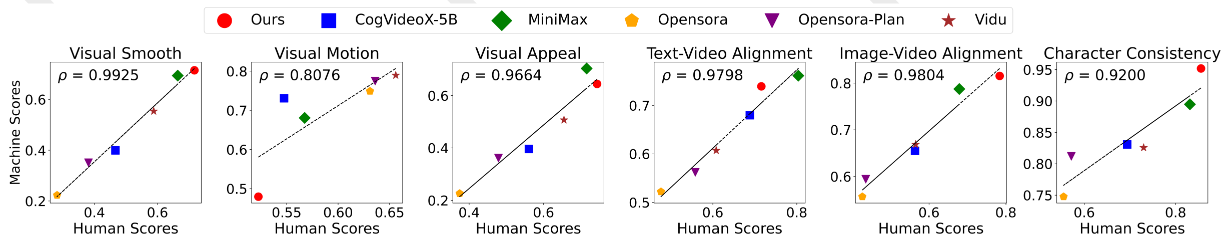

Figure 5: Human Evaluation and Benchmark Results Alignment.

of control. This may due to the effective prompt-based guidance for the main subject. Motion mask guidance examples are shown in the appendix.

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

Table 2: Comparison of motion mask precision.

6.3 Animation Video Training

2D and 3D Animation Analysis using QWEN2 [Wang *et al.*, 2024] shows that 2D samples account for 85% of our data set, yet the quality of 3D animation generation consistently exceeds that of 2D. Benchmark evaluations in the appendix confirm 3D animations demonstrate superior visual appearance and consistency, a phenomenon unique to 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.

Multi-Task Learning 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 the appendix, 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.

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 demonstrated in the appendix, 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.

7 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. Besides, we are going to evaluate more models on our benchmark, providing valuable insights for model optimization.

Despite promising results, some artifacts and flickering are still present in the generated videos. In future work, we plan to integrate reinforcement learning with our evaluation benchmark to generate higher-quality videos.

Contribution Statement

Yudong Jiang, Baohan Xu, and Siqian Yang contributed equally to this work and share first authorship. Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, and Jing Liu contributed the main ideas and execution of the research and are considered the main contributors. All authors participated in discussions, contributed to the writing, and approved the final manuscript.

Acknowledgments

We sincerely thank Qiang Zhang, Xinwen Zhang, and Xingyu Zheng for their valuable support and insightful suggestions on improving the manuscript.

References

[Alexey, 2020] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[Baek *et al.*, 2019] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *CVPR*, 2019.

[Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[Breakthrough, 2024] Breakthrough. Pyscenedetect. https://github.com/Breakthrough/PySceneDetect, 2024.

[Chen *et al.*, 2024] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024.

[christophschuhmann, 2022] christophschuhmann. improved-aesthetic-predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[Dai *et al.*, 2023] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv e-prints*, 2023.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, et al. Generative adversarial nets. *NeurIPS*, 27, 2014.

[Gupta *et al.*, 2023] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, et al. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[Huang *et al.*, 2024] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

[Kim *et al.*, 2022] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, et al. Revisiting image pyramid structure for high resolution salient object detection. In *ACCV*, 2022.

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

[Lab and etc., 2024] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, April 2024.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[Liu *et al.*, 2023] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, et al. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[Minimax,] Minimax. https://www.minimaxi.com.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[Polyak *et al.*, 2024] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[princeton vl, 2020] princeton vl. Raft. https://github.com/princeton-vl/RAFT, 2020.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, et al. Sam 2: Segment anything in images and videos, 2024.

[Ren *et al.*, 2024] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, et al. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.

[Salimans and Ho, 2022] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[Siyao *et al.*, 2021] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, 2021.

[Vaswani, 2017] A Vaswani. Attention is all you need. *NeurIPS*, 2017.

[Vidu,] Vidu. https://www.vidu.studio.

[Wang *et al.*, 2021] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[Wang *et al.*, 2023] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wu *et al.*, 2022] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. In *NeurIPS*, 2022.

[Xing *et al.*, 2024] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, et al. Tooncrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024.

[Yang *et al.*, 2024] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[Yu *et al.*, 2023] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[Zeng *et al.*, 2024] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024.

[Zheng *et al.*, 2024] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, et al. Open-sora: Democratizing efficient video production for all, March 2024.

[Zhenglin Pan, 2024] Yuxuan Mu Zhenglin Pan, Yu Zhu. Sakuga-42m dataset: Scaling up cartoon research. *arXiv preprint arXiv:2405.07425*, 2024.