WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Text and Image to Video Generation using Stable Diffusion and RIFE Frame Interpolation

Sheela Verma[1], Dharini Sonwane [2], Himanshu Sahu[3], Neelesh Kumar Agashe[4], Nishant Sakesh[5]

[1]Assistant Professor, Department of Computer Science, Bhilai Institute of Technology Raipur, Chhattisgarh, India
[2,3,4,5]Student, Department of Computer Science, Bhilai Institute of Technology Raipur, Chhattisgarh, India

ABSTRACT :

This project examines a new approach to generating short animated video sequences from a single static image. This is guided by the input request of text and uses the feature of stable diffusion for translation into images (real-time intermediate flow estimation) to improve temporal coherence and smoothness. The central challenge is taking still photos with dynamic narrative sheets only through AI-controlled manipulation and textual instructions. Our methodology uses a multi-stage process: segment arbitrary models (SAM) are used to isolate the main subjects, followed by iterative frame generation via the IMG2IMG function of stable diffusion. These generated frames are conditioned against many development text demands generated by custom algorithms to effectively introduce subtle and consistent changes in motion, perspective, and environmental contexts. To alleviate the inherent temporal contradictions available in stable diffusion-based animation, we recommend that pregnancy techniques be synthesized and life can be integrated to achieve more liquid transitions and smooth visual experiences. The generated video sequences are qualitatively assessed for narrative consistency and visual loyalty, which is the focus of range influence on perceived smoothing. This study provides a pipeline to transform static images into dynamic video narratives with accessible and powerful AI tools. This highlights the possibilities for creative storytelling and the challenges of achieving top-class temporary video editions with dedicated video models. Rife integration is a key innovation in combating smoothing restrictions on stable diffusion in this context

1. Introduction

The advent of powerful generative models of artificial intelligence has revolutionized the creation of digital content, particularly in the field of image and video integration. While text-to-image and dedicated video generation models have garnered significant attention, the task of transforming a static image into a dynamic video narrative, guided by textual instructions, presents a unique set of challenges and opportunities. This project delves into this intriguing domain, exploring a methodology that leverages the strengths of established image generation models, specifically Stable Diffusion, and integrates a state-of-the-art frame interpolation technique, RIFE (Real-Time Intermediate Flow Estimation), to generate compelling animated video sequences from single still images and descriptive text prompts.

The ability to animate still images holds considerable potential across various applications. In storytelling, it offers a means to bring photographs and illustrations to life, adding a layer of dynamism and engagement that static visuals often lack. This allows you to create subtle movements and develop scenes from a single source image. Furthermore, animated images that arise from existing images can improve understanding in areas such as education and marketing, and can more effectively attract viewers.

However, generating coherent and visually appealing animations from still images using AI is not a trivial task. Existing methodologies often struggle with maintaining temporal consistency, resulting in jerky or unnatural transitions between generated frames. While Stable Diffusion has proven highly effective in image generation and image-to-image translation, its direct application to video creation from a single image and evolving prompts can lead to visual discontinuities.

To address these limitations, this project proposes a novel methodology that combines the creative power of Stable Diffusion with the temporal smoothing capabilities of RIFE. Through it, we want to create a smoother, visually adjacent video board based on carefully manufactured text conferences, and then interpolating interpolation interpolation frames. This approach attempts to bridge the gap between the static nature of individual images and the dynamic requirements of video counting.

2. Literature Review

2.1. Overview of Text and Image to Video Generation.

Text and image to video generation involves creating dynamic video content from static prompts such as natural language descriptions or images. This field combines natural language processing, computer views, and generative models. The typical pipeline begins with generating one or more images that align with the text prompt, followed by synthesizing intermediate frames to form a temporally smooth and visually coherent video. Recent advances in

deep learning, especially in diffusion models and neural frame interpolation, have significantly improved the feasibility and quality of such video synthesis.

2.2. Key Components from Text and Image to Video System.

- Text-to-Image Generation: Models like Stable Diffusion generate high-quality images based on text prompts using latent diffusion techniques, balancing quality and computational efficiency.
- Consistency between images: Maintaining consistency between frameworks requires immediate engineering (consistent input requests for frame generation) or fine-tuning models for time recognition.
- Frame Interpolation: RIFE (Real-Time Intermediate Flow Estimation) is used to generate smooth transitions between keyframes by synthesizing intermediate frames, enabling fluid video output from sparse image inputs.

2.3. Application and use case of Text and Image to Video Generation This

technology has growing applications across various domains:

- Content Creation: Automating content video in storytelling, marketing, and social media
- Education and Training: Generating visual explanations or animated educational content from text input.
- Film Pre-Visualization: Help creators to quickly plot visual scenes from scripts and storyboards.
- Virtual Reality and Gaming: Enabling dynamic scene generation based on player input or narrative prompts.

3. Methodology

The proposed system for animating a single static image using text input requests integrates object segmentation, frame generation, and time interpolation.

3.1 Initial Object Segmentation with SAM

The process begins with segmenting the primary object using the Segment Anything Model
(SAM), which produces a high-quality binary mask. This segmentation, guided by user input (clicks, bounding boxes, or coarse masks), ensures focused manipulation and visual consistency throughout the animation.

3.2 Iterative Frame Generation via Stable Diffusion and Prompt Evolution

Use a stable diffusion IMG2IMG pipeline for repeated frame synthesis. All new frameworks are conditioned on previous editions and a set of text requests generated by custom strategies. This strategy includes: Manage coherent output length and voting for parameters.

3.3 Temporal Smoothing with RIFE

To address temporal inconsistencies, we apply RIFE for high-quality frame interpolation. By generating intermediate frames between Stable Diffusion outputs, we achieve smoother transitions. We experiment with varying interpolation factors to balance quality and computational cost.

3.4 Final Video Assembly

Frames, both generated and interpolated, are compiled into a video using OpenCV, encoded into formats like MP4 (H.264/HEVC) with optimized frame rates and quality settings.

3.5 Challenges and Mitigation

We anticipate challenges including long-term coherence, occlusion handling, and computational efficiency. Solutions include advanced prompt engineering, potential inpainting for occlusion correction, and GPU-accelerated processing to optimize performance.

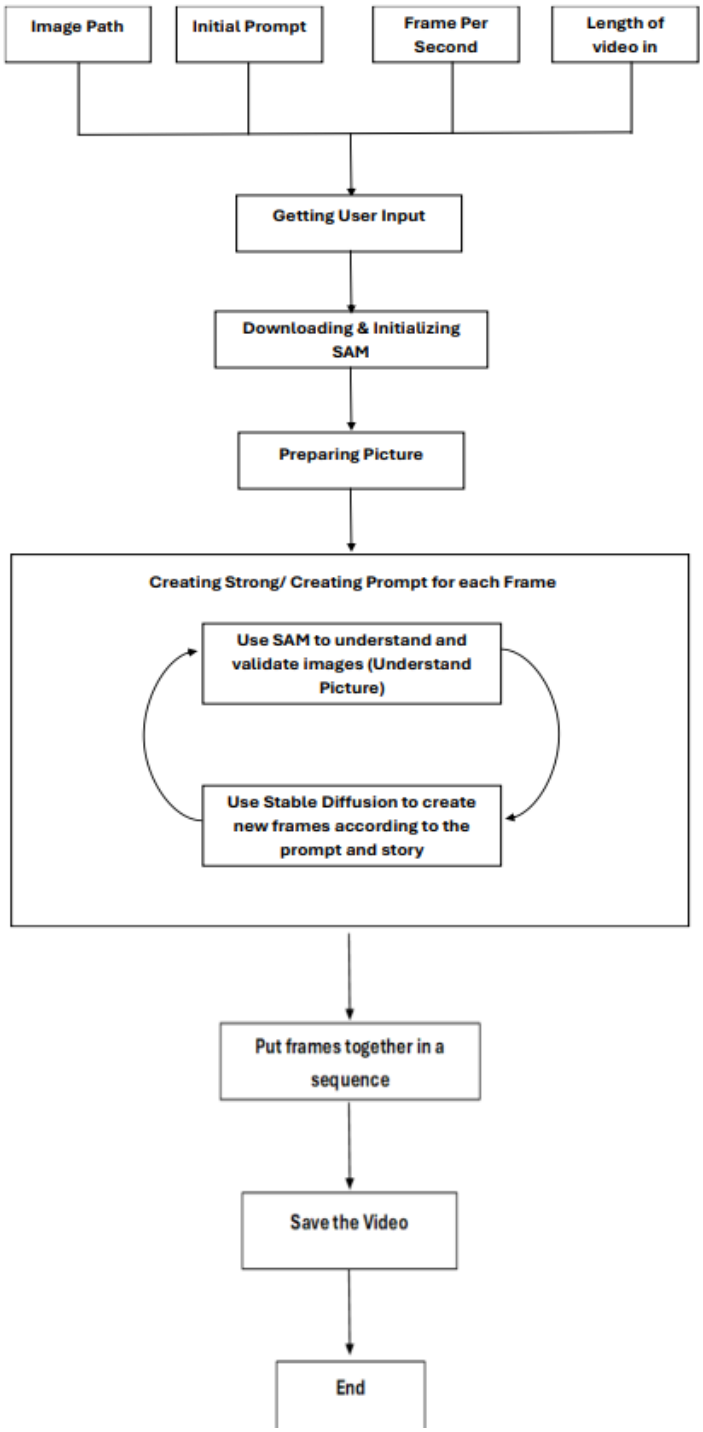

Fig1. Workflow

4. Results and Analysis

The proposed pipeline was tested with a variety of static images combined with narrative textual tasks to assess their ability to generate smooth, visually coherent sequences. The assessment focused on three key criteria: visual quality, temporal consistency, and narrative consistency.

4.1. Qualitative results

The animation sequences demonstrated strong preservation of object identity and stylistic coherence across frames, particularly when using well-defined masks from the SAM segmentation phase. The evolving prompts effectively guided subtle transformations in perspective, motion, and environmental conditions, enabling the creation of mini visual narratives from a single input image.

Before applying RIFE, the animations exhibited slight frame-to-frame jitter and visual artifacts— common challenges with iterative generation using Stable Diffusion's img2img. These inconsistencies are more clear with some transitions that have been made longer, although minimally in a minimal sequence.

After integration of Rife, a significant improvement in smoothness was observed. Interpolation of the intermediate frame reduced liquid movement and sudden transitions, which resulted in a more cinematic and natural experience. Improved continuity allows for subtle effects (leaf drift, slow camera push-in) that are thought to be realistic and attractive.

4.2 Comparative Visual Analysis

Metric	**Without RIFE**	**With RIFE**
Frame-to-frame jitter	Moderate	Low
Temporal smoothness	Fair	High
Narrative consistency	Medium-High	High
Visual artifacts	Occasional	Rare
Viewer feedback (subjective)	"Impressive but slightly choppy"	"Smoother and more immersive"

Visual inspection confirms that RIFE successfully bridges the temporal gaps that occur in traditional generated frame methods.

4.3 Performance Consideration

Additional framework interpolation steps increase computing costs due to compromise, but compromise is justified by improved perception of smoothness and continuity. Modular type pipelines allow flexibility in performance and quality development by adapting interpolation coefficients.

Output

Fig. 2

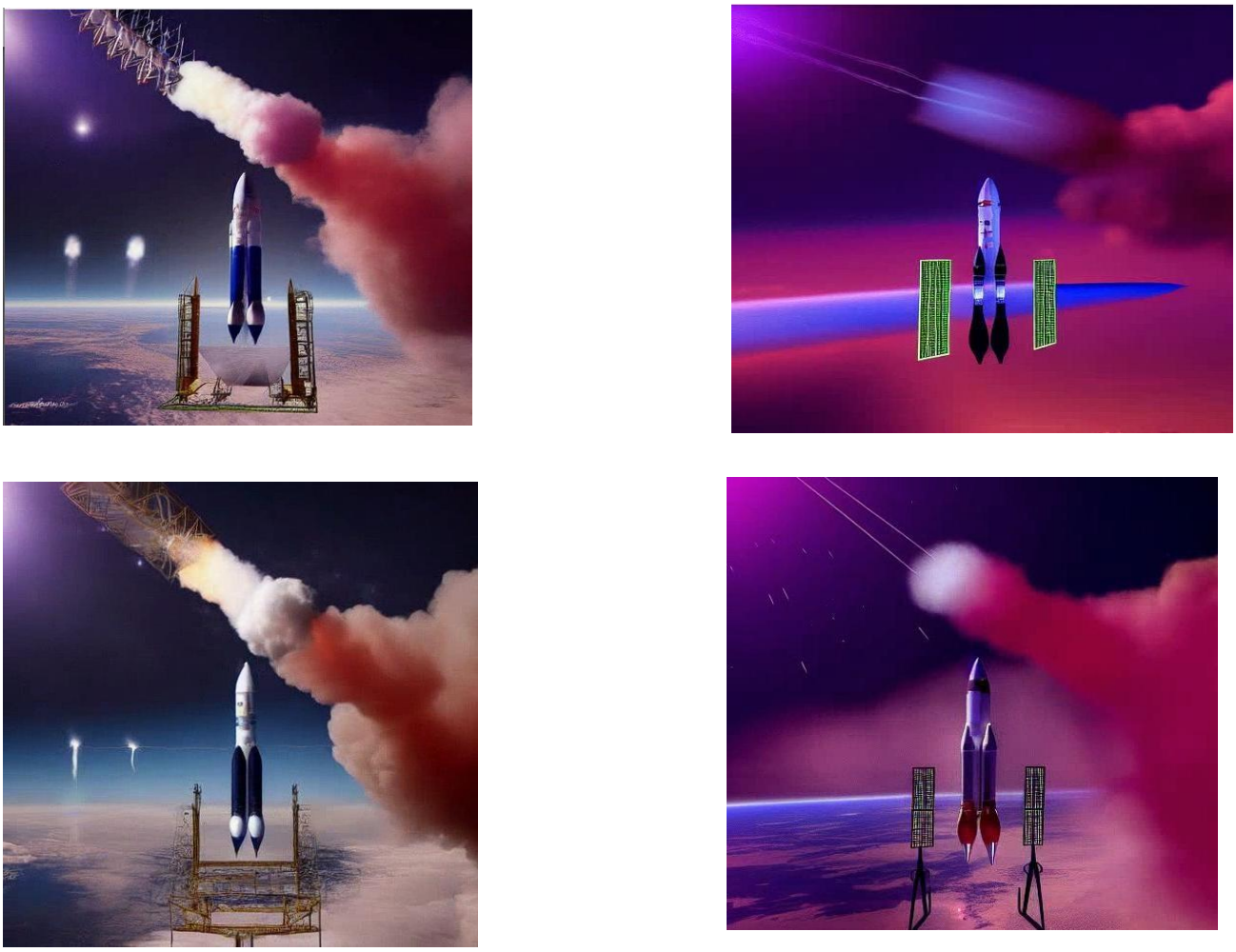

Fig 3

5. Conclusion

This project presents a new and effective methodology for utilizing static images that help develop text input requests. This is a stable spread of iterative frame generation combined by SAM for accurate object segmentation and time interpolation. The proposed pipeline handles the most important limitations of the current image animation approach, particularly with the challenge of maintaining temporal and narrative consistency. Input prompt strategies under development lead to smooth, semantically induced visual transitions, but significantly improve the continuity of intense movements, leading to liquid and attractive video sequences. The pipeline demonstrates the powerful possibilities of applications such as digital storytelling, artistic creation, visual communication, and offers a low but powerful alternative to complex models for video generation. The project highlights the creative possibilities that will be unlocked in innovative ways by a combination of existing AI tools and form the basis for future research into AI-controlled video integration.

1. Improved input prompt generation: Integration of enhanced NLP, user tracks, and instant feedback loops for richer stories.
2. Temporary coherence and visual quality improvements: Improve the use of temporal diffusion models, interpolation of the extension framework, object tracking, and improved visual consistency.
3. Advances in creative possibilities: Multi-object animation, style transmission, interactive generation, and audio synchronization for richer storytelling.
4. Optimization and Efficiency: Model and distributed processing inspections for improved performance and accessibility.
5. Evaluation and Benchmarks: Benchmarks for quantitative metric development, implementation of user research, and other methods for assessing quality and coherence generation animation.

REFERENCES

1. "DALL-E: Creating Images from Text" by OpenAI.
2. "Zero-Shot Text-to-Image Generation" by Aditya Ramesh et al.
3. "High-Resolution Image Synthesis with Latent Diffusion Models" by Robin Rombach et al.
4. "RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation" by Xiangyu Xu et al.

5. "Image Super-Resolution Using Deep Convolutional Networks" by Chao Dong et al.
6. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution" by Justin Johnson et al.
7. "Generative Adversarial Networks" by Ian Goodfellow et al.
8. "Attention Is All You Need" by Ashish Vaswani et al.
9. "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation" by Tao Dai et al.
10. "Video Frame Interpolation via Adaptive Separable Convolution" by Simon Niklaus et al.