**Introduction**

Generating videos from a single image and a text description is a challenging task in computer vision, requiring both realistic image synthesis and temporal coherence across frames. This project explores an approach based on **Stable Diffusion** to create videos from an initial image and a textual description, incorporating techniques for interpolation and sequential image generation.

The proposed method follows three main steps. First, **key images** are generated using a diffusion model: each image is produced based on the previous image and a progressive evolution of the initial text, ensuring a coherent scene transition. Next, an **interpolation** step is applied between these key images to create smooth transitions, followed by refinement using **Stable Diffusion** to maintain a natural appearance. Finally, special attention is given to text structuring: the text is decomposed into multiple sentences, each serving as a prompt to generate a new image from the previous one, allowing the visual theme to evolve gradually throughout the video.

This project aims to explore the potential of diffusion models for video generation from multimodal inputs, focusing on visual quality, temporal consistency, and semantic alignment with the text.

**1. Generation of Key Images**

The first step in our video generation pipeline consists of generating **key images** that serve as anchor points for the final video. This process leverages **Stable Diffusion**, a powerful generative model capable of synthesizing high-quality images from text and visual inputs.

**1.1 Methodology**

The generation of key images follows an iterative process, where each new image is created based on the previous one and an evolving textual description. The process can be broken down into the following steps:

1. **Initial Image Generation**:
   * The first key image is generated using **Stable Diffusion**, conditioned on both the given starting image and the input text.
   * This ensures that the first generated image aligns with the semantic content of the text while preserving relevant features from the input image.
2. **Sequential Image Generation**:
   * Instead of generating all frames independently, we adopt a **progressive approach** to ensure continuity.
   * Each subsequent key image is generated using **the previous key image and the next segment of the text** as input to Stable Diffusion.
   * This step-by-step process allows for a smooth transition between different visual states while maintaining coherence with the evolving narrative.

**1.2 Objective and Challenges**

The key image generation step is crucial for ensuring **semantic consistency** and **visual quality** in the final video. However, several challenges must be addressed:

* **Maintaining Coherence:** Ensuring smooth transitions between key images without abrupt changes in scene composition or style.
* **Balancing Consistency and Variation:** Each image should evolve according to the text while preserving key visual features from the previous image.
* **Minimizing Artifacts:** Preventing distortions, inconsistencies, or unwanted transformations introduced by the generative model.

To improve continuity and realism, the next step of our pipeline applies **interpolation** to refine transitions between key images.

**2. Interpolation and Generation of Intermediate Images**

Once the key images have been generated, the next step is to create smooth transitions between them. To achieve this, we use an **interpolation method** followed by a refinement step with **Stable Diffusion**, ensuring that each frame in the video remains visually meaningful.

**2.1 Interpolation Process**

To generate an intermediate image between two key images, we first compute the **pixel-wise average** of both images. However, this naive interpolation often results in an image that appears unnatural or unrecognizable.

To address this issue, we apply **Stable Diffusion** to the interpolated image using the prompt *"a natural image"*. This process enhances the visual coherence of the image, transforming the raw interpolation into a more realistic and meaningful frame.

We then recursively refine the transitions by **iteratively adding new interpolated images** between existing frames. At each iteration:

1. A new interpolated image is generated between each pair of consecutive images.
2. Stable Diffusion is applied to each interpolated image to refine its appearance.
3. This process is repeated, progressively increasing the number of frames and ensuring a smooth transition.

**2.2 Direct Frame Interpolation and its Limitations**

An alternative method was tested, where multiple interpolated images were generated first, and Stable Diffusion was then applied to each frame in a single step. However, this approach resulted in less coherent videos due to:

* **Uncontrolled distortions**: Stable Diffusion modifies each frame independently, leading to inconsistencies between consecutive images.
* **Accumulation of discrepancies**: Each interpolation step introduces noise and deviation, causing successive images to drift apart, leading to discontinuities in the video.
* **Loss of temporal consistency**: Since Stable Diffusion introduces randomness in each transformation, consecutive frames may diverge unpredictably.

However, this approach has a **computational advantage**: by applying Stable Diffusion to all interpolated images at once, the process can be **parallelized**, making it significantly faster than the progressive refinement method. While this improves efficiency, it comes at the cost of reduced video continuity and coherence.

**2.3 Latent Space Interpolation Using a VAE**

An alternative approach to generating intermediate images is to perform interpolation directly in a **latent space** using a **Variational Autoencoder (VAE)**. This method eliminates the need for Stable Diffusion in the interpolation process while maintaining smooth transitions.

**Methodology**

1. **Encoding Images into Latent Space**:
   * Each key image is first mapped into a lower-dimensional latent representation using the encoder of a **VAE**. This latent space captures the high-level features of the image while reducing noise and redundancy.
2. **Latent Space Interpolation**:
   * Instead of performing pixel-wise interpolation, we interpolate between the **latent vectors** of two consecutive key images using **linear or spherical interpolation (LERP or SLERP)**.
   * Since the latent space of the VAE is structured, this ensures smoother and more semantically meaningful transitions.
3. **Decoding Back to Image Space**:
   * The interpolated latent vectors are then passed through the **VAE decoder** to reconstruct the corresponding intermediate images.

**Advantages and Challenges**

* **Advantages**:
  + **Better Continuity**: Since interpolation occurs in a structured feature space, transitions between images are more coherent.
  + **No Need for Stable Diffusion**: This reduces computational cost and eliminates the randomness introduced by diffusion models.
  + **Faster Processing**: Encoding, interpolating, and decoding using a VAE is computationally more efficient than performing interpolation followed by Stable Diffusion.
* **Challenges**:
  + **Quality Dependence on the VAE**: The interpolated images depend on how well the VAE captures important visual details. A poorly trained VAE may introduce artifacts or blurring.
  + **Loss of High-Frequency Details**: Since VAEs often produce slightly smoothed outputs, some fine details may be lost in the reconstructed images.

This approach offers a promising alternative by leveraging the **latent space of a VAE**, enabling efficient and smooth transitions while avoiding costly post-processing steps.