Zero-Shot Text-to-Video Generation: Leveraging Efficient Text-to-Image Synthesis for Dynamic and Context-Preserving Video Synthesis

Adarsh Saripalli* **Akshay Sooria Ajukumar Rekha*** **Josue Cortez*** **Poojah Ganesan***
*Arizona State University

***Abstract*—We propose a new approach, Zero-Shot Text-to-Video generation, revolutionizing video creation without the need for extensive training or large datasets. Our method, leveraging text-to-image synthesis techniques such as Stable Diffusion, is tailored for minimal computational cost. Key enhancements include integrating motion dynamics into generated frames and optimizing cross frame-level attention. This ensures high-quality and temporally consistent videos, marking a significant advancement in video synthesis.**

***Index Terms*—Zero-Shot Text-to-Video generation, Stable Diffusion, motion dynamics, cross frame-level attention.**

I. Problem Statement

We initially aimed to create an innovative video summarizing system using DiffuSum. However, DiffuSum primarily focuses on text processing, which was misaligned with our objectives. To address this, we shifted our focus to adapt Stable Diffusion for video generation from text descriptions.

In the realm of generative AI and diffusion models for text-to-image synthesis, many methods in text-to-video generation require extensive labeled data and training. Our approach is a zero-shot text-to-video synthesis system that operates without additional training by leveraging a pre-trained text-to-image model (Stable Diffusion). This adaptation ensures temporal consistency, resulting in high-quality, time-consistent video generations. Our contributions encompass novel techniques for temporal consistency and versatile applications in video synthesis.

II. Implementation Overview

The process initiates with the selection of a random latent code, representing the video's initial state. Through the backward DDIM process, utilizing a pre-trained Stable Diffusion model (SD) [1], this results in a modified latent code that incorporates motion dynamics for temporal consistency. Following this, the forward DDPM process generates latent codes for each frame, providing flexibility in object motion. Fig. 1 shows Stable Diffusion implementation with three components: 1. Auto-Encoder – Converts the original image into smaller latent representations. 2. UNet – Predicts a denoised image representation of the noisy latent images. 3. Text-Encoder – Transforms the input text prompt into an embedding space, which is inputted to Unet.

- Initial Input Image, (Fig.1) Z_0: latent images encoded from training data.
- Output Image, (Fig.1) Z_t: Noisy image generated by diffusion algorithm.

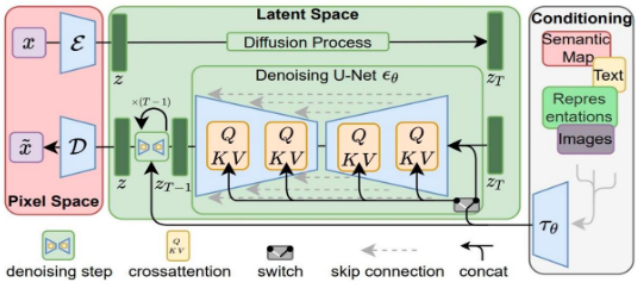

Fig. 1. Stable Diffusion trains to predict the added noise in the noisy image with time step 't', text prompt, and image diffusion algorithms.

Conventional SD methods risk random image generation. To address this, we redefine SD to work with sequences of latent codes, introducing motion dynamics for global scene time consistency and cross-frame attention which will enhance temporal consistency by focusing on the first frame.

III. Proposed Solution

Our solution employs Stable Diffusion's text-to-image synthesis for zero-shot text-to-video generation [2]. Enriching latent codes with motion dynamics and cross-frame attention ensures efficient, high-quality video synthesis. It extends to conditional and specialized video generation, including Video Instruct-Pix2Pix.

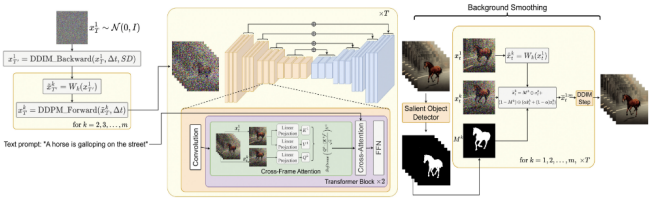

Fig. 2. Flowchart illustrating the comprehensive framework of our proposed solution for zero-shot text-to-video synthesis. The process encompasses latent code enrichment, motion dynamics integration, and cross-frame attention reprogramming. The implementation details cover conditional and specialized video generation, Video Instruct-Pix2Pix, and the seamless integration of Stable Diffusion for enhanced control and performance.

To adapt Stable Diffusion for videos, we redefine its operation, working with sequences of latent codes. Initially, m latent codes undergo independent sampling, followed by

DDIM to obtain tensors. To avoid random image generation, we introduce motion dynamics between latent codes for global scene consistency. A cross-frame attention mechanism preserves appearance and identity of foreground objects.

Algorithm 1 Motion dynamics in latent codes

Require: $\Delta t \geq 0, m \in \mathbb{N}, \lambda > 0, \delta = (\delta_x, \delta_y) \in \mathbb{R}^2$, Stable Diffusion (SD)

$x_T^1 \sim \mathcal{N}(0, I)$ ▷ *random sample the first latent code*

$x_{T'}^1 \leftarrow$ DDIM_Backward$(x_T^1, \Delta t, SD)$ ▷ *perform Δt backward steps by SD*

for all $k = 2, 3, \ldots, m$ **do**

  $\delta^k \leftarrow \lambda \cdot (k-1)\delta$ ▷ *computing global translation vectors*

  $W_k \leftarrow$ Warping by δ^k ▷ *defining warping functions*

  $\tilde{x}_{T'}^k \leftarrow W_k(x_{T'}^1)$

  $x_T^k \leftarrow$ DDPM_Forward$(\tilde{x}_{T'}^k, \Delta t)$ ▷ *DDPM forward for more motion freedom*

return $x_T^{1:m}$

Fig. 3. Algorithmic depiction of our motion dynamics for zero-shot text-to-video synthesis, we randomly sample the first frame's latent code. Using DDIM backward steps with the Stable Diffusion (SD) model, a global motion direction generates translation vectors for subsequent frames, creating motion flow. The sequence, enriched with motion dynamics, enhances temporal consistency for the global scene and background. However, initial latent codes may lack details, potentially causing temporal inconsistencies, especially for foreground objects.

[2]

Our motion dynamics algorithm involves random sampling of the first frame's latent code, DDIM backward steps (which is a deterministic process), computation of global translation vectors, and DDPM forward steps (which is a probablistic process). While enhancing temporal consistency, modified latent codes may lack specificity, potentially causing temporal inconsistencies. Cross-frame attention reprogramming, a key aspect, replaces traditional self-attention layers, focusing on the first frame to enhance temporal consistency across generated frames.

$$\text{Cross-Frame-Attn}(Q^k, K^{1:m}, V^{1:m}) = \text{Softmax}\left(\frac{Q^k (K^1)^T}{\sqrt{c}}\right) V^1$$

Fig. 4. This formula replaces traditional self-attention layers. It focuses on the first frame, enhancing temporal consistency throughout the generated frames. Where 1:m are the latent codes for 1^{st} frame to m^{th} frames respectively. Q, K, V are learnable weight matrices and c is dimension of the key vectors

[2]

Background smoothing combines masked latent codes for video generation without guidance. Our solution uses ControlNet [3] for conditional and specialized text-to-video generation, providing powerful control using edges, poses, masks, and depths. Our modifications, along with ControlNet, enhance video consistency compared to traditional methods. The implementation includes Video Instruct for improved edited videos [4]. The base model employs Stable Diffusion, generating m=8 frames with a 512x512 resolution, allowing flexibility in frame count. Conditional generation and specialized models like DreamBooth [5] enhance task-specific synthesis.

IV. Experiment Results

Our proposed method demonstrates the capability for zero-shot video generation across diverse scenarios:

1) Text to Video:
   Utilizing solely a textual prompt, generates compelling videos. The system adeptly translates textual descriptions into visually appealing and contextually coherent video sequences.[2]

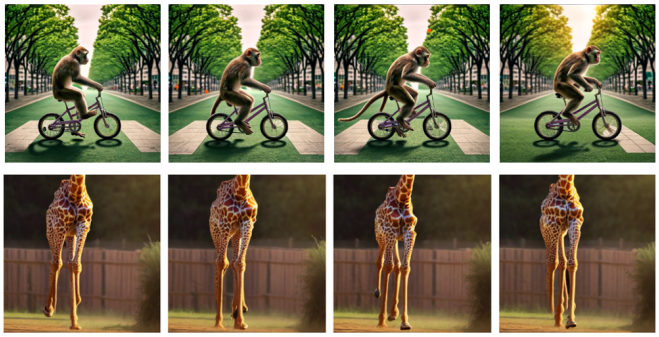

Fig. 5. Text to Video Generation: (a.) Input: "A monkey riding a bike on a park" (b.) Input: "An giraffe running on a zoo"

2) Text to Video with Pose Guidance:
   Extending the capabilities of our approach, we explored video generation with additional pose guidance. The integration of pose information resulted in videos with enhanced realism and improved alignment with specified poses.[2]

Fig. 6. Text to Video with Pose Guidance: (a.) Input: "A kid playing dodgeball" (b.) Input: "A student practicing karate in a dojo"

3) Text to Video with Edge Guidance:
   Further expanding the application scope, we examined the impact of incorporating edge guidance into the video generation process. The addition of edge information contributed to sharper and more defined video content.[3]

Fig. 7. Text to Video with Edge Guidance: (a.) Input: "A wolf walking in the woods"

4) Video Instruct Pix2Pix:
To validate the adaptability of our method to existing video editing algorithms, we integrated Video Instruct Pix2Pix. The combination showcased the seamless incorporation of our zero-shot text-to-video synthesis into established video editing frameworks.[4]

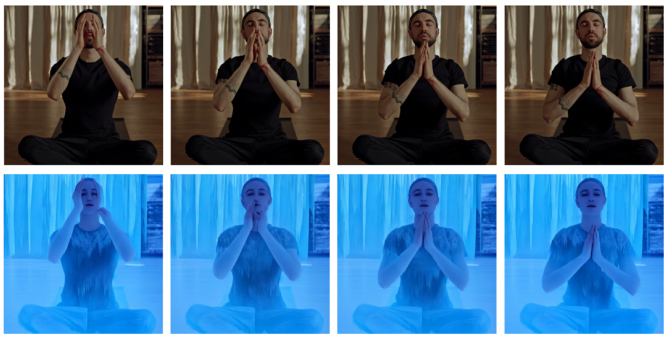

Fig. 8. Video Instruct Pix2Pix: (a.) Input: "Make the environment freezing cold"

5) Text to Video with Edge Guidance and Dream-Booth Specialisation:
Incorporating both edge guidance and Dream-Booth specialization, our method achieved videos with a heightened level of detail and specificity. The combination of edge guidance and Dream-Booth specialization resulted in videos that not only captured textual descriptions but also exhibited unique visual characteristics.[5]

V. Conclusion

Our approach represents a notable stride in the pursuit of dynamic and context-preserving video synthesis, making a meaningful contribution to the field of generative AI. While the project successfully demonstrates the efficacy of our proposed approach, it also lays the groundwork for further advancements and enhancements in text-guided video synthesis methodologies. The findings suggest a promising avenue for future exploration and refinement in this evolving domain.

Fig. 9. Text to Video with Edge Guidance and Dream-Booth Specialization: (a.) Input: "A Business man with glasses, arcane style"

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models (a.k.a. ldm & stable diffusion)," *Journal Name*, 2023.

[2] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," *Journal Name*, 2023.

[3] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *Journal Name*, 2023.

[4] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2022, pages 3, 7, 8, 21.