Efficient and consistent zero-shot video generation with diffusion models

Ethan Frakes, Umar Khalid, and Chen Chen

University of Central Florida, 4000 Central Florida Blvd., Orlando, FL

ABSTRACT

Recent diffusion-based generative models employ methods such as one-shot fine-tuning an image diffusion model for video generation. However, this leads to long video generation times and suboptimal efficiency. To resolve this long generation time, zero-shot text-to-video models eliminate the fine-tuning method entirely and can generate novel videos from a text prompt alone. While the zero-shot generation method greatly reduces generation time, many models rely on inefficient cross-frame attention processors, hindering the diffusion model's utilization for real-time video generation. We address this issue by introducing more efficient attention processors to a video diffusion model. Specifically, we use attention processors (i.e. xFormers, FlashAttention, and HyperAttention) that are highly optimized for efficiency and hardware parallelization. We then apply these processors to a video generator and test with both older diffusion models such as Stable Diffusion 1.5 and newer, high-quality models such as Stable Diffusion XL. Our results show that using efficient attention processors alone can reduce generation time by around 25%, while not resulting in any change in video quality. Combined with the use of higher quality models, this use of efficient attention processors in zero-shot generation presents a substantial efficiency and quality increase, greatly expanding the video diffusion model's application to real-time video generation.

Keywords: Attention processor, zero-shot, diffusion model, Stable Diffusion, Flash attention, Real-time video generation, Real-time video editing, cross-frame attention

1. INTRODUCTION

Generative diffusion models[1,2,3] have seen rapid advancement within the last few years. With the introduction of diffusion-based Text-to-Image (T2I) models and their general accessibility, AI-generated imagery and artwork continue to accumulate rapid mainstream appeal. T2I diffusion models have already been applied to a multitude of applications, including art, photo editing, and industrial applications. T2I models such as the open-source Stable Diffusion[2] and the closed-source Midjourney[3] have seen wide adoption not just among the computer vision field, but also in the arts. While T2I models rapidly improve, Text-to-Video (T2V)[4,5,6,7,8,9,10] diffusion models are also becoming more advanced. Previous works such as fine-tuned one-shot models[4,5] attempted to extend the success of the T2I model by applying them to the video domain. While their implementation drastically reduces complexity by using a T2I model for video generation, eliminating the need for a separate T2V diffusion model, their one-shot fine-tuning implementation leads to long runtimes for video generation and heavy computation.

Zero-shot video diffusion models,[6,7,8] which implement Stable Diffusion or a similar T2I model and generate multiple frames of output without fine-tuning, address the computation issue. While efficiency is drastically improved, they still struggle with frame consistency, text prompt consistency, and optimization. To address the issue of frame and text consistency/quality, we propose applying more advanced T2I models[11] to a preexisting T2V model. And to address the issue of optimization, we propose using attention processors which optimize for reducing complexity and increasing hardware parallelization.

Our objective and motivation with this work is to highlight how more efficient processing and higher quality diffusion models have a profound effect on generation time and output quality. As T2V models become increasingly more researched, their speed and quality continue to improve. Our research helps to validate this improvement and supply faster and better video diffusion models for general use.

Further author information: (Send correspondence to E.F.)
E.F.: E-mail: et250818@ucf.edu
U.K.: E-mail: Umar.Khalid@ucf.edu
C.C.: E-mail: chen.chen@crcv.ucf.edu

Real-Time Image Processing and Deep Learning 2024, edited by Nasser Kehtarnavaz, Mukul V. Shirvaikar,
Proc. of SPIE Vol. 13034, 1303407 · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3013575

2. RELATED WORK

2.1 Text-to-Image Diffusion Models

Research into text-to-image diffusion models has increased exponentially within recent months and years. First proposed in 2015,[1] rapid advancements in both transformers[12] and attention processing have allowed for diffusion models to become widely available. Their robust nature also allows for their application in a variety of tasks, such as text-to-image and image-to-image generation, as well as image denoising. DALL-E,[13] for example, introduced a zero-shot text-to-image generator, later improved with DALL-E 2[14] by utilizing CLIP[15] for text-image encodings. Stable Diffusion[2] is an open-source T2I diffusion model with all model parameters available online. Its robustness and modifiability have lent it popular appeal. Stability AI, one of the co-authors of the original Stable Diffusion, has since released more advanced models such as Stable Diffusion XL.[11]

2.2 Text-to-Video Diffusion Models

As the text-to-image diffusion field progresses, early research into diffusion's applicability into the video domain are promising. Utilizing a pretrained video diffusion model trained from a video dataset may sound like the optimal choice at first; however, one must consider the large size of these datasets in comparison to their image counterparts. Moreover, training a separate video diffusion model is not an optimal choice when considering training time and memory requirements for a large video dataset. Some research has been performed in one-shot T2V models[45] that fine-tune a Stable Diffusion model on a single video input. Methods such as Tune-A-Video[4] and Video-P2P[5] can produce videos identical to the input video in motion while gaining a new aesthetic style from the encoded text prompt. While one-shot models have promise in the video-editing field, they are incapable of generating novel videos from text input alone. Additionally, their one-shot nature requires fine-tuning the base model, which results in long generation times for one video.

Alternatively, zero-shot video diffusion models[67810] do not require a video to fine-tune with and are instead capable of generating video frames using the pretrained model weights from Stable Diffusion.[211] Several methods exist for generating coherent frames of video; cross-frame attention[45678910] is often employed to utilize the query, key, and value tensors from multiple frames. Utilizing cross-frame attention and applying cross-frame attention processors is of core focus in this paper.

3. BACKGROUND

We begin this section by giving a brief overview of the architecture of a T2I diffusion model such as Stable Diffusion. We then give an overview of applying a T2I model to the video domain for zero-shot novel video synthesis and video-editing.

3.1 Stable Diffusion

Stable Diffusion[2] is a latent diffusion model which is contained within an autoencoder, in this case $x \sim \mathcal{D}(\mathcal{E}(x))$, where x is an image, $\mathcal{E}$ is an image encoder, and $\mathcal{D}$ is an image decoder. When image x is encoded, a clean latent x_0 is generated with dimensions $\mathbb{R}^{h \times w \times c}$, where h, w, and c are the height, width, and number of channels respectively. Equivalently, x_0 can be represented as $x_0 = \mathcal{E}(x)$. After generating the clean latent, the forward process then progressively adds Gaussian noise to the encoded latent in T number of timesteps. During training, the model generates noisy latents x_t, where $t = 1, \ldots, T$. With these latents, the Stable Diffusion's U-Net,[16] which is composed of alternating convolution and transformer/attention blocks containing self- and cross-attention layers, backpropagates by learning how to denoise the noisy latents $x_t = T, \ldots, 1$ as close to their clean counterpart as possible. After the model is trained and the backward process learned, we can then apply a deterministic sampling process, in our case DDIM sampling,[17] to remove the Gaussian noise from the latents over T timesteps. The DDIM sampling process can be represented as:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta^{(t)}(x_t), + \sigma_t \epsilon_t \tag{1}$$

where $t = T, ..., 1$ and x_{t-1} is the denoised latent extrapolated from x_t.[17]

3.2 Text-to-Video with Stable Diffusion

To apply Stable Diffusion to the video domain, we must first consider that there will be multiple latents at timestep t for each corresponding frame f, where $f = 1, \ldots, F$, and F is the total number of frames. We must then consider that each latent will now be 4-dimensional, with dimensions $\mathbb{R}^{F \times h \times w \times c}$. To generate a multi-frame video, we could sample each latent code x_T^f for $f = 1, \ldots, F$, then apply DDIM sampling to receive their clean latent counterparts x_0^f. However, this presents a problem: how can a novel video with coherent frames be generated if the self-attention function utilized in the Stable Diffusion U-Net is completely independent of any other image? In other terms, if we were to generate a video with self-attention, all generated frames would not possess any coherence, leading to a generated product more akin to an image collage rather than a true video.

Self-attention,[12] as the name implies, computes attention only on a single sequence. The formula for self-attention, in particular scaled dot-product attention, is

$$Attention(Q, K, V) = \sigma(\frac{QK^T}{\sqrt{d_k}})V, \tag{2}$$

where Q, K, and V are the query, key, and value tensors respectively, d_k is the dimension of the key and query tensors, and σ is the non-linear softmax function.[12] As we can see, however, self-attention only accounts for the query, key, and value weights of the current latent frame, rather than all frames. To address this, the self-attention in the SD U-Net[16] can be reprogrammed into cross-frame attention,[45678910] where the Q, K, and V weights of frame f are factored into the attention equation. There are many different forms of how to compute cross-frame attention: frame-attention,[569] for example, utilizes the key and value weights from only the first frame on the query weight Q^f; this can be represented as

$$\begin{gathered} Attention(Q, K, V) = \sigma(\frac{Q^f (K^1)^T}{\sqrt{d_k}})V^1, \\ f = 1, \ldots, F, \end{gathered} \tag{3}$$

where K^1 and V^1 are the key and value weights from the first frame, and Q^f are the query weights from all frames.[69] By expanding attention across all frames of the video to the first frame, the model will produce a video containing coherently structured frames with similar visuals, motion dynamics, foreground objects, and background objects.

There are other forms of cross-frame attention as well, such as spatial-temporal attention,[10] which utilizes the key and value weights from all frames, not just the first. This, however, greatly increases the memory and runtime complexity, as the number of FLOPs becomes exponentially higher as you increase the number of frames calculated during attention.[9] For all experiments performed in this paper, we use frame-attention.

4. METHOD

In this section, we detail the method we used to improve the results of T2V diffusion models in the domains of efficiency and quality.

4.1 Attention Processing

One of the key challenges to generating videos efficiently is the reduction and optimization of processing and memory, particularly attention processing. Normally, attention requires a quadratic memory complexity of $\mathcal{O}(n^2)$,[18] which significantly limits the amount of attention processing at any given time. This renders large numbers of weights (in our case, a large number of frames) infeasible, even on high-end hardware such as the NVIDIA H100 GPU. Meta's xFormers[19] addressed this memory overlay by implementing a memory-efficient attention mechanism requiring $\mathcal{O}(\sqrt{n})$ memory complexity.[18]

Another method for significantly increasing the efficiency of attention processors is in hardware optimization. One of the main challenges in processing attention is inefficient hardware communication, such as the high-performance SRAM of the GPU vs the relatively slow high-bandwidth memory (HBM) of the GPU. FlashAttention[20] addressed this issue by eliminating reading and writing the attention matrix to and from the HBM,

attempting to instead perform the attention calculation on the SRAM, then writing the output to the HBM. FlashAttention-2[21] further increased efficiency by reducing the number of non-matrix multiplication FLOPs and parallelizing the attention computation across multiple GPU thread blocks. Additionally, HyperAttention[22] implemented a modular design that introduced two parameters that calculate and reduce the attention's time complexity, achieving near linear time complexity and integration with other efficient processors like FlashAttention.

Performing low-end hardware optimizations to attention processing greatly reduces generation time. This increase in efficiency is highly applicable to the video generation domain, as generation time for videos, especially in the case of long videos, is greatly reduced. We illustrate later in this paper that utilizing attention processors such as xFormers,[19] FlashAttention-2,[21] and HyperAttention[22] does indeed greatly reduce generation time for videos.

4.2 Frame Quality and Frame/Text Consistency

While reduction in unnecessary processing is useful for reducing runtime, this does not result in any improvements to the quality of the videos themselves, only the efficiency. In this paper, we illustrate that the choice of the SD model used can bottleneck video quality, and utilizing fine-tuned or larger diffusion models can generate videos with better quality.

Recent advancements in image diffusion models, either fine-tuned or larger in parameter size, have proved to generate higher quality results. For example, Stable Diffusion XL (SDXL),[11] a high-parameter SD model from Stability AI, has increased the parameter size from SD 1.5's 860 million to 2.6 billion. In addition, it also uses the OpenCLIP ViT-bigG[23] text encoder in addition to CLIP ViT-L.[15] SDXL also features an optional refiner model that can further refine image quality through image-to-image diffusion. User preference for SDXL is generally far higher than SD 1.5 or SD 2.1,[2] with the base model far outperforming both and, with the inclusion of the refiner model, further improving user preference.[11] Through experimentation, we illustrate that frame quality and text consistency can be improved by utilizing SDXL in place of SD 1.5.

4.3 Testing Efficiency and Consistency

To properly test both the efficiency and consistency of the video generation, the proposed method tests several different attention processors, as well as different Stable Diffusion models, to accurately gauge their generation speed and output quality. As shown in Fig. 1, we have two inputs: a text prompt and an optional video. If using a video, Stable Diffusion's ControlNet[24] extension is used, which can "control" the output of an image based on input conditions such as a Canny edges or a depth map. These conditions are extracted from the input video, and its and the text prompt's embeddings are extracted. ControlNet has a separate U-Net architecture that copies neural network blocks from the primary U-Net.[24]

Using the text prompt and ControlNet embeddings, as well as the latent codes, we then run the model's U-Net, which is swapped between different models, such as SD 1.5[2] and SDXL.[11] We also swap between different cross-frame attention processors, measuring their generation time from start to finish to compare their speeds. After the video with complete frames is generated, we extract the runtimes, as well as use CLIP[15] to calculate the frame consistency by measuring the average consistency between two consecutive frames. We similarly measure the average between the encoded text prompt and each encoded frame, outputting this as the text consistency.

5. EXPERIMENTAL RESULTS

To accurately compare efficiency and quality, we use Text2Video-Zero,[6] a zero-shot video diffusion model that utilizes frame attention and has ControlNet[24] integration, as our base. As illustrated by Fig. 1, we perform two primary experiments: (1) replace the attention processor utilized by Text2Video-Zero and test with ControlNet enabled, and (2) test SD 1.5 as well as SDXL 0.9 and 1.0 to compare video quality. All experiments were performed on an NVIDIA H100 GPU with 80GB of VRAM. Full samples of results from both ControlNet and novel generation are avaiable at https://drive.google.com/drive/folders/18MpnO0Q3uRXK3H2yyJPmqP7kdrocZX6-?usp=sharing

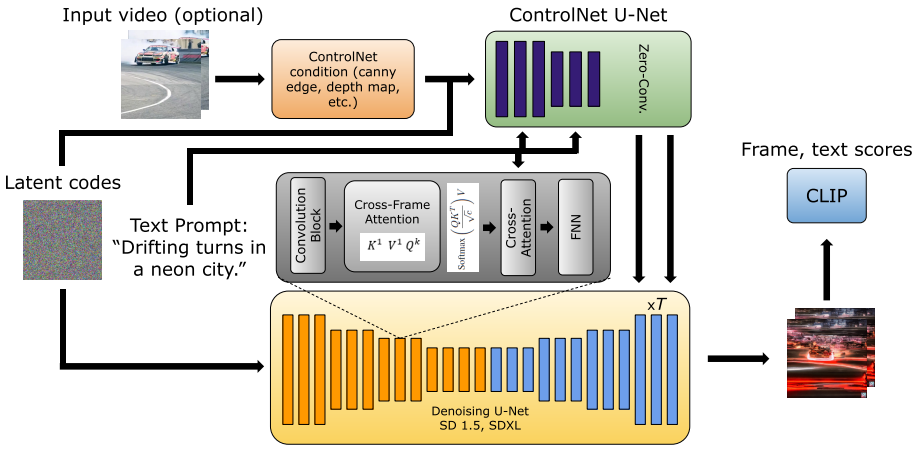


Figure 1. General framework of our experiments. In Text2Video-Zero,[6] latent codes are denoised in the Stable Diffusion U-Net, with the model of choice varying between SD 1.5, SDXL 0.9, and SDXL 1.0. Within the U-Net, the cross-frame attention processor is modified with either xFormers, FlashAttention-2, or HyperAttention used. An optional input video can be used to capture its canny edges, depth map, etc., which can then be utilized by ControlNet[24] for direct zero-shot video editing. After the denoised frames are generated, the final result's frame and text consistency score, as well as its total generation time in seconds, are calculated.

5.1 Attention Processor

To compare attention processors, we test four different attention processors: Text2Video-Zero's,[6] xFormers's memory efficienct attention processor,[19] FlashAttention-2,[21] and HyperAttention.[22] They key and value tensors are preprocessed using frame attention (using key and value weights from first frame), then each processor calculates the hidden states using the four processors' unique attention calculation function. Each is tested with Text2Video-Zero's ControlNet[24] extension, using both SD 1.5 and SDXL as the base model, and 50 videos from DAVIS 2016[25] are used as the testing dataset. Each video from DAVIS is cropped to 480-by-480 resolution to reduce generation time and produce higher quality videos. To generate complimentary text prompts for each video, the names of each video were given to the GPT-3[26] large language model. GPT-3 was instructed to generate a prompt 10 words or fewer, each with a randomized art or aesthetic style. Two videos for each DAVIS input video were generated: one using canny edges and one using a depth map. Each video was generated four times; one time per processor. Each video was generated at a resolution of 512-by-512 for SD 1.5 and 1024-by-1024 for SDXL. The time in seconds to generate each video was calculated, and the mean and total times were then calculated for each attention processor. All quantitative results are shown in tables 1, 2, 3, and 4.

In both cases, xFormers and FlashAttention-2 far outperform both the processor used with Text2Video-Zero, and slightly outperform HyperAttention. Between Text2Video-Zero and the best performing processor, generation time for the same video is reduced by approximately 25%. This demonstrates that reducing memory and time complexity, as well as optimizing hardware communication and parallelization does have a significant impact on processing efficiency. Therefore, advances in attention processing have a significant impact on the rate of transformer-based attention processing, allowing for faster training and generation times.

5.2 Stable Diffusion Model

To compare Stable Diffusion models, we tested three different models for generating both novel and ControlNet-guided videos: SD 1.5,[2] SDXL 0.9,[11] and SDXL 1.0.[11] For generating the novel videos, each model was benchmarked without any motion dynamic warping or background smoothing; each model was tested with their only major modification being their use of cross-frame attention instead of self-attention. All three models were tested for novel generation, while SD 1.5 and SXDL were tested for ControlNet integration. After the 50 videos are generated for each of the three models, their frame consistency score is calculated by taking two consecutive frames, computing their CLIP[15] embeddings, normalizing the embeddings, and multiplying the two embeddings to receive a consistency score. This score is then averaged for all frames in the video, and then all average frame

Table 1. Stable Diffusion 1.5 ControlNet Canny Edge Mean & Total runtimes for all processors in seconds

Attn Processor	T2VZ Base	xFormers	FlashAttn-2	HyperAttn
Mean Time (s)	56.78	42.16	**41.82**	43.04
Total Time (s)	2839.12	2108.24	**2091.20**	2152.19

Table 2. Stable Diffusion 1.5 ControlNet Depth Map Mean & Total runtimes for all processors in seconds

Attn Processor	T2VZ Base	xFormers	FlashAttn-2	HyperAttn
Mean Time (s)	58.13	**43.98**	44.38	45.60
Total Time (s)	2906.50	**2199.11**	2219.19	2280.15

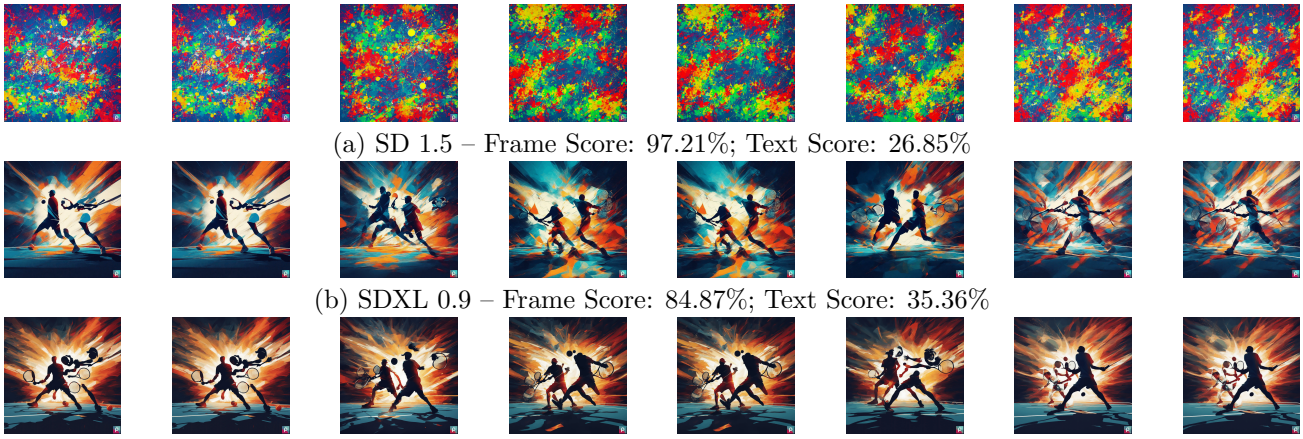
Table 3. Stable Diffusion XL ControlNet Canny Edge Mean & Total runtimes for all processors in seconds

Attn Processor	T2VZ Base	xFormers	FlashAttn-2	HyperAttn
Mean Time (s)	291.50	231.75	**225.16**	232.70
Total Time (s)	14575.09	11587.72	**11257.92**	11635.24

Table 4. Stable Diffusion XL ControlNet Depth Map Mean & Total runtimes for all processors in seconds

Attn Processor	T2VZ Base	xFormers	FlashAttn-2	HyperAttn
Mean Time (s)	289.20	**224.34**	230.70	234.68
Total Time (s)	14459.93	**11216.98**	11535.10	11733.84

scores from all videos are averaged to give a mean frame consistency score. Likewise, the text consistency score measures the consistency between the text prompt and frame output. The text prompt is tested against each frame, and then this score is averaged with all frames, like with frame consistency. Examples of qualitative results and mean quantitative results are displayed in Fig. 2, 3, 4, and 5, and tables 5, 6, and 7.

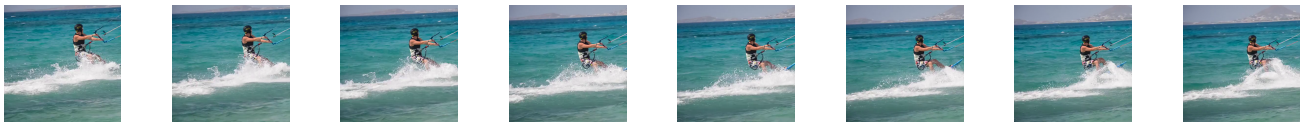


(a) SD 1.5 – Frame Score: 97.21%; Text Score: 26.85%

(b) SDXL 0.9 – Frame Score: 84.87%; Text Score: 35.36%

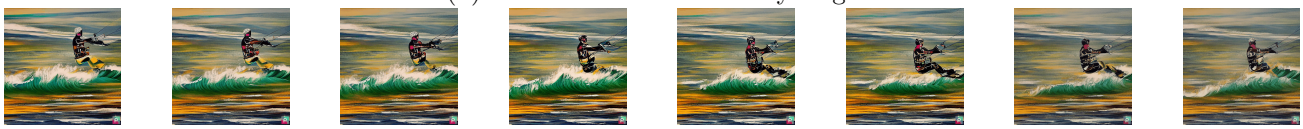
(c) SDXL 1.0 – Frame Score: 91.49%; Text Score: 36.31%

Figure 2. Our research shows that different diffusion models have a large impact on the output quality of video generative models. For this example, three different diffusion models were tested: Stable Diffusion 1.5, Stable Diffusion XL 0.9, and Stable Diffusion XL 1.0. Each had varying results, with some models having better textual consistency while others having greater frame consistency. The prompt for this example was "Animated tennis match in abstract art style."



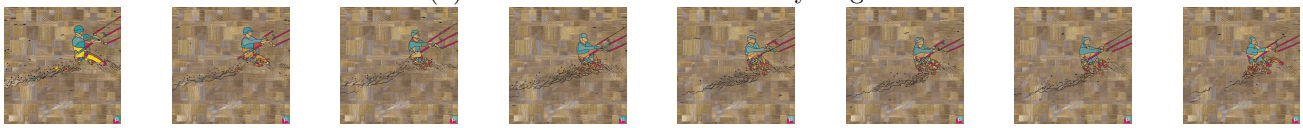
(a) Input video: "kite-surf.mp4" – Prompt: "Kite-surfing in abstract art."

(b) SD 1.5 ControlNet Canny Edge



(c) SD 1.5 ControlNet Depth Map

(d) SDXL 1.0 ControlNet Canny Edge



(e) SDXL 1.0 ControlNet Depth Map

Figure 3. Example of using ControlNet to apply a text prompt with an input video. GPT-3 was given the video name "kite-surf.mp4" and instructed to create a prompt 10 words or less with a varying art style. The resulting prompt, "kite-surfing in abstract art," is tested with both a preprocessed canny edge and depth map from the input video, using both SD 1.5 and SDXL.

Table 5. Mean Novel Video Consistency Scores

SD Model	SD 1.5	SDXL 0.9	SDXL 1.0
Frame Consistency	**94.23%**	90.22%	94.12%
Text Consistency	29.05%	**31.23%**	31.09%

Table 6. Mean ControlNet Canny Edge Video Consistency Scores

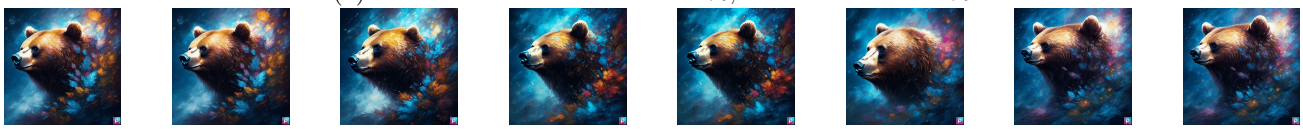
SD Model	SD 1.5	SDXL 0.9	SDXL 1.0
Frame Consistency	96.02%	96.40%	**96.61%**
Text Consistency	28.95%	**29.10%**	28.31%

Table 7. Mean ControlNet Depth Map Video Consistency Scores

SD Model	SD 1.5	SDXL 0.9	SDXL 1.0
Frame Consistency	96.29%	96.94%	**97.13%**
Text Consistency	**30.35%**	27.49%	27.33%

For novel generation, although the frame consistencies between each model are negligible, SDXL 0.9 and 1.0 show an improved text prompt consistency with the video output. For ControlNet guidance, SDXL outperforms SD 1.5 for frame quality and consistency, while SD 1.5 outperforms or is similar in terms of text consistency, especially when utilizing a depth map. Visually, the results from SDXL display a closer resemblance to the text prompt than SD 1.5. For example, the SDXL Depth Map frames from Fig. 3 display a result more akin to abstract artwork than in SD 1.5, and the frames from the SDXL Depth Map in Fig. 5 are also more akin to a cubist art style. These results highlight that more advanced Stable Diffusion models can generate more coherent and text-consistent videos depending on the guidance, qualitative measures, and quantitative measures.

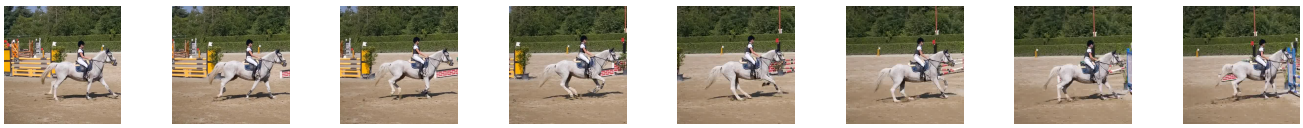
(a) SD 1.5 – Frame Score: 90.09%; Text Score: 29.55%



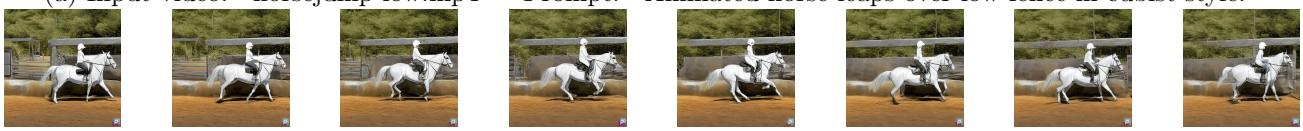
(b) SDXL 0.9 – Frame Score: 96.91%; Text Score: 32.90%

(c) SDXL 1.0 – Frame Score: 98.50%; Text Score: 34.24%

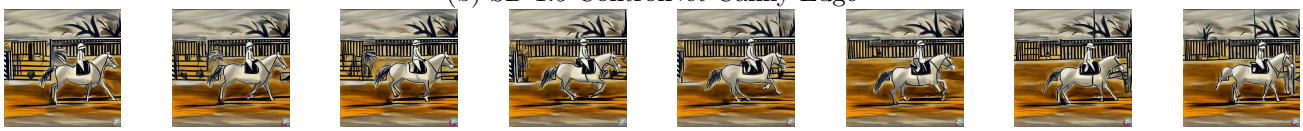
Figure 4. Example of frames from novel video generation comparison between SD 1.5, SDXL 0.9, and SDXL 1.0. For all three, the input prompt was "Fantasy bear painting."



(a) Input video: "horsejump-low.mp4" – Prompt: "Animated horse leaps over low fence in cubist style."

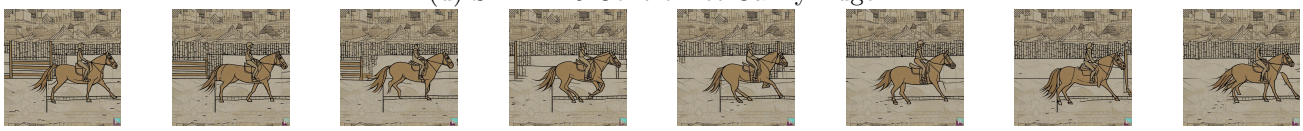


(b) SD 1.5 ControlNet Canny Edge



(c) SD 1.5 ControlNet Depth Map

(d) SDXL 1.0 ControlNet Canny Edge



(e) SDXL 1.0 ControlNet Depth Map

Figure 5. Example of using ControlNet to apply a text prompt with an input video. The video name is "horsejump-low.mp4," and the GPT-generated prompt is, "Animated horse leaps over low fence in cubist style."

6. CONCLUSION

In this paper, we addressed two major issues in zero-shot video generation: efficiency and quality. Our approach for increasing both is easily applicable to most zero-shot models, and they can be easily updated and modified with efficient attention processing and diffusion models. While video efficiency appears to be making steady progress, however, temporal coherence still underperforms in comparison to its image counterpart. While applying frame smoothing techniques such as motion warping or background/foreground smoothing[6] could increase consistency,

and while using ControlNet with advanced models such as SDXL could also increase quality, more research could prove effective. Some future areas of research we plan to explore are: the potential use of a video's optical flow to control the video diffusion process, fine-tuning advanced SD models such as SDXL to produce more specialized imagery (e.g. realistic, animated, etc.), and exploring other quantitative methods for measuring video quality such as human feedback, rather than consistency scores alone. Although progress can still be made, we believe that our results show a promising direction for video-generative diffusion models, making them more effective for use and accessible to the general public.

ACKNOWLEDGMENTS

The authors of this manuscript would like to thank the UCF Center for Research in Computer Vision (CRCV) for their assistance and guidance in the creation of this paper, including Dr. Mubarak Shah, Dr. Niels Lobo, Dr. Tanvir Ahmed, and Nayoun Ham. The authors would also like to thank the UCF Advanced Research Computing Center (ARCC), member of the Institute for Simulation and Training (IST), for providing computing resources necessary for performing all experiments. The authors would finally like to thank the National Science Foundation REU program for initiating the research project.

REFERENCES

[1] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S., "Deep unsupervised learning using nonequilibrium thermodynamics," in [*Proceedings of the 32nd International Conference on Machine Learning*], Bach, F. and Blei, D., eds., *Proceedings of Machine Learning Research* **37**, 2256–2265, PMLR, Lille, France (07–09 Jul 2015).

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., "High-resolution image synthesis with latent diffusion models," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 10684–10695 (June 2022).

[3] Midjourney. https://www.midjourney.com (2022). (Accessed: 3 April 2024).

[4] Wu, J., Ge, Y., Wang, X., Lei, S., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in [*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*], 7589–7599, IEEE Computer Society, Los Alamitos, CA, USA (October 2023).

[5] Liu, S., Zhang, Y., Li, W., Lin, Z., and Jia, J., "Video-P2P: Video editing with cross-attention control," *arXiv preprint arXiv:2303.04761* (2023).

[6] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H., "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," *arXiv preprint arXiv:2303.13439* (2023).

[7] Yang, S., Zhou, Y., Liu, Z., , and Loy, C. C., "Rerender a video: Zero-shot text-guided video-to-video translation," in [*ACM SIGGRAPH Asia Conference Proceedings*], (2023).

[8] Zhang, Y., Wei, Y., Jiang, D., Shang, X., Zuo, W., and Tian, Q., "ControlVideo: Training-free controllable text-to-video generation," *arXiv preprint arXiv:2305.13077* (2023).

[9] Karim, N., Khalid, U., Joneidi, M., Chen, C., and Rahnavard, N., "SAVE: Spectral-shift-aware adaptation of image diffusion models for text-guided video editing," *arXiv preprint arXiv:2305.18670* (2023).

[10] Wang, W., Xie, k., Liu, Z., Chen, H., Cao, Y., Wang, X., and Shen, C., "Zero-shot video editing using off-the-shelf image diffusion models," *arXiv preprint arXiv:2303.17599* (2023).

[11] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R., "SDXL: Improving latent diffusion models for high-resolution image synthesis," in [*The Twelfth International Conference on Learning Representations*], (2024).

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," in [*Advances in Neural Information Processing Systems*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., **30**, Curran Associates, Inc. (2017).

[13] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I., "Zero-shot text-to-image generation," in [*Proceedings of the 38th International Conference on Machine Learning*], Meila, M. and Zhang, T., eds., *Proceedings of Machine Learning Research* **139**, 8821–8831, PMLR (18–24 July 2021).

[14] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M., "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125* (2022).

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., "Learning transferable visual models from natural language supervision," in [*International conference on machine learning*], 8748–8763, PMLR (2021).

[16] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing, Cham (2015).

[17] Song, J., Meng, C., and Ermon, S., "Denoising diffusion implicit models," in [*International Conference on Learning Representations*], (2021).

[18] Rabe, M. N. and Staats, C., "Self-attention does not need $\mathcal{O}(n^2)$ memory," *arXiv preprint arXiv:2112.05682* (2022).

[19] Lefaudeux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S., Labatut, P., Haziza, D., Wehrstedt, L., Reizenstein, J., and Sizov, G., "xformers: A modular and hackable transformer modelling library." https://github.com/facebookresearch/xformers (2022). (Accessed: 3 April 2024).

[20] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C., "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in [*Advances in Neural Information Processing Systems*], (2022).

[21] Dao, T., "FlashAttention-2: Faster attention with better parallelism and work partitioning," *arXiv preprint arXiv:2307.08691* (2023).

[22] Han, I., Jarayam, R., Karbasi, A., Mirrokni, V., Woodruff, D., and Zandieh, A., "HyperAttention: Long-context attention in near-linear time," *arXiv preprint arXiv:2310.05869* (2023).

[23] Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L., "OpenCLIP." https://doi.org/10.5281/zenodo.5143773 (July 2021).

[24] Zhang, L., Rao, A., and Agrawala, M., "Adding conditional control to text-to-image diffusion models," in [*IEEE International Conference on Computer Vision (ICCV)*], (2023).

[25] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A., "A benchmark dataset and evaluation methodology for video object segmentation," in [*Computer Vision and Pattern Recognition*], (2016).

[26] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., "Language models are few-shot learners," in [*Advances in Neural Information Processing Systems*], Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., eds., **33**, 1877–1901, Curran Associates, Inc. (2020).