LateNtMovies: High-Resolution Video Completion with Pseudo-3D Latent Diffusion Models

Melody Halbert Nayoun Ham Ethan Legum Julie Wan
University of Central Florida
{melody.halbert, nayoun.ham, ethanlegum}@knights.ucf.edu juliewan@zoho.com

Abstract

Diffusion Models are promising content creators and yet remain unaccustomed to video applications. Predicting plausible sequential moving image frames requires temporally-coherent cognition. We present LateNtMovies (Late Night Movies), which incorporates a pre-trained Text2Image Latent Diffusion Model and fuses motion understanding through 1D axial convolutions on unsupervised video. The sensible combination of knowledge transfer, perceptual compression, and pseudo-3D simulation observes contemporary resource constraints, inherits the prior's diversity and salience, and gains frame completion capabilities on intra-domain scenes. We demonstrate the effectiveness of the model in a range of experiments including CLIP benchmarks and a human evaluation study. Pending acceptance, our model will be available https://github.com/mhalbert/AdvCVFinalProject.

1. Introduction

Diffusion models (DM) have meteorically evinced their image-generating performance, eliciting scrutiny of their applicability in new modalities, e.g. video [2, 4, 7, 15, 16]. With an etymological and empirical decomposition to a sequence of moving images, videos remain unfeasible to compute and memorize due to their unlimited storytelling–plots can unfold in a number of ways. Looking forward in time is to consider exponentially increasing possible futures. And yet, enhancing video completion, i.e., predicting frames with or without partial observation, is crucial for honing autonomous driving, events forecasting, and safety-critical decision making (Hoppe et al., 2022; Voleti et al., 2022) [9, 15]. Our aim is to further the monumental task of video synthesis, wherein contemporary improvements remain comparatively modest compared to those in image synthesis.

In addition to static image sampling's desiderata of visual and contextual fidelity, moving image sampling mires temporal consistency, i.e., depicting the same content with plausible motion. Tian et al. considered video frames as a meaningful consecutive trajectory of latent codes [14]. As the costs associated with video modeling grows quadratically in pixel space compared to image modeling (Qiu et al. 2017) [11], estimating trajectories in latent space would ground compute liftoff. In the first place, DM's mode-covering behavior devotes excessive computes on repeated function evaluations and gradients in high-dimensional RGB space on imperceptible details. Inferencing is similarly expensive, the same repetitive evaluations are performed on just the noised version of the input space. This substantial resource budget previously restricted DM to a fraction of the field, however Latent Diffusion Models (LDM)–Stable Diffusion–democratized exploration (Rombach et al., 2022) [12]. Latent models scale more gracefully to higher-dimensions by autoencoding lower-dimensional representations that are perceptually-equivalent to the image space. The complexity reduction from abstracting away high-frequency, imperceptible details in compression also realizes efficient single network pass reconstructions. The generations are, most importantly, fidelous and salient, attributable to the Transformers' semantic structure retention and generalizability and to the DM's detail preservation.

In addition to latent exploration, we employ text-conditioning to the advantage of our generator. Hoppe et al. and Ho et al. underscore conditioning for the synthesized output's harmonization with the input frame [6, 9]. Lastly, and most formatively, we eliminate the exigency of an exhaustive semantic video dataset for intuiting motion.

While billions of text-image pairs are available, text-video pairs are insufficient, e.g. the largest annotated video dataset stands at just 41,250 samples (Hong et al., 2022) [8]. Singer et al. conceded the limited replicability of text-to-image (T2I) breakthroughs in text-to-video (T2V) due to few paired videos and the prohibitive erudition of from-scratch T2V models [13]. Unlabelled videos, i.e. without caption and thus more readily accessible, are sufficient to teach movement and interaction. We gain inspiration from the spatiotemporally factorized diffusion-based Make-

A-Video's leveraging of T2I models' text-image correspondences and prior amplification to motion through unsupervised ingestion of unlabeled videos. Singer et al.'s adoption of pseudo-3D layers was motivated by previous video and 3D vision works' impression that it best inherits the T2I predecessor and fuses temporal information when compared to Video Diffusion Models' (VDM) 3D U-Net (Ho et al., 2022) [7, 13].

2. Related Works

2.1. Diffusion-based Video Generation

The recent application of diffusion to video, especially text-driven, synthesis has attained recognition and inspired expeditious evolutions, e.g. Video Diffusion Models (Ho et al., 2022) and Imagen Video (Ho et al., 2022) [6, 7]. VDM introduced space-time separable attention to 2D U-Nets to accommodate and learn 3D blocks of temporally-dependent frames. The hollow conversion of 2D, $h \times w$, to space-only 3D convolutions, $f \times h \times w$–where f represents the video frame index, heeds memory constraints during their training from scratch. The leading authors of VDM additionally contributed Imagen Video which cascades VDM's Video U-Nets and evaluates on multiple frames simultaneously to generate temporally-cognitive 1024 x 1024 samples. They assert the indispensability of and condition on embeddings from a large frozen language model. However, this landmark publication employed an internal, and thus inaccessible, 14 million paired text-video dataset.

The first open-source large-scale-trained CogVideo, similar to our effort, recognized the unaffordable computes of training from scratch and sought to inherit from their pretrained CogView2. In our experiments, we determined that its samples are of lesser quality and/or unable to converge after its allotted training, i.e. blurry.

2.2. Diffusion-based Video Editing Variants

Our work leverages Stable Diffusion's Inpainting mechanism, and we compare our method to Dreamix, a diffusion-based video editor (Molad et al., 2023) [10]. In text-guided editing, the original sample is provided along with a prompt which describes the desired modification. The edited outcome should exemplify alignment, i.e. semantic conformity, and fidelity, i.e. connectedness or semblance to the input. LateNtMovies accepts as input a single frame and interpolates subsequent frames from noise, infilling the whole region with generalized motion that semantically follows the text guide. Dreamix demonstrates higher linguistic finesse in subject-driven tasks, however to reconstruct its partially degraded input, it is preliminarily fine-tuned on out-of-order high-resolution frames. Ours ships inference-ready for universal movement while Dreamix is still memorizing its specimen!

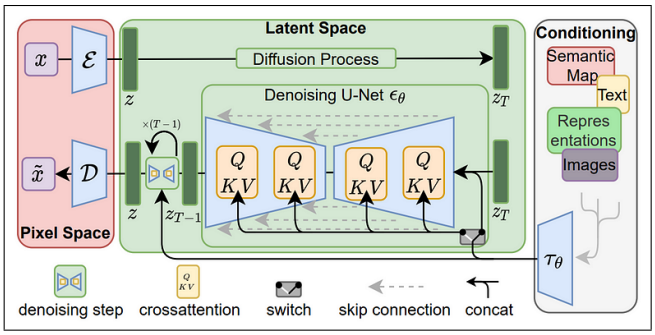

Figure 1. Latent diffusion archetype that perceptually compresses from pixel to latent space and consigns diffusion processes to lower dimensional computes [12]. Interleaving cross-attention mechanisms tend to conditioning modalities. A denoising network reconstructs the latent vectors to return to image space. Unpictured as above, our model additionally integrates 1D layers to separately assimilate time.

Esser et al. avoid the per-video re-training expense by pre-loading content pedagogy through a large-scale corpus of uncaptioned videos and labeled images, much alike earlier video generator authors [3]. We aimed to circumvent this inaccessible extravagance and resource waste by repurposing a pre-trained image model. Ultimately, their Structure and Content-Guided Video DM succumbs to fine-tuning as well, albeit on an image subset, for further customization.

3. Method

We present LateNtMovies, which hijacks Stable Diffusion's inpainting pipeline to generate temporally consistent multi-image frames. Inpainting is the task of filling masked regions of an image to replace existing but undesired content, i.e. the static portions should be dynamic in subsequent frames (Rombach et al., 2022). We mask the entire 512 x 512 input canvas and infill the subsequent frames conditioned on the input frame with prompt-guidance, at any specified number of frames and frames-per-second. We fine-tune a pretrained LDM which utilizes a conditional denoising autoencoder, guided by text and low- or same-resolution inputs that are projected to condensed intermediate representations. The underlying U-Net has cross-attention augmentations to receive and attend to various conditioning modalities (Rombach et al., 2022) [12]. At model initialization, spatial layers are extended with 1D attention modules to instantly transfer previous T2I wisdom and thus accelerates our model's command of temporal dynamics (Singer et al., 2022) [13].

3.1. Conditional Diffusion for Video

Let $x_0 \in \mathbb{R}^d$ be a sample from the data distribution p_{data}. During forward diffusion, x_0 is directly corruptible to x_t for

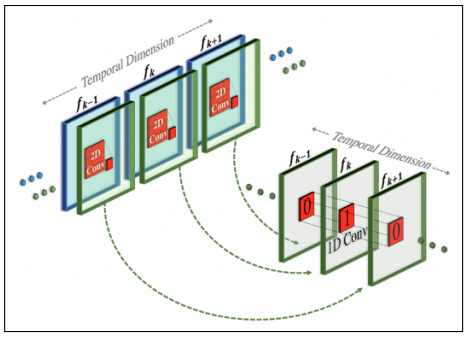

Figure 2. Pseudo-3D convolutional and attention layers' architecture and initialization layout that enables immediate adoption of pre-learnt visuals and seamless transition to the temporal axes. These axial convolutions achieve temporal fusion [13].

some arbitrary time step t using the accumulated kernel

$$q(x_t|x_0)=\mathcal{N}(x_t;\sqrt{\overline{\alpha}_t}x_0,(1-\overline{\alpha}_t)\mathbf{I}) \implies x_t=\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\epsilon \quad (1)$$

where $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\epsilon \sim (\mathbf{0}, \mathbf{I})$

In the reverse, x_0 given x_t is unknown and is estimated through $\hat{x}_0 = (x_t - \sqrt{1-\overline{\alpha}_t}\epsilon)/\sqrt{\overline{\alpha}_t}$ where $\epsilon_\theta(x_t|t)$ estimates ϵ using a time conditional neural network parameterized by θ. The loss function of the neural network that recovers data from noise is

$$L(\theta)=\mathbb{E}_{t,x_0\sim p_{data},\epsilon\sim(\mathbf{0},\mathbf{I})}[||\epsilon-\epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\epsilon|t)||_2^2] \quad (2)$$

3.2. Pseudo-3D Convolutional Layers

The funneling building block structure is straightforward: an identically-initialized 1D convolution for learning temporal coherence is inserted after each 2D convolution. The "pseudo" architectural appendage institutes information transfer between the spatial and temporal axes while circumventing exorbitant 3D convolutional burdens. Because the 1D convolutions are distinct from the 2D convolutions, temporal understanding can be separately trained from scratch while effortlessly retaining rich and heterogeneous knowledge from the image domain [11, 13].

The spatiotemporal-approximating Pseudo-3D convolutional layer is defined as

$$Conv_{P3D}(h):=Conv_{1D}(Conv_{2D}(h)\circ T)\circ T \quad (3)$$

where the input tensor $h \in^{B\times C\times F\times H\times W}$ has dimensions indicating the batch index, channels, number of frames, height, and width, and where the transpose operator facilitates interdimensional swaps.

3.3. Pseudo-3D Attention Layers

The dimension decomposition strategy is extended to the attention layers, wherein temporal attention layers are stacked behind spatial attention layers. Again, for smooth spatiotemporal initialization, the 2D attentional layer is initialized from pre-trained T2I model weights and the 1D attentional layer is identically-initialized [11, 13].

We define flatten as the condensing of the height and widths into $h' \in^{B\times C\times F\times HW}$, and unflatten as the inverse operator. The Pseudo-3D attention layers is defined as

$$ATTN_{P3D}(h)= \\ unflatten(ATTN_{1D}(ATTN_{2D}(flatten(h))\circ T)\circ T) \quad (4)$$

In contrast to VDM's unflattened 1x3x3 convolution filters, Pseudo-3D Attention applies an additional 3x1x1 convolution projection to construct temporal connections on adjacent features in time (Qiu et al. 2017) [11].

3.4. Frame Rate Conditioning

A conditioning frames-per-second parameter addresses limited available training videos volume and provides additional control during inference [13].

4. Training

Independent video frames are first extracted and compressed to semantically fundamental embeddings using RunwayML's released Stable Diffusion variational autoencoder. Prompt embeddings that are most representational according to least cosine distance are derived from OpenAI's CLIP (Contrastive Language-Image Pre-training) ViT-Large/14 text encoder.

Fine-tuning is performed for 1800 to 3600 steps on just the temporal layers, i.e. the spatial layers are frozen by setting `requires_grad` to false. At each step, a stack of 5 consecutive frame embeddings' marginal latents are proportionally noised to some arbitrary time step and processed through the conditional U-Net with the accompanying text embedding to return the noise vector. Mean squared error loss, known to result in smoother visuals, was elected. Lion (Evolved Sign Momentum) optimizer, more performant and with less memory overhead than Adam, computed updates of uniform magnitude (Chen et al., 2023) [1] at a learning rate of 0.00003.

5. Experiments

5.1. Video Selection and Preprocessing

Our efforts are focused on fine-tuning individual tasks and videos. Fine-tuning training occurs on several videos varying in content, length and quality sourced from different online video databases. LateNtMovies targets video synthesis in the natural image realm, e.g. cloudscapes, sunset/sunrise time-lapses, and moving water. Each input video is paired with a user-suggested text prompt that is transformed using a frozen CLIP language model to produce the

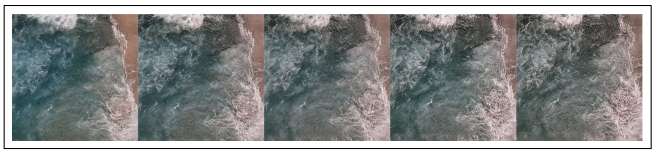

Figure 3. 3600 training steps at 5 frames per step and 15 inferencing steps to synthesize the above 5 frames.

appropriate text embeddings to then feed into training. We generate samples that abide by a similar content distribution as our fine-tuning footage, e.g. if shown ocean waves, inference on still photos of water surface.

5.2. Quantitative Results

Evaluation on CLIP score To ensure that the videos comply with the specified text conditions, we collected a set of natural images to use for verification. This involved carefully selecting a diverse range of images that captured the intended content and mood of each video. We then compared the performance of our approach using the [5] CLIP score with that of CogVideo [8], a widely used benchmark for video synthesis. Our results showed that our approach achieved a score of around 0.80, which was significantly higher than that of CogVideo. This suggests that our approach is more effective at synthesizing videos that align with the given text prompts, and demonstrates the potential of using advanced language models such as CLIP for video synthesis applications.

Evaluation on FID We calculated the Fréchet Inception Distance (FID) of our best sample. The reference video used during training had a resolution of 1920x1029 and consisted of 1060 frames, while our sample had a resolution of 512x512 and only 5 frames. We got the 164 FID scroe for our samples, and got 384 for the samples from CogVideo. Less FID score means our model successfully generated similar distributions than the CogVideo.

Method	Preferences	Realism	Resolution	CLIPSIM(↑)
CogVideo	21%	24%	480x480	0.7522
LateNtMovies(ours)	79%	76%	512x512	0.7597

Table 1. Video generation evaluation of LateNtMovies and CogVideo

5.3. Qualitative Results

Examples of LateNtMovies are shown in Figure 4, and are compared with the output from CogVideo. The wave simulations generated by LateNtMovies exhibit a high level of realism and quality, outperforming those produced by CogVideo. Our model is able to generate detailed waves that closely resemble those in real videos, whereas CogVideo fails to capture this level of detail. Additionally, the cat generated by our model is more realistic, with detailed fur that closely resembles that of a real cat. In contrast, the cat generated by CogVideo lacks these details, making it easily noticeable that it is not a real video. These examples demonstrate the superiority of LateNtMovies over CogVideo in terms of generating high-quality, realistic videos.

Figure 4. Comparison between LateNtMovies and CogVideo

Human evaluation Although the CLIP score for LateNtMovies is similar to that of CogVideo, we have demonstrated that the output from LateNtMovies is more realistic than that of CogVideo. To confirm this, we conducted a comparison between LateNtMovies and CogVideo using the same dataset that was trained on LateNtMovies. In this evaluation, we focused on assessing preferences and realism between the two methods. For the preference test, we asked human evaluators to select their preferred videos from LateNtMovies and CogVideo. Results showed that 78% of evaluators selected LateNtMovies as their preferred output. For the realism test, evaluators were asked to rank the quality of the videos from LateNtMovies and CogVideo. We found that 74% of evaluators ranked LateNtMovies as being closer to real videos, while only 30% selected CogVideo. These results indicate that LateNtMovies outperformed CogVideo in terms of generating realistic videos, especially for natural images. Our approach successfully captures how the real world moves, which is critical for applications such as video synthesis

6. Discussion

We will discuss the setbacks and limitations we encountered during the development of the LateNtMovies model, as well as future research directions to address these issues. Initially, our research showed us that all the components needed to develop a super-resolution video diffusion model existed, and we simply had to combine them. However, as our project progressed beyond the research phase we found that most of these components lacked official codebases and were unrealistic to train with our available time and resources. As a result, we had to pivot in a new direction, which had significant overlap with our existing work ultimately leading to the development of LateNtMovies. Throughout the project, we faced severe memory

constraints on the Newton server, limiting the length of our video generation. Generating high-resolution frames takes up a significant portion of the available memory limiting the number of frames we are able to generate resulting in only short video generation. We also identified some limitations of the current LateNtMovies model including sensitivity to camera movement in the fine-tuning video, the significant impact of alignment between the fine-tuning video and the conditioning image on the realism of the output, and the poor cropping of subjects in certain fine-tuning videos.

Some of the future directions for research to improve the limitations of the current LateNtMovies models include finding a better fine-tuning video dataset with static cameras and clearly framed moving subjects. This remains a critical limitation in the development of video synthesis research, as most research is the result of the acquiring or possession of vast private datasets. Another direction to explore is a more robust alignment method between the fine-tuning video and conditioning image or finding new methods of conditioning the model. Finally to fix the cropping issue an auxiliary function can be used to crop the video based on the subject(s). Lastly, it would be interesting to explore the effects of a cosine noise scheduler–we used a linear beta schedule in our training.

7. Conclusion

Overall, LateNtMovies presents a promising latent diffusion approach to video synthesis that can significantly reduce the resource budget required for video modeling while maintaining high visual and contextual fidelity. As inspired by Make-A-Video's approach, the use of text-conditioning combined with pseudo-3D convolutional and attention layers enables the accelerated generation of temporally coherent, and semantically well-structured high-resolution videos, curtailing the need for large semantic video data sets for intuiting motion.

References

[1] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In arXiv:2302.06675v2, 2023. 3

[2] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. In arXiv:2209.04747v4, 2023. 1

[3] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In arXiv:2302.03011, 2023. 2

[4] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In arXiv:2205.11495v3, 2022. 1

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In arXiv preprint arXiv:2104.08718., 2022. 4

[6] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. In arXiv:2210.02303, 2022. 1, 2

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In arXiv:2204.03458v2, 2022. 1, 2

[8] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In arXiv:2205.15868, 2022. 1, 4

[9] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. In arXiv:2206.07696v3, 2022. 1

[10] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. In arXiv:2302.01329, 2023. 2

[11] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In arXiv:1711.10305, 2017. 1, 3

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In arXiv:2112.10752v2, 2022. 1, 2

[13] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In arXiv:2209.14792, 2022. 1, 2, 3

[14] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In arXiv:2104.15069, 2021. 1

[15] Vikram Voleti and Alexia Jolicoeur-Martineau. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In arXiv:2205.09853v4, 2023. 1

[16] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. In arXiv:2203.09481v5, 2022. 1