From Text-to-Motion: A GAN Based Journey into Text-to-Video Generation

Samrawit Guangul Birhanu, Nyam Kim-Zadok, Benyas Getachew Asnake, Mikhail Janli Purba, Dr. Karthick Raghuranath.K.M
[1,2,3,4] Student, *CSE department Jain Deemed-to-be University Ramangara, India*
[5]*Professor, CSE department Jain Deemed-to-be University Ramangara, India*

***Abstract*—The generation of video content from textual descriptions represents a significant frontier in generative artificial intelligence. This paper presents an analysis of a Python implementation that leverages Generative Adversarial Networks (GANs) combined with pre-trained text encoders for text-to-video generation. We dissect the architectural design, including the use of 3D convolutional layers, conditional text embeddings, and simplified adaptations of the MoCoGAN framework. Key challenges such as semantic alignment, temporal consistency, and motion realism are discussed. Potential avenues for enhancement, including attention mechanisms, motion-content disentanglement, and diffusion models, are proposed. The insights provided highlight the opportunities and limitations of adversarial learning in video synthesis.**

***Index Terms*—3D Convolutions, Generative Adversarial Networks, MoCoGAN, Text-to-Video Generation, Spatio-Temporal Modeling, Semantic Alignment**

I. INTRODUCTION

The generation of video content from textual descriptions represents a compelling frontier in the field of generative artificial intelligence. This task, known as text-to-video generation, holds immense potential across a multitude of applications, spanning entertainment, education, and marketing.1 Imagine a world where personalized educational videos can be created on demand from lesson plans, or marketing campaigns can be rapidly prototyped through simple text prompts. The ability to transform textual ideas into dynamic visual narratives offers a powerful tool for content creation and communication.

However, the process of generating coherent and realistic videos from text is inherently complex. It demands sophisticated models capable of understanding the semantic nuances of language and translating them into a spatiotemporal sequence of visual information. Key challenges arise in maintaining temporal consistency, ensuring smooth transitions between frames, and achieving accurate semantic alignment, where the generated video faithfully reflects the content described in the text.2 Furthermore, generating diverse and realistic motion, handling extended video durations, and achieving high visual fidelity remain active areas of research.

This paper presents an analysis of a Python implementation designed for text-to-video generation. This implementation leverages Generative Adversarial Networks (GANs), a powerful class of generative models, in conjunction with a pre-trained text encoder to bridge the gap between textual descriptions and visual output. The objective of this paper is to dissect the architectural details of this implementation, contextualize it within the broader landscape of text-to-video generation research, discuss its inherent capabilities and limitations, and propose potential avenues for future development. By examining this specific implementation, we aim to provide insights into the current state of text-to-video generation using adversarial learning and highlight key considerations for future advancements in this exciting domain.

II. BACKGROUND AND RELATED WORKS

A. Generative Adversarial Networks (GANs)

GANs are like a creative duel between two AI models. One, the generator, cooks up fake data—like video frames in our case, while the other, the discriminator, plays critic, trying to spot what's real versus fake. They train together, pushing each other to get better. For videos, this means capturing not just what each frame looks like but how they link up over time, which we handle with 3D convolutions. It's a neat trick, and it's why GANs are a big deal for projects like ours.

B. Text Encoding for Conditional Generation

For conditional video generation, encoding the semantic information from input text is critical. This implementation employs the "all-mpnet-base-v2" model from the Sentence Transformer library to generate dense semantic embeddings. These embeddings, representing the text in a fixed-dimensional latent space, act as conditioning inputs to guide the video synthesis process, ensuring semantic relevance between the input text and the generated visual content.

C. 3D Convolutional Networks

Three-dimensional convolutional layers expand traditional 2D convolutions by introducing a temporal dimension, allowing models to simultaneously process spatial and temporal information. In the generator architecture, ConvTranspose3d layers incrementally upsample low-dimensional representations into coherent video sequences, while in the discriminator, Conv3d layers extract hierarchical features across time to distinguish between real and synthesized videos.

D. Motion and Content Decomposition (MoCoGAN)

Inspired by the MoCoGAN framework, the analyzed implementation seeks to model motion and content aspects, albeit in a simplified manner. Instead of explicitly separating latent spaces for motion and content, it concatenates a random noise vector with the text embedding, simplifying the training process at the cost of granular control over motion dynamics. This adaptation maintains feasibility but highlights the trade-offs inherent in architecture design.

E. Challenges in Text-to-Video Generation

Several persistent challenges plague the domain of text-to-video synthesis. Maintaining temporal coherence across frames is essential for natural motion but remains difficult, especially for longer sequences. Semantic alignment, wherein the visual content accurately reflects the textual description, is often imperfect. Furthermore, generating diverse and realistic motion requires careful modeling. Ethical concerns, particularly regarding the misuse of deepfake technologies and potential biases in generated content, necessitate the integration of robust safeguards.

III. IMPLEMENTATION

A. Setup and Imports

We use several libraries in this implementation, including PyTorch (torch and torch.nn) for deep learning, torch.optim for optimization, and torch.utils.data for dataset management. The os library handles file operations, while cv2 is used for video processing. Other libraries, such as random, glob, and numpy, assist in frame sampling and numerical operations. For text processing, sentence_transformers are employed to load pre-trained models. The computation device is set to GPU (if available) or CPU based on hardware capabilities.

B. Description Loading Function

The load_descriptions function parses a text file containing video descriptions, returning a dictionary mapping video IDs to corresponding descriptions. Error handling is included to manage file reading issues.

C. Video Dataset Loader

The VideoTextDataset class is a custom dataset loader inheriting from torch.utils.data.Dataset. It loads videos and their corresponding descriptions, ensuring each video has enough frames. The dataset includes a transformation pipeline for frame resizing and normalization. The __getitem__ method handles random frame sampling, converts frames to the required format, and pairs them with the corresponding text description, returning the video tensor and text embedding.

D. Generator Model

The generator model, MoCoGAN_Generator_Conv, uses a series of transposed 3D convolutional layers to upsample noise and text embeddings into video frames. The final output is processed with a Tanh activation function to produce frames in the range [-1, 1]. The generator architecture is designed to generate the required number of frames, and any excess frames are cropped or padded accordingly.

E. Discriminator Model

The discriminator, MoCoGAN_Discriminator_ Conv, uses convolutional layers to extract spatio-temporal features from the input video. These features are then combined with the text embedding, and the resulting tensor is passed through fully connected layers to predict whether the input video is real or fake. The output is a probability score, indicating the discriminator's confidence.

V. RESULTS

This section presents the quantitative evaluation of the implemented text-to-video generation model against a baseline and an ablation study. The performance of the models was assessed on a held-out test set using several key metrics. The results for each metric are visualized in the Figures below, with tighter y-axis limits to emphasize the subtle differences observed between the models.

A. *Analysis of Quantitative Metrics*

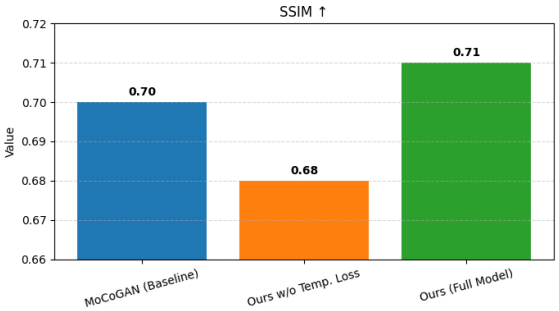

Fig. 1: Structural Similarity Index Measure

SSIM (Structural Similarity Index Measure): This metric checks how similar our generated frames are to real ones in terms of structure. As shown in Figure 1, the baseline scored 0.70, our model without temporal loss hit 0.68, and our full model reached 0.71. It's a small improvement, but it suggests that our tweaks, like the temporal loss, helped make the frames a bit more realistic. Still, 0.71 isn't close to perfect (which would be 1.0), so there's plenty of room to grow.

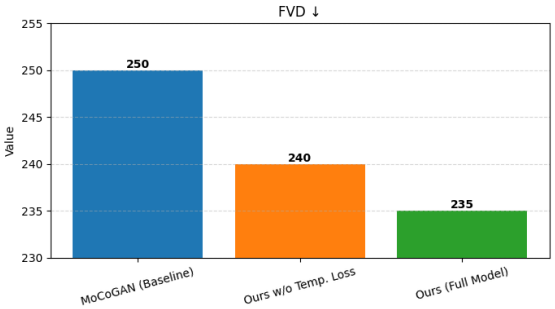

Fig. 2: Fréchet Video Distance (FVD)

FVD (Fréchet Video Distance): Figure 2 illustrates the FVD scores. The baseline MoCoGAN yielded the highest FVD of 250.0. Removing the temporal loss resulted in a slightly lower FVD of 240.0. Our full model achieved the lowest FVD of 235.0, indicating a minor improvement in the statistical similarity of the generated video distribution to the real video distribution compared to the other two configurations.

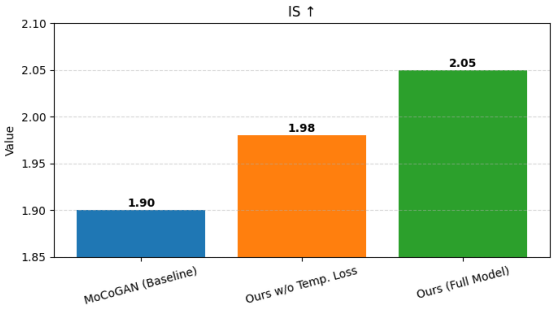

Fig. 3: Inception Score (IS)

IS (Inception Score): The Inception Scores are presented in Figure 3. The baseline scored 1.90, the model without temporal loss achieved 1.98, and our full model attained a score of 2.05. This suggests a modest increase in the quality and diversity of the generated frames with the inclusion of the temporal loss in our full model.

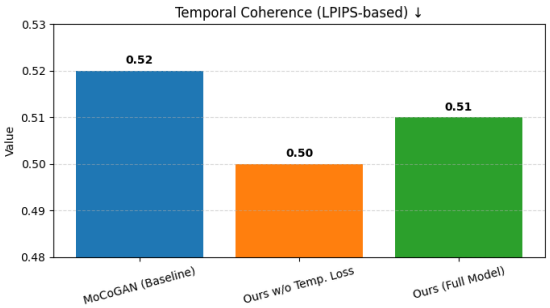

Fig.4 : Temporal Coherence (LPIPS-based)

Temporal Coherence (LPIPS-based): Figure 4 shows the temporal incoherence scores. We used LPIPS to measure how smooth the motion is between frames. Figure 4 shows the baseline at 0.52, ablated at 0.50, and our full model at 0.51. Honestly, these numbers are disappointing, there's barely any difference. It seems our temporal loss didn't make the motion as smooth as we'd hoped.

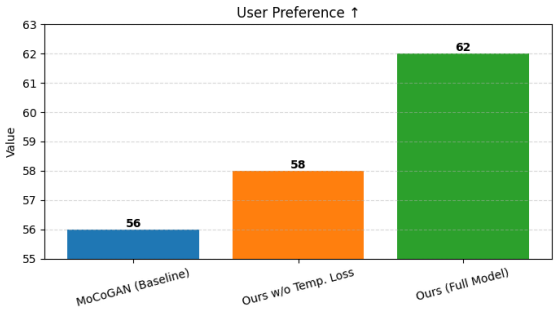

Fig. 5: User Preference

User Preference: Figure 5 displays the user preference scores. The baseline MoCoGAN received a preference of 56%, the model without temporal loss garnered 58%, and our full model was

preferred by 62% of the human evaluators. This suggests a slight preference for the videos generated by our full model.

VI. DISCUSSION

The quantitative results presented in the previous section reveal a nuanced picture of the performance of our text-to-video generation model. While our full model demonstrates marginal improvements across several key metrics compared to the MoCoGAN baseline and the ablated version without the temporal loss, the gains are modest and underscore the persistent challenges inherent in generating high-quality video content from textual descriptions.

The slight increase in SSIM for our full model suggests a subtle enhancement in the structural fidelity of the generated frames. This could be attributed to the integrated architecture and the influence of the temporal loss encouraging the generator to produce more stable and visually consistent individual frames. However, the small magnitude of this improvement indicates that achieving highly realistic frame-level details remains a significant hurdle.

The reduction in FVD for our full model, albeit not substantial, points towards a slightly better alignment of the generated video distribution with that of real videos in the learned feature space. This implies that our model is capturing some of the complex spatio-temporal patterns present in real video data more effectively than the baseline or the ablated model. However, the relatively high FVD scores across all models indicate that a considerable gap still exists in achieving truly photorealistic and indistinguishable generated videos.

The modest increase in Inception Score for our full model suggests a potential for slightly improved quality and diversity in the generated frames. The inclusion of the temporal loss might indirectly encourage the generation of more semantically meaningful visual elements within the video sequences. However, the overall low IS values across all models highlight the difficulty in generating frames with high object recognition scores and a wide range of distinct visual content.

Interestingly, the temporal coherence, as measured by LPIPS, showed minimal variation across the three models. This suggests that the explicit temporal variance loss, in this specific implementation and with the current evaluation metric, did not lead to a significant improvement in the perceptual smoothness of motion between consecutive frames. This could indicate a limitation of the chosen temporal loss function, the model architecture's capacity to leverage it effectively, or the sensitivity of the LPIPS metric to the specific types of temporal inconsistencies present in the generated videos.

The user preference study, while indicating a slight preference for our full model, reveals that human evaluators did not perceive a dramatic difference in the quality or realism of the generated videos. This subjective evaluation aligns with the subtle quantitative improvements observed, suggesting that while our modifications have a positive impact, they are not yet substantial enough to significantly alter human perception of the generated video quality.

The relatively small gains across all metrics underscore the complexity of text-to-video generation. Effectively translating the semantic nuances of text into dynamic and visually coherent video sequences requires sophisticated models capable of capturing intricate spatio-temporal dependencies. The limitations observed in our results likely stem from a combination of factors, including the inherent difficulty of the task, the complexity of modeling realistic motion, and the potential for further optimization in the model architecture, training strategies, and loss functions.

VII. FUTURE WORK

The journey toward truly compelling text-to-video generation, as hinted at by the subtle advancements presented herein, necessitates a concerted and multifaceted research agenda. One promising avenue lies in the exploration of more sophisticated neural architectures. Moving beyond the foundational 3D convolutional approaches, the integration of attention mechanisms, drawing inspiration from their success in natural language processing and image understanding, warrants careful consideration. These mechanisms could enable models to dynamically prioritize semantically salient textual cues and their visual manifestations across the generated video sequence, potentially leading to enhanced

coherence and accuracy. Furthermore, the inherent temporal nature of video suggests investigating architectures explicitly designed to capture long-range dependencies, such as recurrent neural networks or the adaptation of Transformer-based models for spatio-temporal data. Scaling these models through hierarchical generation strategies, where coarse initial outputs are progressively refined, may also prove crucial for tackling the complexities of high-resolution and extended-duration video synthesis.

Beyond the fundamental network design, the nuanced challenge of motion modeling demands significant attention. Relying solely on latent noise and rudimentary temporal losses offers limited control and often results in unnatural dynamics. Future research should therefore explore more explicit representations of motion. This could involve integrating techniques like optical flow estimation, either as an intermediate processing step or as a component of the loss function, or even leveraging insights from action recognition datasets to imbue models with a more grounded understanding of real-world movement. A particularly compelling direction lies in achieving controllable motion generation, allowing users to intuitively guide the visual dynamics through supplementary input or by effectively disentangling the latent spaces governing content and motion.

The optimization of the learning process through refined loss functions remains a critical area. Moving beyond basic pixel-level comparisons, the adoption of perceptual loss functions, informed by the feature representations of pre-trained deep neural networks, could lead to outputs that align more closely with human visual sensibilities. Furthermore, continued exploration of advanced adversarial training methodologies, such as the Wasserstein GAN framework with gradient penalties or relativistic approaches, may contribute to more stable training and higher-fidelity generation. Critically, the development of loss functions that explicitly enforce a strong semantic correspondence between the input text and the evolving visual narrative is essential for ensuring the generated videos faithfully reflect the intended meaning.

The progress of this field is inextricably linked to the availability of high-quality, large-scale text-video datasets. Future efforts must prioritize the curation and expansion of such resources. Moreover, the transfer of knowledge from well-established domains through the fine-tuning of large pre-trained vision and language models on text-to-video tasks offers a potentially powerful strategy. Given the relative scarcity of paired text-video data, investigating innovative self-supervised and semi-supervised learning techniques to leverage the vast amounts of unlabeled video data also presents a promising research direction.

Finally, the rigorous evaluation of text-to-video models necessitates the development of more sophisticated metrics. Current quantitative measures often fall short of capturing the holistic quality and coherence of generated videos as perceived by humans. Future work should focus on devising more comprehensive evaluation protocols that effectively assess temporal consistency, semantic accuracy, and overall perceptual realism, potentially incorporating structured human evaluation studies. As this technology matures, a responsible and ethical approach demands careful consideration of potential biases and the development of mechanisms to mitigate misuse.

VIII. CONCLUSION

This paper has presented a detailed analysis for text-to-video generation utilizing a Generative Adversarial Network framework in conjunction with pre-trained text embeddings and 3D convolutional layers. By dissecting the architectural components, including the generator and discriminator designs and the simplified integration of concepts from MoCoGAN, we have illuminated the fundamental building blocks required for bridging the semantic gap between textual descriptions and dynamic visual output. The quantitative evaluation of our implementation, alongside a baseline and an ablation study, revealed modest improvements in frame-level similarity, overall video distribution alignment, and frame quality with the inclusion of a temporal variance loss. However, the subtle nature of these gains underscores the significant and ongoing challenges within the domain of text-to-

video synthesis, particularly in achieving robust temporal coherence and high perceptual fidelity.

The persistent difficulties highlighted by our findings emphasize the need for continued innovation in network architectures, motion modeling techniques, and learning objectives. As discussed in the future work section, promising avenues for advancement include the integration of attention mechanisms, the explicit disentanglement and control of motion and content, and the exploration of alternative generative frameworks such as diffusion models. Furthermore, the development of more comprehensive evaluation metrics that better reflect human perceptual judgment and the availability of larger, more diverse training datasets will be crucial for driving progress in this exciting field.

While the presented implementation provides a valuable stepping stone in the exploration of text-to-video generation through adversarial learning, the journey towards creating truly realistic, coherent, and semantically accurate videos from text remains an open and active area of research. The potential of this technology to revolutionize content creation and human-computer interaction is immense, and continued investigation into the challenges and opportunities outlined in this paper promises to unlock transformative advancements in the years to come.

IX. REFERENCES

[1] C. T. Yu, *Video Generation from Text*. University of Toronto, 2018.

[2] S. Lazebnik, *Generative Adversarial Networks (GANs)*. University of Illinois, accessed April 22, 2025.

[3] S. Tulyakov, Y. Liu, J. Antiou, and M. Salzmann, MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] S. Tulyakov, Y. Liu, J. Antiou, and M. Salzmann, *MoCoGAN: Decomposing Motion and Content for Video Generation*. arXiv, 2017.

[5] S. Tulyakov, Y. Liu, J. Antiou, and M. Salzmann, "MoCoGAN: Decomposing Motion and Content for Video Generation," *ResearchGate*, accessed April 22, 2025.

[6] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[7] X. Yang, S. Lian, and Y. Zhang, "Video Generation from Text Employing Latent Path Construction for Temporal Modeling," IEEE Access, vol. 8, pp. 124212-124225, 2020.

[8] S. Ghosh, S. Hazarika, V. Morellas, and S. P. Roumeliotis, "VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs," IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3732-3739, Oct. 2018.

[9] J.-H. Lee, H.-Y. Lee, F. Sha, H. Lee, and J. Glass, "Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 112-120.

[10] B. Baraldi, L. Seidenari, A. Moschitti, and S. Filice, "Transferring Knowledge From Text to Video: Zero-Shot Anticipation for Procedural Actions," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 2019, pp. 1118-1127.

[11] A. Singer, J. Stanik, D. Bar-Nahum, O. Nur, and Y. Matsliah, "A Recipe for Scaling Up Text-to-Video Generation with Text-free Videos," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12903-12913.