Machine Learning

ELEC-8900

Project Proposal

Zero-Shot Video Generation

Group Members:

1.	Amey Mahendra Thakur	110107589
2.	Jithin Gijo Varghese	110120947
3.	Ritika Agarwal	110121443

Group Number: 19

Date of Submission: October 01, 2023

University of Windsor

Department: Faculty of Engineering

Instructor: Dr. Yasser M. Alginahi

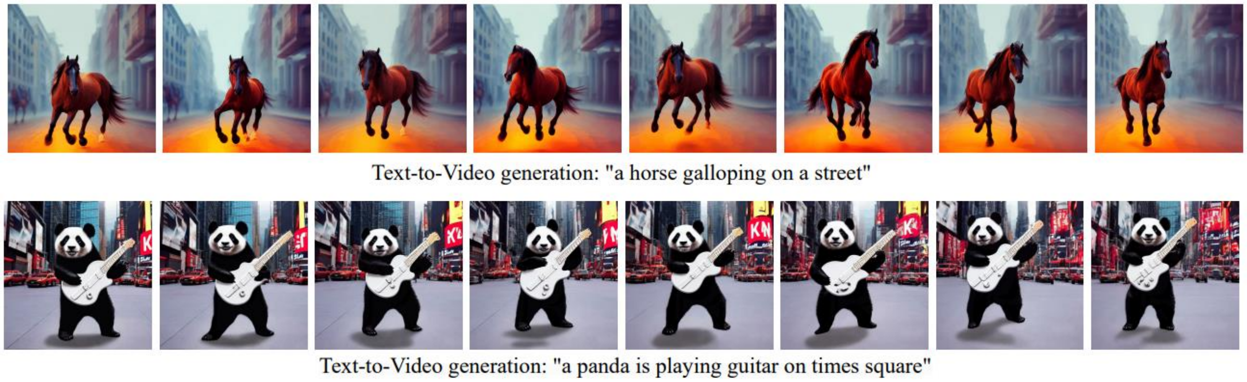
List of Figures

List of Tables

Table of Contents

I. Introduction

The field of artificial intelligence continually seeks to break barriers between different forms of media. At the forefront of this endeavor stands the research titled "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators" conducted by Picsart AI Research Lab [1]. This study introduces a method that converts textual descriptions directly into videos, marking a significant advancement in the integration of natural language processing and computer vision. Such a development not only addresses the growing demand for dynamic visual content but also showcases the potential of machines to interpret and render human language in a visual format. By offering a solution to the challenge of text-to-video synthesis, the research sets a new benchmark for interdisciplinary studies in artificial intelligence.



Text-to-Video generation: "a horse galloping on a street"

Text-to-Video generation: "a panda is playing guitar on times square"

Figure 1: Text-to-Video generation [1]

II. Project Motivation

In an era where visual storytelling is paramount, the ability to convert textual narratives into dynamic videos holds transformative potential. As platforms and audiences increasingly favor visual content, the research from Picsart AI Research Lab, titled "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators," emerges as a timely and innovative response to this demand [1]. Delving into its significance reveals:

A. ***Blending Words with Vision:*** The initiative by Picsart AI Research Lab offers a novel approach, merging the realms of text interpretation and visual representation. This is not just about generating images; it's about crafting a coherent visual story based on textual cues.

B. ***A New Era of Content Creation:*** With the digital landscape being saturated with content, differentiation becomes key. A tool that can take textual descriptions and produce videos offers a unique edge, streamlining content creation and offering bespoke visual outputs.

C. ***Making Learning More Visual:*** In education, the value of a tool that can translate textual concepts into visual content is immeasurable. It offers a tangible way to represent abstract ideas, catering to a broader spectrum of learners.

D. ***Handling Big Data Challenges:*** The emphasis on utilizing large image datasets signifies the project's ambition to operate at scale, ensuring that vast amounts of data can be processed without compromising on the quality of the generated videos.
E. ***Prioritizing User Experience:*** By integrating a user interface, the project underscores its commitment to accessibility. It's a nod to the importance of ensuring that such groundbreaking technology is usable and beneficial to a wide audience.

In summary, the research by Picsart AI Research Lab is not just academically intriguing; it addresses a contemporary need, offering a solution that aligns with the evolving preferences of today's digital consumers.

III. Literature Review

The journey towards the synthesis of textual narratives into visual content has been paved by several groundbreaking works, each contributing a piece to the puzzle. The following literature chronicles this progression, leading up to the innovative "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators" by Picsart AI Research Lab:

Table 1: Literature Review

Year	Research Title	Authors	Contribution	Description
2014	Generative Adversarial Networks (GANs) [2]	Ian Goodfellow et al.	Introduction of GAN architecture	- Introduced the concept of two competing neural networks - Laid the foundation for image generation advancements [2]
2014	Large-Scale Video Classification with Convolutional Neural Networks [3]	Karpathy et al.	Insights into large-scale video generation	- Delved into intricacies of video data - Emphasized need for optimized computational processes [3]
2016	Generating Videos with Scene Dynamics [4]	Vondrick et al.	Exploration of video generation from cues	- Explored generation of short video clips - Provided insights into text-to-video synthesis challenges [4]
2021	DALL·E [5]	OpenAI	Advancements in text-to-image synthesis	- Generated diverse images from textual prompts - Bridged language and visuals [5]

2021	Diffusion Models [6]	Prafulla Dhariwal and Alex Nichol	Emphasis on iterative refinement	- Demonstrated effectiveness in high-quality image production - Highlighted potential of iterative techniques [6]

The following timeline visually represents the sequence of foundational works providing a clearer understanding of the research progression.

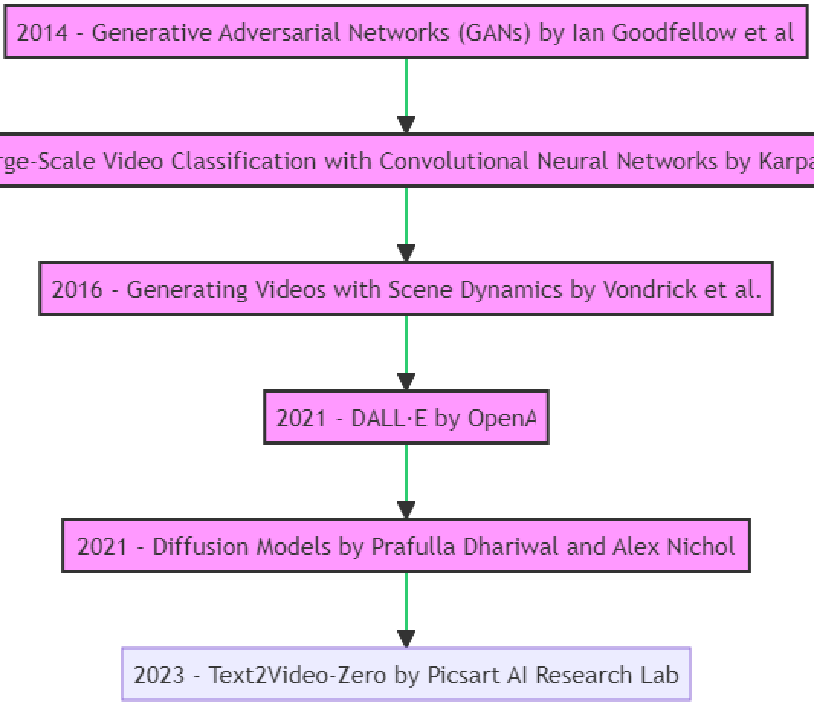

2014 - Generative Adversarial Networks (GANs) by Ian Goodfellow et al
2014 - Large-Scale Video Classification with Convolutional Neural Networks by Karpathy et al.
2016 - Generating Videos with Scene Dynamics by Vondrick et al.
2021 - DALL·E by OpenA
2021 - Diffusion Models by Prafulla Dhariwal and Alex Nichol
2023 - Text2Video-Zero by Picsart AI Research Lab

Figure 2: Research Progression

IV. Project implementation

The primary objective of this project is to implement and evaluate the "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators" model. This model promises to convert textual descriptions directly into videos, a significant advancement in the realm of AI-driven content generation.

The flowchart illustrates the sequential steps involved in the project:

A. ***Data Collection & Preprocessing:*** The initial stage involves sourcing and preparing the dataset.
B. ***Feature Extraction:*** This step focuses on extracting relevant features from both textual and visual data.
C. ***Model Implementation & Training:*** Here, the "Text2Video-Zero" model is adapted and trained.

D. ***Evaluation:*** The model's performance and the quality of generated videos are assessed.
E. ***User Interface Development:*** The project concludes with the creation of a user-friendly interface.

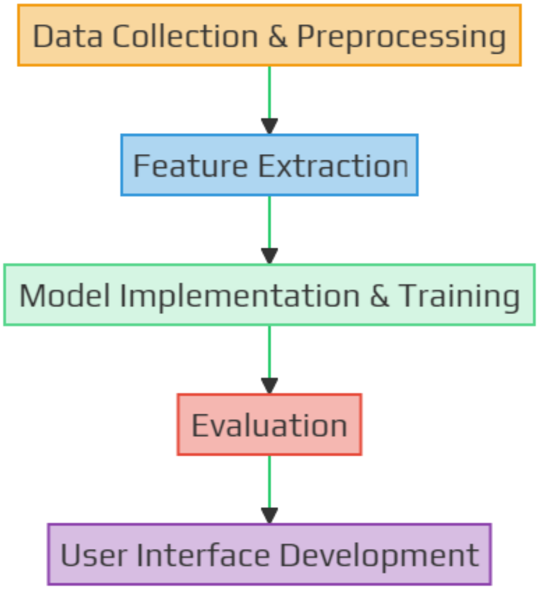

Data Collection & Preprocessing
Feature Extraction
Model Implementation & Training
Evaluation
User Interface Development

Figure 3: Flowchart

V. Methodology

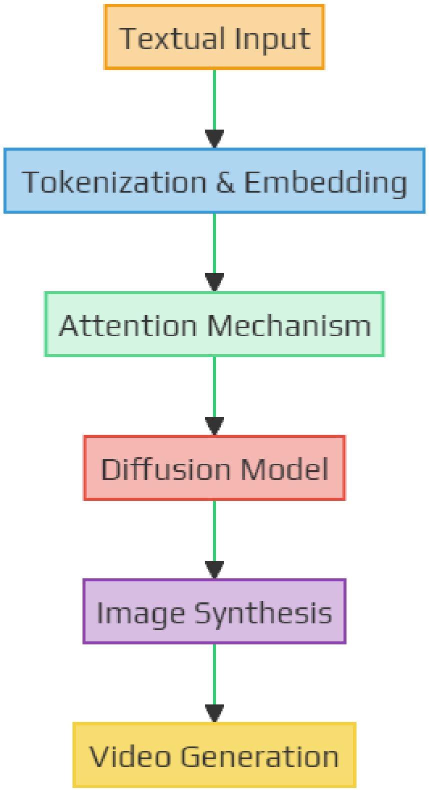


Figure 4: High-level architecture of the "Text2Video-Zero" model [1]

Table 2: Components of high-level architecture of the "Text2Video-Zero" model

Component	Description	Function
Textual Input	The starting point where a textual description or narrative is provided to the model.	Acts as the primary source of information that the model will use to generate visual content.
Tokenization & Embedding	The textual input is broken down into smaller chunks or tokens and then converted into numerical vectors.	Facilitates the model's understanding of the textual content by representing words or phrases in a format suitable for processing.
Attention Mechanism	A technique that allows the model to focus on specific parts of the textual input that are more relevant for the current task.	Enhances the model's capability to generate coherent visual content by emphasizing important textual cues.
Diffusion Model	The core of the "Text2Video-Zero" approach, responsible for the iterative refinement of the generated content.	Enables the model to produce high-quality images by refining the generated content over multiple iterations.
Image Synthesis	The model generates a static image based on the refined content from the diffusion model.	Acts as an intermediary step before video generation, ensuring that the initial frame or image aligns well with the textual description.
Video Generation	The model extends the synthesized image into a coherent video sequence, adding dynamic elements to the visual content.	Produces the final output, a video that visually represents the provided textual narrative.

VI. Dataset

A. ***Source:*** While there are several datasets available for image generation tasks, for this project, datasets like COCO (Common Objects in Context) or ImageNet could be considered due to their vastness and diversity. Additionally, datasets specifically designed for video tasks, like UCF101 or Kinetics, might be explored to understand temporal dynamics.

B. ***Key Features of the Dataset:***

1) *Diversity:* The dataset will include images from various categories, ensuring that the model can handle a wide range of textual prompts.
2) *High-Resolution:* To generate quality videos, the dataset will prioritize high-resolution images.

3) *Annotated Data:* Each image in the dataset will be paired with textual descriptions, aiding in supervised training.
4) *Temporal Consistency:* For video generation, the dataset will also include sequences of images that showcase movement or change over time.

VII. ML Libraries

The successful implementation of the "Text2Video-Zero" model requires a combination of specialized tools and libraries, each tailored to handle specific tasks within the project. The following table provides a comprehensive overview of these essential components, detailing their specific roles and features:

Table 3: Library/Tool used

Component	Library/Tool	Specific Role & Features
Programming Language	Python	- Core language for the project. - Offers a wide range of libraries and frameworks for machine learning and data processing.
Deep Learning Framework	TensorFlow [7]	- Build, train, and deploy ML models. - TensorFlow Extended (TFX) for end-to-end platform components. - TensorFlow Hub for reusable model components.
	PyTorch [8]	- Dynamic computation graph for flexibility. - TorchVision for computer vision tasks. - Native support for neural network architectures.
Image & Video Processing	OpenCV [9]	- Image and video capture, processing, and manipulation. - Feature extraction and object detection. - Integration with deep learning frameworks.
Natural Language Processing	HuggingFace's Transformers [10]	- Pre-trained models like BERT, GPT-2, and T5. - Tokenization and embeddings for various languages. - Pipelines for common NLP tasks.
User Interface Development	Streamlit [11]	- Rapid development of data applications. - Interactive widgets for user input. - Data visualization and plotting.
	Flask [12]	- Micro web framework for Python. - Flask-RESTful for building APIs. - Flask-Bootstrap for integrating Bootstrap with the application.

Dataset Handling	Pandas [13]	- Data manipulation and analysis. - DataFrame structure for tabular data. - Data cleaning, aggregation, and visualization.
	NumPy [14]	- Numerical operations and computations. - Array and matrix data structures. - Mathematical functions for linear algebra and statistics.

VIII. Timeline

F. Timeline

Table 4: Project Timeline

Date	Task
Oct 1	Project Description/Proposal Submission
Oct 2 - Oct 6	Conference with Instructor
Oct 7 - Oct 15	Data Preprocessing
Oct 16 - Oct 23	Model Development
Oct 24 - Oct 31	Training Phase I & Progress Report
Nov 1 - Nov 10	Training Phase II & UI Development
Nov 11 - Nov 15	Testing & Debugging
Nov 16 - Nov 19	Finalization, Documentation & Submissions
Nov 22	PPT File Submission
Nov 23 - Nov 24	Project Presentations

G. Gantt Chart

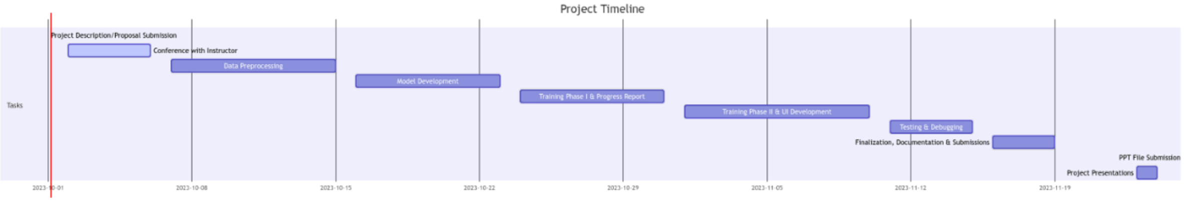


Figure 5: Gantt Chart

IX. Scope Limitations of the Text2Video-Zero Implementation

While the ambition of this project is to faithfully implement the "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators" model, several factors might limit its scope. The quality and diversity of the vast image dataset can significantly influence the model's performance, with inconsistent or biased data potentially leading to suboptimal results [1]. The computational demands of the model, especially during the training phase, might be constrained by available resources, potentially extending training durations or limiting the model's complexity.

Additionally, the model's generalization capabilities could vary based on the intricacy and ambiguity of the input text. The user interface, designed for seamless interaction, might face real-time video generation latencies due to server loads or input complexities. Lastly, the project's reliance on specific machine learning libraries means that unforeseen updates or changes in these libraries could introduce challenges or deviations in the intended functionality.

X. Conclusion

Project aims to implement the "Text2Video-Zero" model marks a pivotal step in AI-driven content creation. By transforming text into video, we address both technical challenges and the broader implications for the future of machine learning applications. As we progress, our focus remains on precision, user experience, and setting new standards in text-to-video synthesis.

References

[1] L. Khachatryan et al., “Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators,” *arXiv:2303.13439,* March 23, 2023, https://arxiv.org/abs/2303.13439 [Accessed: October 1, 2023].

[2] I. J. Goodfellow et al., “Generative Adversarial Networks,” *arXiv:1406.2661,* 2014, https://arxiv.org/abs/1406.2661 [Accessed: October 1, 2023].

[3] A. Karpathy et al., “Large-Scale Video Classification with Convolutional Neural Networks,” *2014 IEEE Conference on Computer Vision and Pattern Recognition,* June 2014, https://doi.org/10.1109/cvpr.2014.223 [Accessed: October 1, 2023].

[4] C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating Videos with Scene Dynamics,” *arXiv:1609.02612,* October 26, 2016, https://arxiv.org/abs/1609.02612 [Accessed: October 1, 2023].

[5] OpenAI Research Lab, “DALL·E: Creating images from text,” https://openai.com/research/dall-e [Accessed: October 1, 2023].

[6] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” *arXiv:2105.05233,* June 2021, https://arxiv.org/abs/2105.05233 [Accessed: October 1, 2023].

[7] M. Abadi et al., "TensorFlow: An end-to-end open source machine learning platform," *TensorFlow Documentation,* 2021, https://www.tensorflow.org/api_docs [Accessed: October 1, 2023].

[8] A. Paszke, S. Gross, and G.S. Chintala, "Pytorch: An Open Source Machine Learning Framework," *PyTorch Documentation,* 2016, https://pytorch.org [Accessed: October 1, 2023].

[9] Opencv, “Opencv/opencv: Open source computer vision library,” *GitHub,* https://github.com/opencv/opencv [Accessed: October 1, 2023].

[10] T. Wolf et al., “HuggingFace’s Transformers: State-of-the-art Natural Language Processing,” *arXiv:1910.03771,* February 2020, https://arxiv.org/abs/1910.03771 [Accessed: October 1, 2023].

[11] Streamlit Documentation, “Streamlit Docs,” https://docs.streamlit.io/ [Accessed: October 1, 2023].

[12] Flask Documentation, “Welcome to Flask — Flask Documentation (3.0.x),” https://flask.palletsprojects.com/en/3.0.x/ [Accessed: October 1, 2023].

[13] Pandas Documentation, “Pandas: powerful Python data analysis toolkit — pandas 0.9.1 documentation,” https://pandas.pydata.org/pandas-docs/version/0.9.1/index.html#:~:text=pandas%20is%20a%20Python%20package [Accessed: October 1, 2023].

[14] Numpy, “NumPy,” *Numpy Documentation,* 2009, https://numpy.org/ [Accessed: October 1, 2023].