Zero-Shot Video Generation

Team Members:

Amey Mahendra Thakur,

Jithin Gijo Varghese,

Ritika Agarwal

Instructor: Dr. Yasser Alginahi

Date: November 22, 2023

University of Windsor

AGENDA

- Introduction
- Problem Statement
- Dataset Overview
- Model Description
- References

INTRODUCTION

Bridging Text and Video with AI

- **Innovative AI Research:** "Text2Video-Zero" project by Picsart AI Research Lab.
- **Textual to Visual Transformation:** Converting text descriptions into dynamic videos.
- **Interdisciplinary Fusion:** Merging natural language processing and computer vision.
- **Meeting Modern Demand:** Addressing the growing need for dynamic visual content.
- **Visual Language Interpretation:** Enabling machines to render human language visually.
- **Setting AI Benchmarks:** Advancing interdisciplinary AI studies.

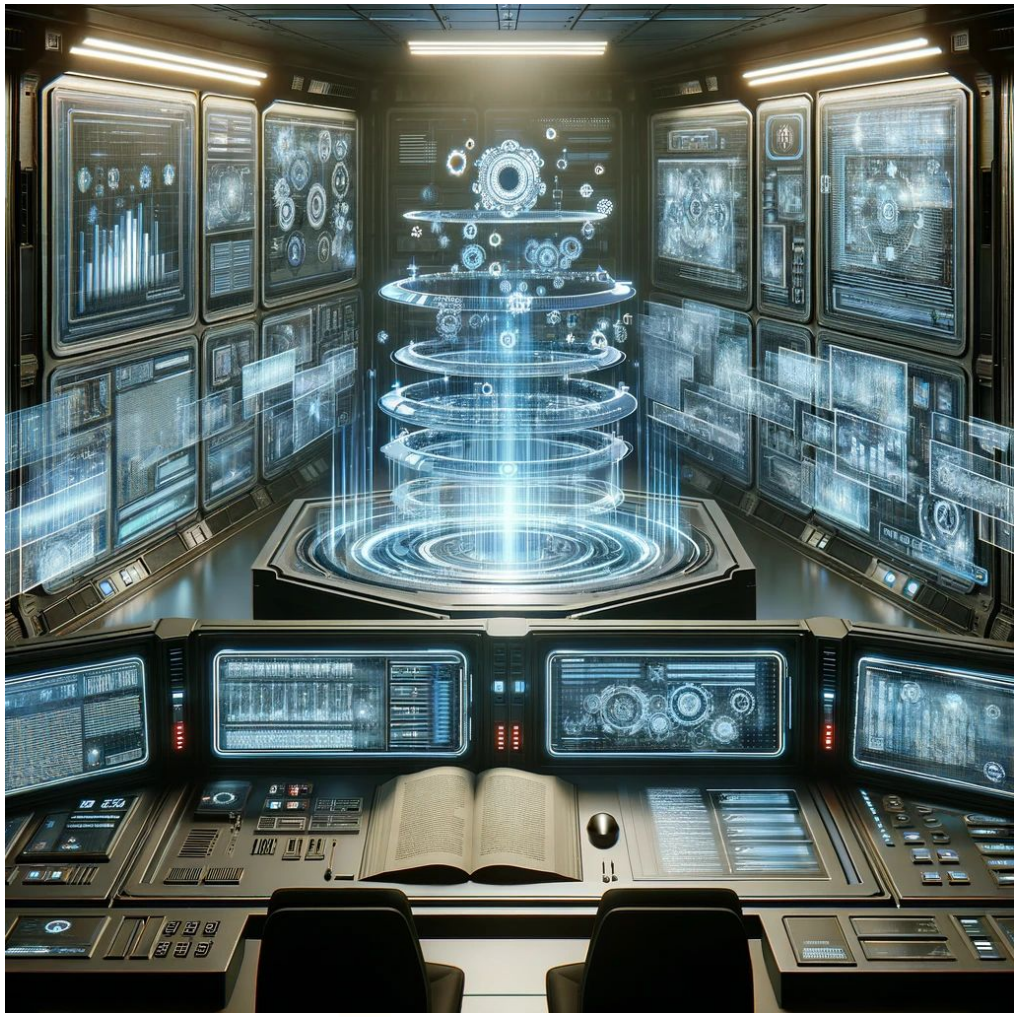
Text-to-Video Generation: A New Frontier

- **Emerging Research Field:** Text-to-video synthesis with autoregressive transformers and diffusion processes.
- **Notable Innovations:**
  - **NUWA:** Introduces a 3D transformer for text-to-image and video generation.
  - **Phenaki:** Utilizes a bidirectional masked transformer for generating long videos from text.
  - **CogVideo:** Adapts CogView 2 model with a training strategy aligning text and video.
  - **Video Diffusion Models (VDM):** Extends image diffusion models to video.
  - **Imagen Video:** Creates high-resolution, time-consistent videos using video diffusion models.
  - **Make-A-Video:** Builds on text-to-image models, using video data unsupervisedly.
  - **Gen-1:** Proposes a structure and content-guided video editing method.
  - **Tune-A-Video:** Focuses on one-shot video generation by tuning on a single reference video.
- **Approach:** Training-free, affordable video generation accessible to everyone, distinct from existing methods which often require significant computational resources.

PROBLEM STATEMENT

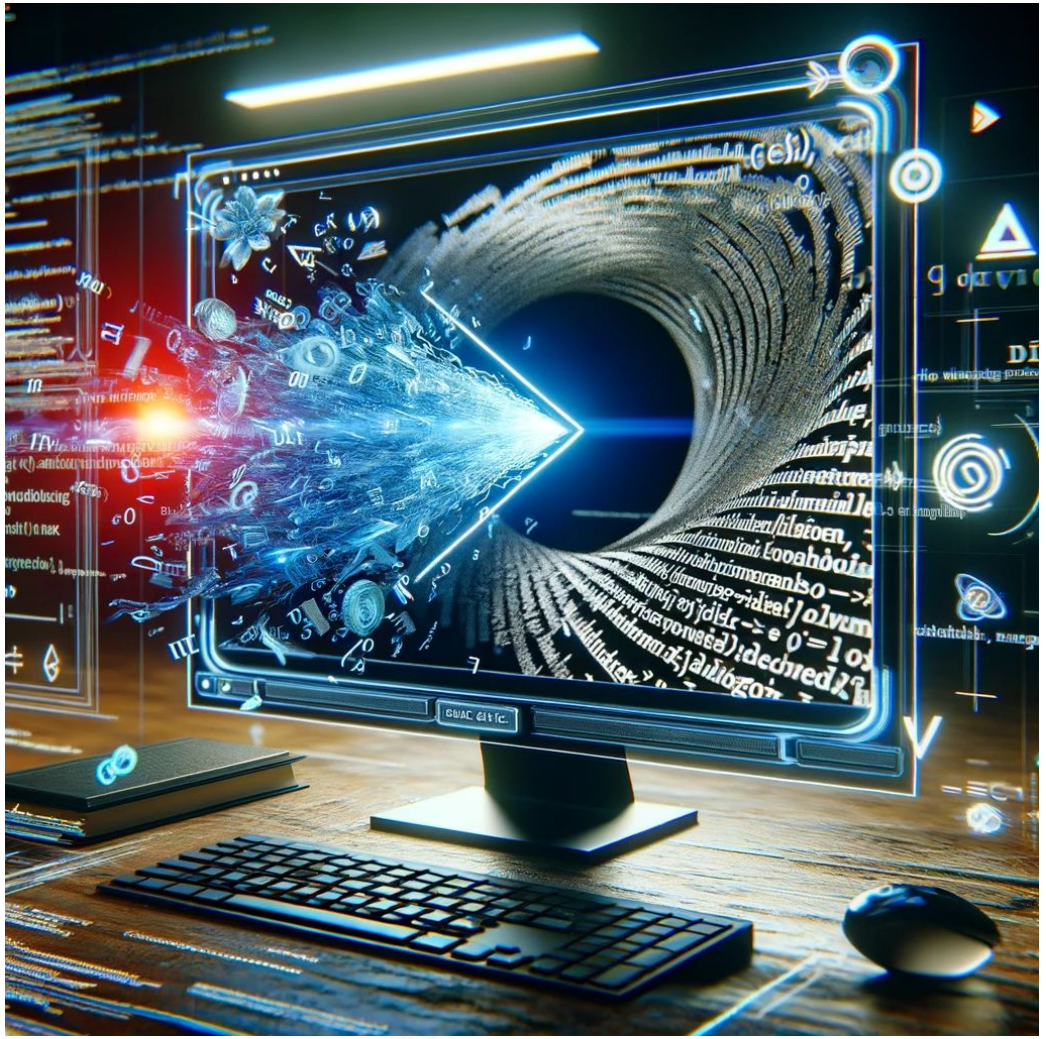
Revolutionizing Visual Storytelling

- **Emerging Needs:** Growing preference for visual content in the digital era.
- **Innovative Solution:** "Text2Video-Zero" by Picsart AI Research Lab, merging text and visuals.
- **Content Differentiation:** Unique edge in the crowded digital content landscape.
- **Educational Transformation:** Converting text prompts into visual educational tools.
- **Big Data Utilization:** Managing large datasets for high-quality video output.
- **User-Centric Design:** Focusing on accessibility with an intuitive interface.



The Impact of Text2Video-Zero

- **Blending Words and Vision:** Transforming textual cues into coherent visual stories.
- **Revolutionizing Content Creation:** Streamlining the process with customized visual outputs.
- **Enhancing Visual Learning:** Rendering abstract educational concepts tangible.
- **Balancing Scale and Quality:** Efficiently processing large datasets for quality videos.
- **Focusing on Accessibility:** Making technology user-friendly for diverse audiences.



DATASET OVERVIEW

9

Building Blocks: Datasets in Focus

- **Rich Datasets for Images:** Utilizing COCO and ImageNet for their vastness and diversity in image data.
- **Video-Specific Datasets:** Exploring UCF101 and Kinetics to understand motion and temporal dynamics.
- **Diversity in Data:** Ensuring the model's capability to interpret a wide array of textual prompts.
- **High-Resolution Priority:** Selecting datasets with high-resolution images for superior video quality.
- **Annotated for Accuracy:** Leveraging datasets with textual descriptions for effective supervised learning.
- **Temporal Elements:** Including sequences to capture movement and change, crucial for video synthesis.

A Glimpse into the Dataset

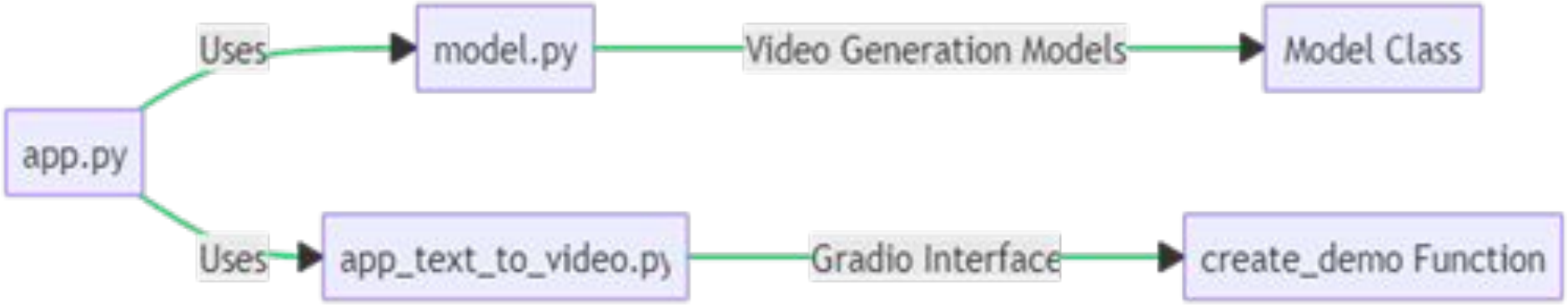
- **COCO & ImageNet Samples:** Showcasing diverse images from these datasets.
- **UCF101 & Kinetics Snippets:** Illustrating temporal dynamics with video sequences.
- **High-Resolution Focus:** Emphasizing the quality in model training images.
- **Annotated Data Showcase:** Displaying images alongside their textual descriptions.
- **Temporal Sequence Visualization:** Demonstrating consistency in video generation.

University of Windsor

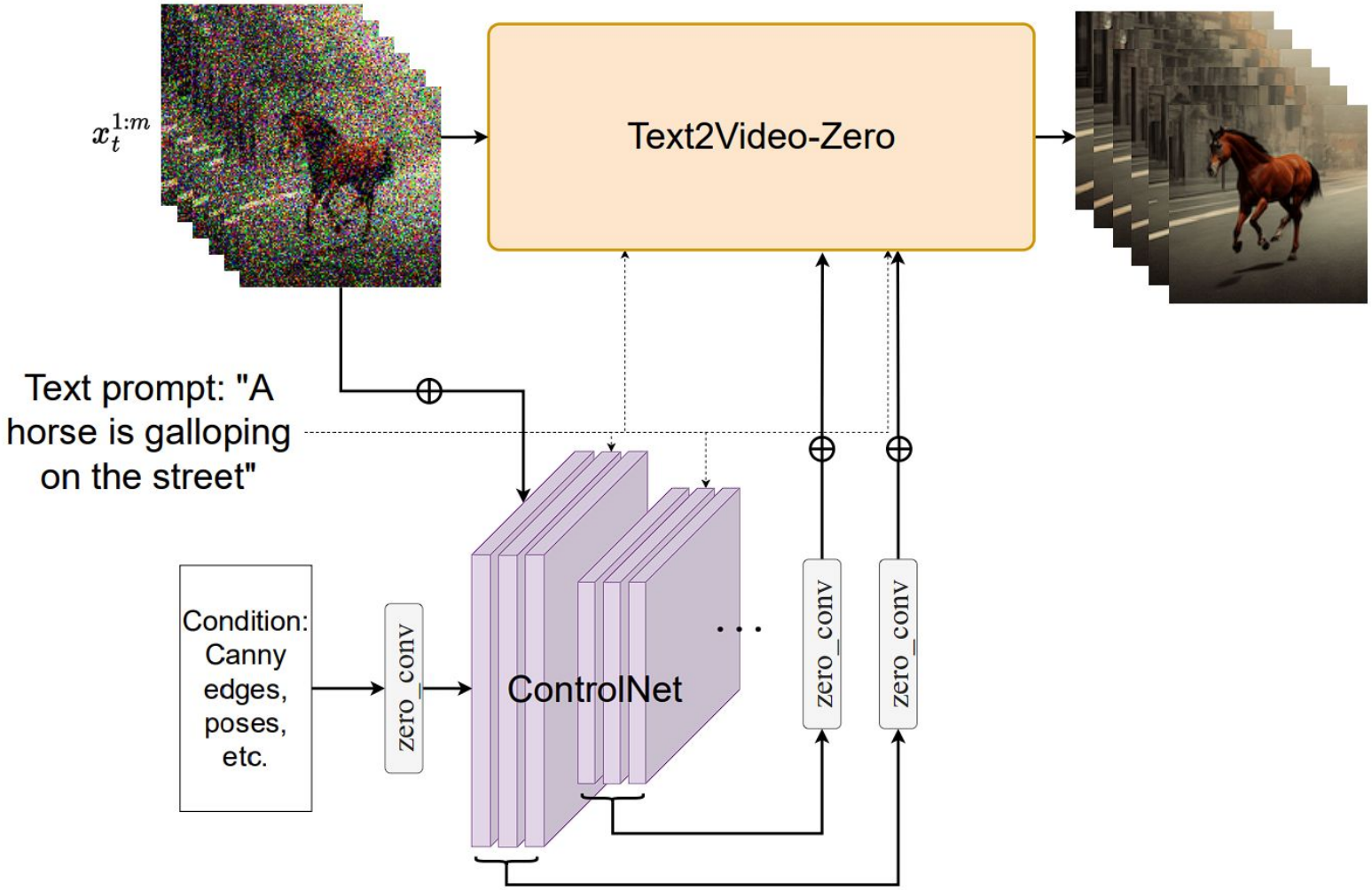
MODEL DESCRIPTION

Zero-Shot Video Generation: An Overview

- **Web-Based AI Application:** Facilitating text-to-video creation via a web interface.
- **Text-to-Image Model Integration:** Utilizing diffusion models for video synthesis.
- **Structured Implementation:** Organized through app.py, model.py, app_text_to_video.py.
- **User-Friendly Design:** Emphasizing intuitive and interactive controls for ease of use.



Diving into the Model Architecture



University of Windsor

Converting Words into Visual Narratives

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

- **Textual Input Processing:** How the system interprets and processes text.
- **Diffusion Model Integration:** Utilizing advanced models for video synthesis.
- **Visual Output Creation:** Transforming text into coherent video sequences.

Text to Video Generation: Features

No Motion in Latents
No Cross-Frame Attention

Motion in Latents
No Cross-Frame Attention

No Motion in Latents
Cross-Frame Attention

Motion in Latents
Cross-Frame Attention

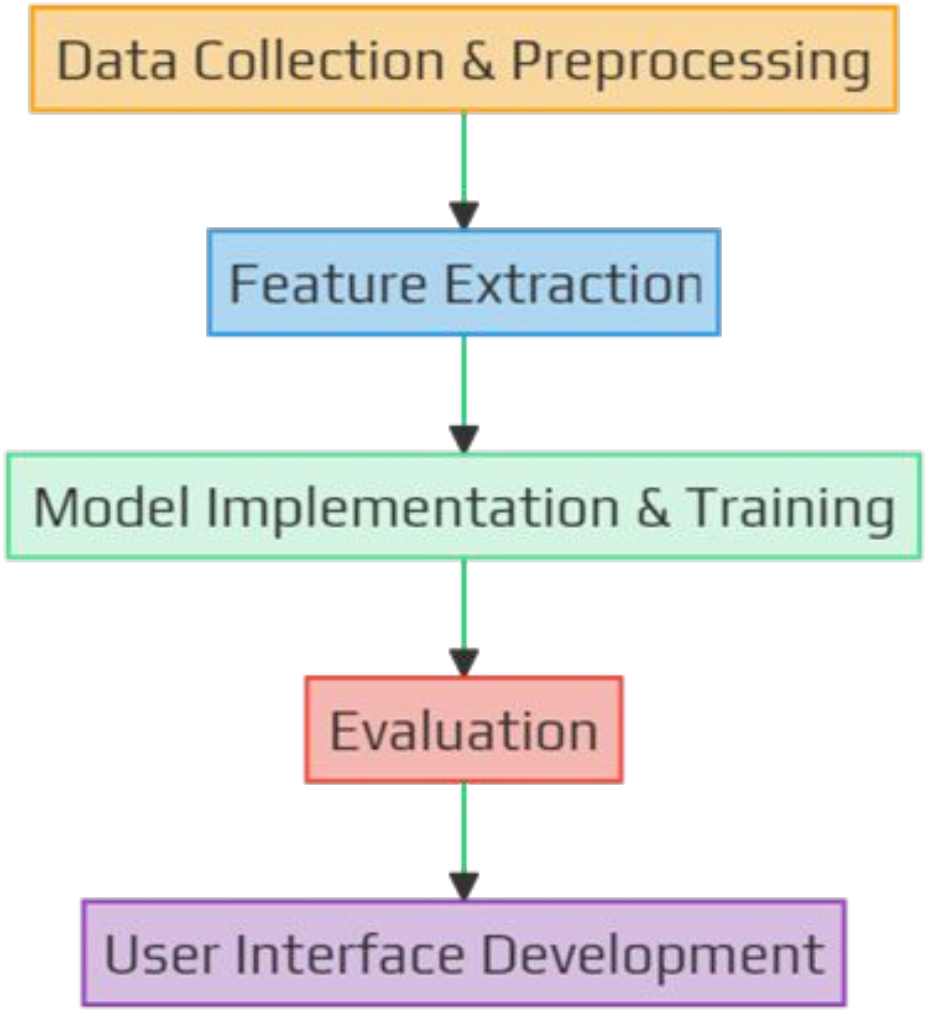
A demonstration on how unique features enhance text-to-video generation and text-guided video editing.

Key ML Libraries Powering Model

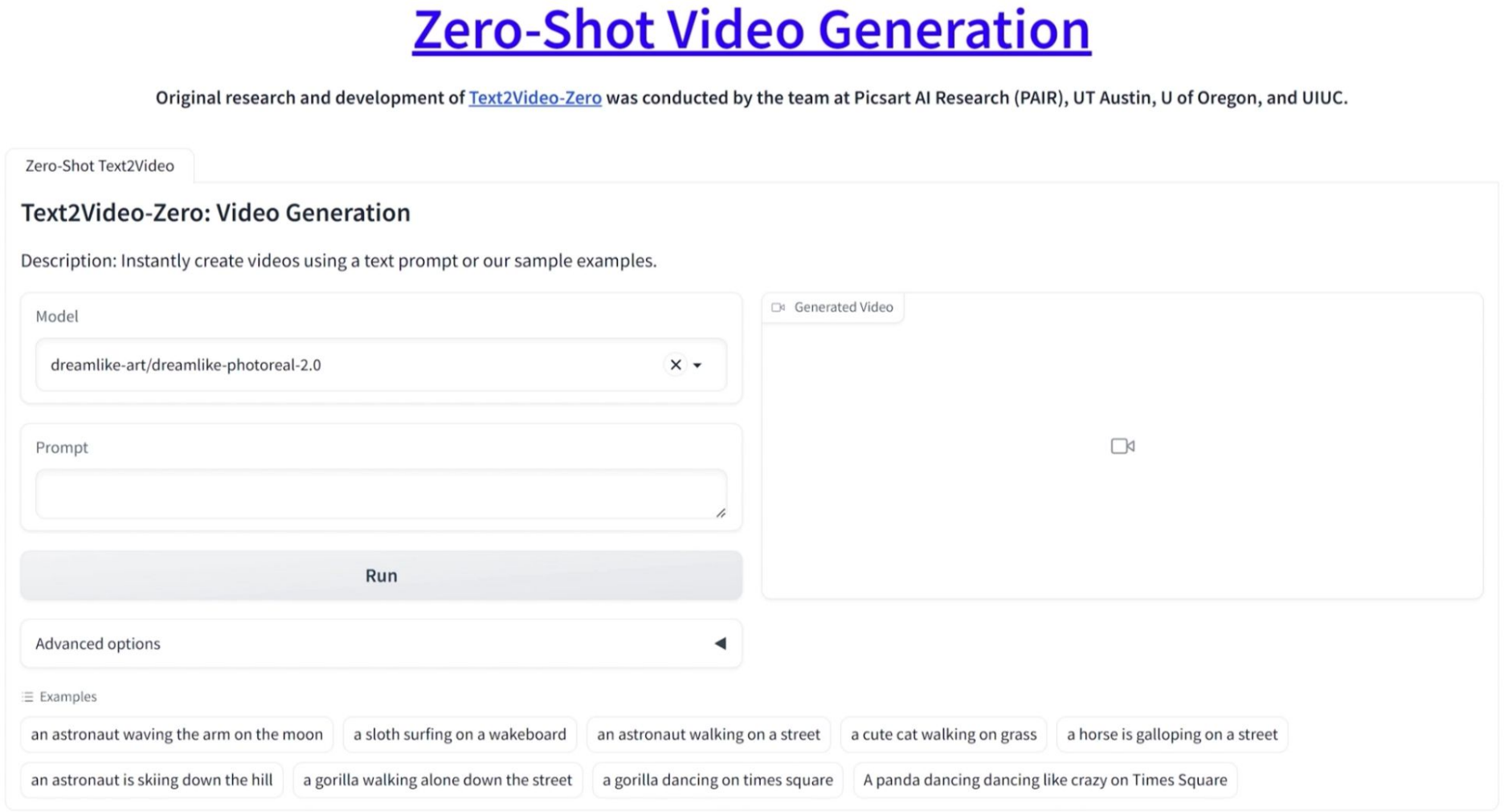
- **PyTorch:** For building and training the neural network. Chosen for its flexibility and ease of use in designing custom models.
- **Gradio:** Utilized to create the web interface. It simplifies the process of building interactive interfaces for our model.
- **OpenCV:** Employed for image and video processing tasks. Offers robust tools for handling visual data.
- **Numpy:** Integral for numerical computations. Aids in efficient handling of large datasets and mathematical operations.
- **Diffusers:** Specifically used for implementing and managing diffusion models, which are central to our text-to-video conversion process.
- **Imageio:** For reading and writing a wide range of image and video formats, crucial in the generation of output videos.

Flowchart

- **Data Collection & Preprocessing:** The initial stage involves sourcing and preparing the dataset.
- **Feature Extraction:** This step focuses on extracting relevant features from both textual and visual data.
- **Model Implementation & Training:** Here, the "Text2Video-Zero" model is adapted and trained.
- **Evaluation:** The model's performance and the quality of generated videos are assessed.
- **User Interface Development:** The project concludes with the creation of a user-friendly interface.



User Interface: Bridging Users and AI



Beyond the Lab: Real-World Applications

- **Content Creation:** Revolutionizing digital storytelling and media production.
- **Educational Tools:** Enhancing learning experiences with visual aids.
- **Industry Implications:** Potential impact across various sectors including marketing, education, and entertainment.

University of Windsor

REFERENCES

References

[1] L. Khachatryan et al., “Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators,” arXiv.org, Mar. 23, 2023, https://arxiv.org/abs/2303.13439 [Accessed: Nov. 19, 2023].

[2] C. Wu et al., “N\"UWA: Visual Synthesis Pre-training for Neural visUal World creAtion,” arXiv:2111.12417 [cs], Nov. 2021, https://arxiv.org/abs/2111.12417 [Accessed: Nov. 19, 2023].

[3] R. Villegas et al., “Phenaki: Variable Length Video Generation From Open Domain Textual Description,” arXiv.org, Oct. 05, 2022, https://arxiv.org/abs/2210.02399 [Accessed: Nov. 19, 2023].

[4] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, “CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers,” May 2022, https://doi.org/10.48550/arxiv.2205.15868 [Accessed: Nov. 19, 2023].

[5] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video Diffusion Models,” arXiv (Cornell University), Apr. 2022, https://doi.org/10.48550/arxiv.2204.03458 [Accessed: Nov. 19, 2023].

[6] J. Ho et al., “Imagen Video: High Definition Video Generation with Diffusion Models,” Oct. 2022, https://doi.org/10.48550/arxiv.2210.02303 [Accessed: Nov. 19, 2023].

References

[7] U. Singer et al., “Make-A-Video: Text-to-Video Generation without Text-Video Data,” Sep. 2022, https://doi.org/10.48550/arxiv.2209.14792 [Accessed: Nov. 19, 2023].

[8] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, “Structure and Content-Guided Video Synthesis with Diffusion Models,” Feb. 2023, https://doi.org/10.48550/arxiv.2302.03011 [Accessed: Nov. 19, 2023].

[9] J. Z. Wu et al., “Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation,” arXiv (Cornell University), Dec. 2022, https://doi.org/10.48550/arxiv.2212.11565 [Accessed: Nov. 19, 2023].

[10] V. Subramanian, Deep Learning with PyTorch. Packt Publishing Ltd, 2018, https://t.ly/Izg92 [Accessed: Nov. 19, 2023].

[11] Vivian Genaro Motti, D. Raggett, Sascha Van Cauwelaert, and J. Vanderdonckt, “Simplifying the development of cross-platform web user interfaces by collaborative model-based design,” Sep. 2013, https://doi.org/10.1145/2507065.2507067 [Accessed: Nov. 19, 2023].

Thank You