POST 9 - GEN AI

UNDERSTANDING GENERATIVE AI VIDEO GENERATION: A STEP-BY-STEP GUIDE MADE EASY

Fundamentals - Part 4

GENERATIVE AI FOR ALL

DINESH LAL
(DIRECTOR, DATA SCIENCE)

DINESHLAL

- In this document we will go through the step by step process of video generation by Generative AI
- We will take an example of A woman walking on Tokyo Street.
- This is a video generated by passing the prompt to SORA, AI model from OpenAI. Video can be seen here
- The guide is in two sections, one looking it as non technical lens, the other section focuses with technical perspective

THE OUTPUT

CLICK ON THE LINK TO SEE THE OUTPUT

HTTPS://CDN.OPENAI.COM/SORA/VIDEOS/TOKYO-WALK.MP4

Input Text:

- Imagine you’re writing a cool description for a movie scene. You’re like the director, telling the AI what you want to see.
- Here’s your text: “A stylish woman walks down a Tokyo street filled with warm glowing neon signs. She’s rocking a black leather jacket, a long red dress, and black boots. Her sunglasses and red lipstick add to her chic look. She walks confidently, like she owns the city. The street is wet from rain, and it reflects the colorful lights. Lots of people are out and about.”

Text Analysis:

- The AI reads your text and figures out the important stuff:
  - Character: Our stylish woman.
  - Setting: The busy Tokyo street with neon signs.
  - Clothes: Her cool outfit—black leather jacket, red dress, and boots.
  - Actions: She struts confidently.
  - Vibe: The street is wet and shiny, with lots of people.

Scene Generation:

- Think of this like planning a comic book. The AI imagines different scenes:
  - **Scene 1:** The woman steps onto the neon-lit street.
  - **Scene 2:** We zoom in on her jacket and red dress.
  - **Scene 3:** Neon signs flicker in the background.
  - **Scene 4:** She adjusts her sunglasses and keeps walking.
  - **Scene 5:** People pass by, making the street feel alive.

Visual Representation:

- The AI draws the characters and the street. It's like creating paper dolls but cooler. The woman gets her leather jacket, red dress, and shades. The street gets neon signs, puddles, and reflections.

Animation:

- Now, the AI makes the pictures move! The woman walks smoothly, her hair swaying. Neon signs blink on and off. People shuffle past, like in a bustling city.

Audio Generation:

- Imagine adding sound effects. Raindrops pitter-patter, footsteps tap, and there's a jazzy tune playing. It's like turning up the volume on your imagination.

Video Rendering:

- The AI puts everything together. It's like making a flipbook. Each page shows a tiny bit of movement. When you flip the pages fast, it becomes a video!

Output:

- **Voilà! You' ve got a mini movie! Share it with friends, and they' ll be like, "Whoa, you made this? Awesome!"**

TECHNICAL EXPLANATION

Input Text:

- **We start with a text prompt, which serves as our creative seed. Think of it as a sequence of words that describe what we want to visualize. In our example, your detailed description of the stylish woman in Tokyo serves as this prompt.**
- **Prompt - "A stylish woman walks down a Tokyo street filled with warm glowing neon signs. She's rocking a black leather jacket, a long red dress, and black boots. Her sunglasses and red lipstick add to her chic look. She walks confidently, like she owns the city. The street is wet from rain, and it reflects the colorful lights. Lots of people are out and about."**

Text Analysis:

- The AI system employs natural language processing (NLP) techniques to dissect the input text:
  - Tokenization: Each word becomes a token, forming a sequence.
  - Named Entity Recognition (NER): It identifies entities like “woman,” “Tokyo,” and “neon signs.”
  - Dependency Parsing: Understanding how words relate to each other (e.g., “walking down” implies an action).

Scene Generation:

- **Here's where the magic begins. We create a semantic representation of the text:**
  - **Graphs or Trees: We organize the entities (woman, street, neon signs) and their relationships (walking, wearing, reflecting).**
  - **Attention Mechanisms: These help the AI focus on relevant parts of the text, attending to crucial details.**

Visual Representation:

- The AI translates the semantic representation into visual elements:
  - Vector Representations: Each entity (woman, neon sign, rain) becomes a vector in a high-dimensional space.
  - Style Embeddings: Capturing the vibe—chic, urban, rainy—encoded as additional vectors.

Animation:

- Now we add motion dynamics:
  - Temporal Models: LSTM or Transformer-based networks predict how the woman walks, how neon signs flicker, and how people move over time.
  - Keyframes: These are like snapshots of the video at different time points. We interpolate between keyframes to create smooth motion.

Audio Generation:

- **Videos need sound! The AI generates:**
  - **Sound Effects:** **Raindrops pitter-patter, footsteps tap, and distant city hum.**
  - **Voice Synthesis:** **If the woman talks, we can create her voice using text-to-speech models.**

Video Rendering:

- This is where we put it all together:
  - Frame Synthesis: We render each frame by combining visual elements, applying motion transformations, and adding audio.
  - GPU Acceleration: GPUs crunch the numbers fast, parallelizing computations for efficiency.

Output:

- Voilà! We have a video tensor—a 3D array of pixels and audio samples. Each pixel encodes color, brightness, and position. You can save it, stream it, or analyze it further using tools like FFmpeg or OpenCV.

Thank You

SPECIAL CREDITS TO OPENAI, CHATGPT,
DALL-E FOR THE CONTENT SUPPORT

DINESHLAL