Video Generierung mit Basismodellen der KI

**Was ist T2V?**

Text-To-Video(t2v) ist eine Technologie, die aus einem Text ein Video erzeugt.

Was bringt T2V?

- Videoproduktion ohne Kamera und Schauspieler
- Reduziert Produktionszeit & Kosten (Film, Werbung, Bildung, Social Media)

Wie funktioniert T2V?

1. Text Decoden: KI liest Text und erkennt was vorkommt
2. Noise generieren: KI startet mit Zufallsverrauschtem Bild
3. Schrittweise denoising Schritten: Diffussionsmodell entfernt Noise in vielen kleinen
4. Postprocessing: Farbkorrektur, Upscaling

Unterschiede von T2V zu T2I:

- Zusätzliche Zeitliche Dimension
- Konsistenz zwischen Frames
- Realistische Bewegungen
- Größerer Rechenaufwand

T2V Architekturen:

- VAE (Variational Autoencoder)
- GAN (Generative Adversarial Network)
- DM (Diffusionmodell)
- DiT (Diffusion Transformer)

**DiT (Diffusion Transformer)**

Grundidee:

- Diffusion kümmert sich um die Bildgenerierung
- Transformer versteht Abhängigkeiten zwischen Text und Zeit (Kohärenz & Physik
- Kombination ermöglicht längere und realistischere Videos

Wie funktioniert DiT?

- Es wird ein Text eingegeben
- Der Text wird durch den Transformer verstanden
- Im Latenten Raum wird nach passenden und vorhandenen Informationen verglichen
- Das Video wird denoised
- Das fertige Video wird erstellt

Eigenschaften von DiT:

- Spatiotemporal Understanding (Raum + Zeit)
- Cross-Attention Mechanismus (Text & Video aufeinander abgeglichen)
- Realistische Bewegungen
- Skalierbarkeit (Mehr Daten + Rechenleistung)

**Tools und Beispiele von T2V:**

- Sora
- Pika
- Runaway

**Herausforderungen von T2V:**

- Verständnis komplexer Prompts
- Konsistenz zwischen Frames
- Physikalischer Realismus
- Rechenaufwand & Energieverbrauch
- Daten & Urheberrecht
- Missbrauch & Deepfakes