Introduction to Text-to-Video Generation

CS 294-43 Fall 2024

Presenter: Long (Tony) Lian

11/4/2024

The Field of Text-to-Video Generation is Moving Fast

Video from VDM (2022)

The Field of Text-to-Video Generation is Moving Fast

Video from Sora (2024)

Overview of Text-to-Video Generation Models

- Text-to-video models generate video sequences based on natural language prompts.
- Two prominent families of models: **Diffusion Models** and **Autoregressive Models**.
- Key challenges: **temporal consistency**, **computational complexity**, and **semantic alignment**.

Challenge 1: Temporal Consistency



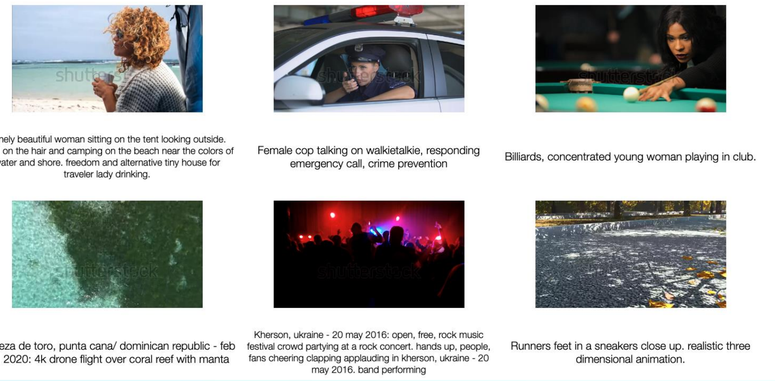

Generating videos is not just generating images.

Challenge 2: High Computational Complexity

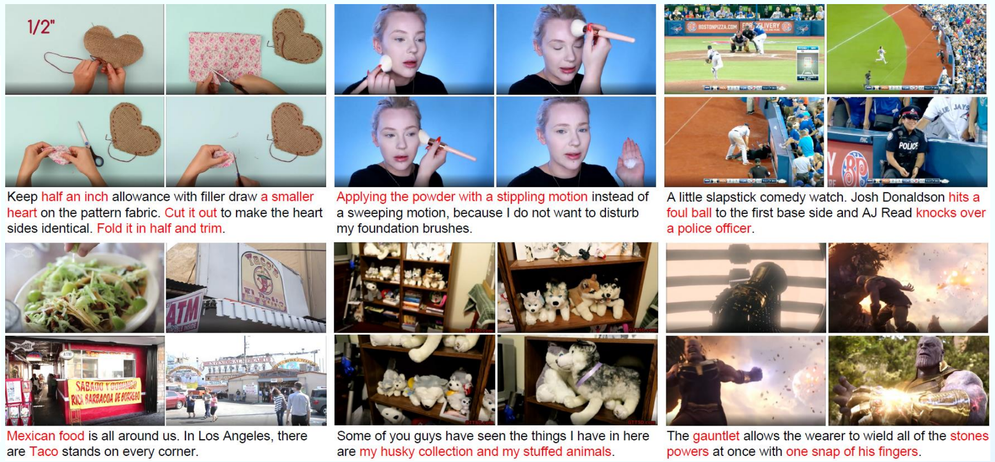
For a 12 fps video, 24 images are just 2 seconds.

60 seconds => **720 frames**

Challenge 3: Lack of High-quality Datasets



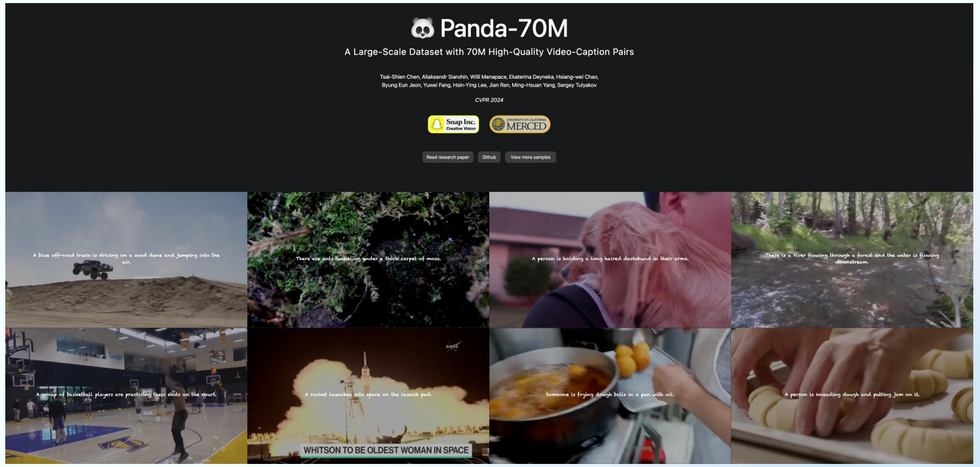
WebVid-10M

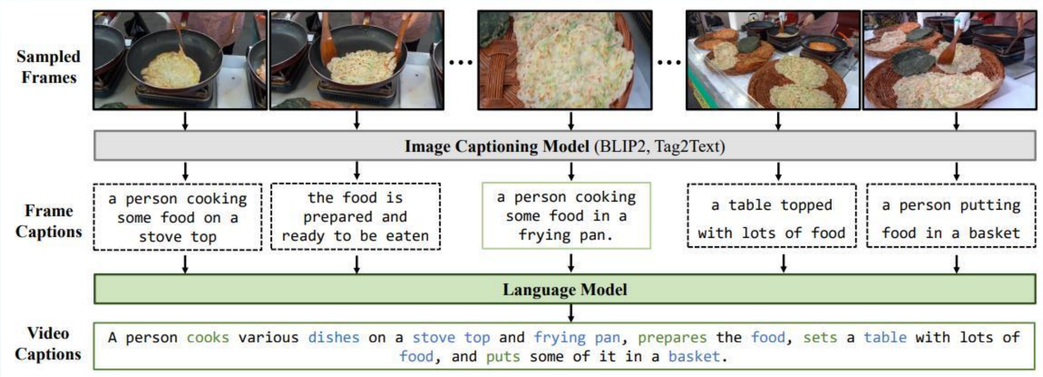


HD-VILA-100M

UCF-101

Recent datasets are of higher quality





InternVid-10M-FLT

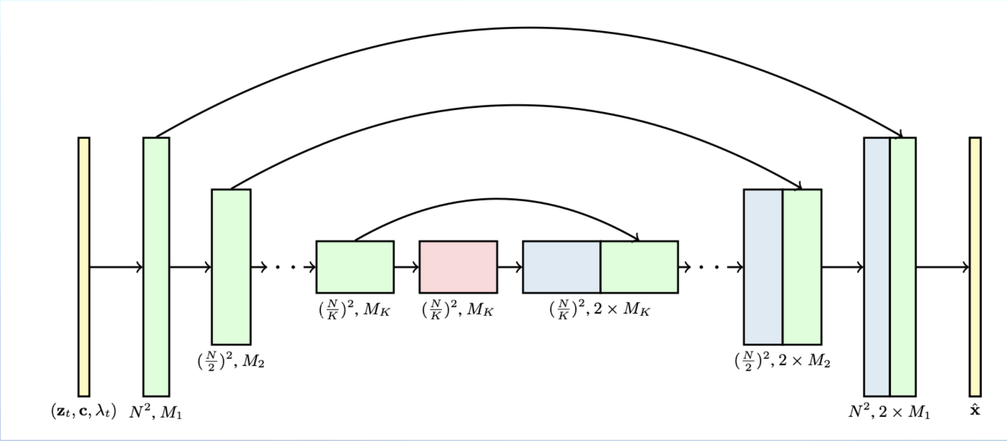
Diffusion Models for Video Generation

- Inspired by denoising diffusion probabilistic models.
- Gradually add noise to the video frames and learn to reverse the process.
- Notably, many models for video generation are re-designed image generation models.
- Example models: **Video Diffusion Models (VDM)**, **Imagen Video**, and **Align Your Latents**

Video Diffusion Models (VDM)

An attempt of spatial -> spatiotemporal generation

- A temporal extension of image-based diffusion models.
- Uses 3D CNNs to model temporal dynamics.
  - Spatial convs are used to process the space
  - Temporal attention layers are used to sync across the time



No 3D conv => high efficiency

Temporal attention => temporal consistency

Some results from VDM

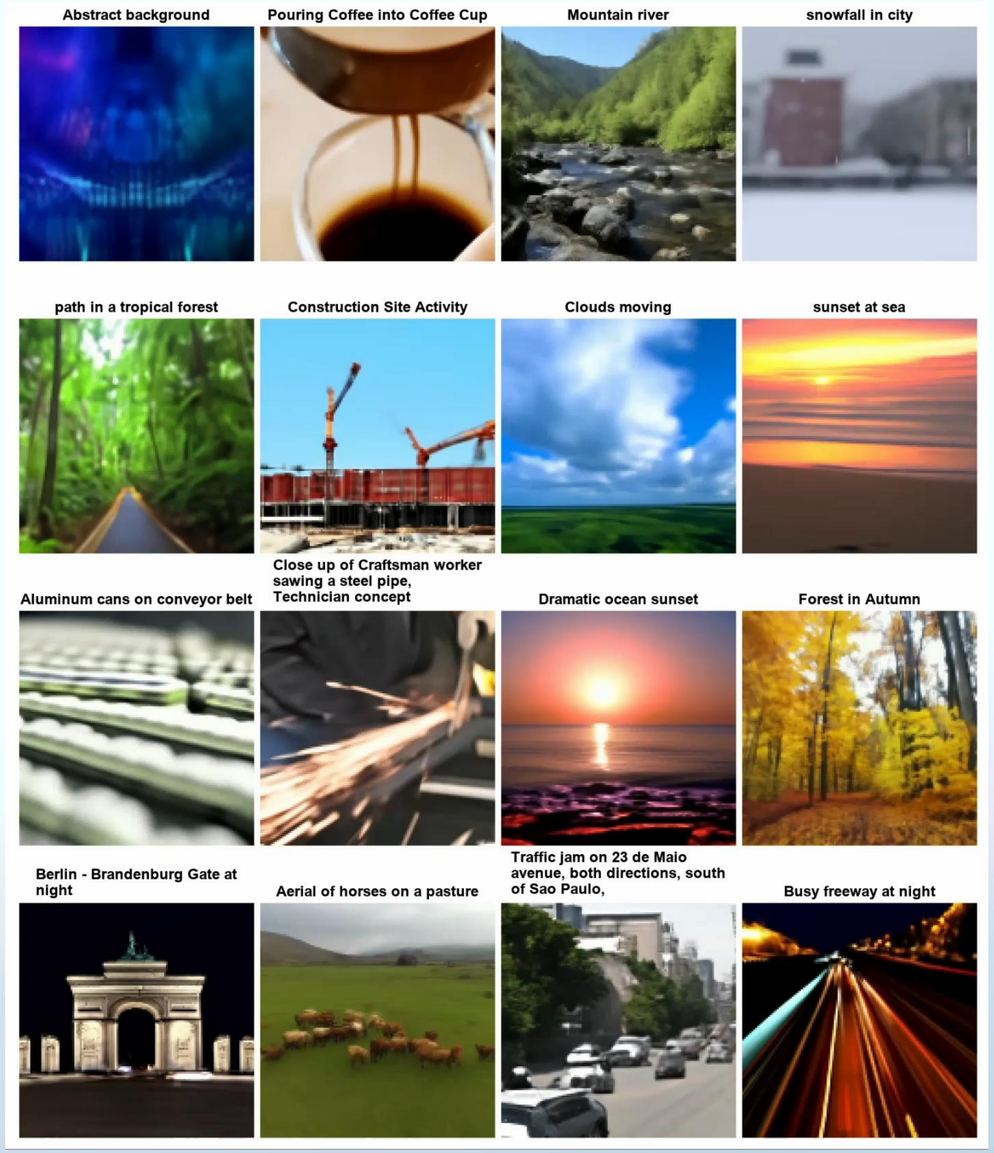
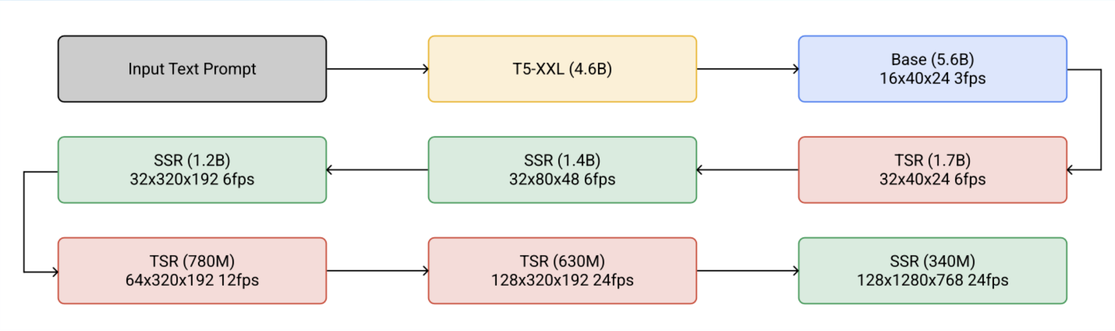
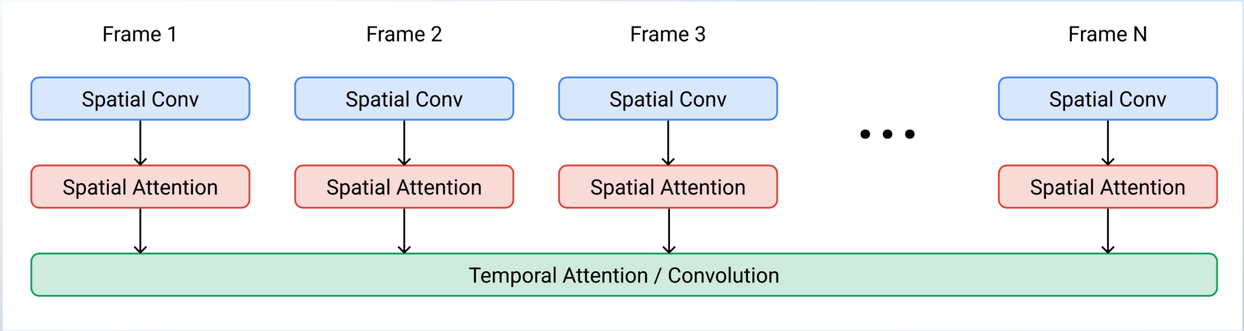


Imagen Video – A Diffusion Model with Hierarchical Generation

Imagen -> Imagen Video: spatial upsampling -> spatiotemporal upsampling

- Hierarchical approach: generates lower-resolution frames first, then refines them.
- Improves coherence across frames with a cascade of diffusion models.
- Specialized for high-resolution video synthesis.



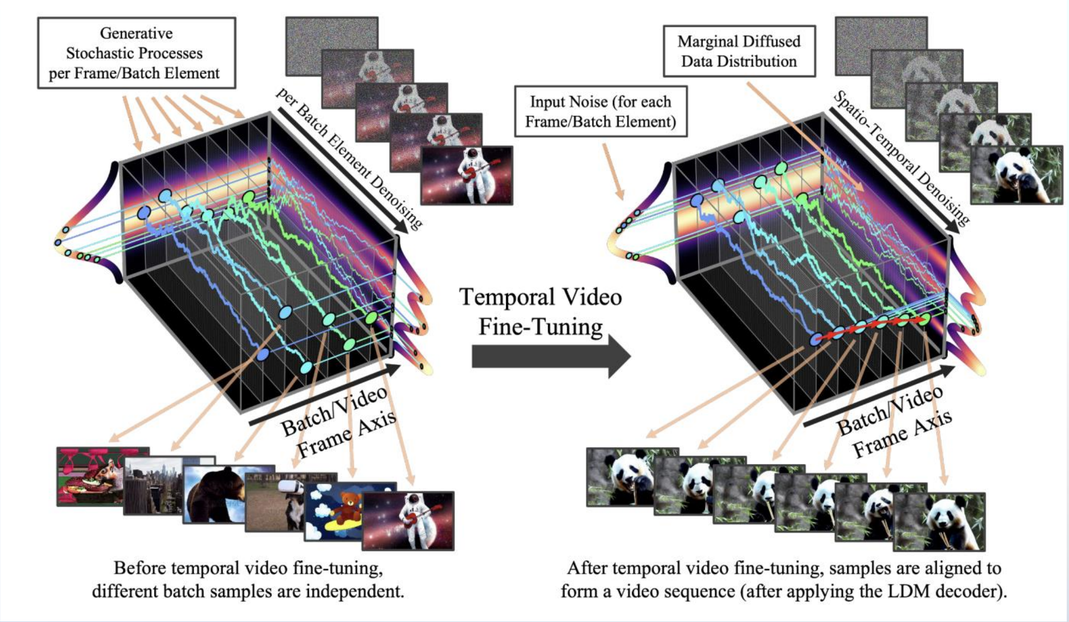


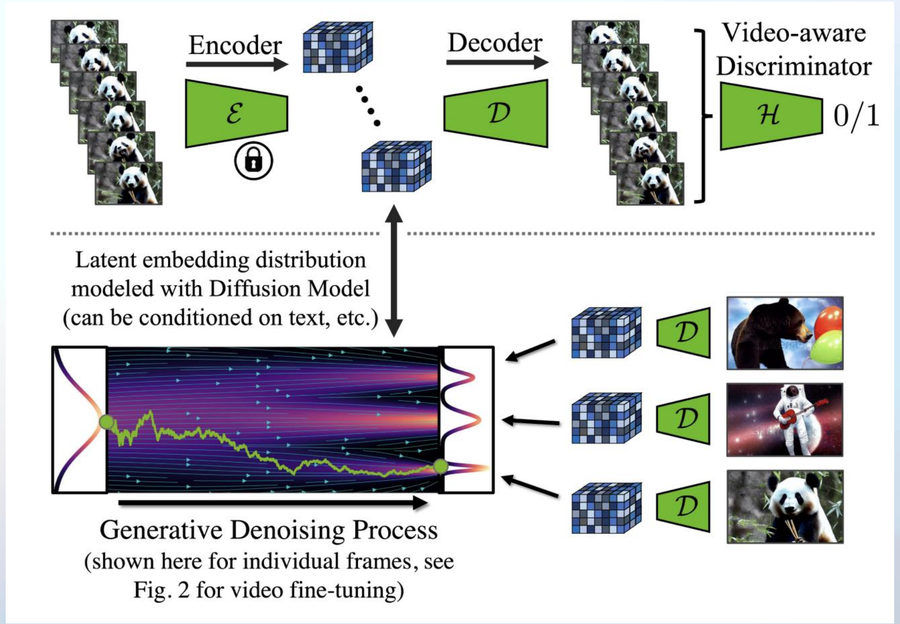
Some results from Imagen Video

Align your Latents – Improving Video Quality with Latent Diffusion

Stable Diffusion -> Stable Diffusion for Videos

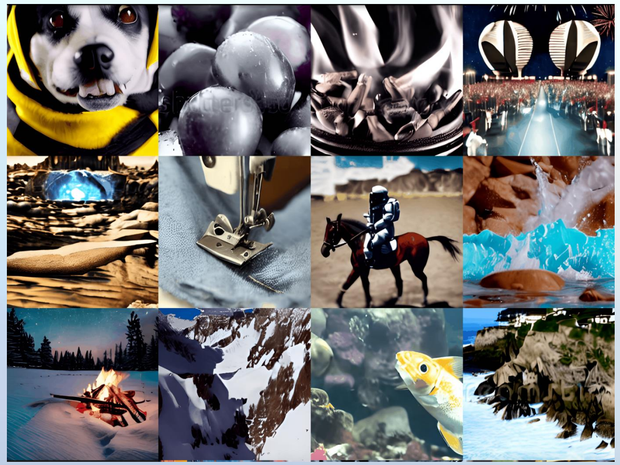
- Ensures temporal coherence by aligning latent representations across frames.
- Alignment is achieved through a video-aware discriminator.





There are so many works in this field

- Phenaki – Enabling Long Video Generation
- Emu Video – Factorizing Text-to-Video Generation into image generation and image-conditioned video generation
- VideoCrafter2 – Addressing Data Limitations in Video Diffusion Models
- MagicVideo – Efficient Video Generation With Latent Diffusion Models



Shutterstock watermark from the examples generated by MagicVideo

Emu Video and VideoCrafter2

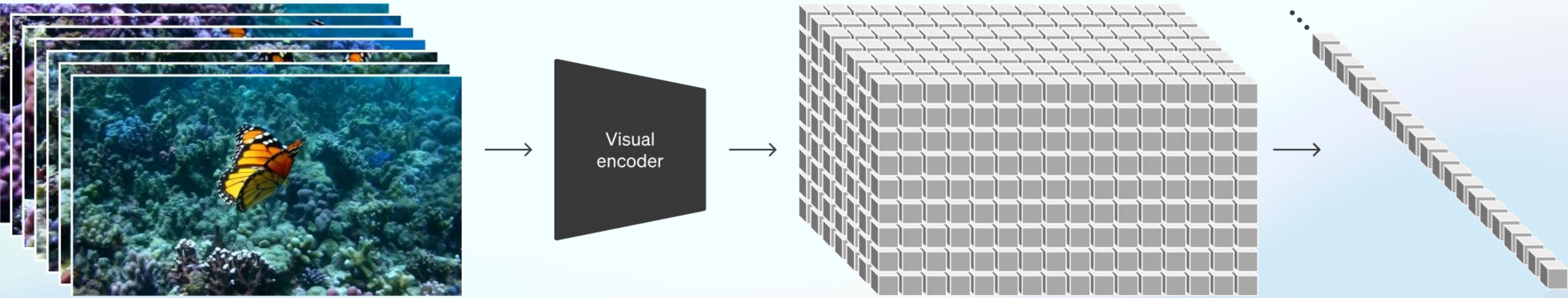
- Ren’s presentation

Sora – OpenAI’s Text-to-Video Generation Model

Here is what we know.

- Capable of generating high-fidelity videos up to one minute in length.
- Trained on diverse datasets with varying durations, resolutions, and aspect ratios.
- Utilizes a transformer architecture operating on spacetime patches of video and image latent codes.
- Aims to serve as a general-purpose simulator of the physical world.

The framework: diffusion transformer



An example

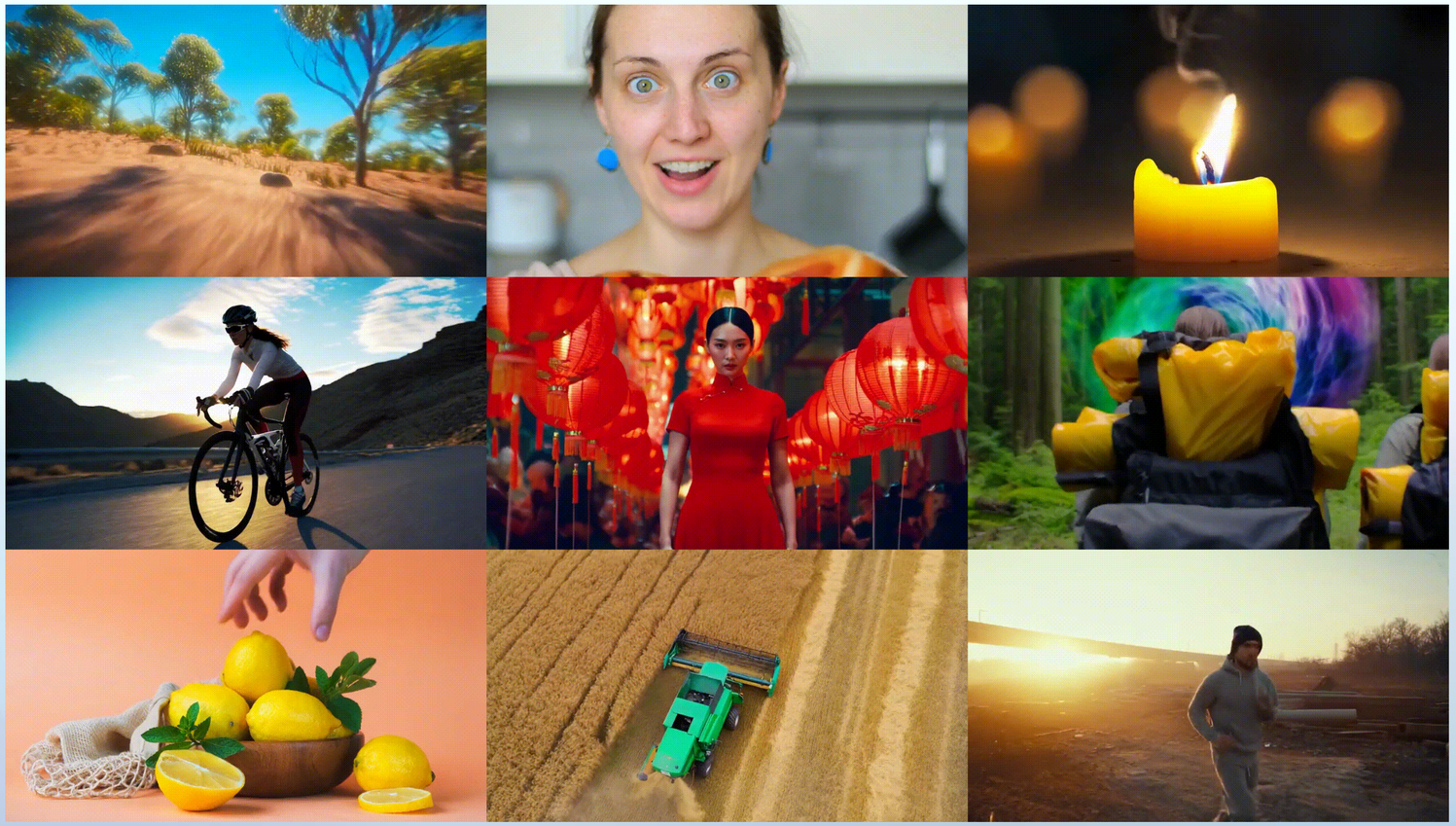
MovieGen – Meta AI’s Media Foundation Framework

Alberto’s Presentation

Recent Advancements: Mochi 1

An open-weights text-to-video generator

- It uses a 10-billion-parameter diffusion model built on the novel Asymmetric Diffusion Transformer (AsymmDiT) architecture



Autoregressive Models for Video Generation

- Diffusion models generate all the frames together.
- Autoregressive models predict video frames one by one, conditioned on previous frames.
- Common issue: propagation of errors (i.e., compounding errors).

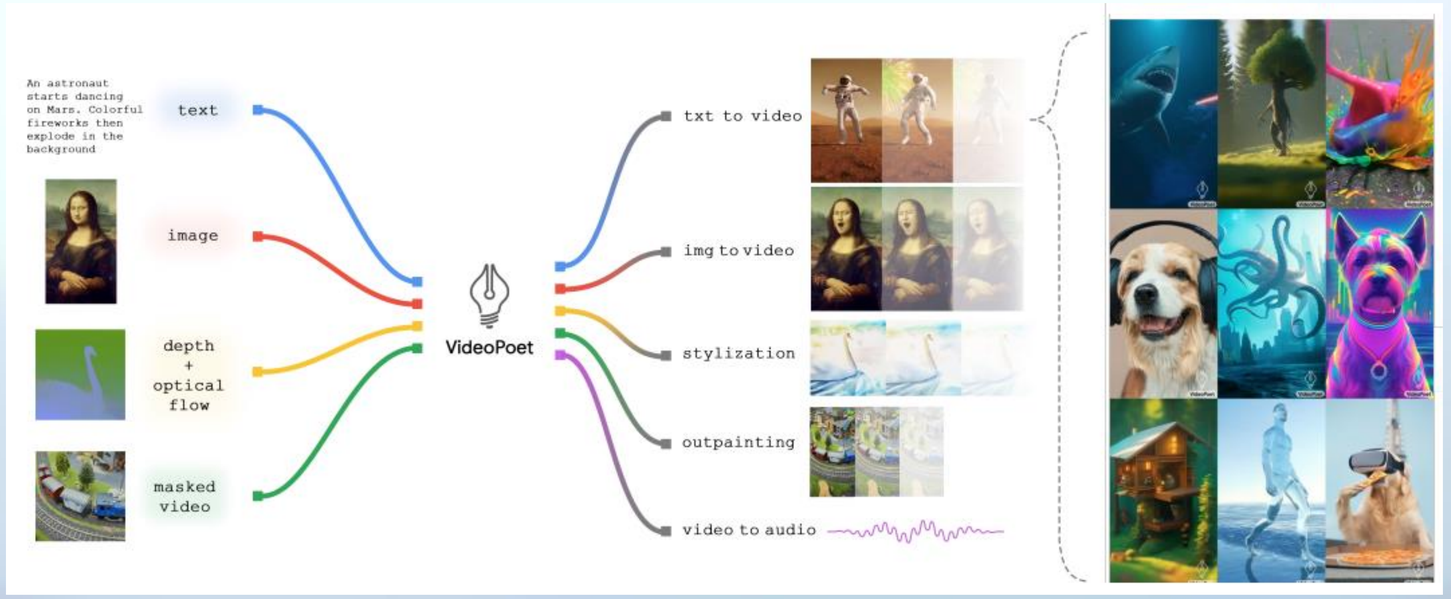
VideoGPT – Frame-wise Autoregressive Model with Transformers

- Uses Vector Quantized-Variational Autoencoders (VQ-VAE) to encode video frames.
- Employs a transformer-based decoder to predict frame sequences.

VideoPoet – Large Language Model for Zero-Shot Video Generation

Zeeshan’s Presentation

- Introduces a large language model (LLM) for zero-shot video generation.
- Learns temporal dynamics without requiring large labeled datasets.
- Uses text embeddings for video synthesis through transformer decoders.



Diffusion vs. Autoregressive Models – Key Differences

- Diffusion Models: Better for high-resolution and long-form video generation, but computationally intensive.
- Autoregressive Models: Faster for short video generation but prone to error propagation.
- Temporal Consistency: Diffusion models handle it better through sequential denoising.

Benchmarking text-to-video generation models

- Importance of Evaluation: Assessing text-to-video (T2V) models is crucial for understanding their performance and guiding future improvements.
- Challenges: Evaluating T2V models involves multiple dimensions, including visual quality, temporal dynamics, and alignment with textual prompts.
- Two important metrics for text-to-video generation are FVD and CLIPScore.
  - Both metrics have their own issues.

Fréchet Video Distance (FVD)

- FVD is similar to FID, except that it’s calculated on video features.
- FVD is the Fréchet Distance between the feature distributions of real and generated videos.

Issues with FVD

- **Non-Gaussian Feature Space:** Assumes a Gaussian distribution in the Inflated 3D Convnet (I3D) feature space, which may not hold true, leading to inaccurate evaluations.
- **Temporal Insensitivity:** I3D features may not effectively capture temporal distortions, undermining the assessment of motion quality.
- **Sample Size Requirements:** Reliable estimation with FVD often necessitates impractically large sample sizes.

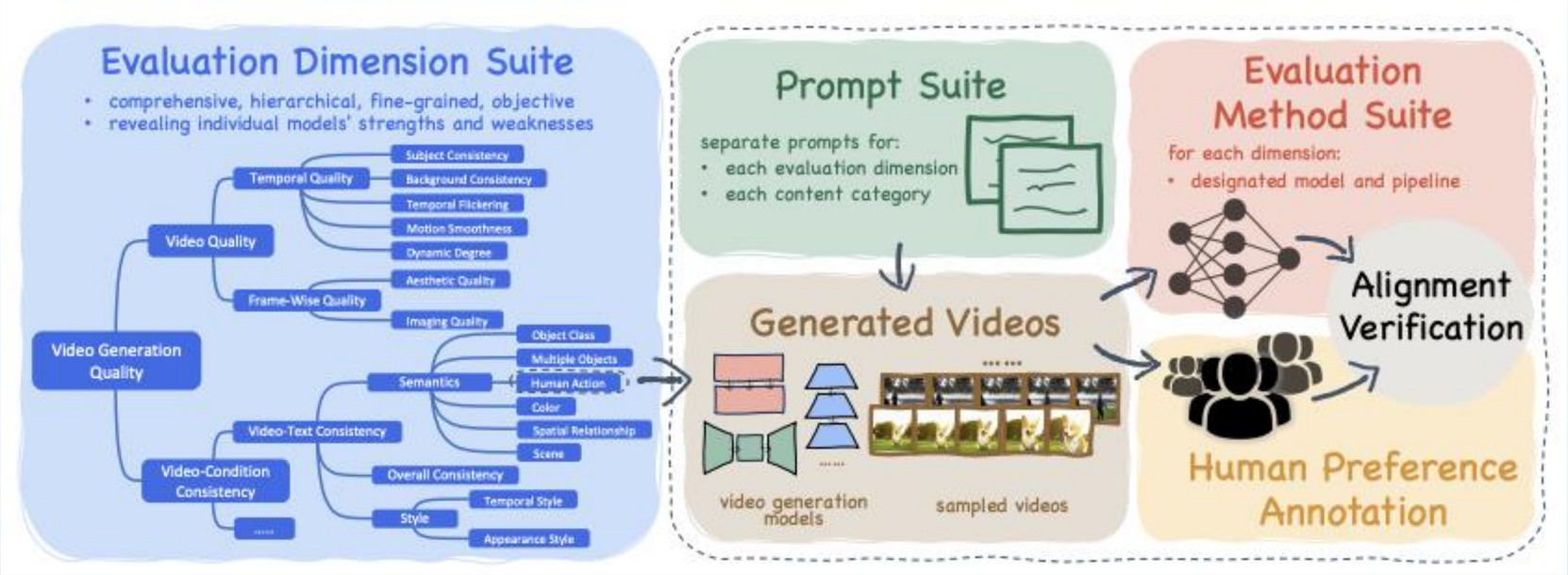
CLIPScore and its issues

- CLIPScore computs the cosine similarity between the image and caption embeddings produced by CLIP.
- It has its own issues:
  - It often struggles to identify fine-grained details in generation.
  - It is insensitive to certain linguistic aspects, such as negation.
- The fact that existing metrics are not good enough for evaluation is the reason for text-to-video benchmarks.

VBench – Comprehensive Benchmark Suite for Video Generative Models

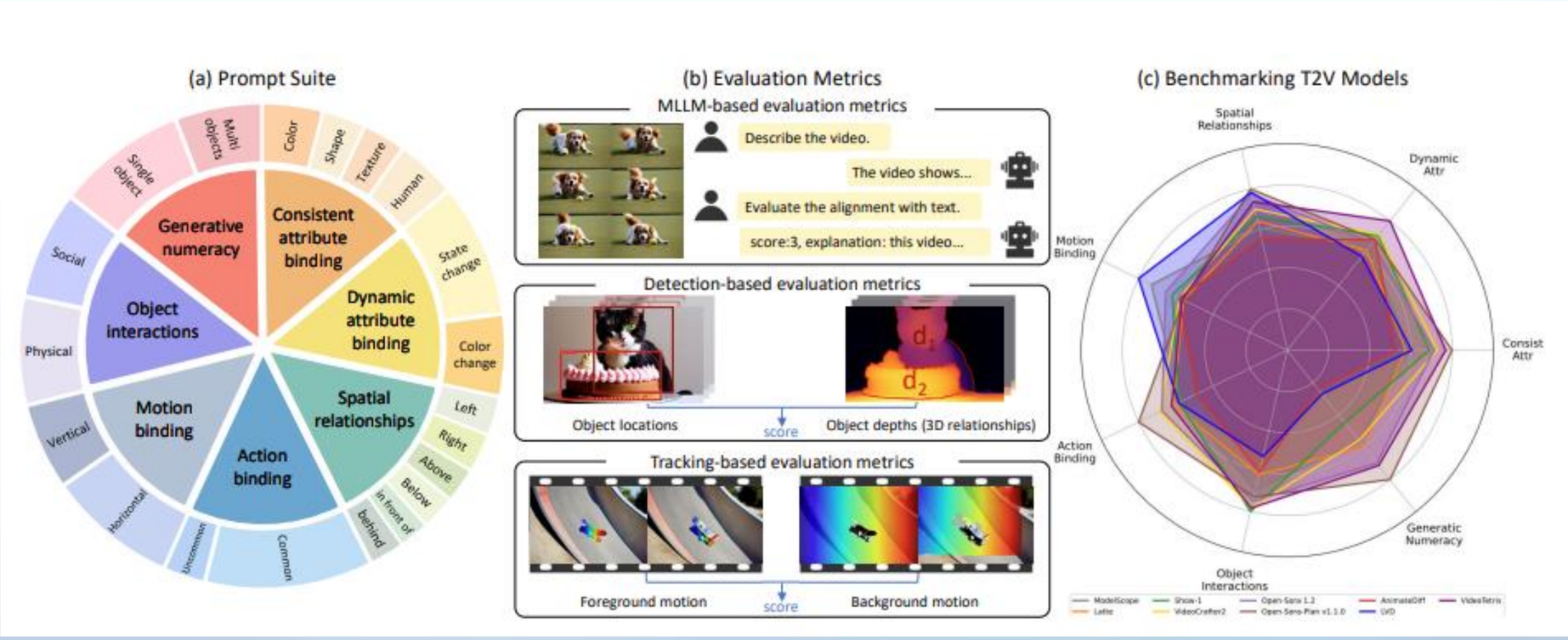
Kazusato’s Presentation

- A benchmarking suite for evaluating video generative models.
- Covers quality, temporal consistency, diversity, and realism metrics.
- Includes both objective measures and human evaluations



T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation

- A work from the same team that proposed T2I-CompBench
- Benchmarking several aspects by leveraging models such as LLaVA



Challenges and Future Research in Text-to-Video

- **Computational Bottlenecks:** Both diffusion and autoregressive models require immense computational power.
- **Scalability:** Need for better models that can handle longer videos with fewer resources.
- **Semantic Drift:** Ensuring consistent alignment with input prompts over time.
- **Model Evaluation:** Lack of standardized metrics for evaluating video quality.

Conclusion – Where Text-to-Video Generation is Heading

- LLM integration will drive more innovative zero-shot models like VideoPoet.
- Factorization strategies (e.g., Emu Video) help simplify generation tasks.
- Synthetic data augmentation (VideoCrafter2) addresses data challenges.
- Large-scale training is needed for content coherence and visual quality, as shown by the high-quality results from MovieGen and Sora.
- Standardized evaluation metrics (VBench) are still needed for fair evaluation.

Discussions

- What should PhD students do in response to the aggressive scaling in text-to-video model training?
- Is scaling all you need for high-quality video generation?