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

Pouring Coffee into Coffee Cup



Video from VDM (2022)

Traffic jam on 23 de Maio avenue, both directions, south of Sao Paulo,



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



Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking.



Cabeza de toro, punta cana/ dominican republic - feb festival crowd partying at a rock concert. hands up, people, 20, 2020: 4k drone flight over coral reef with manta fans cheering clapping applauding in kherson, ukraine - 20



Female cop talking on walkietalkie, responding emergency call, crime prevention



Kherson, ukraine - 20 may 2016: open, free, rock music may 2016. band performing



Billiards, concentrated young woman playing in club.



Runners feet in a sneakers close up. realistic three dimensional animation.



Keep half an inch allowance with filler draw a smaller heart on the pattern fabric. Cut it out to make the heart a sweeping motion, because I do not want to disturb sides identical. Fold it in half and trim.



are Taco stands on every corner.



WebVid-10M



my foundation brushes.



are my husky collection and my stuffed animals.



Applying the powder with a stippling motion instead of A little slapstick comedy watch. Josh Donaldson hits a foul ball to the first base side and AJ Read knocks over a police officer.



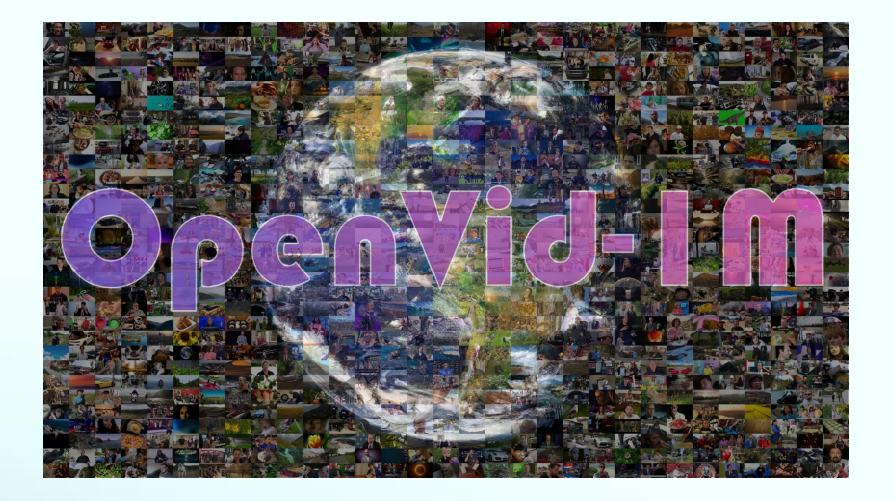
Mexican food is all around us. In Los Angeles, there Some of you guys have seen the things I have in here The gauntlet allows the wearer to wield all of the stones powers at once with one snap of his fingers.

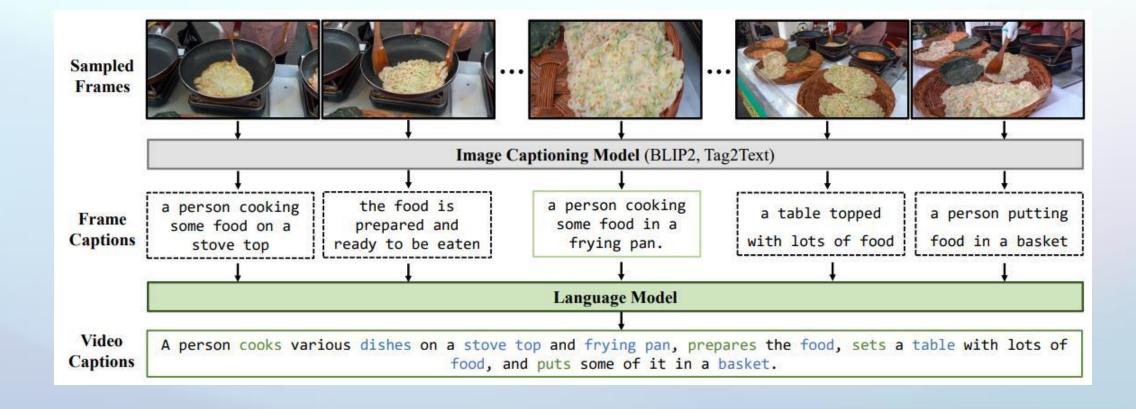


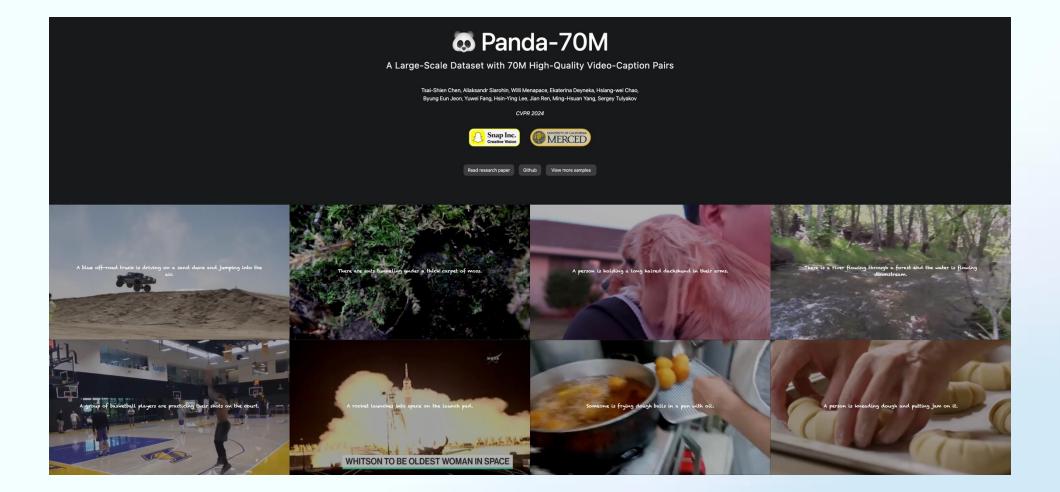
HD-VILA-100M

UCF-101

Recent datasets are of higher quality







InternVid-10M-FLT

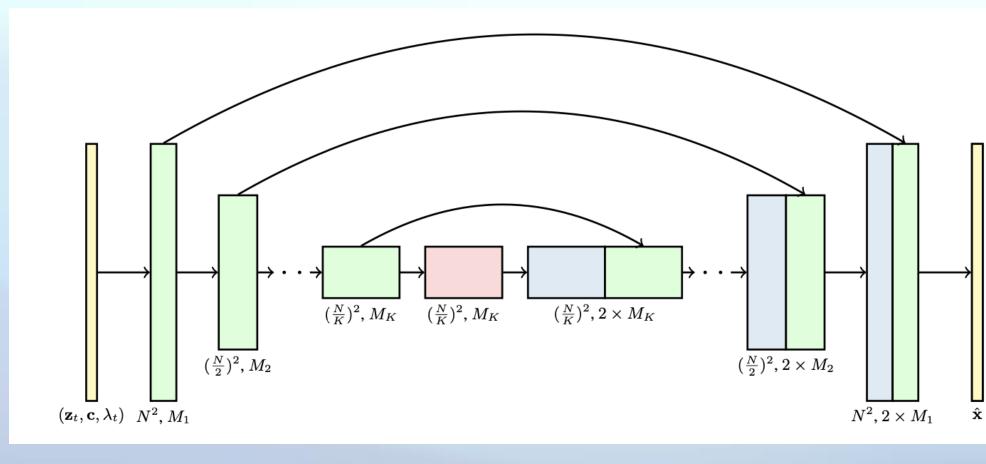
Diffusion Models for Video Generation

- 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

Abstract background

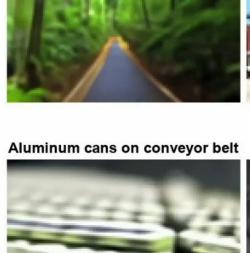




path in a tropical forest



Close up of Craftsman worker







Berlin - Brandenburg Gate at night











Pouring Coffee into Coffee Cup

Mountain river

snowfall in city



Construction Site Activity

Clouds moving





Dramatic ocean sunset

Traffic jam on 23 de Maio avenue, both directions, south of Sao Paulo,







Busy freeway at night

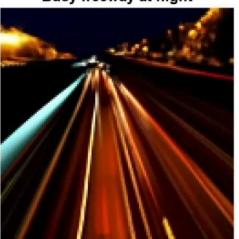
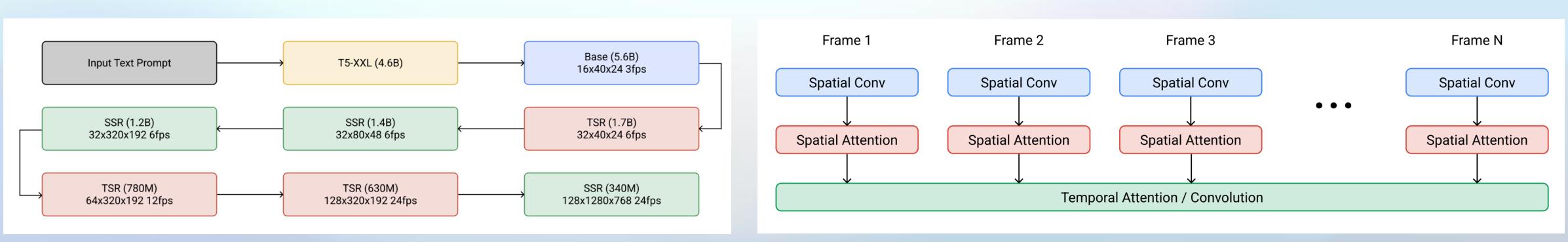
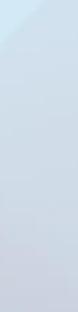


Imagen Video – A Diffusion Model with Hierarchical Generation

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.





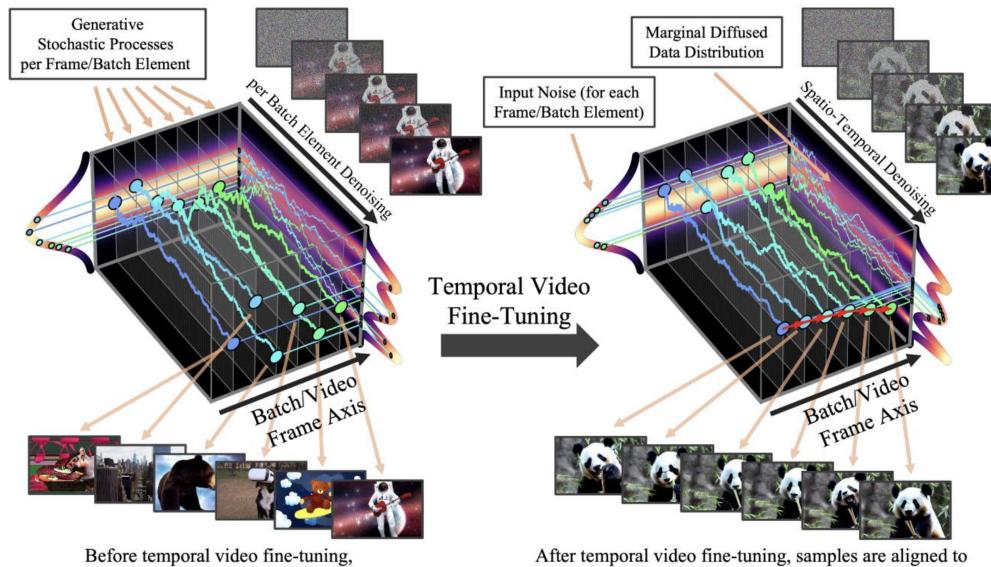
Some results from Imagen Video



Align your Latents – Improving Video Quality with Latent Diffusion

Stable Diffusion -> Stable Diffusion for Videos

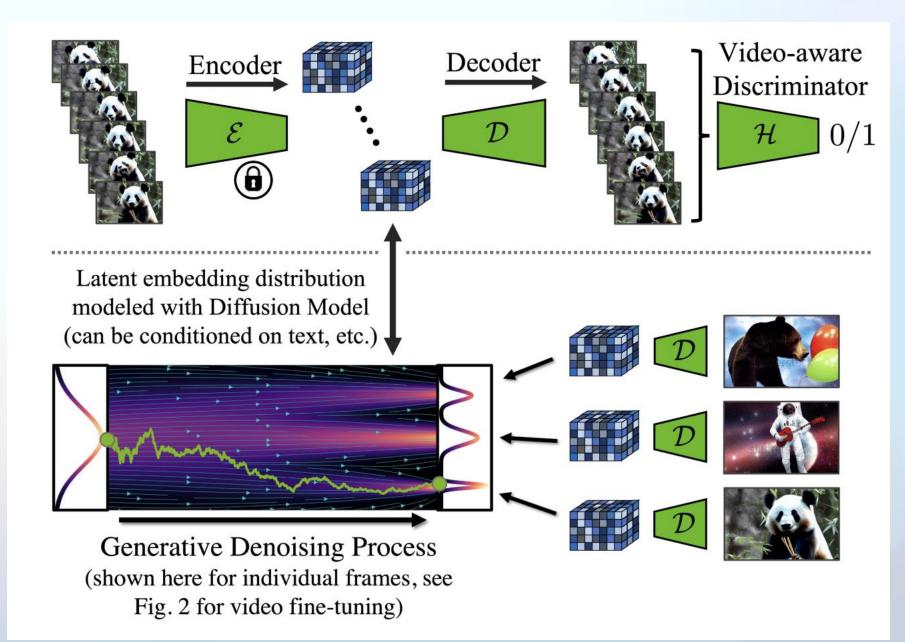
- Ensures temporal coherence by aligning latent representations across frames.
- Alignment is achieved through a video-aware discriminator.



different batch samples are independent.

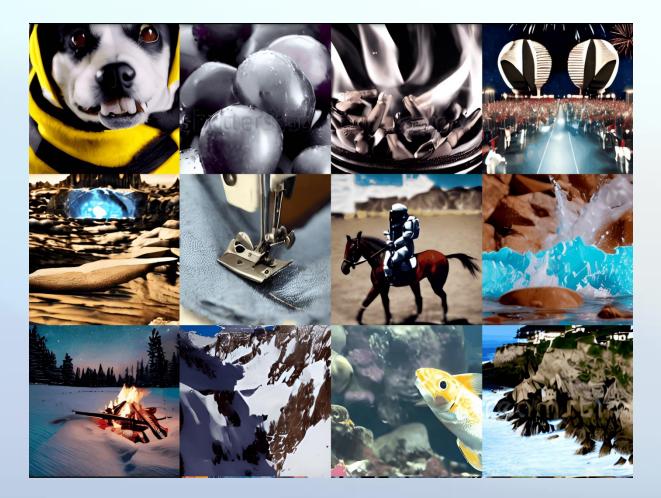
After temporal video fine-tuning, samples are aligned to form a video sequence (after applying the LDM decoder).





There are so many works in this field

- Phenaki Enabling Long Video Generation
- generation
- VideoCrafter2 Addressing Data Limitations in Video Diffusion Models
- MagicVideo Efficient Video Generation With Latent Diffusion Models



• Emu Video – Factorizing Text-to-Video Generation into image generation and image-conditioned video



Shutterstock watermark from the examples generated by MagicVideo

Emu Video and VideoCrafter2

Ren's presentation

Sora – OpenAl's Text-to-Video Generation Model

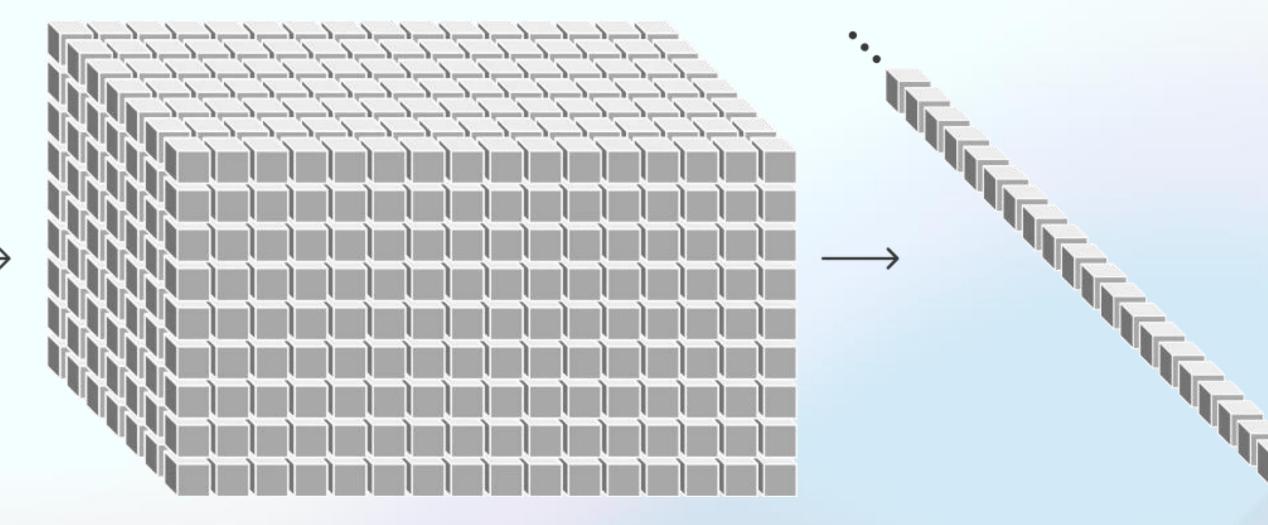
Here is what we know.

- 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

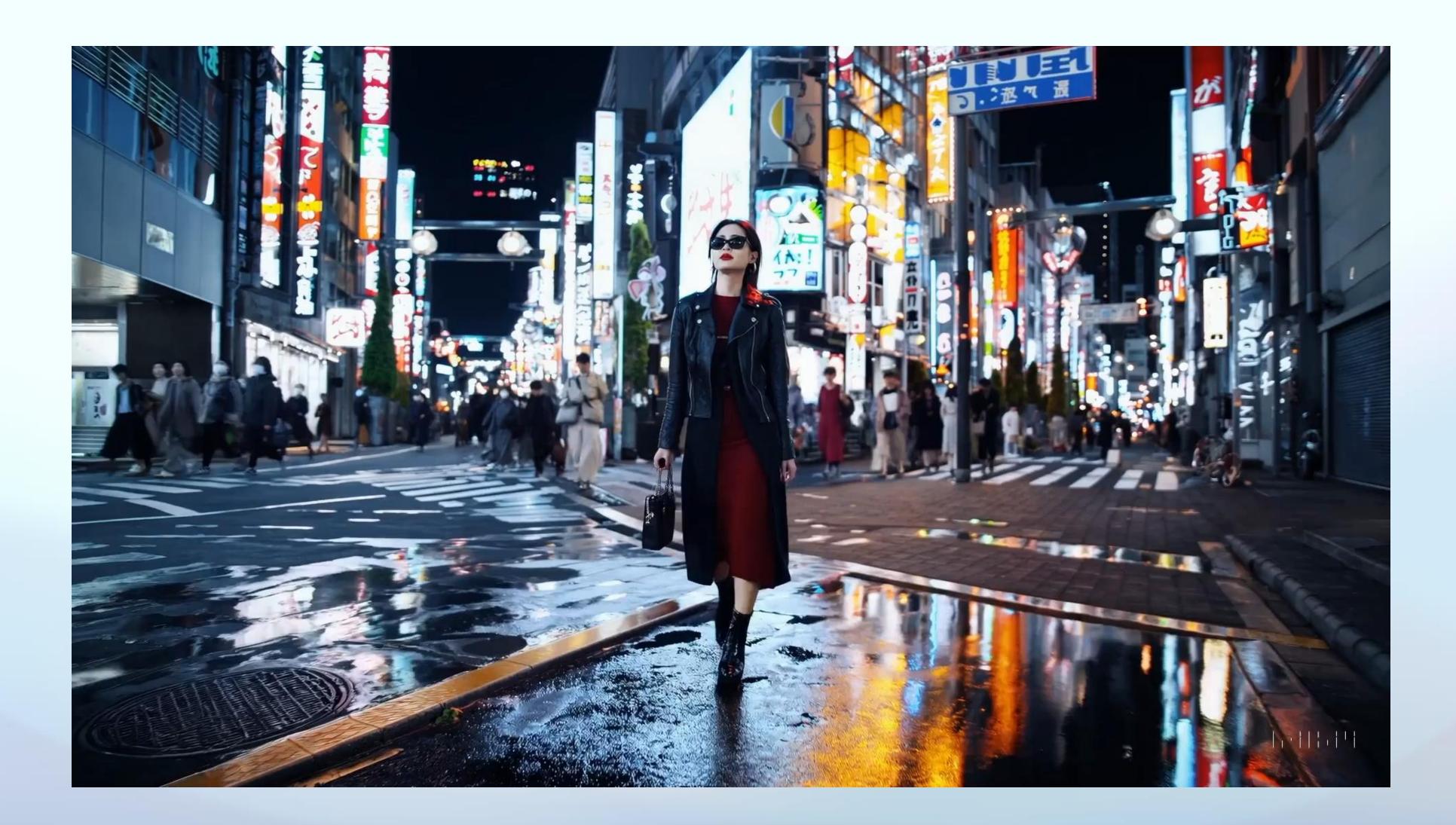


Visual encoder



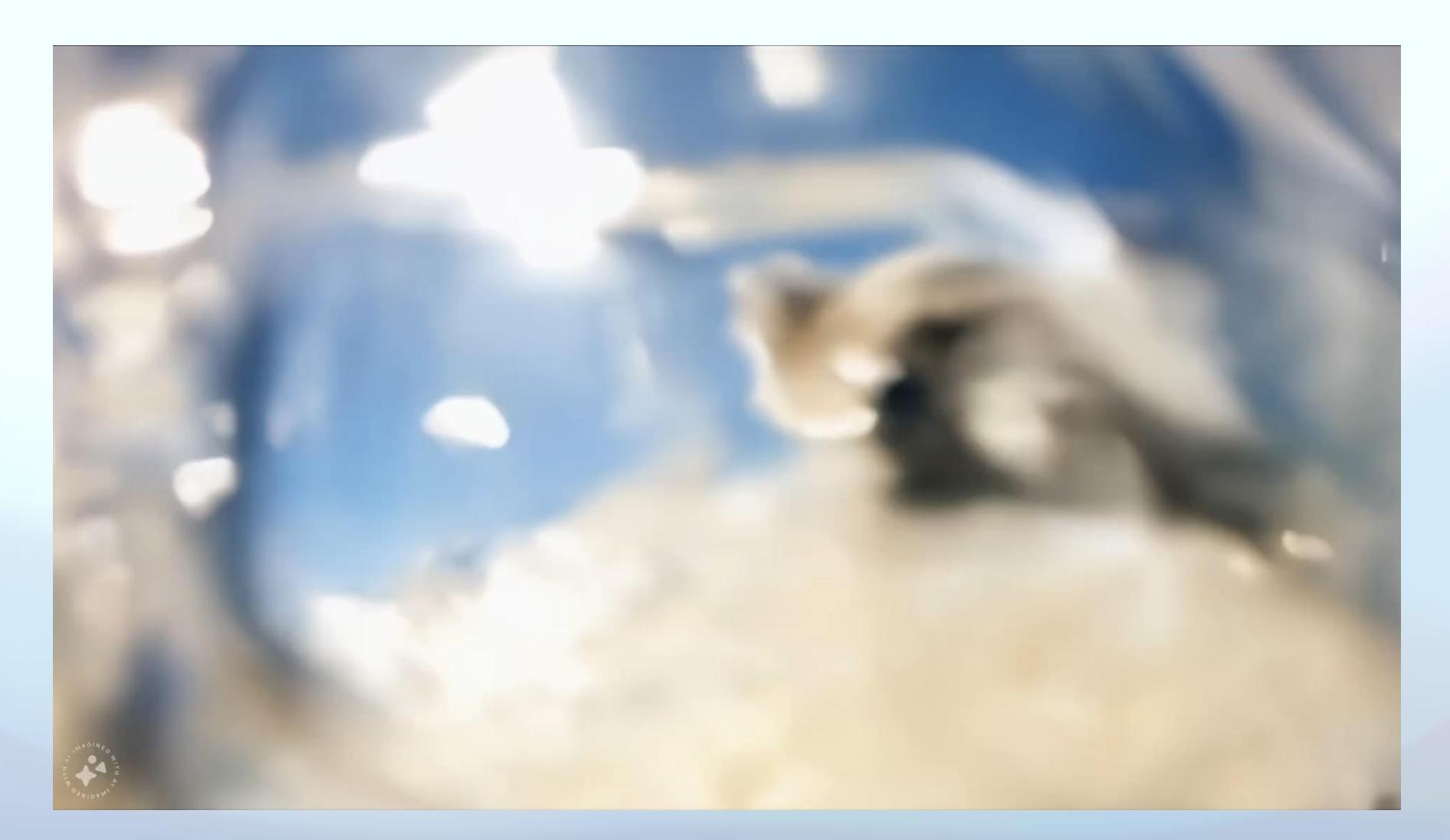


An example



MovieGen – Meta Al's Media Foundation Framework

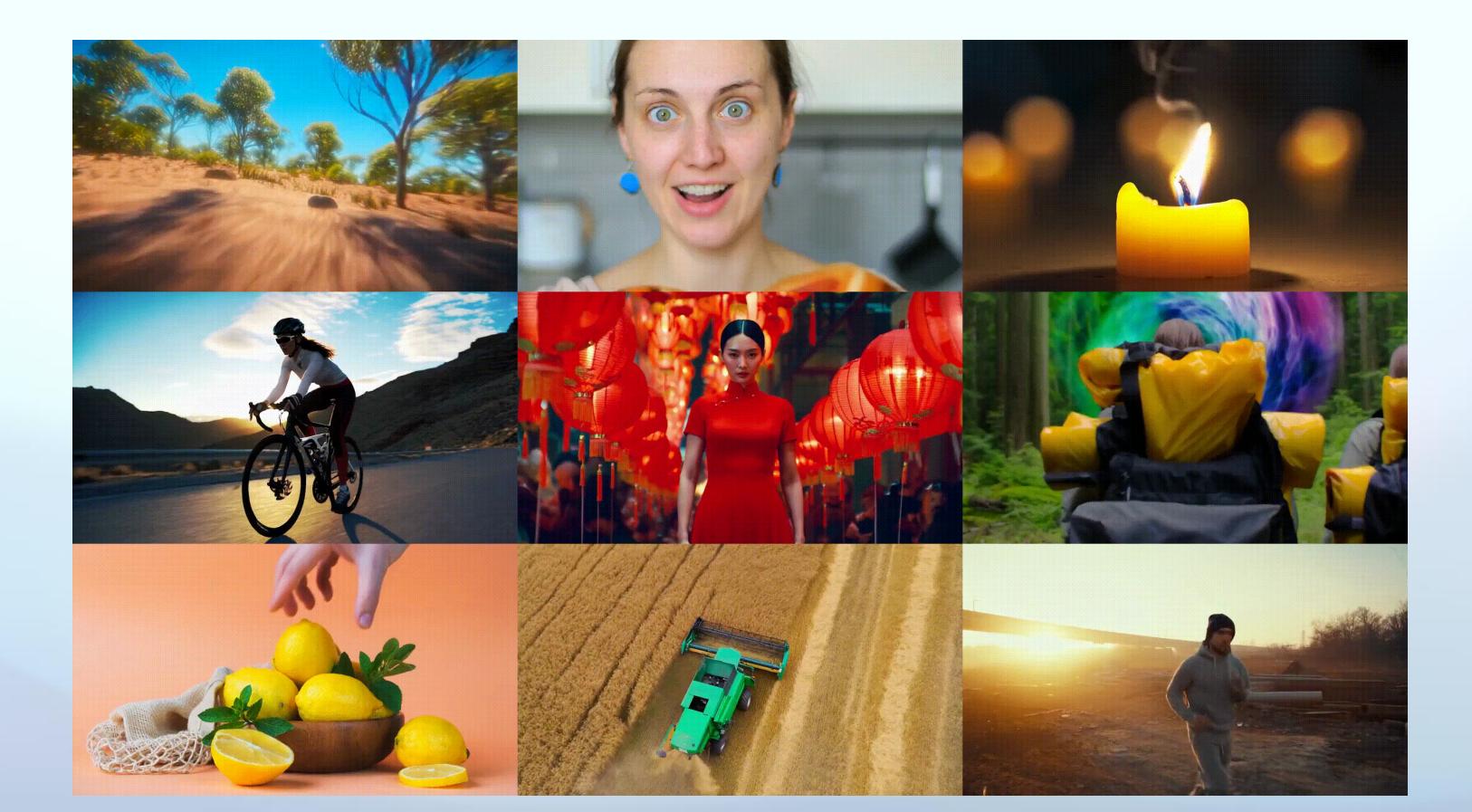
Alberto's Presentation



Recent Advancements: Mochi 1

An open-weights text-to-video generator

 It uses a 10-billion-parameter diffusion model b (AsymmDiT) architecture



• It uses a 10-billion-parameter diffusion model built on the novel Asymmetric Diffusion Transformer

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

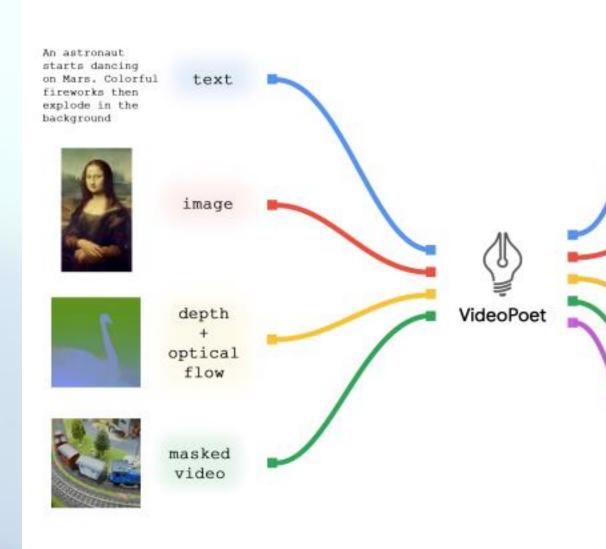


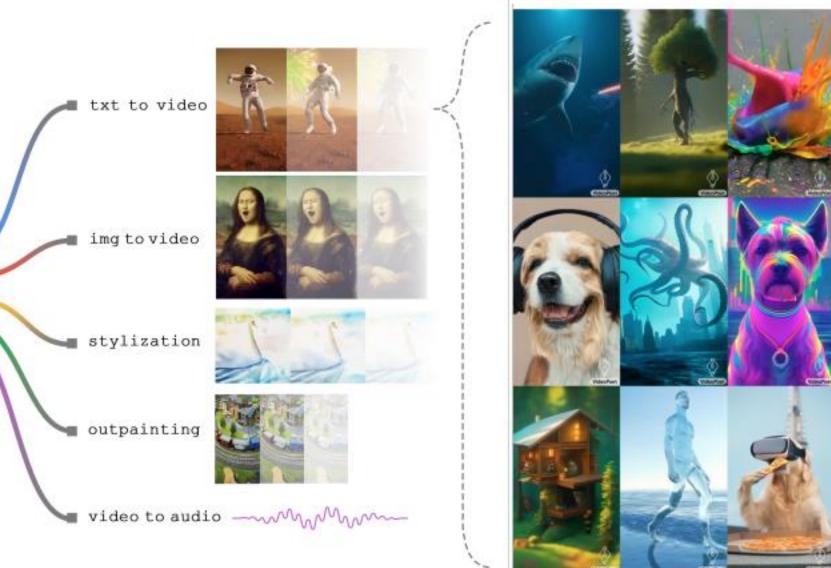
- Uses Vector Quantized-Variational Autoencoders (VQ-VAE) to encode video frames.
- Employs a transformer-based decoder to predict frame sequences.

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

- intensive.
- Autoregressive Models: Faster for short video generation but prone to error propagation.
- Temporal Consistency: Diffusion models handle it better through sequential denoising.

• Diffusion Models: Better for high-resolution and long-form video generation, but computationally

Benchmarking text-to-video generation models

- performance and guiding future improvements.
- dynamics, and alignment with textual prompts.
- Two important metrics for text-to-video generation are FVD and CLIPScore.
 - Both metrics have their own issues.

• Importance of Evaluation: Assessing text-to-video (T2V) models is crucial for understanding their

• Challenges: Evaluating T2V models involves multiple dimensions, including visual quality, temporal

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.

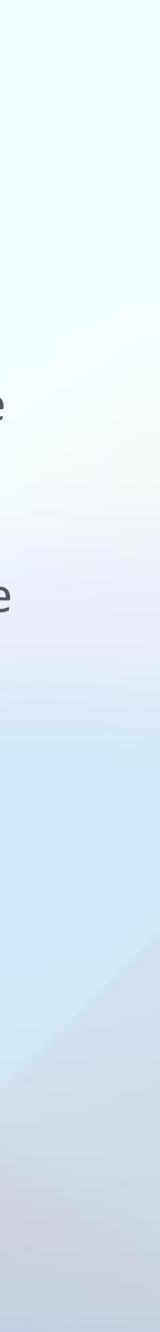
Issues with FVD

- space, which may not hold true, leading to inaccurate evaluations.
- assessment of motion quality.
- sizes.

• Non-Gaussian Feature Space: Assumes a Gaussian distribution in the Inflated 3D Convnet (I3D) feature

• **Temporal Insensitivity:** I3D features may not effectively capture temporal distortions, undermining the

• Sample Size Requirements: Reliable estimation with FVD often necessitates impractically large sample



CLIPScore and its issues

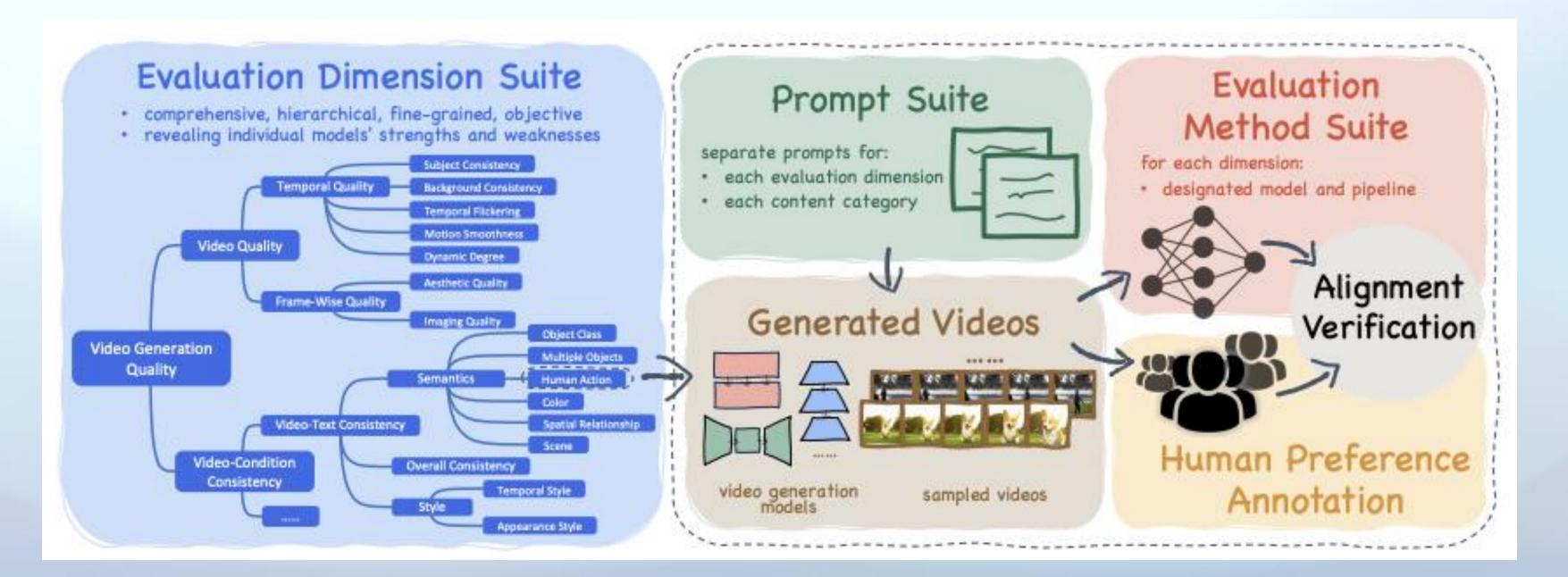
- It has its own issues:
 - It often struggles to identify fine-grained details in generation.
 - It is insensitive to certain linguistic aspects, such as negation.
- The fact that existing metrics are not good enough for evaluation is the reason for text-to-video benchmarks.

• CLIPScore computs the cosine similarity between the image and caption embeddings produced by CLIP.

VBench – Comprehensive Benchmark Suite for Video Generative Models

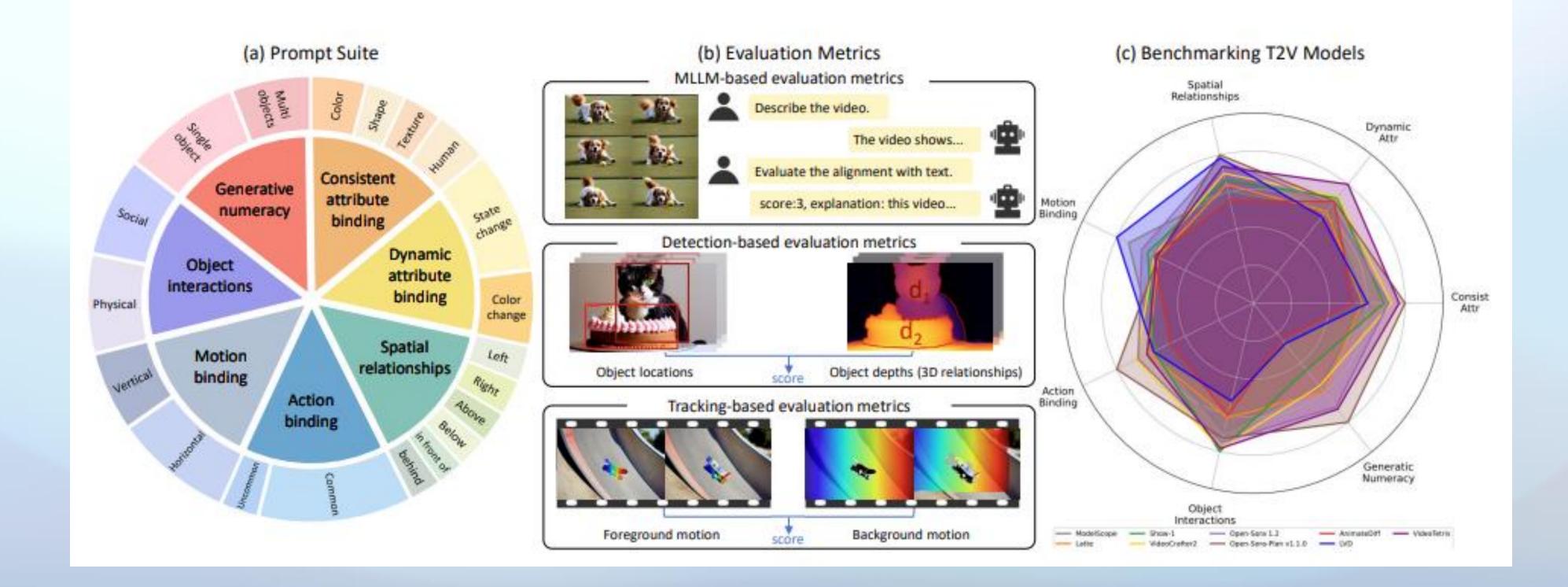
Kazusato's Presentation

- 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

- Is scaling all you need for high-quality video generation?

• What should PhD students do in response to the aggressive scaling in text-to-video model training?