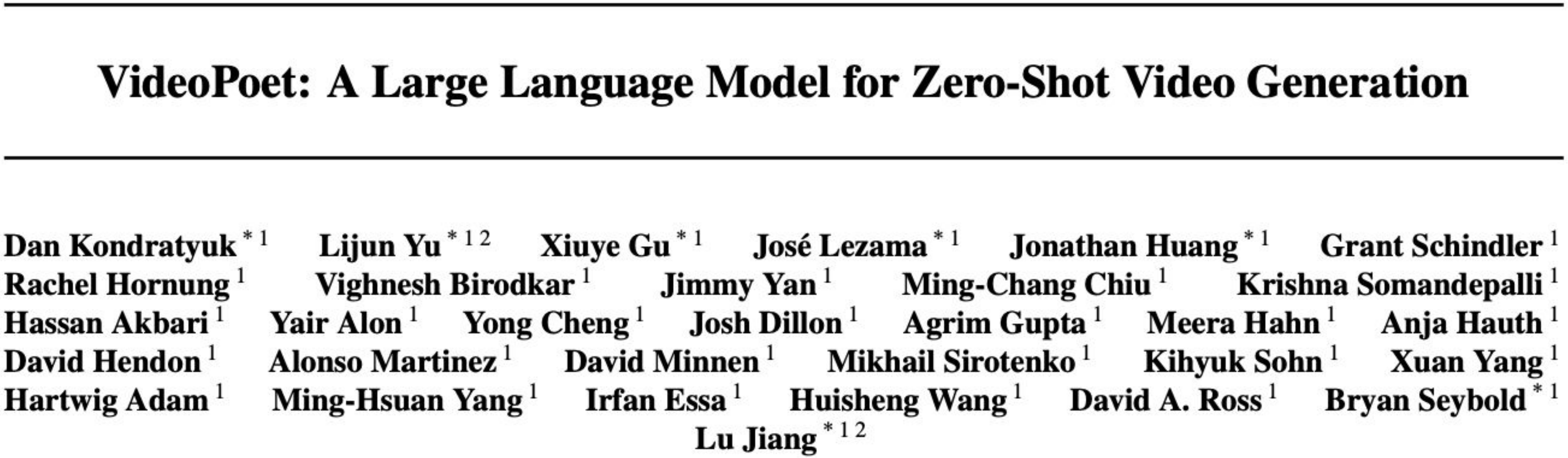


VideoPoet: A Large Language Model for Zero-Shot Video Generation

Dan Kondratyuk[*1] Lijun Yu[*12] Xiuye Gu[*1] José Lezama[*1] Jonathan Huang[*1] Grant Schindler[1]
Rachel Hornung[1] Vighnesh Birodkar[1] Jimmy Yan[1] Ming-Chang Chiu[1] Krishna Somandepalli[1]
Hassan Akbari[1] Yair Alon[1] Yong Cheng[1] Josh Dillon[1] Agrim Gupta[1] Meera Hahn[1] Anja Hauth[1]
David Hendon[1] Alonso Martinez[1] David Minnen[1] Mikhail Sirotenko[1] Kihyuk Sohn[1] Xuan Yang[1]
Hartwig Adam[1] Ming-Hsuan Yang[1] Irfan Essa[1] Huisheng Wang[1] David A. Ross[1] Bryan Seybold[*1]
Lu Jiang[*12]

Presented By: Zeeshan Patel

Zero-Shot Video Generation

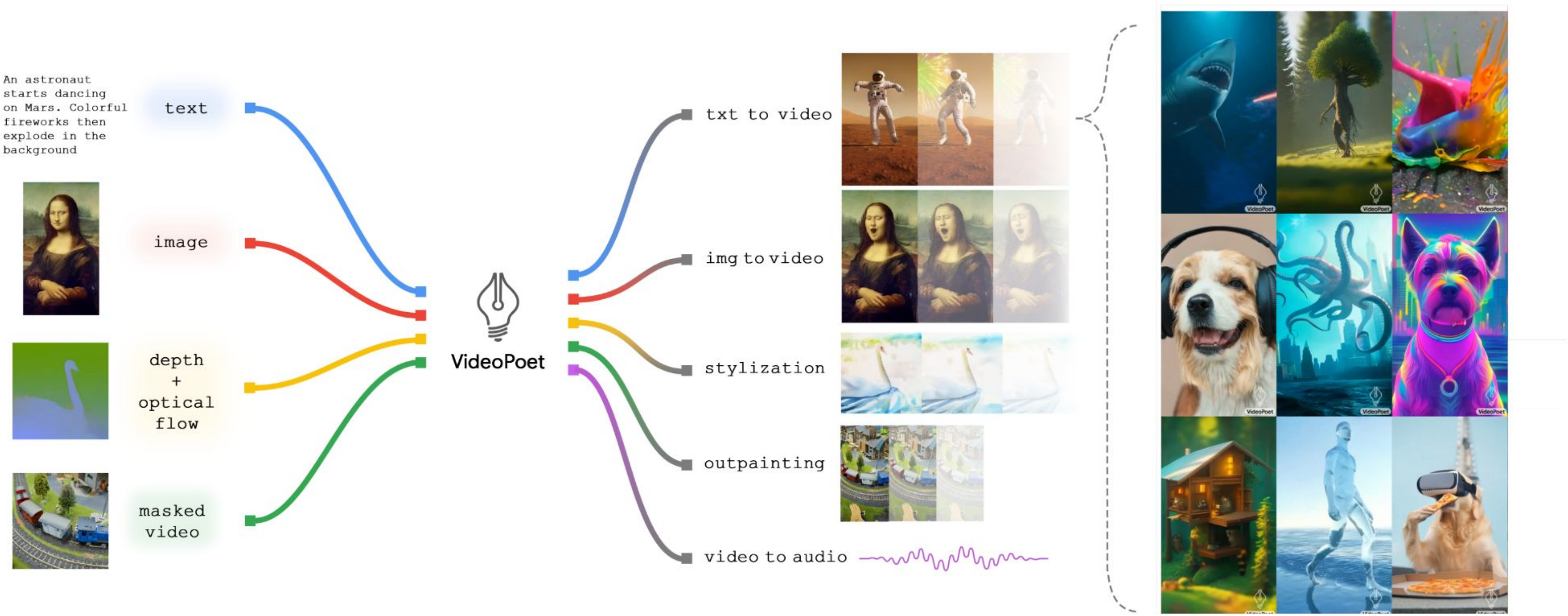


Figure 1: **VideoPoet Overview**: a versatile video generator that conditions on multiple types of inputs and performs a variety of video generation tasks.

Why Use LLMs for Video Generation?

- **Scalable**: Autoregressive LLMs have been proven to scale effectively for language with models such as GPT-4 scaling above 1T parameters
- **Infrastructure**: Can reuse training / inference infra and optimizations for LLMs to scale and serve video generation models
- **End-to-End**: Flexibility in encoding many diverse tasks in the same model, compared to most diffusion models needing architectural changes and adapter modules for more diverse tasks

Model Architecture

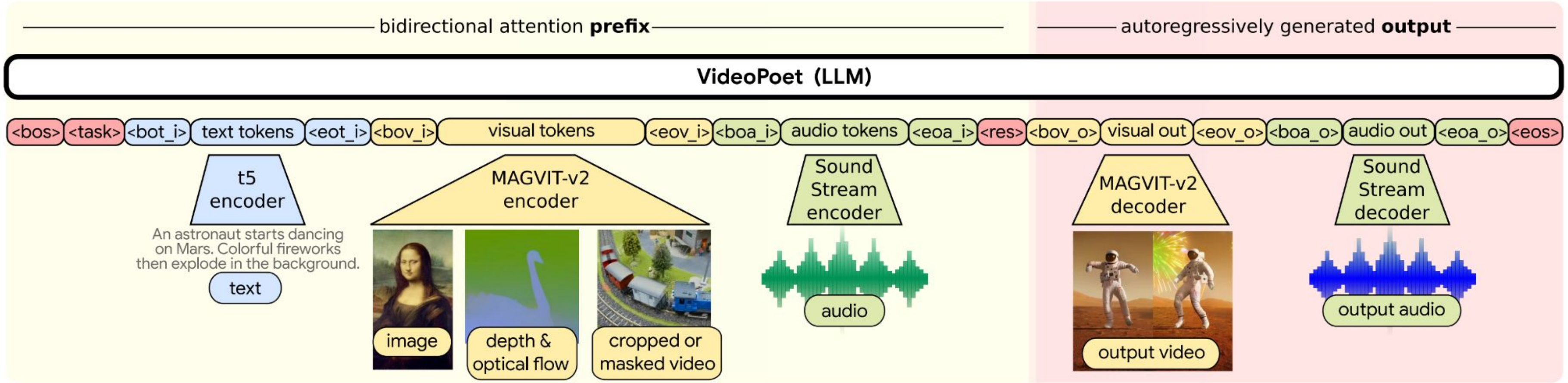
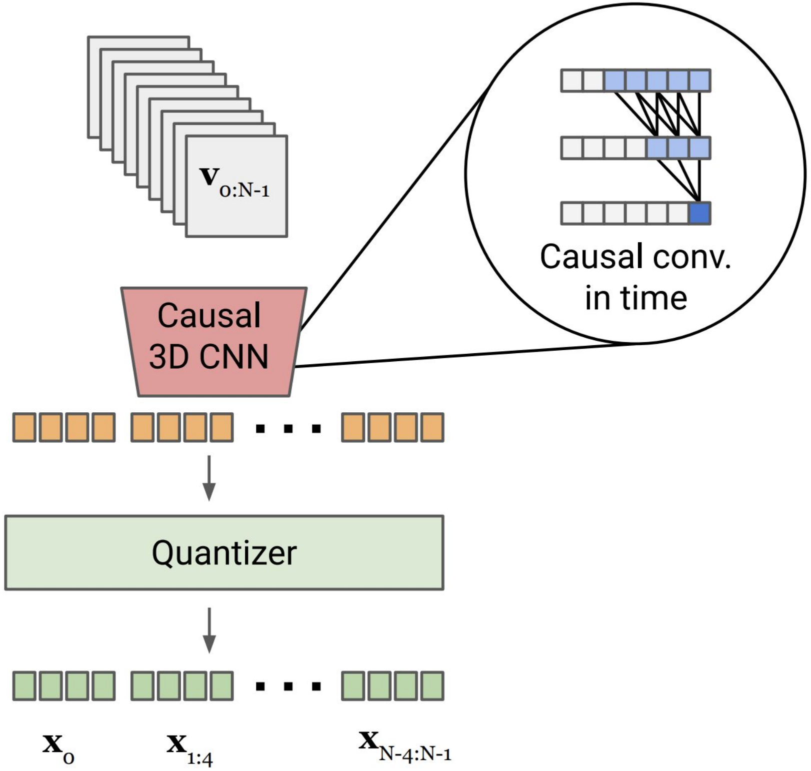


Figure 2: **Sequence layout for VideoPoet**. We encode all modalities into the discrete token space, so that we can directly use large language model architectures for video generation. We denote special tokens in <> (see Table 4 for definitions). The modality agnostic tokens are in darker red; the text related components are in blue; the vision related components are in yellow; the audio related components are in green. The left portion of the layout on light yellow represents the bidirectional prefix inputs. The right portion on darker red represents the autoregressively generated outputs with causal attention.

Video Tokenizer

- MAGViT-v2: Causal 3D CNN tokenizer
- Uses Lookup-Free Quantization to generate discrete tokens without codebook
- Current SoTA tokenizers use a similar approach as MAGViT-v2 but also incorporate attention layers, 3D wavelet transforms for more compact / less redundant video representation, and Finite-Scalar Quantization instead of LFQ



Video Super-Resolution

- Expensive to generate high-res video directly from AR model
- VideoPoet uses video super-resolution transformer to upsample low-res video generated by AR model
- Q: Is this a good way to efficiently generate high-res videos, or does this approach present a new inductive bias that will hinder scalability?

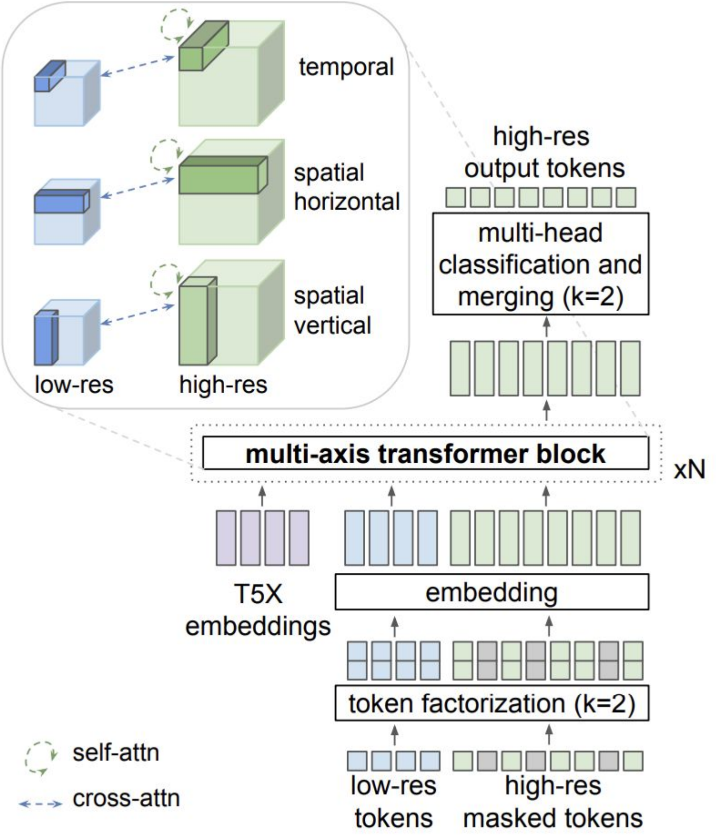
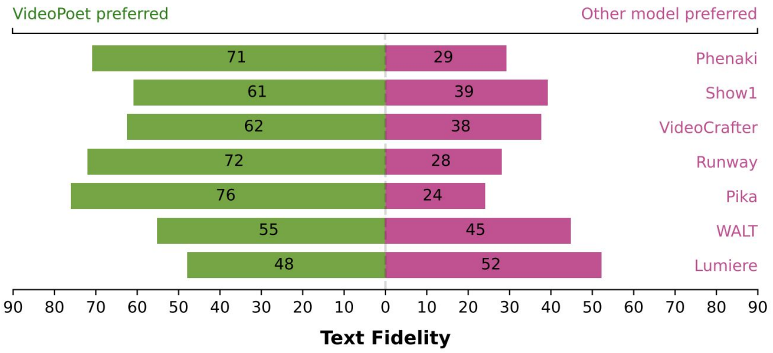


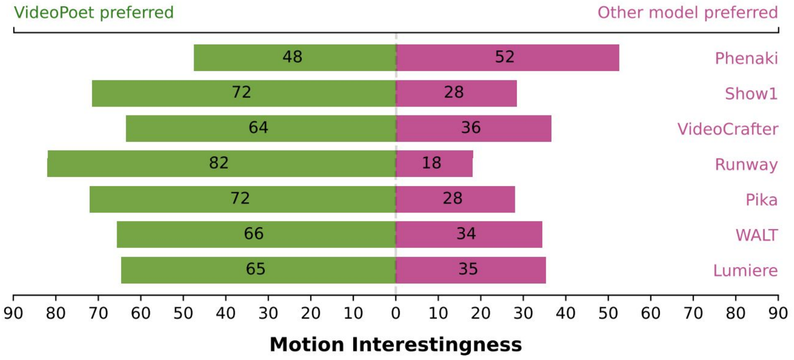
Figure 3: **Custom transformer architecture for video super-resolution.**

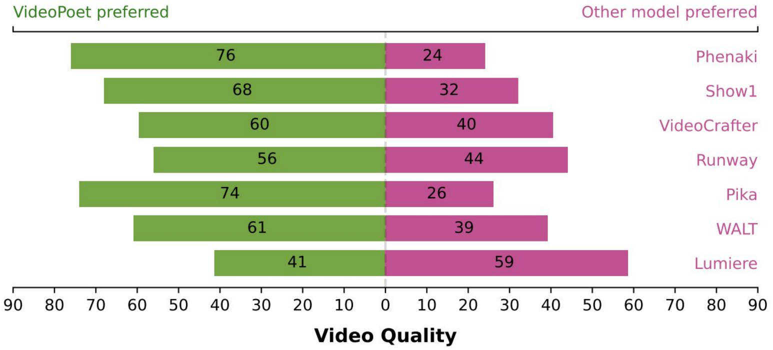
Dataset / Training Strategy

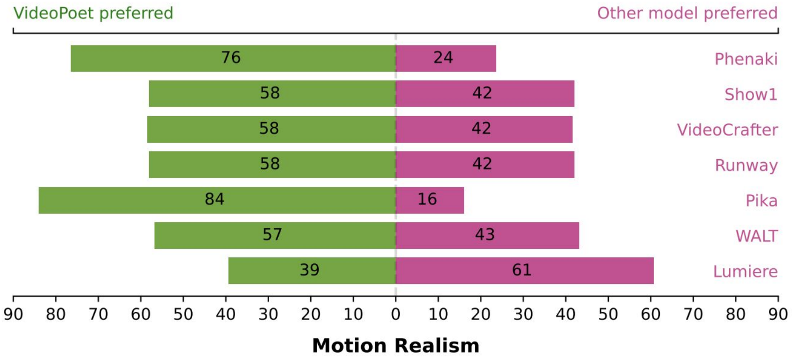
- Mix of images / videos at varying resolutions
- 1B image-text pairs
- ~270M videos
  - ~100M with paired text, of which ~50M used for high quality fine-tuning
  - ~170M with paired audio
- Approx. 2 trillion tokens across all modalities
- Q: The paper doesn’t specify how they curate/filter their dataset. What kinds of things would you consider when creating a pre-training dataset for video vs. text modeling?

Human Evaluation



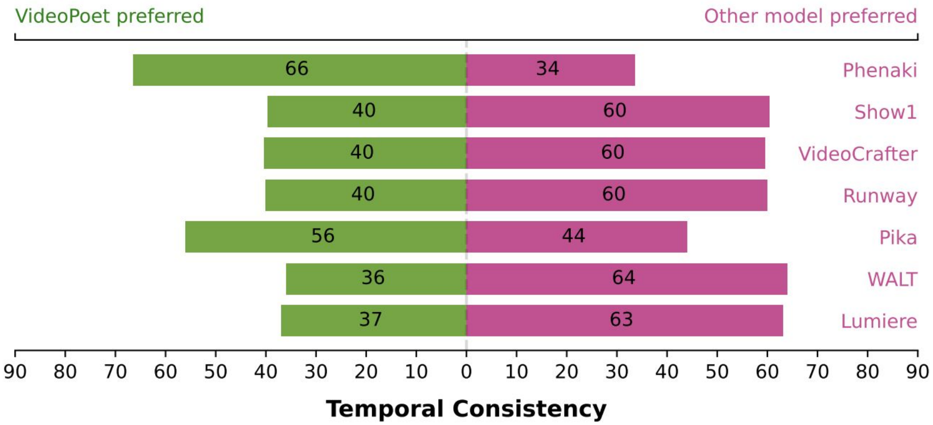






Temporal Consistency

- Authors claim that scaling the model helped with temporal consistency
- Q: How do other methods ensure better temporal consistency and can VideoPoet leverage them? (Patrick)
- Q: Authors note it’s hard to capture fine details for VideoPoet with discrete tokens. Is there a better ARM approach over these tokens, or should we investigate the encoders used to generate the tokens in the first place? (Ryan)



Comparison with Closed Source Models

Sora

MovieGen

Comparison with Open Source Models

Mochi 1 (Diffusion)

Pyramid Flow Matching (AR)

Current SoTA Approaches + Open Problems

- Sora / MovieGen use similar tokenizer as VideoPoet but operate on continuous tokens with a diffusion / flow-matching objective
- Full sequence diffusion vs. AR next token prediction
- Open Problems in Long Video Generation
  - Full sequence diffusion can be costly (i.e. 1M+ tokens for high quality one minute video)
  - AR models typically collapse and have consistency issues over long sequence length due to error accumulation
- Q: What are some limitations of existing approaches in modeling real-world environments? Will video models be useful for closing the sim-to-real gap?