# ICML

International Conference On Machine Learning

Vienna, 2024

27.5% (2610/9473) (144 orals, 191 spotlights and 2275 posters)

# Best Papers

1. VideoPoet: A Large Language Model for Zero-Shot Video Generation (Google Deepmind)
2. Genie: Generative Interactive Environments (Google, generate interactive and playable environments for AI agents)
3. Stealing part of a production language model (Google Deepmind & ETH & OpenAI & McGill, attack algorithms, recovering the embedding projection layer,hidden dimension of gpt-3.5)
4. Debating with More Persuasive LLMs Leads to More Truthful Answers (UCL & Anthropic & FARAI)
5. Probabilistic Inference in Language Models via Twisted Sequential Monte Carlo (UoT & Vector Institute)
6. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution (Stanford & Pika Labs)
7. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis (Stability AI)
8. Information Complexity of Stochastic Convex Optimization: Applications to Generalization, Memorization, and Tracing
9. Position: Considerations for Differentially Private Learning with Large-Scale Public Pretraining
10. Position: Measure Dataset Diversity, Don't Just Claim It

Google DeepMind

ICML
International Conference
On Machine Learning

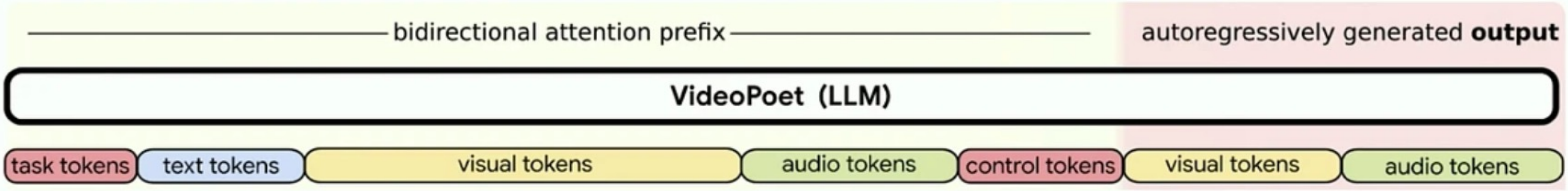
# VideoPoet: A Large Language Model for Zero-Shot Video Generation

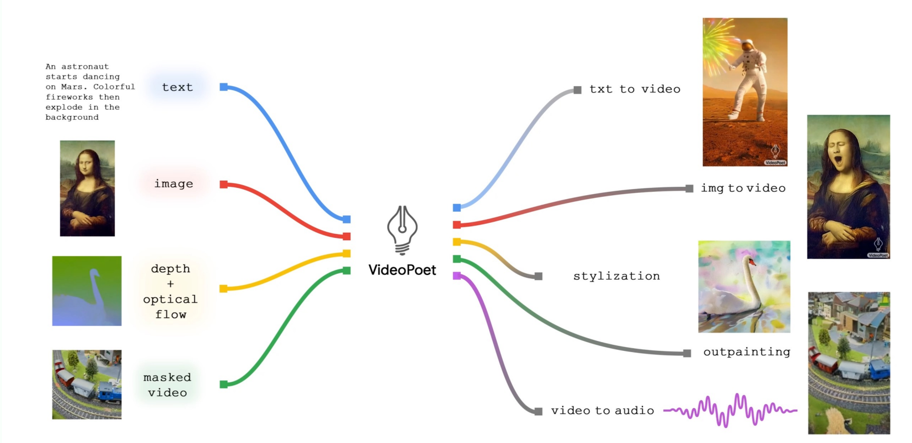
ICML 2024 Best Paper

Presenter: Lijun Yu

https://sites.research.google/videopoet/

# VideoPoet Framework



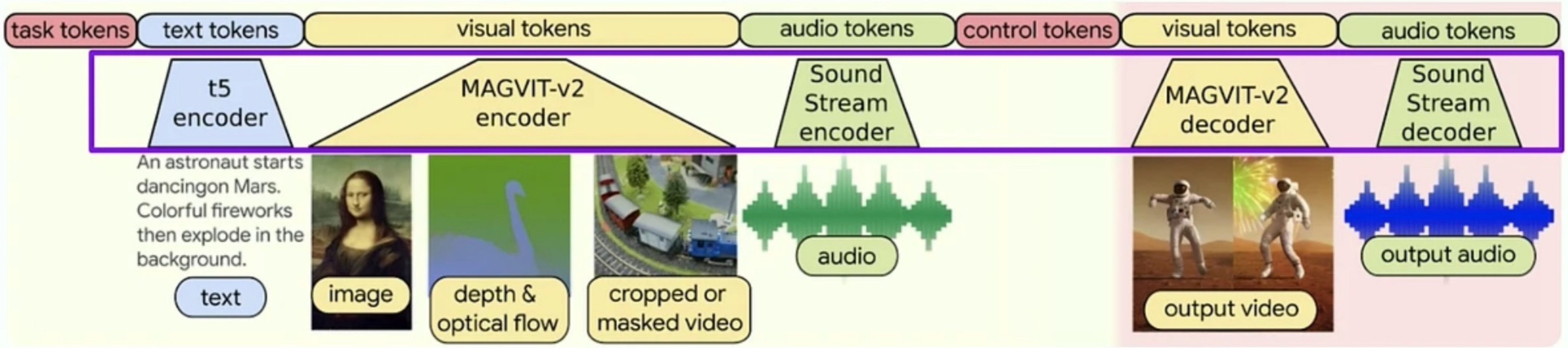


# VideoPoet Framework

Modality-specific tokenizers

Raw signal ➡ Encode & Compress ➡ Token

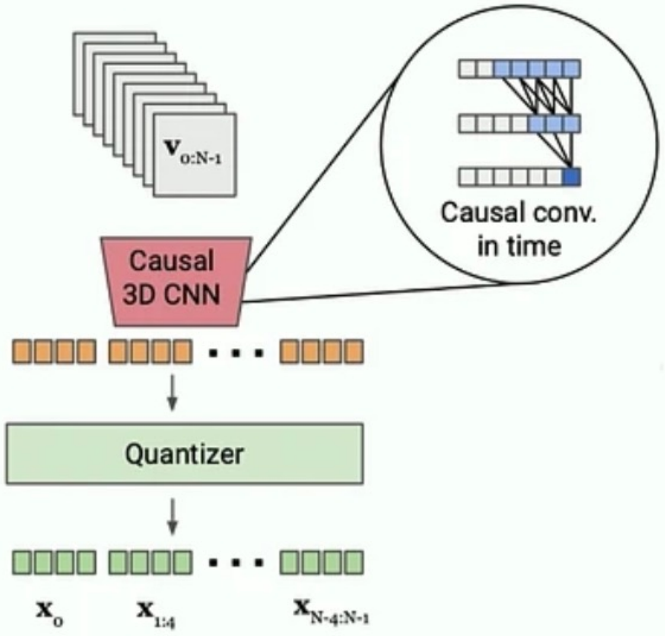
Raw signal ⬅ Decode & Decompress ⬅ Token



# MAGVIT-v2 Video Tokenizer

Defining the visual “language”

- Quantized VAE w/ temporally causal 3D CNN
  - Image as a prefix of video for joint training
  - Seamless support for long videos

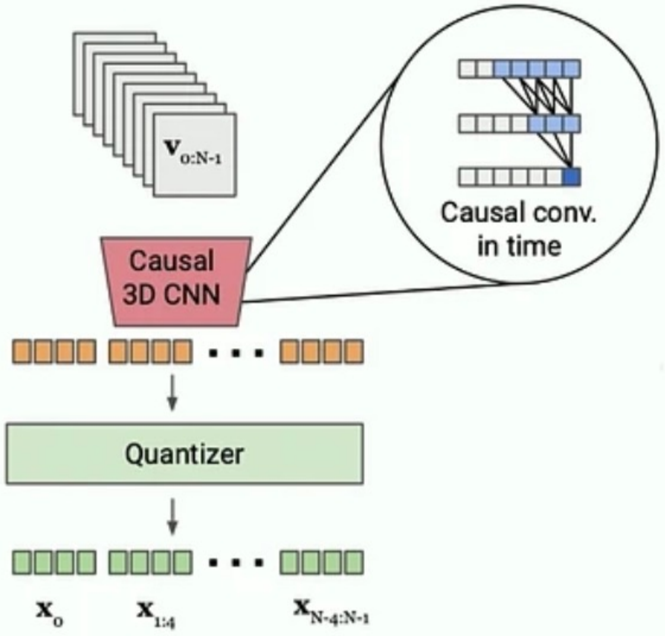


Yu et al. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. In ICLR 2024.

# MAGVIT-v2 Video Tokenizer

Defining the visual "language"

- Quantized VAE w/ temporally causal 3D CNN
  - Image as a prefix of video for joint training
  - Seamless support for long videos



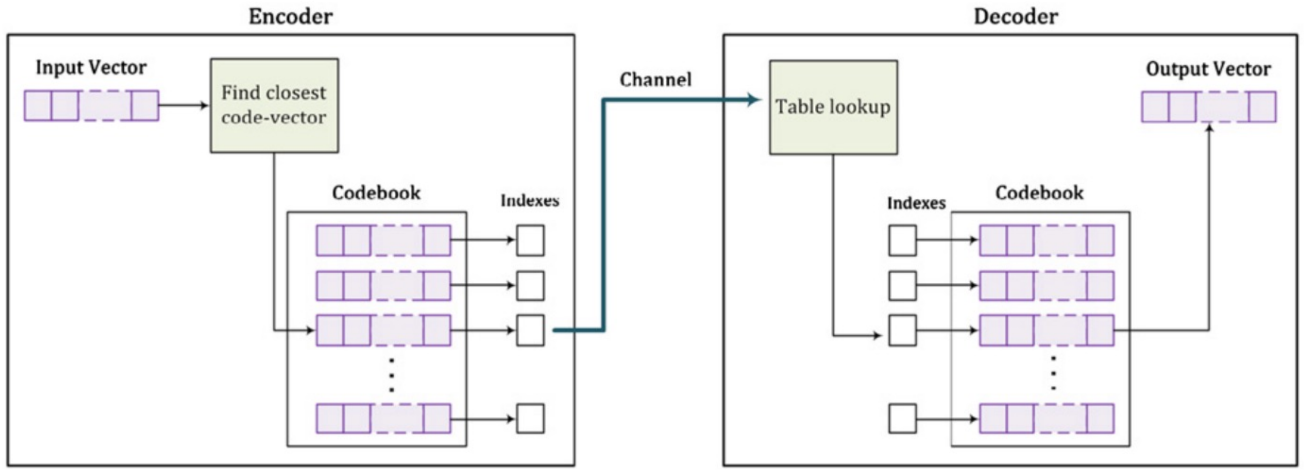


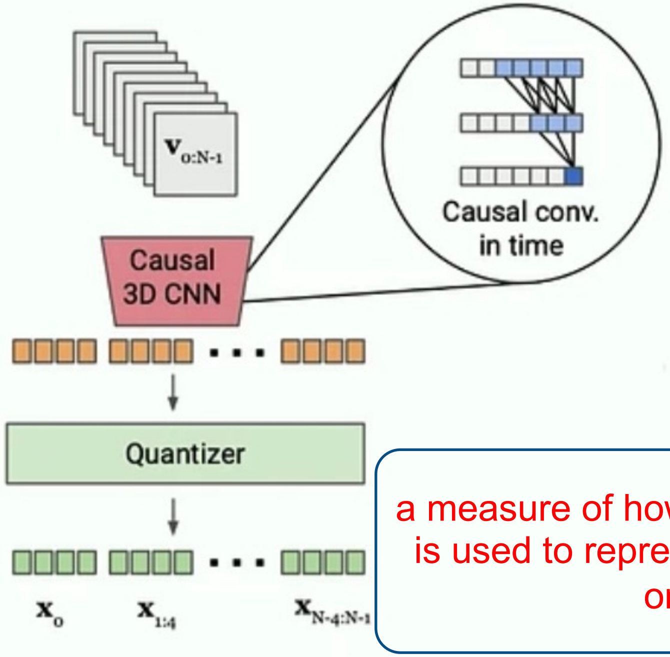
**Fig. 1.** Block diagram for the vector quantization.

Yu et al. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. In ICLR 2024.

# MAGVIT-v2 Video Tokenizer

Defining the visual "language"

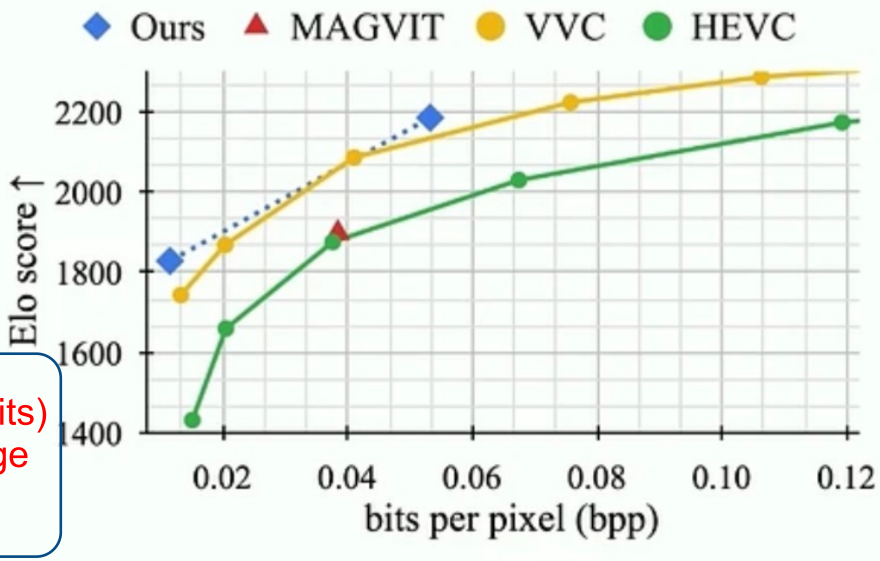
- Quantized VAE w/ temporally causal 3D CNN
  - Image as a prefix of video for joint training
  - Seamless support for long videos



262,144 discrete tokens

- Scalable quantizer w/ $2^{18}$ large vocabulary for higher prediction bandwidth
- Better compression than VVC with reconstructive and adversarial training

H.266



a measure of how much **information** (in bits) is used to represent each pixel in an image or video frame.

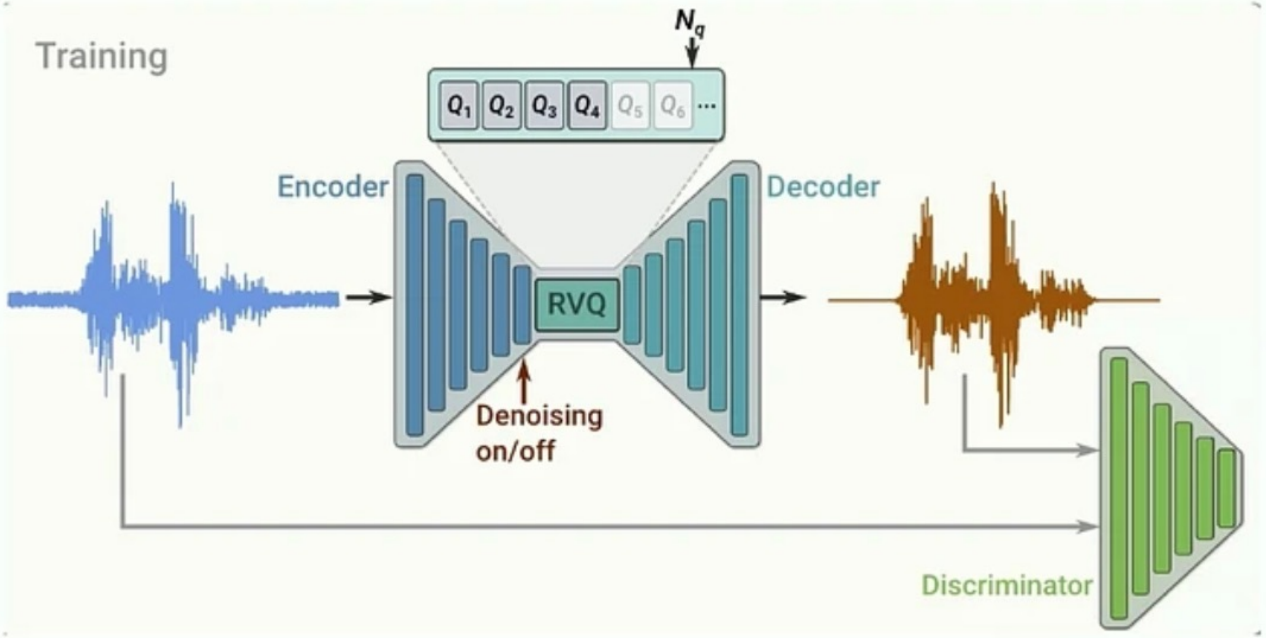
Yu et al. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. In ICLR 2024.

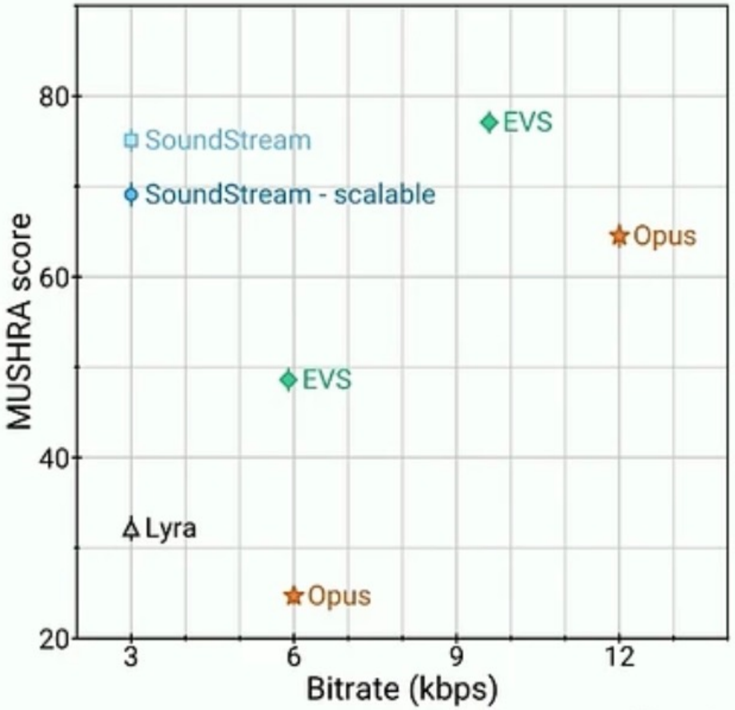
# SoundStream Audio Tokenizer

Defining the audio "language"

- Quantized VAE w/ causal 1D CNN

- Residual vector quantizer
- Better compression than Opus

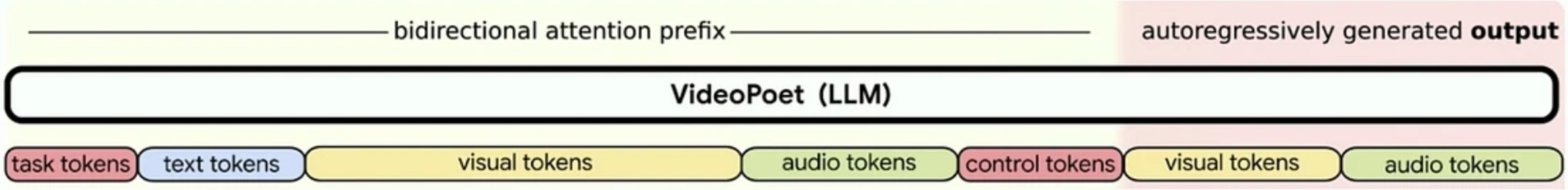




Zeghidour et al. SoundStream: An End-to-End Neural Audio Codec. TASLP 2021.

# VideoPoet Framework

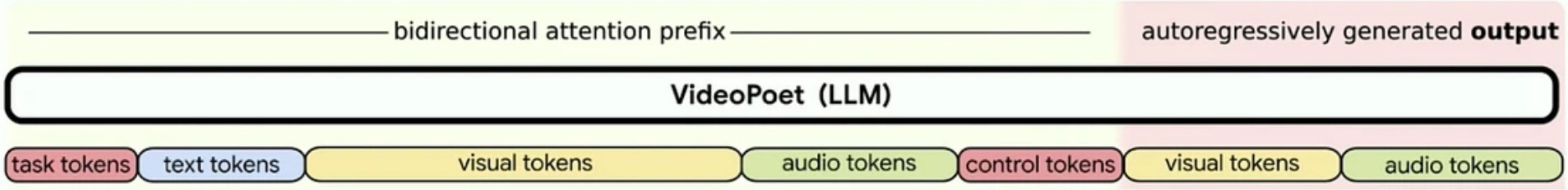
Out-of-the-box LLM transformer on discrete token sequences



# VideoPoet Framework

Out-of-the-box LLM transformer on discrete token sequences

- Flexibility: any to any task setup in a single model
- Training efficiency: learning entire sequence in a single step with causal attention
- Inference efficiency: various acceleration techniques such as caching, where full decoding FLOPs equal to one full forward pass



Google DeepMind

# Training Data

Mixture of pre-existing sources and formats, in two training phases

- 🔥 Pretraining uses everything, including unlabeled and noisy data

| Source | 🎥 Video | 🔊 Audio | 🖼 Image | 🔤 Text | Sample Count | 🔥 Pretrain |
|---|---|---|---|---|---|---|
| A | ✅ | ✅ | | | ~170M | ✅ |
| B | ✅ | ◑ | | ◑ | ~50M | ✅ |
| C | ✅ | | | ✅ | ~50M | ✅ |
| D | | | ✅ | ✅ | ~1B images | ✅ |

# Training Data

Mixture of pre-existing sources and formats, in two training phases

- 🔥 Pretraining uses everything, including unlabeled and noisy data
- 🔧 Task adaptation uses task-specific high quality data

| Source | 🎥<br>Video | 🔊<br>Audio | 🖼<br>Image | 🔠<br>Text | Sample Count | 🔥<br>Pretrain | 🔧<br>Adapt T2V |
|---|---|---|---|---|---|---|---|
| A | ✅ | ✅ | | | ~170M | ✅ | |
| B | ✅ | ◑ | | ◑ | ~50M | ✅ | |
| C | ✅ | | | ✅ | ~50M | ✅ | ✅ |
| D | | | ✅ | ✅ | ~1B images | ✅ | |

# Training Tasks

Self-supervised

| **Output**<br>**Prefix** | Continue ▶️ | **Video** 🎥 | Audio 🔊 | 🎥🔊 | Image 🖼️ |
|---|---|---|---|---|---|
| 🚫 Unconditional | | ✅ | ✅ | ✅ | ✅ |
| 🎥 Video | ✅ | | | | |
| 🔊 Audio | ✅ | | | | |
| 🎥🔊 Video + Audio | ✅ | | | | |
| 🖼️ Image | | ✅ | | | |

# Training Tasks

Self-supervised

| Output<br>Prefix | Continue ▶️ | **Video** 🎥 | Audio 🔊 | 🎥🔊 | Image 🖼️ |
|---|---|---|---|---|---|
| 🚫 Unconditional | | ✅ | ✅ | ✅ | ✅ |
| 🎥 Video | ✅ | ✅++ | ✅ | | |
| 🔊 Audio | ✅ | ✅ | | | |
| 🎥🔊 Video + Audio | ✅ | | | | |
| 🖼️ Image | | ✅ | | | |

# Training Tasks

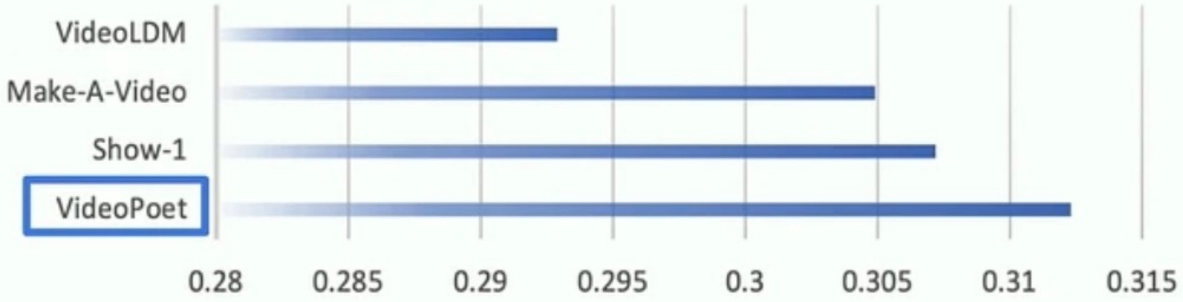
Self-supervised + Supervised

| **Output**<br>**Prefix** | Continue ▶️ | **Video** 🎥 | Audio 🔊 | 🎥🔊 | Image 🖼️ | Style 🎨 |
|---|---|---|---|---|---|---|
| 🚫 Unconditional | | ✓ | ✓ | ✓ | ✓ | |
| 🎥 Video | ✓ | ✓++ | ✓ | | | ✓ |
| 🔊 Audio | ✓ | ✓ | | | | |
| 🎥🔊 Video + Audio | ✓ | | | | | |
| 🖼️ Image | | ✓ | | | | |
| 🔠 Text | | ✓ | | ✓ | ✓ | |

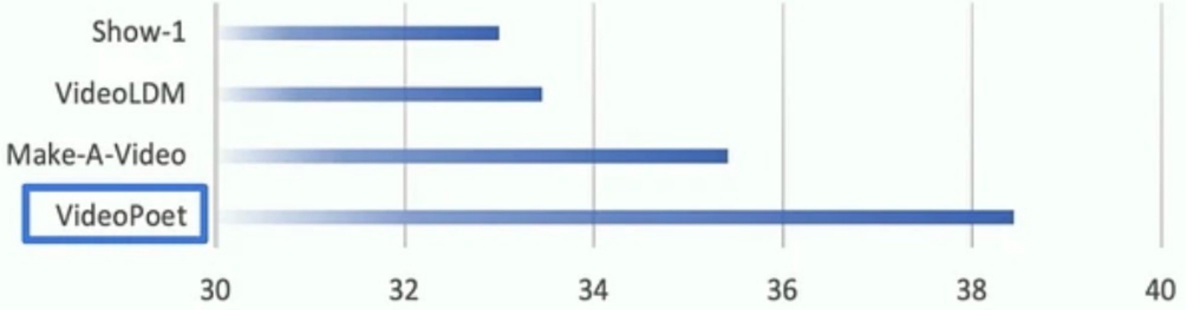
# Automatic Benchmarks

Zero-shot text-to-video generation comparison with state-of-the-art

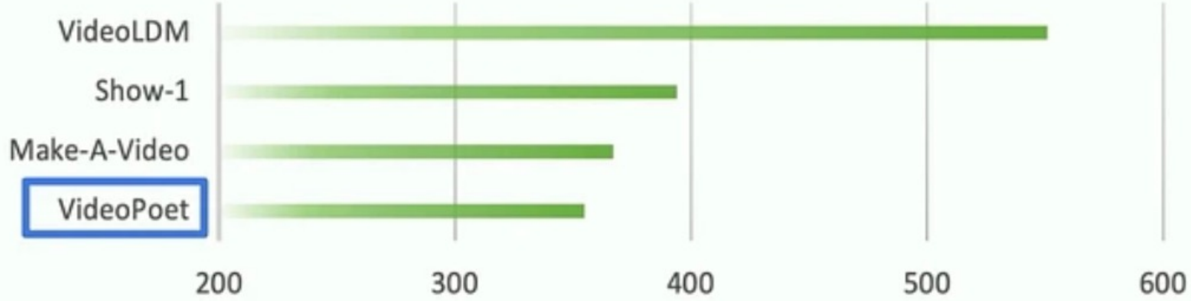
MSR-VTT CLIPSIM↑



UCF-101 IS↑



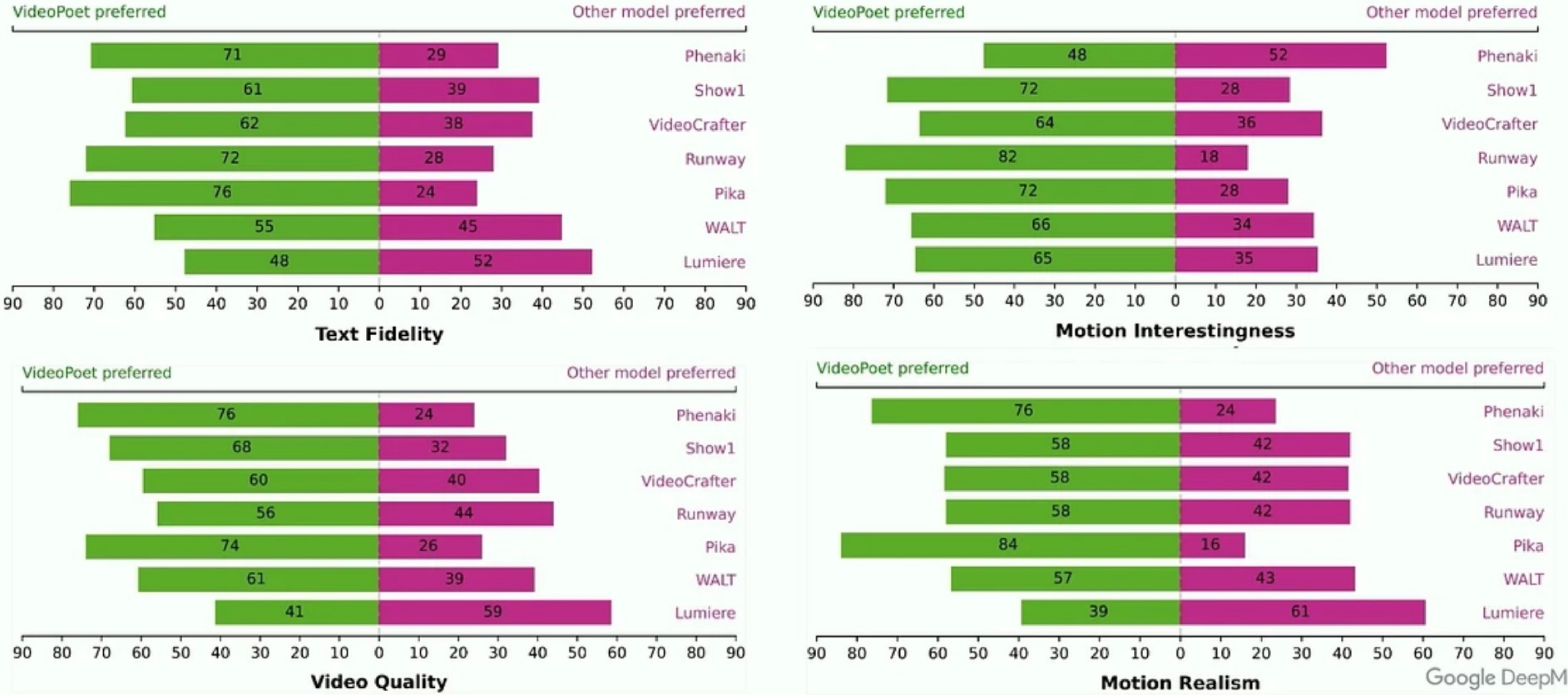
UCF-101 FVD↓



Google DeepMind

# Human Evaluations

Zero-shot text-to-video generation comparison with prior and concurrent works



# Future Research Directions

- Real-time streaming video generation
  interactive neural gaming, neural user interface for OS / APPs
- Universal multimodal generative model
  SOTA generation of text & video (& audio & …) and reasoning
  c.f., human-level machine translation (~18′) -> ChatGPT (23′)

Query: Can you show me how to tie this shoe with a single hand? 🥾

# Summary

VideoPoet represents a distinct approach to video generation

- State-of-the-art quality,
  challenging the diffusion monopoly
- Multi-task flexibility,
  going beyond the text to video translation paradigm
- Video-first foundation model,
  building upon LLM infrastructure for native integration