Language Model Beats Diffusion
– Tokenizer is Key to Visual Generation

Lijun Yu

01/2024

lijun@cmu.edu

A Short Bio of Lijun Yu

- Graduating Ph.D. student at Carnegie Mellon University,
  working with Prof. Alex Hauptmann
- (Former) student researcher at Google, working with Dr. Lu Jiang
- Research focus:
  multi-modal foundation models,
  esp. video generation w/ transformers

- “Computers are scared of me”
- Nice to meet you all!

Motivation

- LMs (e.g., GPT-4) have dominated generative tasks in language
- LMs can also generate images and videos, e.g., DALL·E, MaskGIT
  - But they do not perform as well as diffusion models, e.g., LDM
  - A significant gap exists on the gold standard ImageNet benchmark (FID 3.4 vs. 1.8)

- Why do language models lag behind diffusion models in visual generation?

Here LMs refer to transformer models that learn discrete token sequences

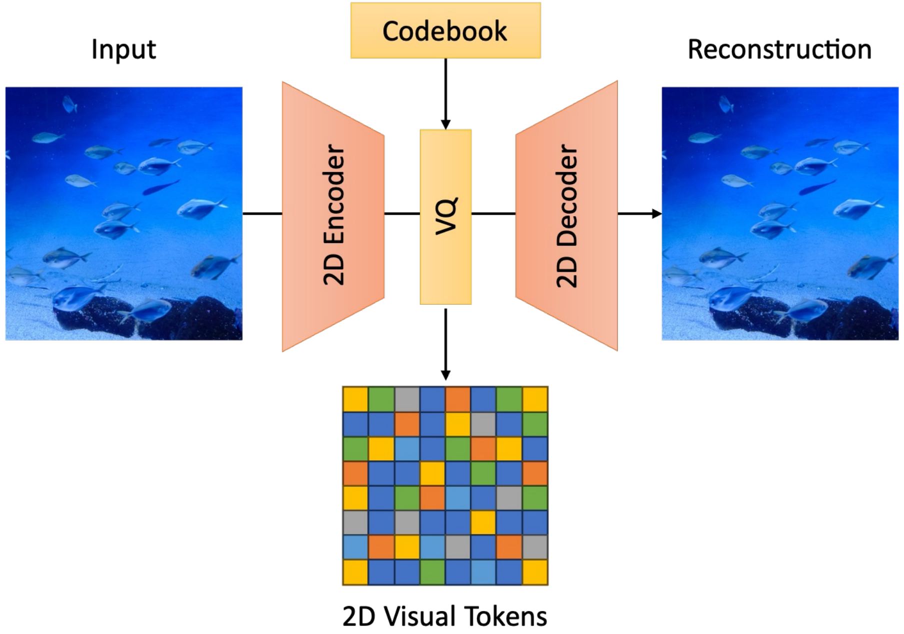
Background: LMs in Visual Generation

- Pixels are mapped into a sequence of discrete tokens by a visual tokenizer, then processed by an LM transformer as if they are lexical tokens.
- Tokenizer remains the key bottleneck that controls sequence length and generation quality.

		Tokenizer	LM Type
Image	ImageGPT	Color clustering	AR-LM & MLM
	DALL·E	dVAE	AR-LM
	Taming transformer	VQGAN	AR-LM
	Parti	ViT-VQGAN	AR-LM
	MaskGIT & Muse	VQGAN	MLM
Video	Phenaki	CViViT VQGAN	MLM
	MAGVIT	3D VQGAN	MLM

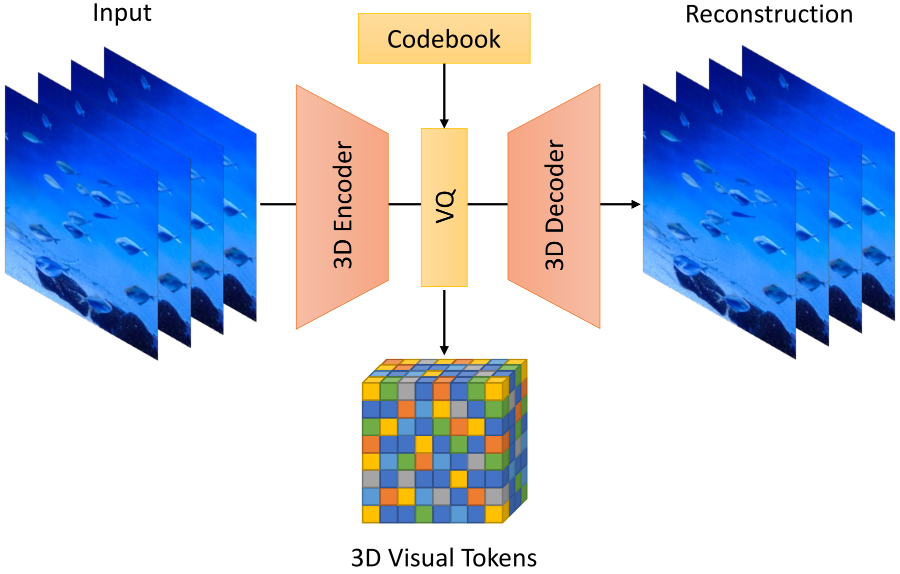
Preliminary: Image Tokenization

- Usually designed around the VQ-VAE framework
  - Autoencoder with a discrete bottleneck
  - Vector quantization with a learned codebook
  - Spatial down sampling with CNN or ViT encoders
- Variants with different setups
  - DALL·E dVAE uses ELB with gumble-softmax
  - VQGAN adds perceptual and GAN losses
  - ViT-VQGAN uses StyleGAN discriminator



Video Tokenization

- Naively: frame-by-frame tokenization
  - Suffer from consistency issues, esp. for VQGAN
  - Long redundant token sequence is a burden
- MAGVIT 3D VQGAN
  – prior best video-native tokenizer
  - Inflated 3D CNN architectures for better motion and consistency
  - Spatial-temporal down sampling to reduce redundancy
  - Losses: L2, perceptual, GAN, commitment, codebook, entropy, LeCAM



Issues with MAGVIT Tokenizer

- AR-LM and MLM do not scale well beyond 2B parameters.
  - Performance mainly bounded by the tokenizer
- Vocabulary is limited around 1-8k, compared to ~200k used in LLMs
  - Larger vocabulary hurts generation performance
- Only supports 16-frame clips, not images or longer videos
  - Convolution padding results in strong implicit temporal encoding
  - Hinders joint training with large-scale image data and long video generation

Introducing MAGVIT-v2 Tokenizer

- Lookup-free quantizer enables scalable vocabulary that helps generation
- Temporally causal 3D CNN jointly supports images and videos of variable length
- A collection of enhancements for visual quality

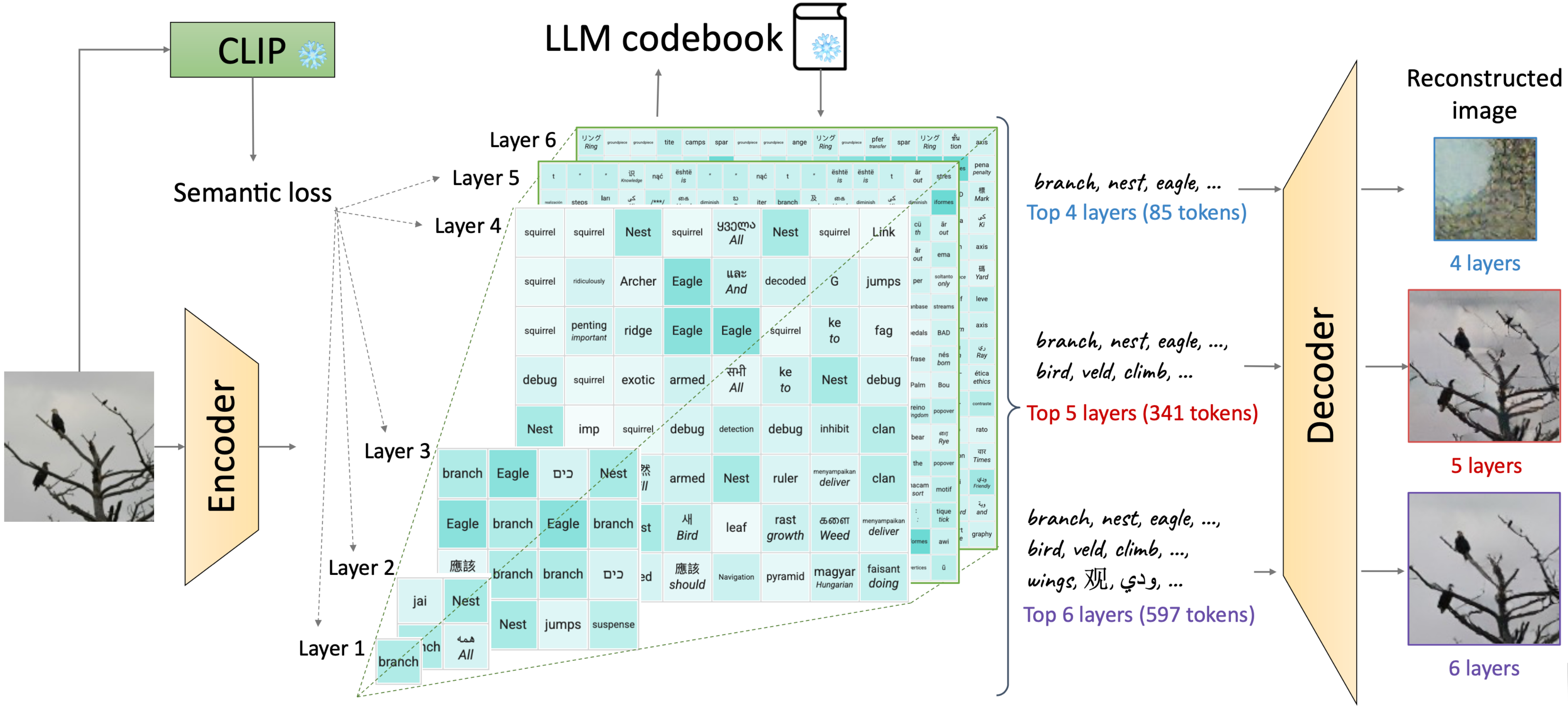
- State-of-the-art image and video generation on standard benchmarks.
- Better video compression than HEVC and VVC
- Stronger video understanding than MAGVIT

- Enabling LMs to scale! E.g., VideoPoet

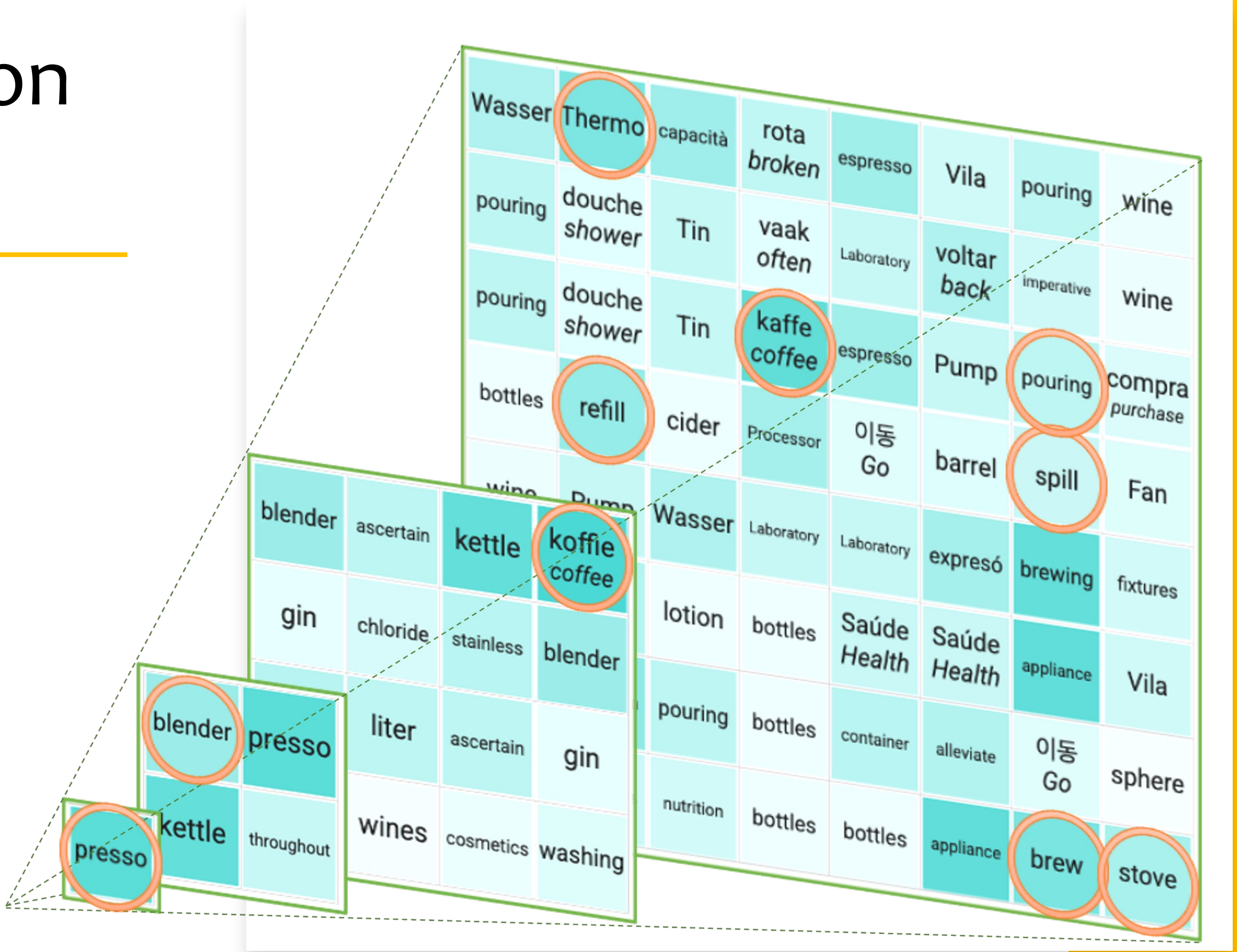
---- Previously ----

SPAE: Semantic Pyramid AutoEncoder

- Anchored frozen language codebook (>65k)
- Hierarchical vector quantization
- Semantic loss

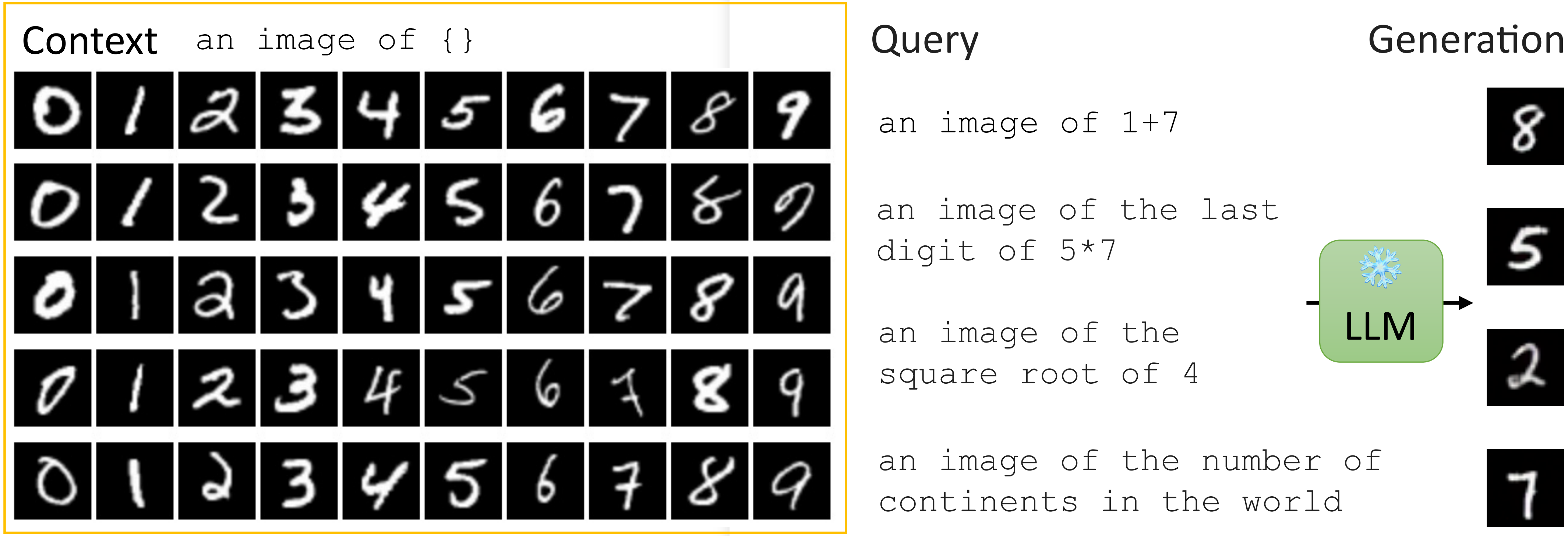


SPAE Tokenization Example



Text to Image with frozen LLMs

The **first time** a frozen LLM generates images without relying on external models, e.g., stable diffusion

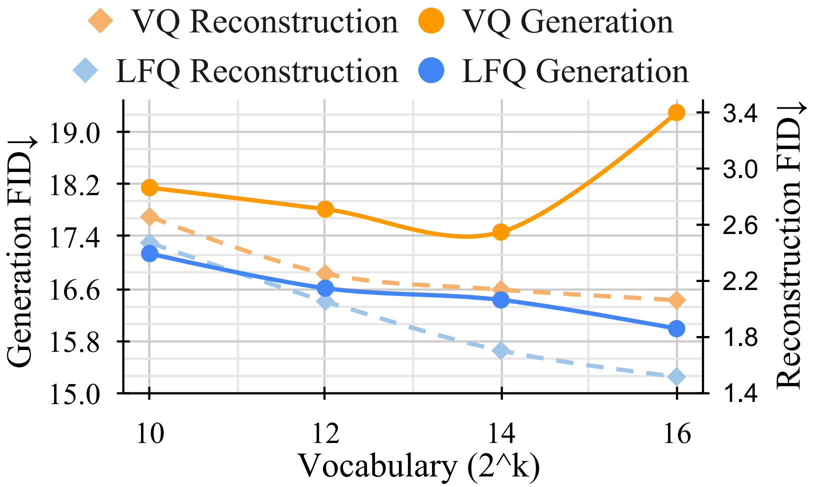


With SPAE, we can transform any LLM into a Gemini-style multimodal model even without tuning.

Lookup-Free Quantization

- Commonly used vector quantization relies on nearest neighbor lookup
  - Suffer from codebook collapse and efficiency issues when scaling to larger vocabulary size
  - Codebook learning may not be necessary, e.g., in SPAE
- Improving tokenizer reconstruction does not guarantee improvement of generation
  - E.g., increasing the VQ vocabulary, so most models use 1-8k vocabulary, <<200k of LLMs
- Reducing the code embedding dimension helps training with a larger codebook
  - Limiting the representational capacity of individual tokens to learn the distribution over a large vocabulary

What if we reduce the embedding dimension to zero?
It becomes lookup-free!
And growing the vocabulary helps generation (even to 2^{40})



Lookup-Free Quantization

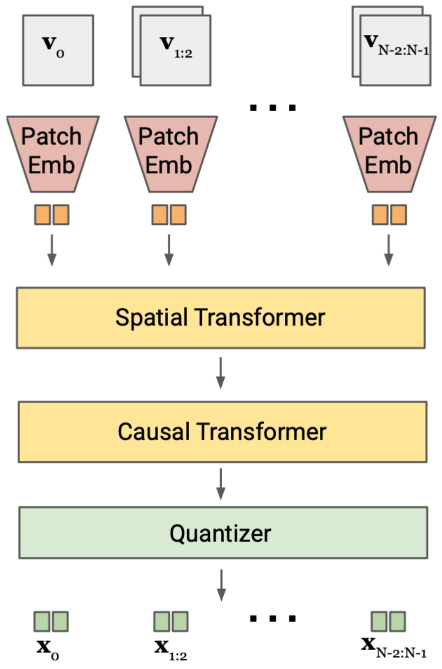
LFQ represents a family of methods in contrast to VQ

We will discuss in a simplest form:
independent codebook dimensions with binary latents

- The latent space is the Cartesian product of K single-dimensional variables $\mathbb{C} = \times_{i=1}^{\log_2 K} C_i$
- Each dimension takes two values: $C_i = \{-1, 1\}$
- With k dimensions, we have an effective vocabulary of 2^k
- An entropy penalty encourages codebook utilization $\mathcal{L}_{entropy} = \mathbb{E}[H(q(\mathbf{z}))] - H[\mathbb{E}(q(\mathbf{z}))]$ which can be factorized for efficient computation with large vocabularies
- Codebook loss is no longer applicable

Joint Image-Video Tokenization

- Utilizing large-scale labeled image data has been shown beneficial for video models
  - E.g., make-a-video, phenaki, etc.
- Native 3D CNNs in MAGVIT face challenges to tokenize single images due to temporal receptive field
- Existing solution: C-ViViT from phenaki ➡️
  - Hard to generalize to different spatial resolutions
  - Worse visual quality
  - Worse spatial causality of tokens

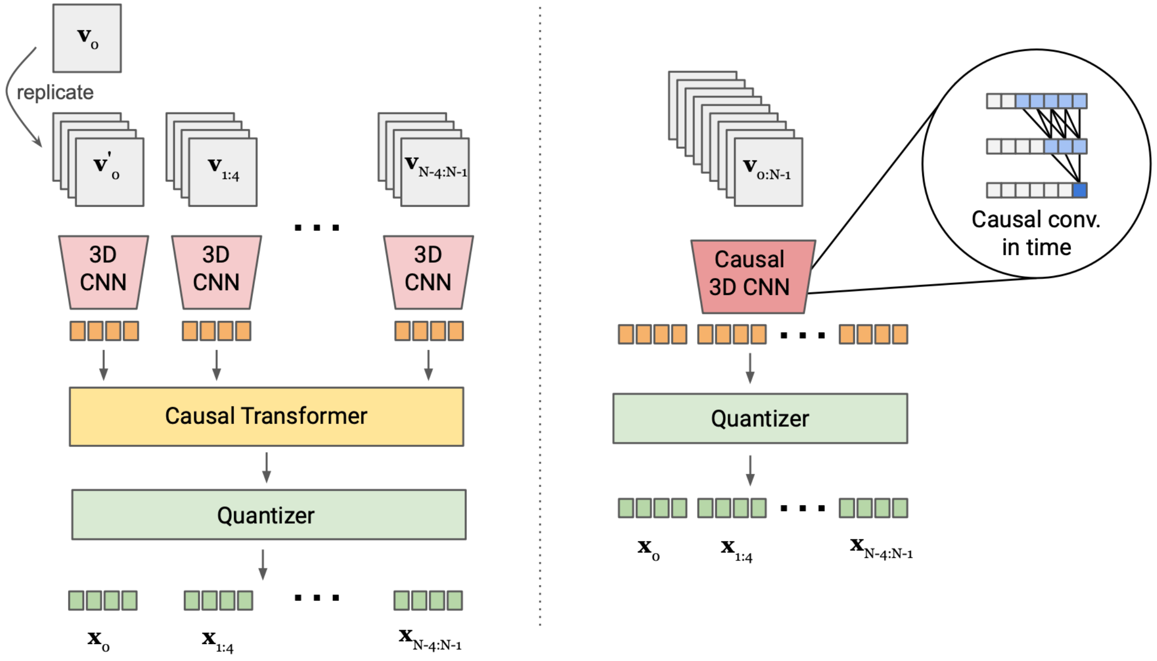


(a) C-ViViT

Joint Image-Video Tokenization

Exploring two new designs:

- Combining C-ViViT and MAGVIT
  - 3D CNNs replace the spatial transformer and process 4-frame blocks.
- **Temporally causal 3D CNN,**
  via custom convolution padding and upsampling
  - The first frame remains independent.
  - Allowing for videos of variable length.



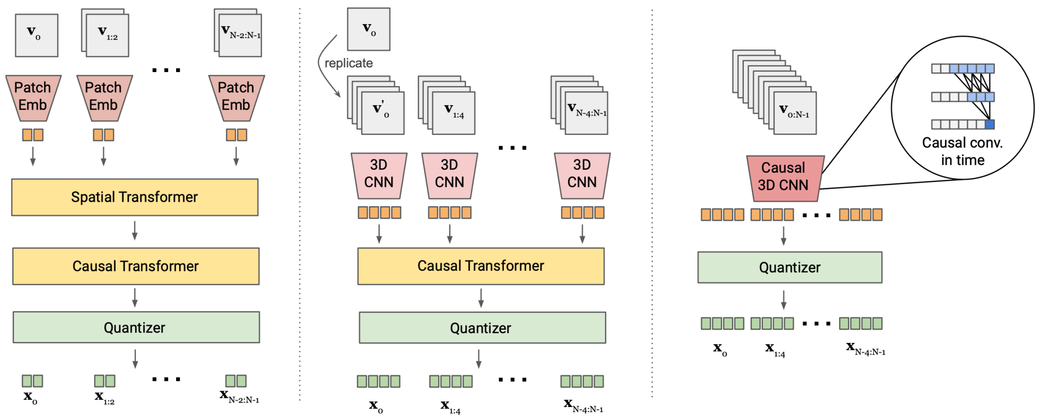
(b) C-ViViT + MAGVIT
(c) Causal 3D CNN

Joint Image-Video Tokenization

Comparing joint/causal tokenization architectures on UCF-101.

FID is calculated on the first frame.

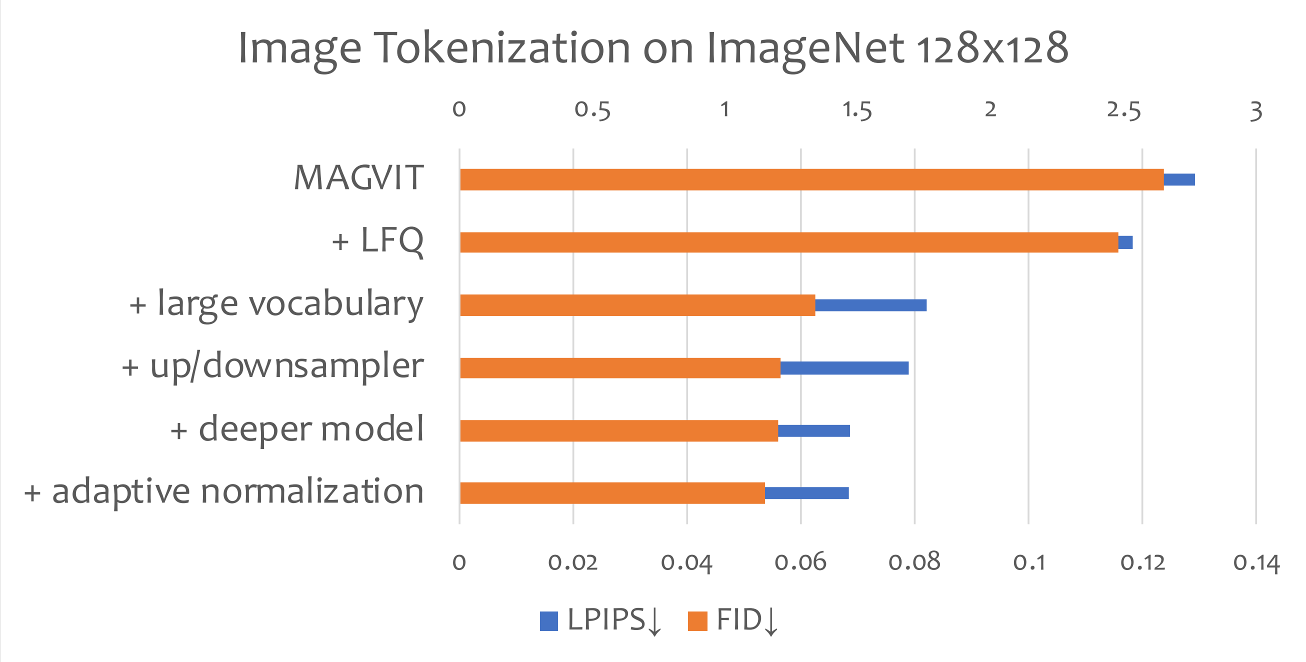
	# Params	FID↓	FVD↓
MAGVIT	39M	n/a	107.15
C-ViViT	90M	28.02	437.54
C-ViViT + MAGVIT	67M	13.52	316.70
MAGVITv2	58M	**7.06**	**96.33**



(a) C-ViViT (b) C-ViViT + MAGVIT (c) Causal 3D CNN

Architecture Ablations

- Quantizer: VQ → LFQ
- Large vocabulary: $2^{10} \rightarrow 2^{18}$
- Downsampler: average pooling → strided convolution
- Upsampler: resize + convolution → depth to space
- Temporal downsample: early → late
- Deeper: residual blocks 2x → 4x
- Decoder adaptive normalization like StyleGAN
- 3D blur pooling for shift invariance



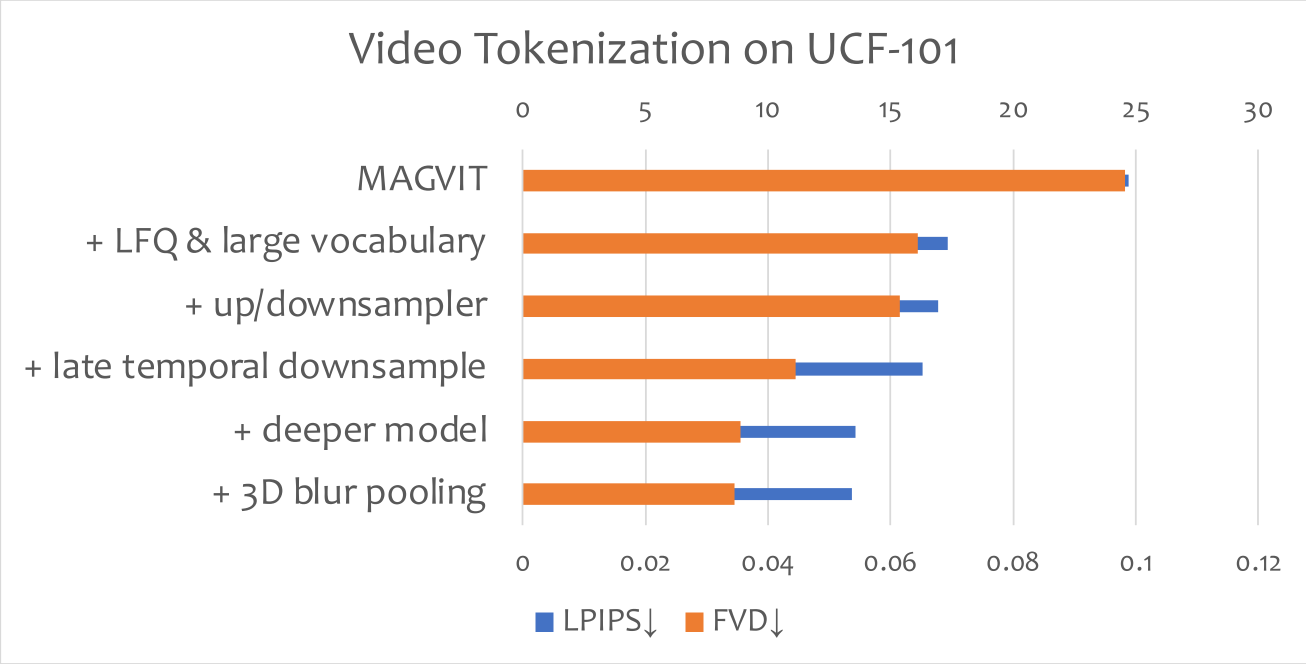
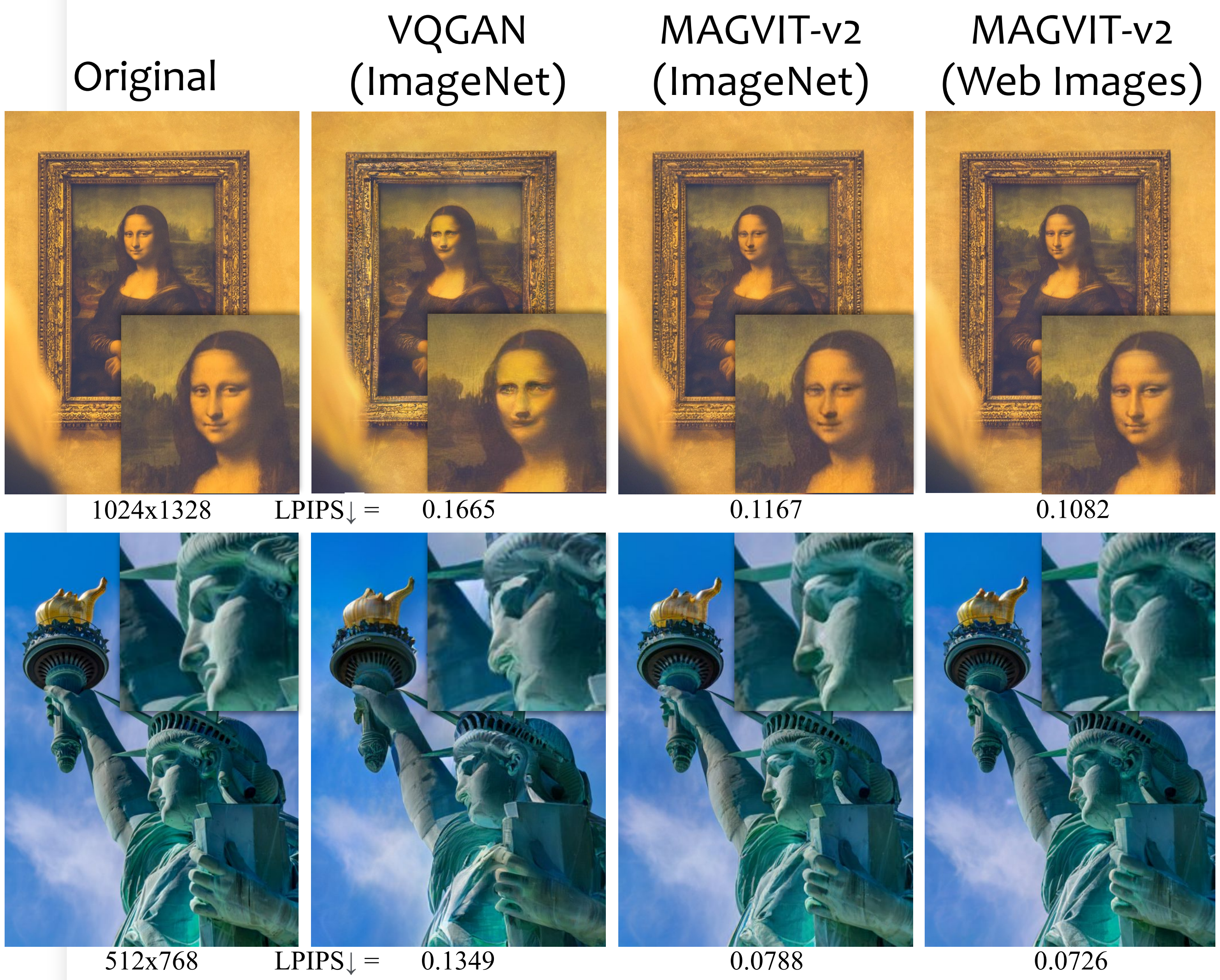


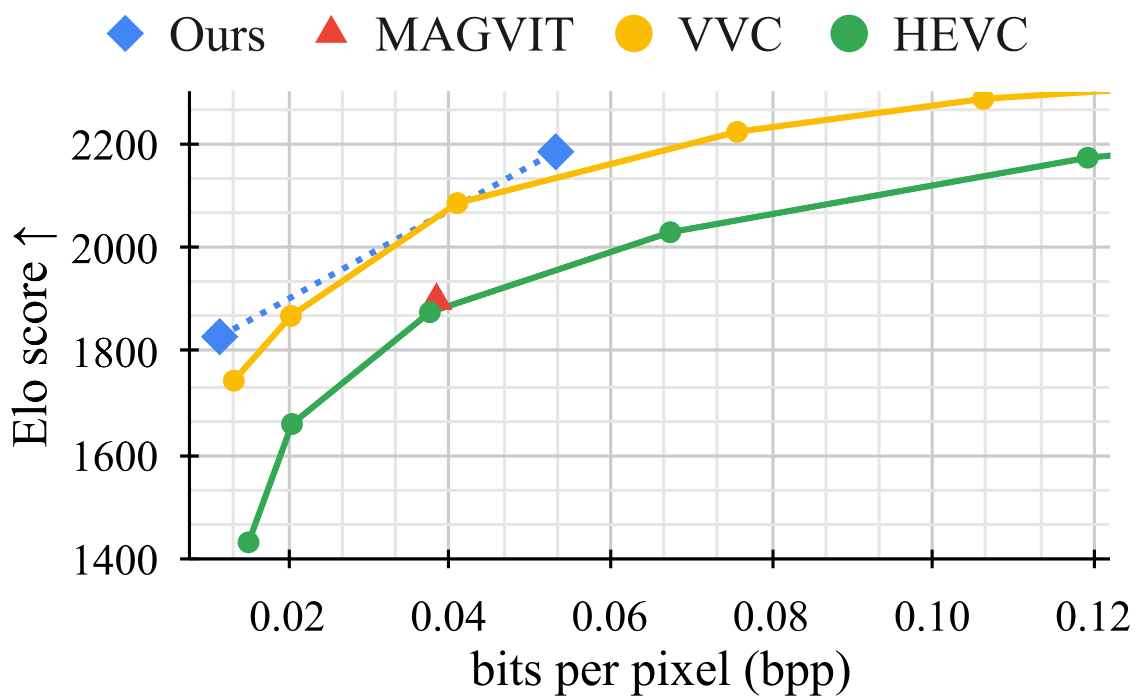
Image Reconstruction

- VQGAN fails to reconstruct facial details
- MAGVIT-v2 does a much better job when trained on the same dataset
  - Much larger vocabulary
  - More powerful decoder
- And it further scales to larger data



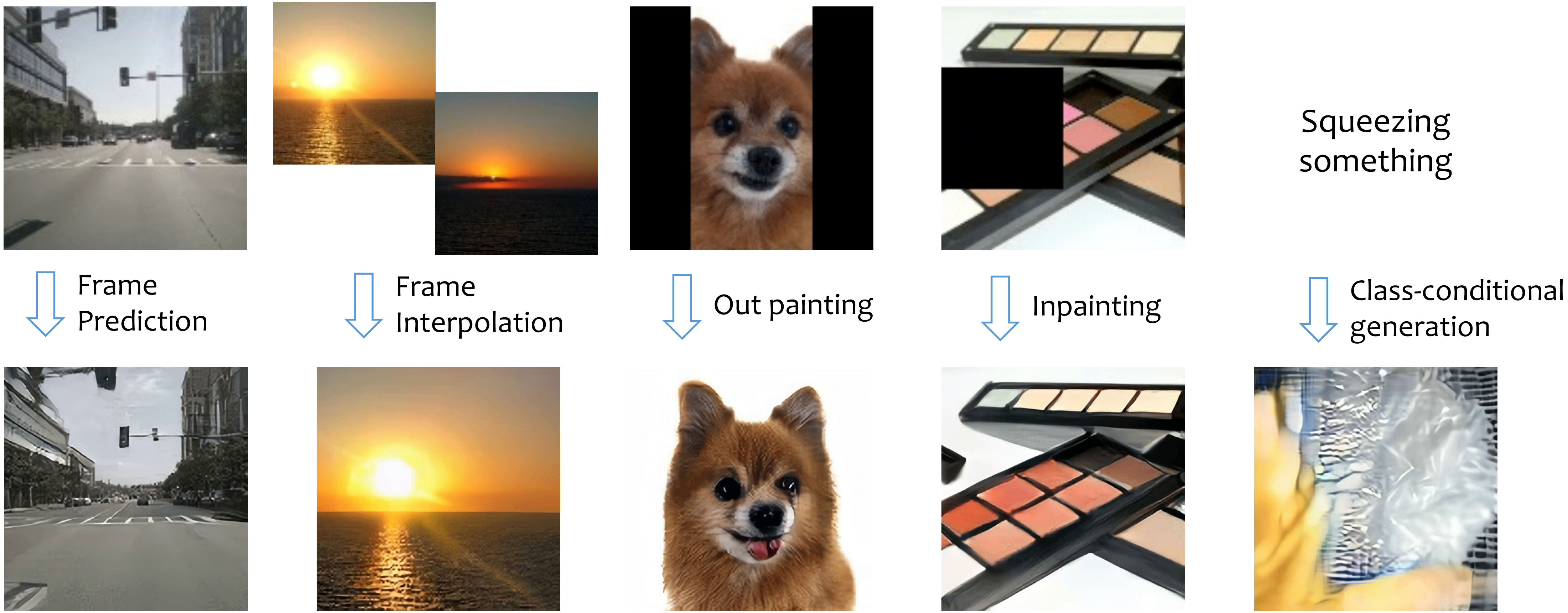
Video Compression

MAGVIT-v2 is preferred over MAGVIT, HEVC (H.265), and VVC (H.266) in subjective rater study.

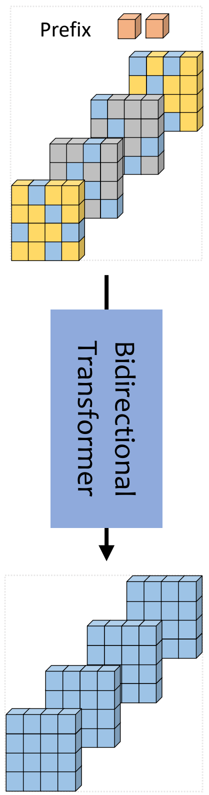


MAGVIT: Masked Generative Video Transformer

The first multi-task masked transformer for video generation,
with state-of-the-art generation quality and efficiency.



Masked Video Synthesis

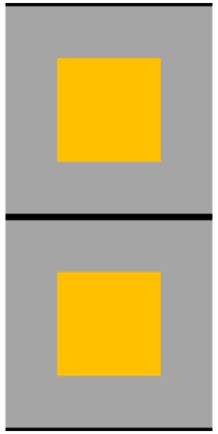


Input video

Generated videos

Here we decode intermediate states for visualization, which does not happen in standard sampling.

Sampling progress →



Masked tokens at each step:
Mask Condition Prediction

- Initial state includes the condition
- Keep some generated tokens at each step
- All output tokens are predicted by the model, including condition

Token Factorization

- Leveraging the independence property of LFQ
- Helpful for smaller transformers predicting in a large vocabulary
  - An MAGVIT-L model has 305M parameters with a 2^{10} vocabulary, but an embedding matrix with 2^{18} entries takes 270M parameters
- E.g., a 2^{18} vocabulary can be factorized into two predictions of 2^9
- Without changing the total sequence length
  - Embedding summation as input
  - Multi-head prediction for output
- Empirically, it also makes the sampling more accurate

Image Generation

- The first evidence suggesting that a language model can outperform diffusion models on ImageNet.
  - In both sampling quality (FID, IS) and inference-time efficiency (sampling steps)
  - Using the same training data, a comparable model size, and a similar training budget.
- Notably, MAGVIT-v2 uses 16×16 latents, much smaller than others

Class-conditional generation on ImageNet 512×512

Type	Method	FID↓	Guided FID↓	# Params	# Steps	Latent
Diff. + VAE*	DiT-XL/2	12.03	3.04	675M	250	64^2
Diffusion	RIN	3.95		320M	1000	
Diffusion	VDM++	2.99	2.65	2B	512	
MLM + VQ	MaskGIT	7.32		227M	12	32^2
MLM + VQ	DPC	3.62		619M	72	32^2
MLM + LFQ	MAGVIT-v2	4.61		307M	12	**16^2**
		3.07	**1.91**		64	

Image Generation

- The first evidence suggesting that a language model can outperform diffusion models on ImageNet.
  - The margin narrows at 256×256 but MLM uses a 50% smaller model and much fewer steps
  - VAE* uses large-scale training data while others are only on ImageNet

Class-conditional generation on ImageNet 256×256

Type	Method	FID↓	Guided FID↓	# Params	# Steps
Diffusion + VAE*	MDT	6.23	1.79	676M	250
Diffusion	RIN	3.42		410M	1000
Diffusion	VDM++	2.40	2.12	2B	512
MLM + VQ	Contextual RQ	3.41		1.4B	72
MLM + VQ	DPC	4.45		454M	180
MLM + LFQ	MAGVIT-v2	3.65	**1.78**	307M	64

Video Generation

- MAGVIT-v2 surpasses all prior arts
- MAGVIT-v2 significantly outperforms MAGVIT
  - Using the same MLM backbone and decoding procedure
  - Highlighting the importance of a good tokenizer

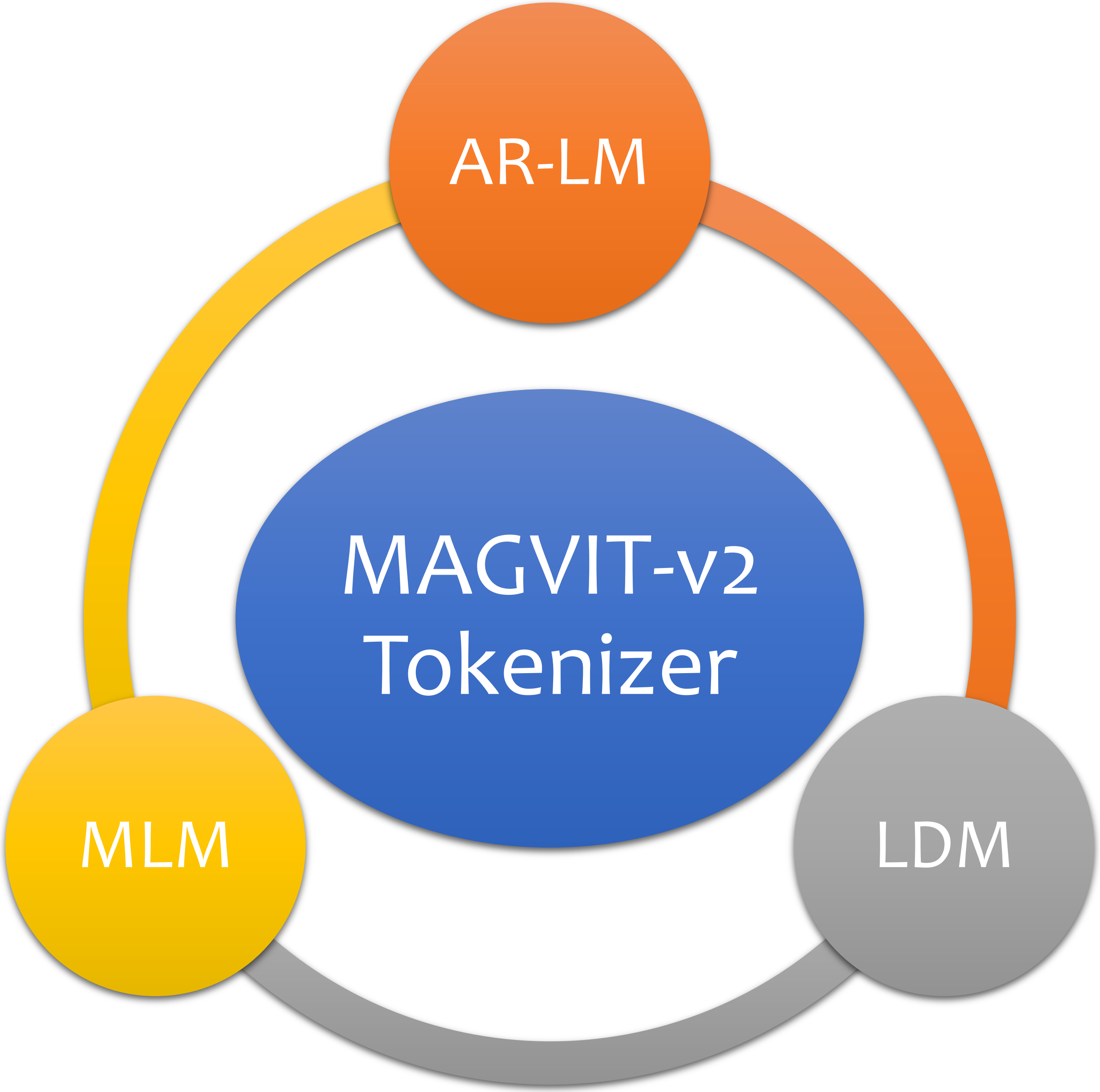
Video Generation: Frame prediction on Kinetics-600 and class-conditional generation on UCF-101

Type	Method	K600 FVD↓	UCF FVD↓	# Params	# Steps
Diffusion	VDM	16.2±0.3		1.1B	256
Diffusion	RIN	10.8		411M	1000
MLM + VQ	MAGVIT	9.9±0.3	76±2	306M	12
MLM + LFQ	MAGVIT-v2	5.2±0.2		307M	12
		4.3±0.1	**58±3**		24

Video Generation

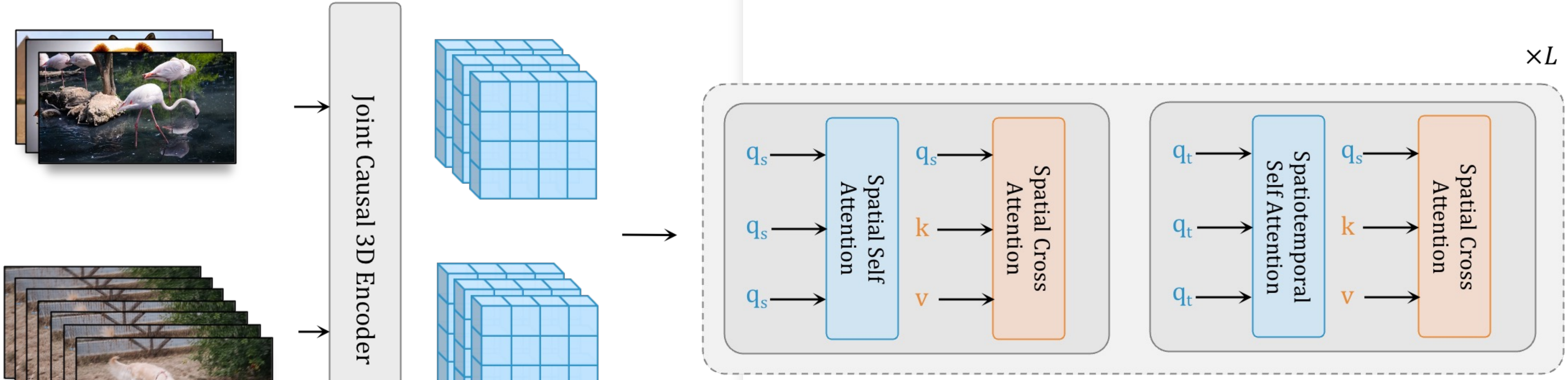
MAGVIT-v2 enables remarkable video generation quality with transformers using various objectives

- MLM – shown so far
- AR-LM – VideoPoet
- LDM – W.A.L.T



W.A.L.T

- Windowed Attention Latent Transformer
- Joint image-video latent from MAGVIT-v2
- Joint diffusion training on image and video

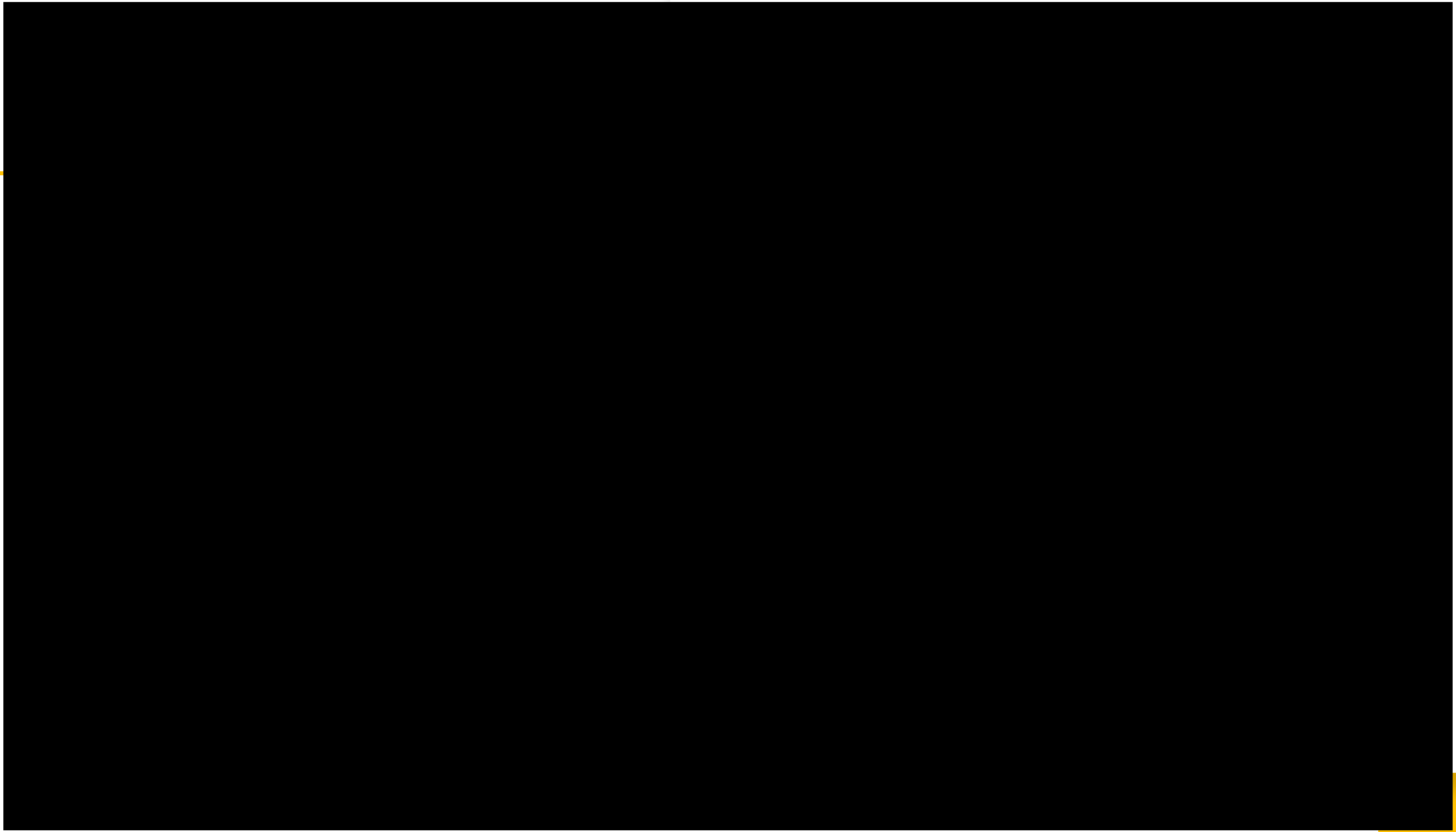


$q_s \in \mathbb{R}^{(B.(1+t))\times 1\times h_p\times w_p\times C}$

$q_t \in \mathbb{R}^{(B.W_n)\times(1+t)\times h'_p\times w'_p\times c}$

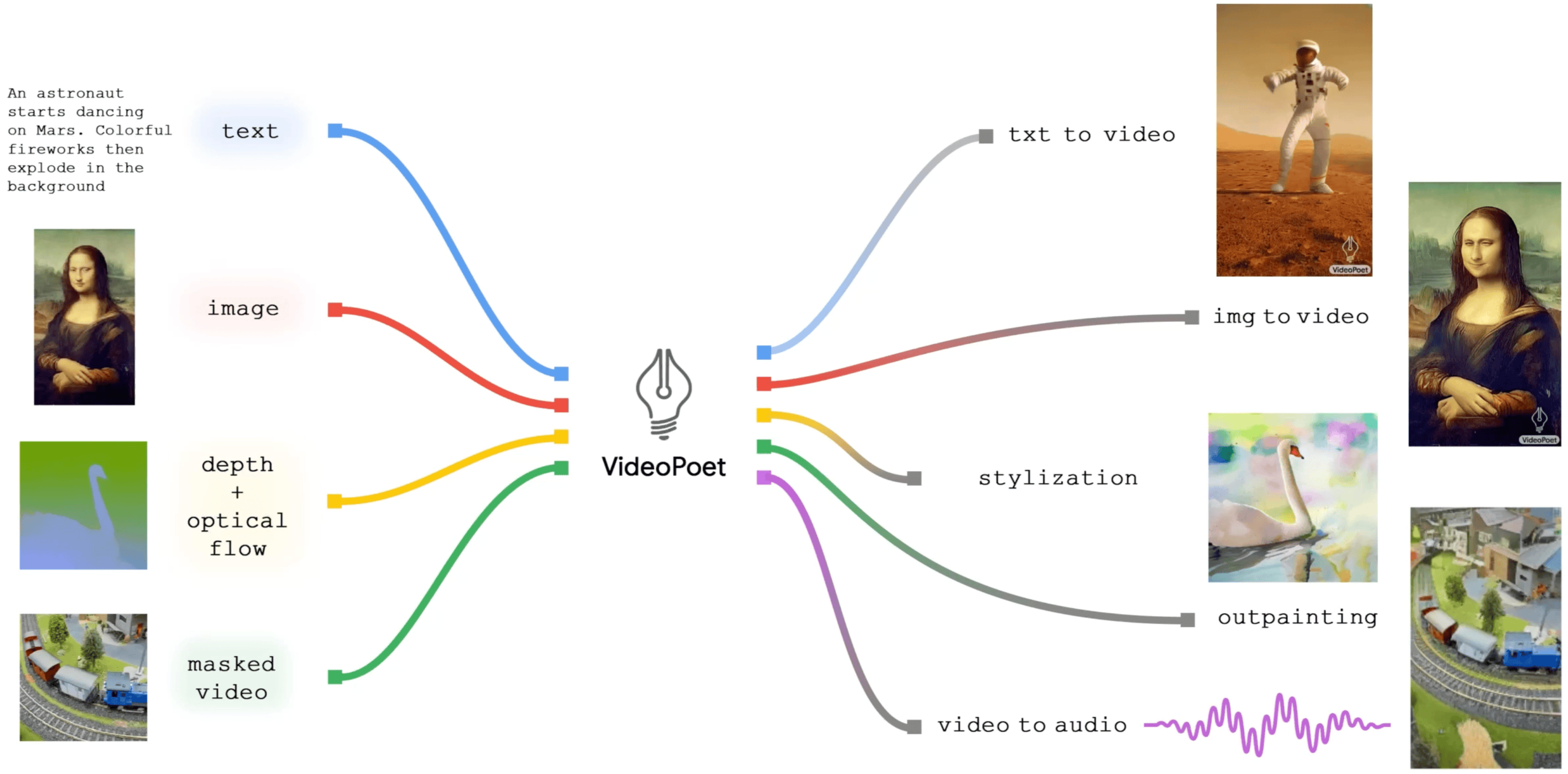
$k = v = \text{concat}[q_s, \text{t5\_xl (text)}]$

W.A.L.T



VideoPoet:

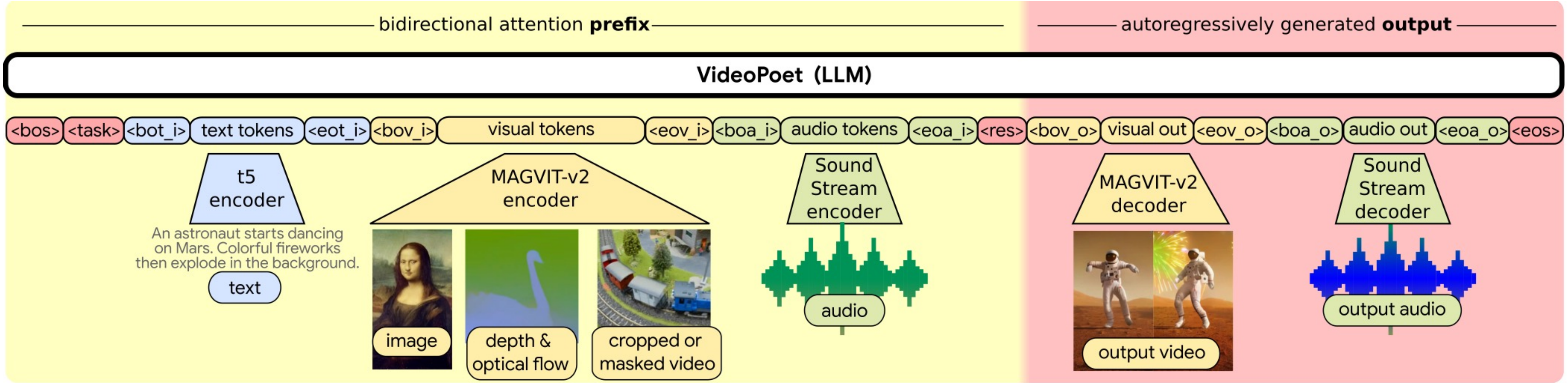
A Large Language Model for Zero-Shot Video Generation



- Synthesis of high-quality video with matching audio, from a large variety of condition signals
- Highlight: high fidelity motion
- A purely token-based approach, without diffusion

VideoPoet

- Multi-modal multi-task in a unified sequence model
  - **MAGVIT-v2 tokenizer** for image/video tokenization
  - SoundStream tokenizer for audio tokenization
  - T5 for text embedding

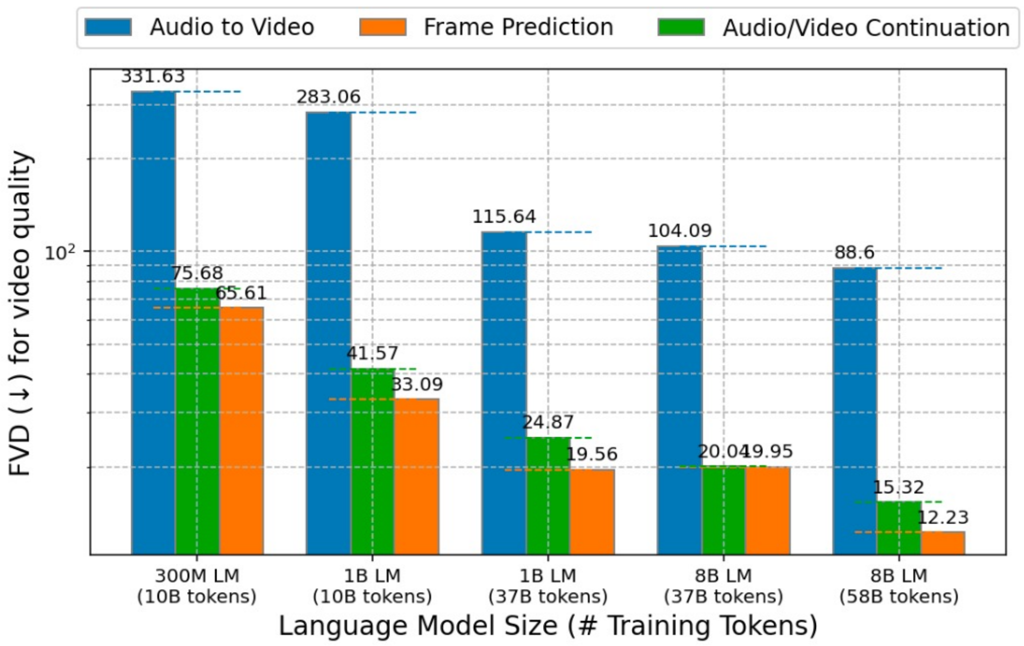


Training

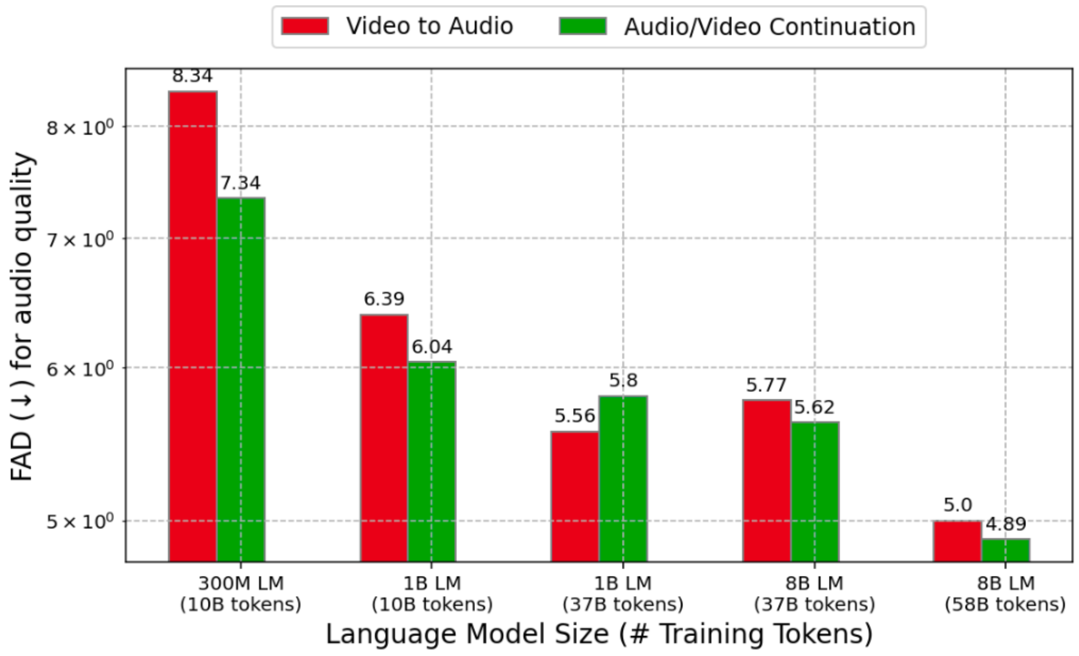
- Prefix LM with UL2-style objective
- A large mixture of tasks with different datasets in a single model
  - Unconditional generation / text-to-image/video
  - Video continuation / image-to-video
  - Video inpainting / outpainting / stylization
  - Video-to-audio / audio-to-video / audio-visual continuation
- Notably, we design sequence formats for easier transfer of capabilities, e.g., text-to-image becomes a prefix of text-to-video

Preliminary Scaling

- Model: 300M, 1B, 8B parameters
- Data: 10B, 37B, 58B tokens



Video Generation

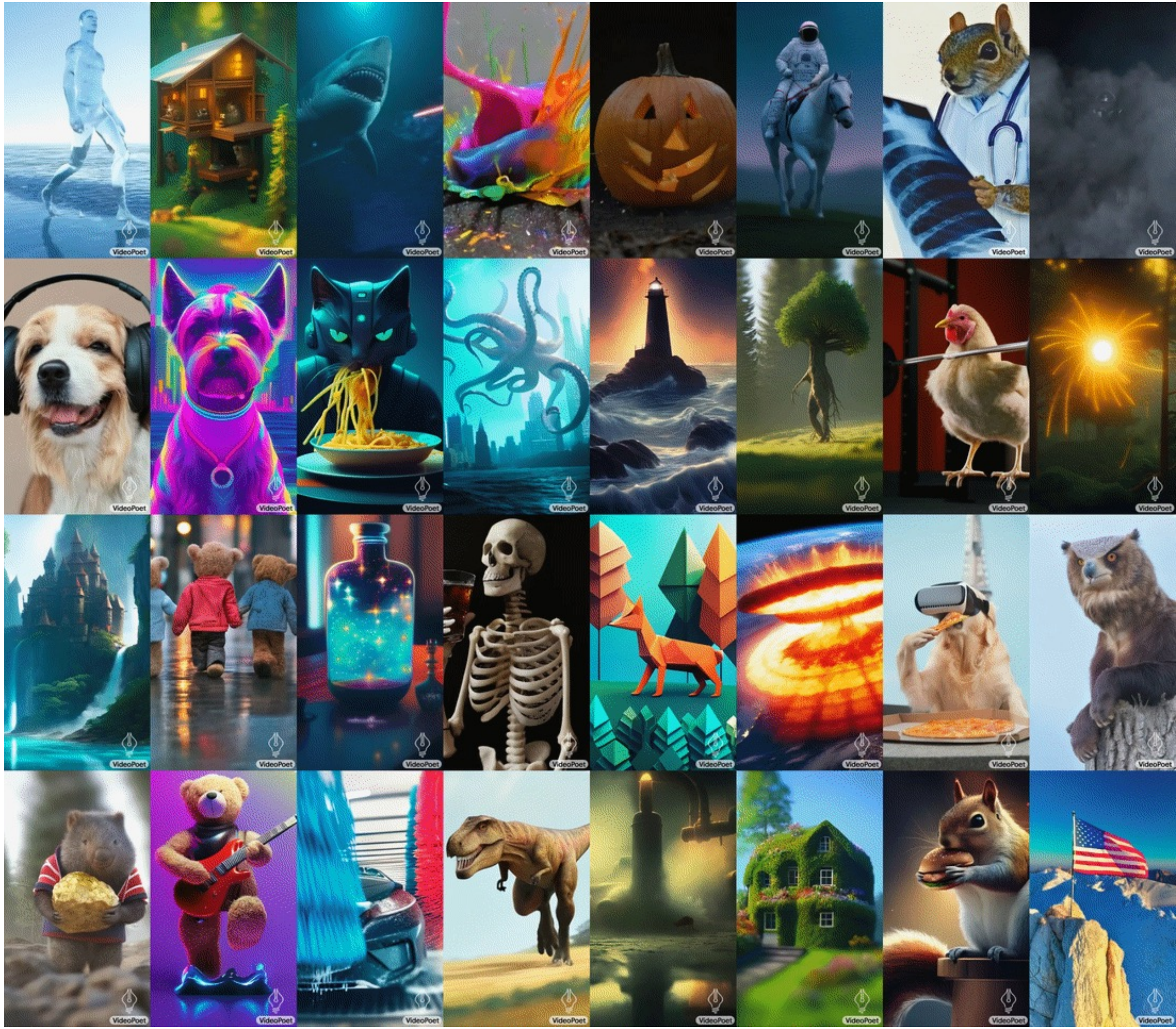


Audio Generation

Text-to-Video

Zero-shot text-to-video evaluation
on MSR-VTT

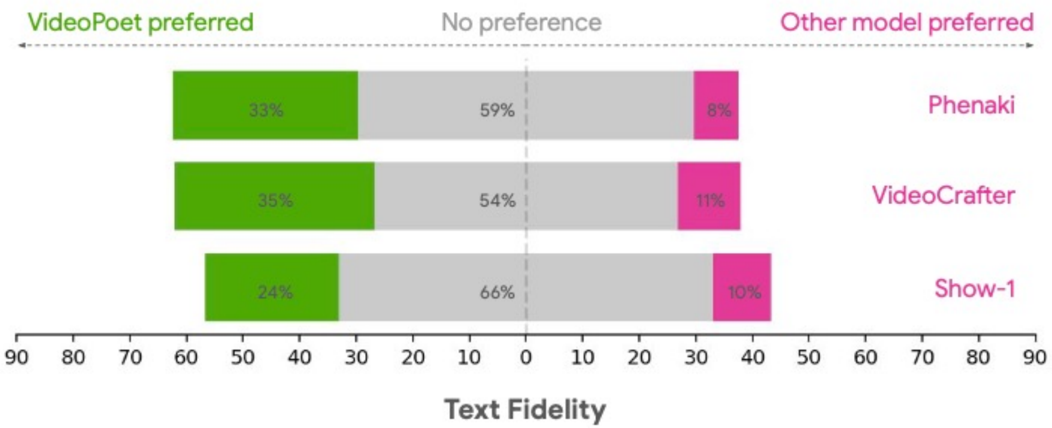
Model	CLIPSIM	FVD
Video LDM	0.2929	
Make-A-Video	0.3049	-
Show-1	0.3072	538
VideoPoet (pretrain)	0.3049	**213**
VideoPoet (task adapt)	**0.3123**	-

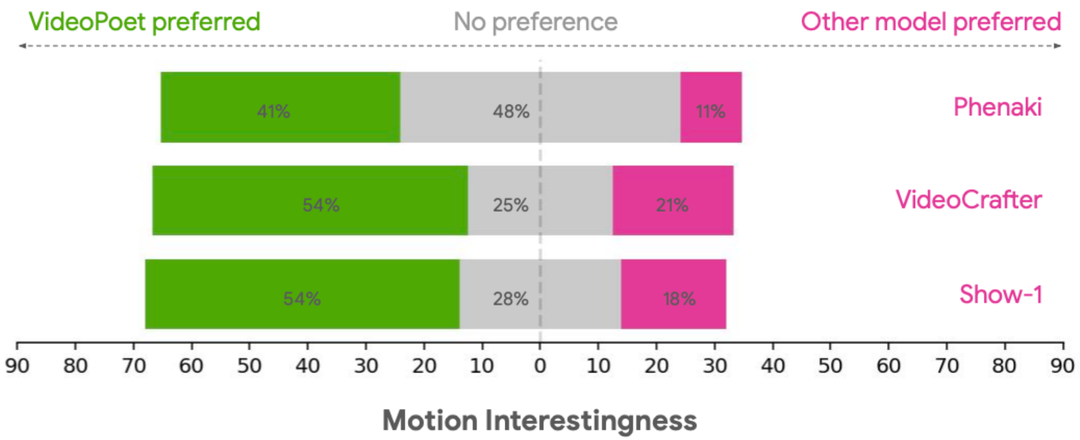


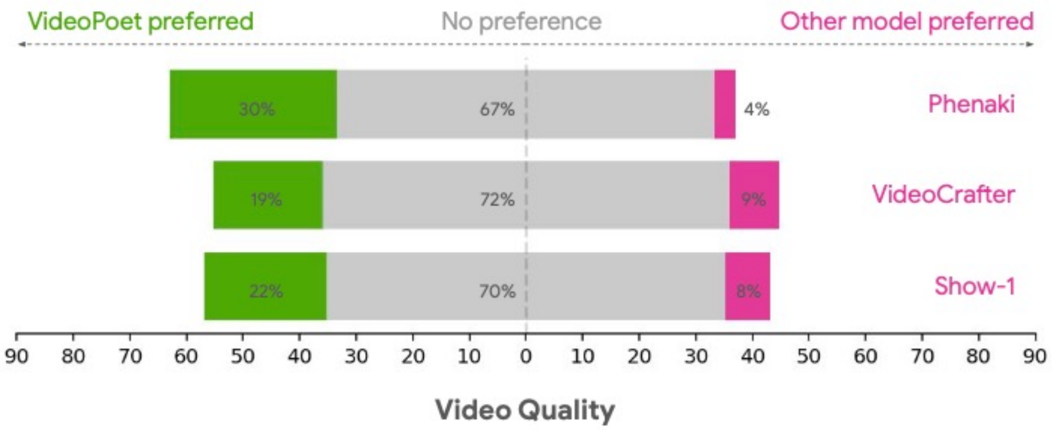
See more: https://sites.research.google/videopoet

Text-to-Video

Human Evaluation







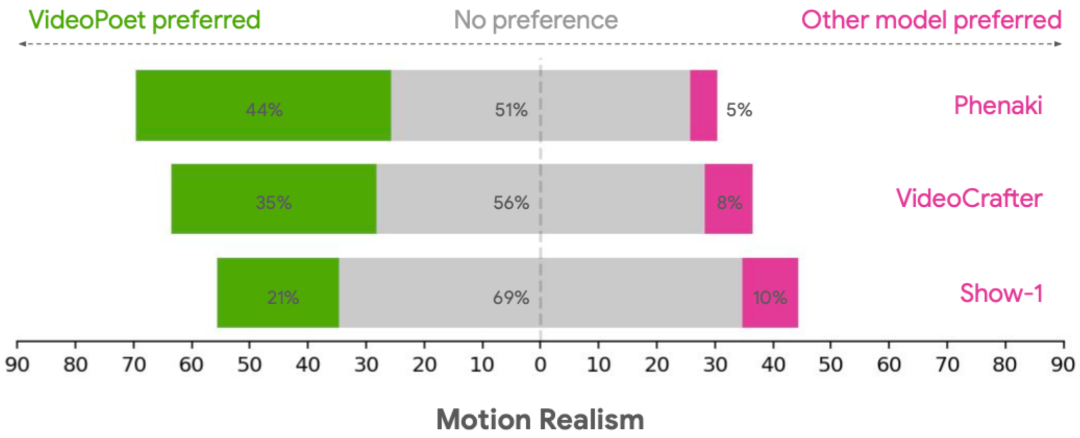
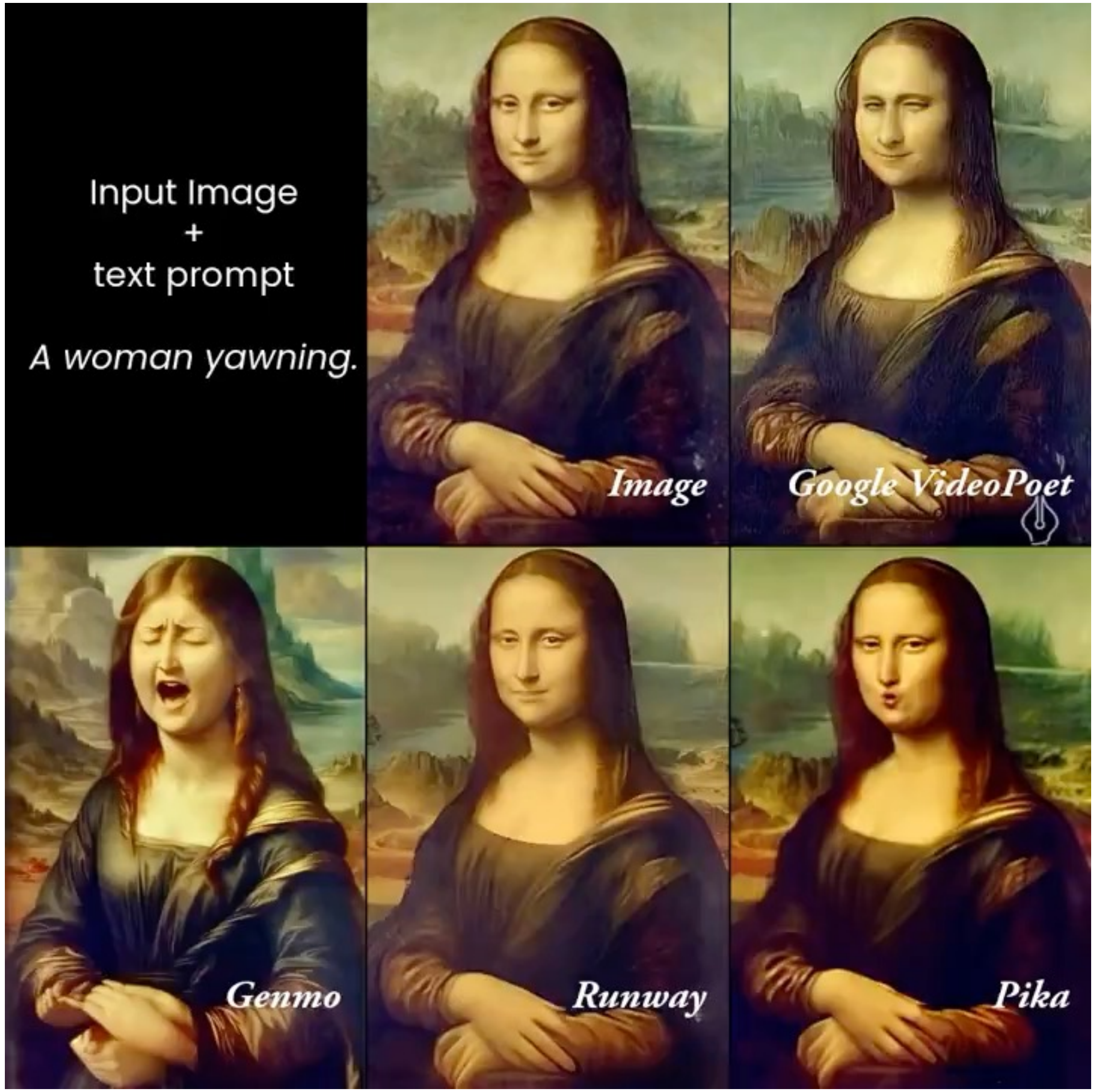
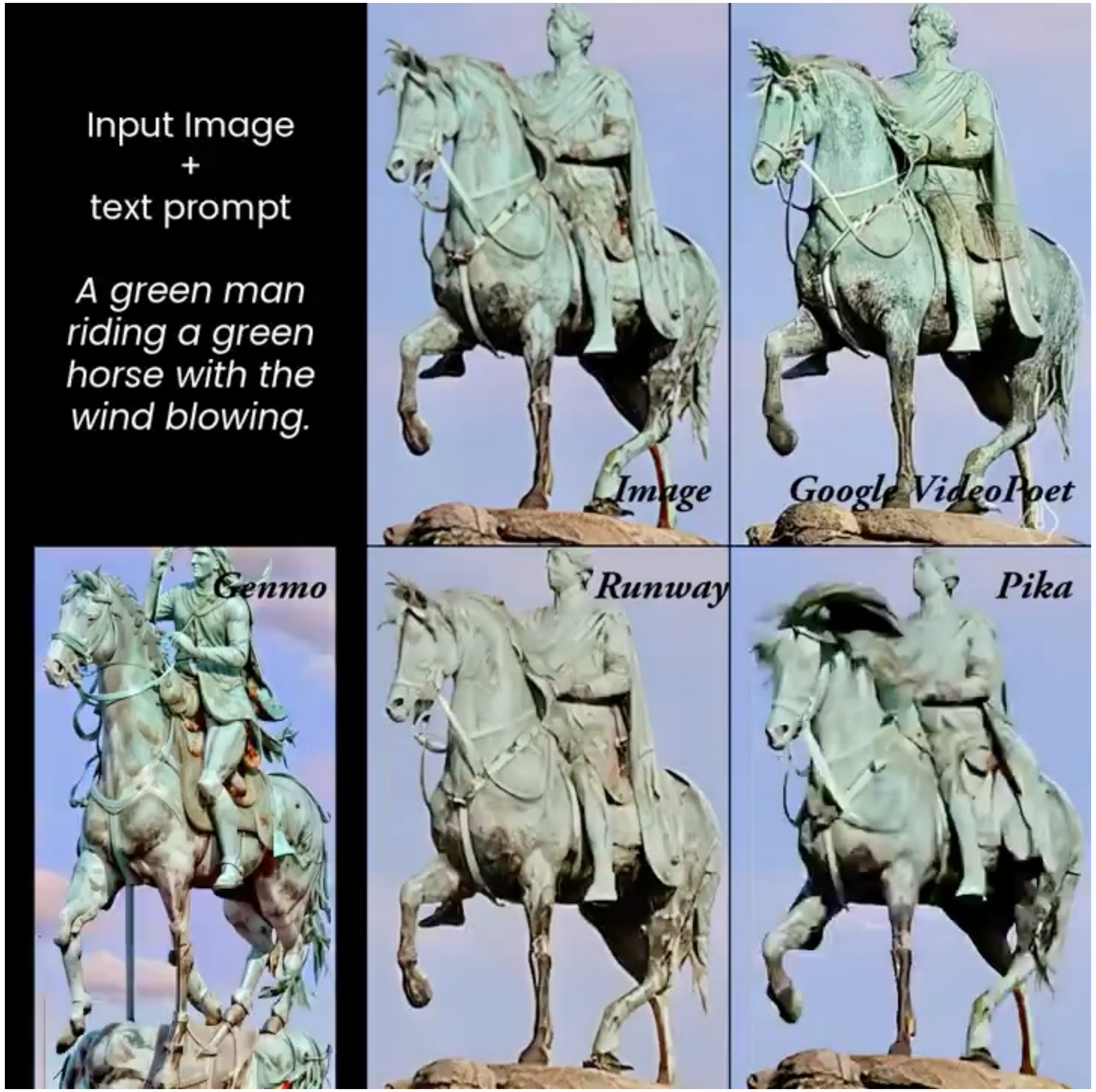


Image-to-Video

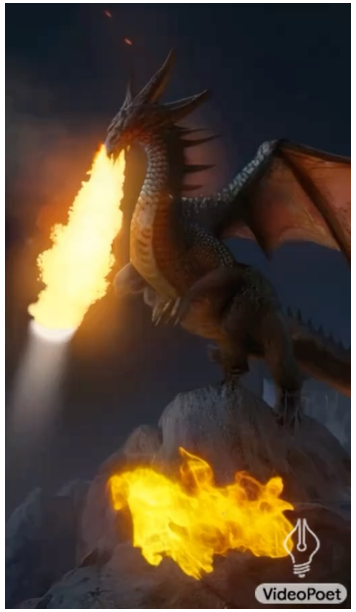


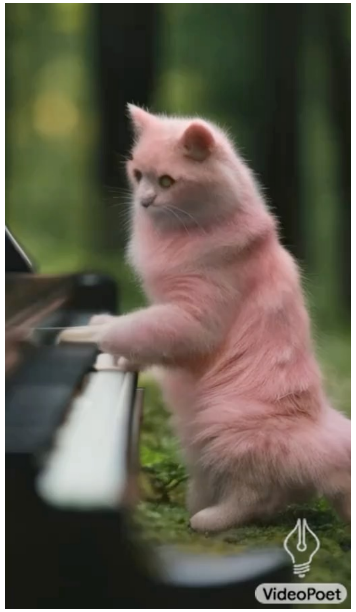


Source: https://twitter.com/AIsonesone

Video-to-Audio

On generated videos





Takeaways

We have covered a lot:

MAGVIT, SPAE, MAGVIT-v2, W.A.L.T, VideoPoet, …

One point that may be noteworthy:

Language models are at least as good as diffusion models on visual synthesis, if a good tokenizer is available.

And language models are much more general and scalable.

Just generating good-looking videos will never be our end goal.

We should build large multi-modal-native models that learn from raw signals and unveils the "truth of the universe" beyond human knowledge as Artificial Super Intelligence.

Given the evidence so far, how should we move towards it?

BTW, maybe intelligence is really all about compression?

Thank you!