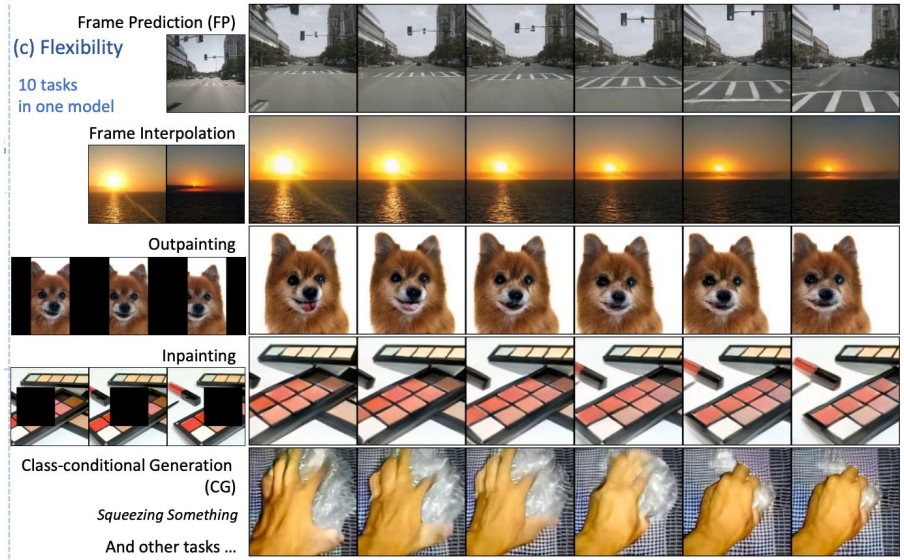
MAGVIT: Masked Generative Video Transformer

Lijun Yu[‡†◇*], Yong Cheng[†], Kihyuk Sohn[†], José Lezama[†], Han Zhang[†], Huiwen Chang[†], Alexander G. Hauptmann[‡], Ming-Hsuan Yang[†], Yuan Hao[†], Irfan Essa[†ǂ], and Lu Jiang[†◇]
[‡]Carnegie Mellon University, [†]Google Research, [ǂ]Georgia Institute of Technology



San Diego Machine Learning
Ryan Chesler

Overview

- Many models have gotten really great at generating images and various different modalities

- Is it possible to train good video models in a similar way?
  - Difficult because of compute
  - 24 frames of video = generating 24 images at the same time
  - Need continuity between frames
- Popular techniques like diffusion, GANs, flow-matching, etc.
  - Quality
  - Memory/Compute
  - Speed

MAGVIT: Two-Stage Approach

- Stage 1: 3D vector-quantized autoencoder
  - Train a model that compresses and decompresses videos really well
  - Need a very compressed in the middle that we can use in the second stage
  - Reducing down to discrete tokens makes it possible to re-use all the transformer innovations
- Stage 2: Transformer
  - Train a masked token model to fill in blanks from the previous stage
  - Can use this to extrapolate or interpolate in many creative ways and then decode with stage 1

Stage 1: 3D Vector-Quantized Autoencoder

- Autoencoder model that heavily compresses and decompresses video
- Bottleneck is discrete tokens, not continuous values
- Compresses 8x on height and width and 4x on frames with a vocabulary size of 1024
  - >256x Compression!!!
- Able to compress very heavily because frames are largely related
- JPEG vs MP4

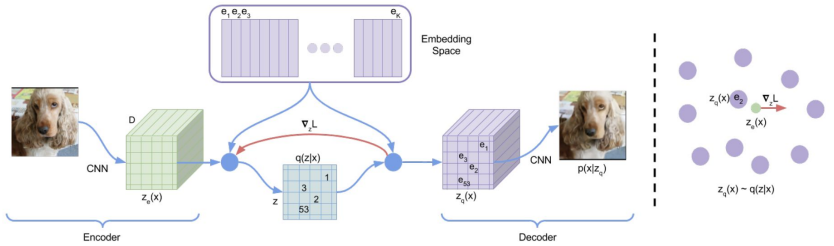
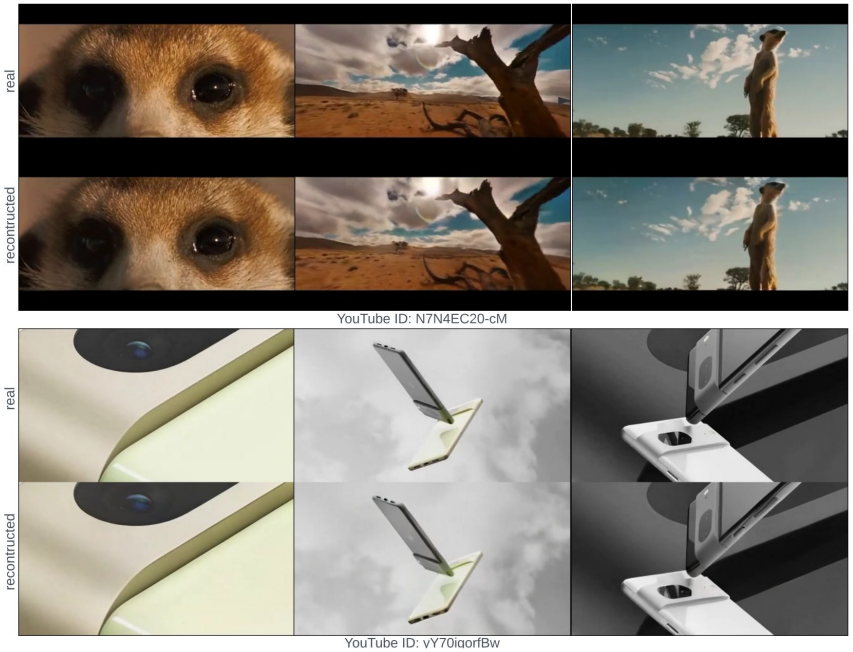


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

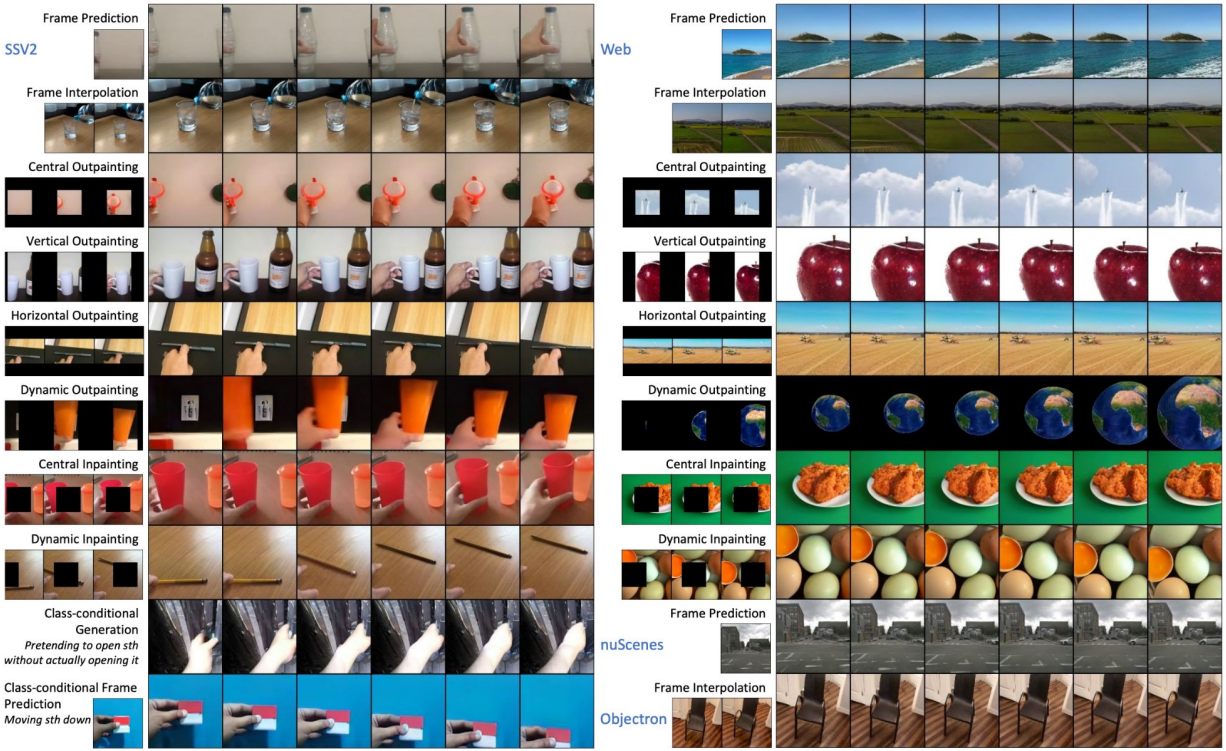
Stage 1: 3D Vector-Quantized Autoencoder

- Autoencoder model that heavily compresses and decompresses video
- Bottleneck is discrete tokens, not continuous values
- Compresses 8x on height and width and 4x on frames with a vocabulary size of 1024
  - >256x Compression!!!
- Able to compress very heavily because frames are largely related
- JPEG vs MP4

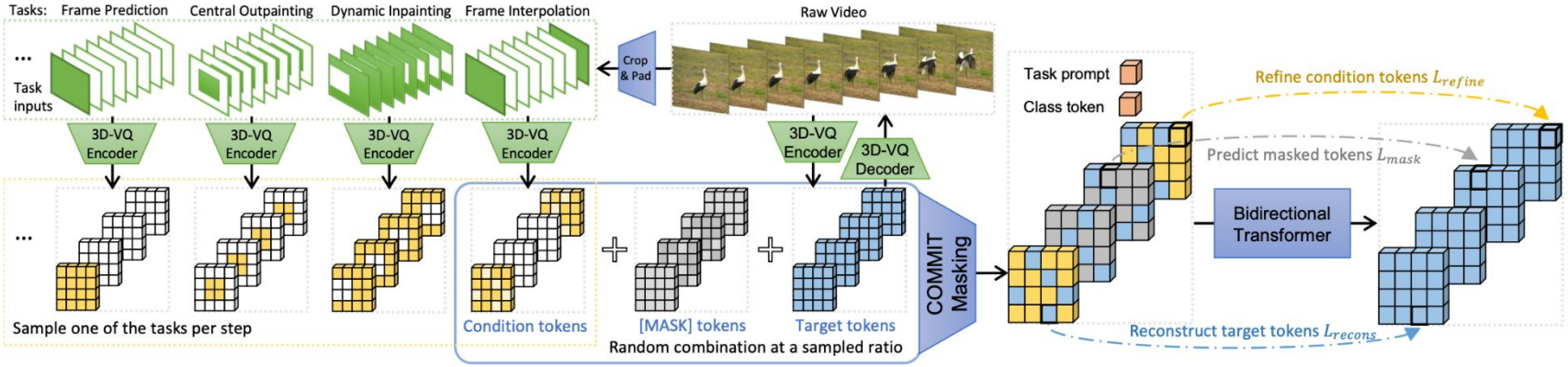


Stage 2:Masked Token Modeling

With compressed representation guess tokens in various different ways

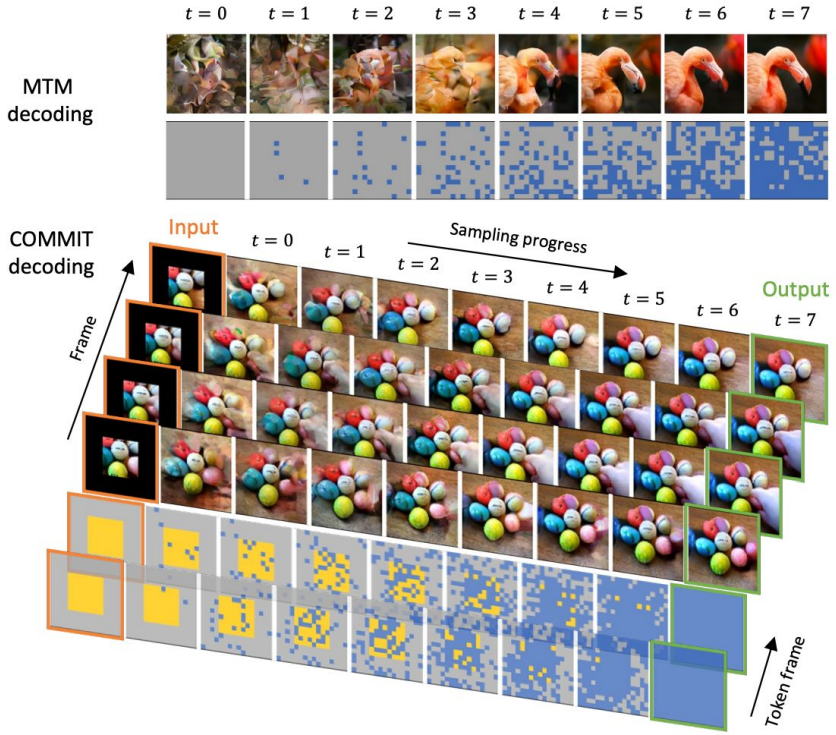


Full-pipeline



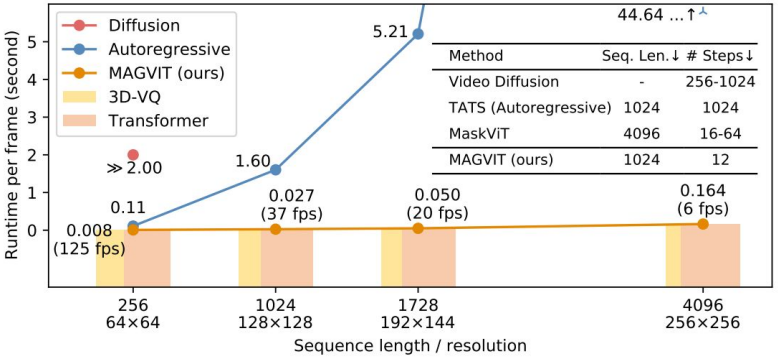
Inference

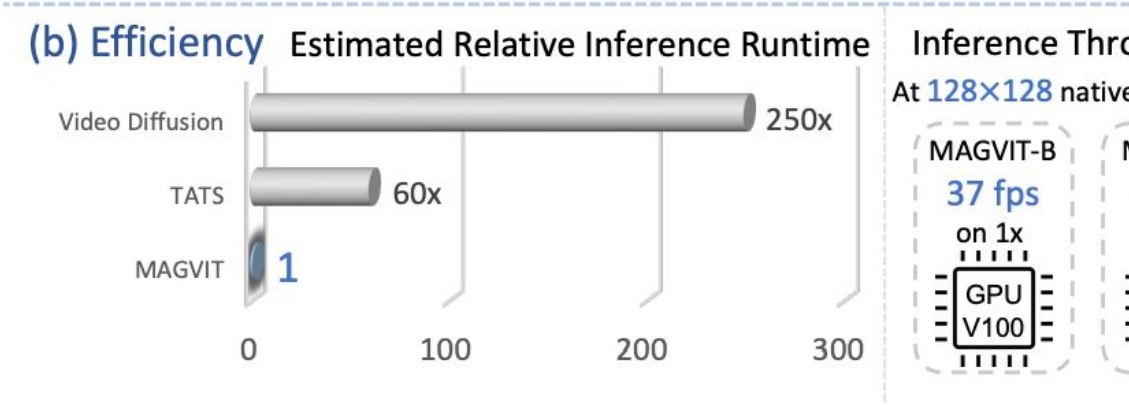
- Second stage decodes in steps similar to diffusion
- Previous approaches have worked autoregressively, much slower



Inference

- Second stage decodes in steps similar to diffusion
- Previous attempts have been autoregressive and much slower





Performance

Method	K600 FVD↓	BAIR FVD↓
CogVideo [34]	109.2	-
CCVS [41]	55.0±1.0	99±2
Phenaki [63]	36.4±0.2	97
TrIVD-GAN-FP [43]	25.7±0.7	103
Transframer [44]	25.4	100
MaskViT [26]	-	94
FitVid [4]	-	94
MCVD [64]	-	90
NÜWA [69]	-	87
RaMViD [35]	16.5	84
Video Diffusion [33]	16.2±0.3	-
MAGVIT-B-FP (ours)	24.5±0.9	76±0.1 (48±0.1)
MAGVIT-L-FP (ours)	**9.9**±0.3	**62**±0.1 (31±0.2)

Table 2. **Frame prediction performance on the BAIR and Kinetics-600 datasets.** - marks that the value is unavailable in their paper or incomparable to others. The FVD in parentheses uses a debiased evaluation protocol on BAIR detailed in Appendix B.3. See Appendix C for more comparisons with earlier works.

Decoding Method	Tokenizer	Type	Param.	Seq. Len.↓	# Steps↓	FVD↓
MaskGIT [12]	2D-VQ	NAR	53M+87M	4096	12	222 (177)
	3D-VQ	NAR	41M+87M	1024	12	122 (74)
MaskViT [26]	2D-VQ	NAR	53M+189M	4096	18	94*
AR	3D-VQ	AR	41M+87M	1024	1024	91 (56)
MAGVIT (ours)	3D-VQ	NAR	41M+87M	1024	12	**76** (**48**)

Table 6. **Comparison of decoding methods** on BAIR frame prediction benchmark. The number of parameters is broken down as VQ + Transformer. NAR is non-autoregressive and AR is autoregressive. FVD and debiased FVD (in parentheses) are reported. * marks the quoted number from their paper.

Performance

Tokenizer	From Scratch FVD↓	From Scratch IS↑	ImageNet [16] Initialization FVD↓	ImageNet [16] Initialization IS↑		
MaskGIT [12] 2D-VQ	240	80.9	216	82.6	-	
TATS [21] 3D-VQ	162	80.6	-		-	
			Average		Central	
MAGVIT 3D-VQ-B (ours)	127	82.1	103	84.8	58	87.0
MAGVIT 3D-VQ-L (ours)	45	87.1	35	88.3	**25**	**88.9**

Table 7. **Comparison of tokenizer architectures and initialization methods** on UCF-101 training set reconstruction results. The 2D-VQ compresses by 8×8 spatially and the 3D-VQ compresses by 4×8×8 spatial-temporally.

Method	Task	**BAIR-MT8↓**	FP	FI	OPC	OPV	OPH	OPD	IPC	IPD	**SSV2-MT10↓**	CG	CFP
MAGVIT-B-UNC	Single	150.6	74.0	71.4	119.0	46.7	55.9	389.3	145.0	303.2	258.8	107.7	279.0
MAGVIT-B-FP	Single	201.1	47.7	56.2	247.1	118.5	142.7	366.3	357.3	272.7	402.9	1780.0	59.3
MAGVIT-B-*MT*	Multi	32.8	47.2	36.0	28.1	29.0	27.8	32.1	31.1	31.0	43.4	94.7	59.3
MAGVIT-L-*MT*	Multi	**22.8**	31.4	26.4	21.3	21.2	19.5	20.9	21.3	20.3	**27.3**	79.1	28.5