LANGUAGE MODEL BEATS DIFFUSION — TOKENIZER IS KEY TO VISUAL GENERATION

- アブスト
  - MAGVIT-v2と呼ばれる動画、画像両方に対する一般的なトークンの辞書を用いた動画のTokenizerを提案
  - このTokenizerによって、LMが拡散モデルに対してImageNetやKineticsなどの画像や動画生成のタスクに対して凌駕することを確認
  - その他にもVVCに匹敵する動画圧縮性能、行動認識タスクのための良い表現が獲得できていることを確認した
- イントロ
  - LLMは動画や画像を生成することができる
    - 画像のピクセルはvisual tokenizerによって離散トークンに圧縮される
    - その後、Transformerに入力され、言語のように扱われ、生成タスクとして処理される
    - 拡散モデルには生成タスクにおいて劣っていた
      - ImageNetにおける画像生成タスクでは、LLM系のSOTAモデルはSOTAの拡散モデルに比べて50%近くの性能しか出ていない
        - 本研究では、visual tokenizerが良い表現が獲得できていないことを理由としてあげている
        - 提案手法によって、同じデータ数、モデルサイズ、計算資源の条件下で、ImageNetにおける生成タスクでLMが拡散モデルを凌駕することを確認
  - 動画と画像を離散トークンに圧縮するMAGVIT-v2を提案
    - VQVAEを使用しているMAGVITを改良

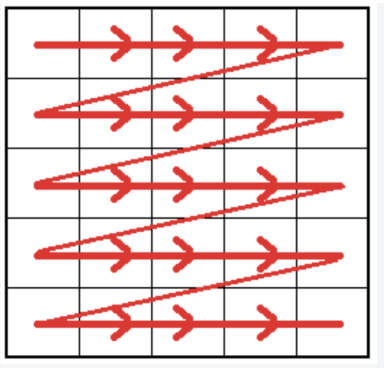
        - lookup-free quantization method
            - 辞書において多くの単語を学習することができる
        - tokenizerを修正することで、同じ辞書で画像と動画両方のトークンを処理することができる
    - 入力：動画（T=1だと画像）

$$\mathbf{V} \in \{\mathbb{R}^{T \times H \times W \times 3}\}$$

    - 離散トークン

$$\mathbf{X} = f(\mathbf{V}) \in \{1, 2, \cdots, K\}^{T' \times H' \times W'}$$

        - Kはvisual tokenizerのcodebook(vocabulary) size
        - Xはラスター走査によって１次元になり、Transformerに入力される



- LM for visual generation
    - the autoregressive LM
        - 前のトークンから次のトークンを予測する
    - the masked LM
        - 一部のトークンをマスクして、マスクしていないトークンから予測する
- Denoising Diffusion Model
    - 連続なトークン
- VQVAE

        - CNN encoder + vector-quantization + CNN decoder
        - 入力

$$\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$$

        - Encoderの出力

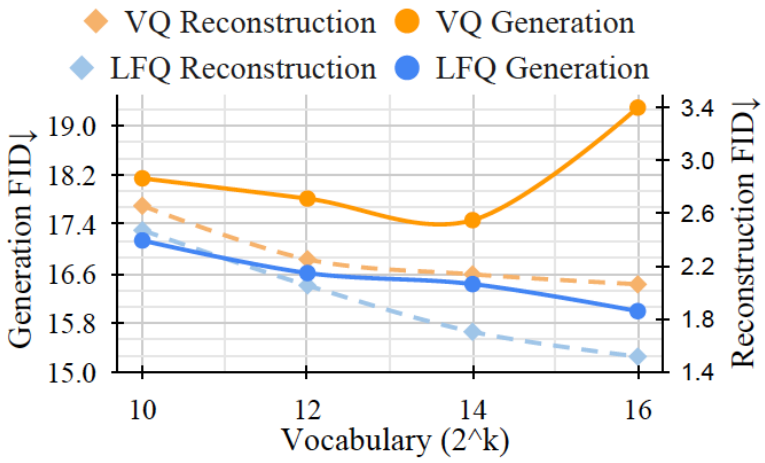
$$\mathbf{Z} = E(\mathbf{V}) \in \mathbb{R}^{T' \times H' \times W' \times d}$$

            - 一つの埋め込みベクトル**z**はvector quantizer qにパスされて、辞書**C**の中で一番距離が近い**c**に埋め込みされる

$$\mathbf{z} \in \mathbb{R}^d, \mathbf{c} \in \mathbb{R}, \mathbf{C} \in \mathbb{R}^{K \times d}$$

$$q(z) = \mathbf{c}_i, \text{ where } i = \underset{j \in \{1,2,\cdots,K\}}{argmin} ||\mathbf{z} - \mathbf{c}_j||_2$$

- メソッド
    - Lookup-free quantizer (LFQ)
        - LMの生成精度はvisual tokenizerの再構成精度と比例しない
            - 辞書の埋め込み数（vocabulary）を増やすと再構成精度は向上する



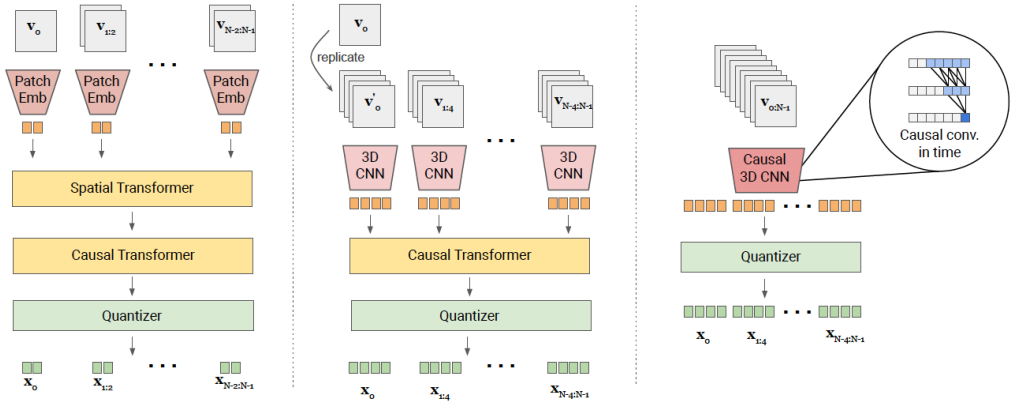
            - 一つの改善方法は、vocabularyを増やした際に辞書の埋め込み次元dを減らす

- LFQ
  - d=0にする

$$\mathbf{C} \in \mathbb{R}^{K \times d}$$を整数集合の$\mathbb{C}$、$|\mathbb{C}| = K$とする

    - これによって、上の図のようにvocabulary sizeを増やしても生成精度も良くなる

- Visual Tokenizer Model Improvement
  - Joint image-video tokenization
    - MAGVITでは、3D CNNを使用しているため画像のtokenizeができなかった
    - Causal 3D CNNに変更



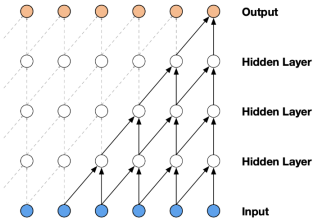
Causal Convolution

Edit

Introduced by Oord et al. in WaveNet: A Generative Model for Raw Audio

Causal convolutions are a type of convolution used for temporal data which ensures the model cannot violate the ordering in which we model the data: the prediction $p(x_{t+1}|x_1, \ldots, x_t)$ emitted by the model at timestep t cannot depend on any of the future timesteps $x_{t+1}, x_{t+2}, \ldots, x_T$. For images, the equivalent of a causal convolution is a masked convolution which can be implemented by constructing a mask tensor and doing an element-wise multiplication of this mask with the convolution kernel before applying it. For 1-D data such as audio one can more easily implement this by shifting the output of a normal convolution by a few timesteps.

Source: WaveNet: A Generative Model for Raw Audio



(a) Causal 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
MAGVIT-v2: Causal 3D CNN	58M	**7.06**	**96.33**

- 実験
  - データセット
    - 動画生成
      - Kinetics-600, UCF-101
    - 画像生成
      - ImageNet
    - Video Compression
      - MCL-JCV
    - Video Understanding
      - Kinetics-400, SSv2
  - 動画/画像生成
    - MAGVITで使用されているMasked Language Modelを使用
    - 動画生成
      - UCF-101
        - クラスを条件付けた生成
      - K600

- 5フレーム条件付けた動画予測

Table 1: **Video generation results**: frame prediction on Kinetics-600 and class-conditional generation on UCF-101. We adopt the evaluation protocol of MAGVIT.

Type	Method	K600 FVD↓	UCF FVD↓	#Params	#Steps
GAN	TrIVD-GAN-FP (Luc et al., 2020)	25.7±0.7			1
Diffusion	Video Diffusion (Ho et al., 2022c)	16.2±0.3		1.1B	256
Diffusion	RIN (Jabri et al., 2023)	10.8		411M	1000
AR-LM + VQ	TATS (Ge et al., 2022)		332±18	321M	1024
MLM + VQ	Phenaki (Villegas et al., 2022)	36.4±0.2		227M	48
MLM + VQ	MAGVIT (Yu et al., 2023a)	9.9±0.3	76±2	306M	12
MLM + LFQ	non-causal baseline	11.6±0.6		307M	12
MLM + LFQ	*MAGVIT-v2 (this paper)*	5.2±0.2 **4.3**±**0.1**	**58**±3	307M	12 24

- 画像生成
  - ImageNet

Table 2: **Image generation results**: class-conditional generation on ImageNet 512×512. Guidance indicates the classifier-free diffusion guidance (Ho & Salimans, 2021). * indicates usage of extra training data. We adopt the evaluation protocol and implementation of ADM.

Type	Method	w/o guidance FID↓	w/o guidance IS↑	w/ guidance FID↓	w/ guidance IS↑	#Params	#Steps
GAN	StyleGAN-XL (Sauer et al., 2022)			2.41	267.8	168M	1
Diff. + VAE*	DiT-XL/2 (Peebles & Xie, 2022)	12.03	105.3	3.04	240.8	675M	250
Diffusion	ADM+Upsample (Dhariwal & Nichol, 2021)	9.96	121.8	3.85	221.7	731M	2000
Diffusion	RIN (Jabri et al., 2023)	3.95	216.0			320M	1000
Diffusion	simple diffusion (Hoogeboom et al., 2023)	3.54	205.3	3.02	248.7	2B	512
Diffusion	VDM++ (Kingma & Gao, 2023)	2.99	232.2	2.65	278.1	2B	512
MLM + VQ	MaskGIT (Chang et al., 2022)	7.32	156.0			227M	12
MLM + VQ	DPC+Upsample (Lezama et al., 2023)	3.62	249.4			619M	72
MLM + LFQ	*MAGVIT-v2 (this paper)*	4.61 3.07	192.4 213.1	**1.91**	**324.3**	307M	12 64

- Video Compression
  - MCL-JCV
    - 640×360
    - Elo score
      - 人の主観評価

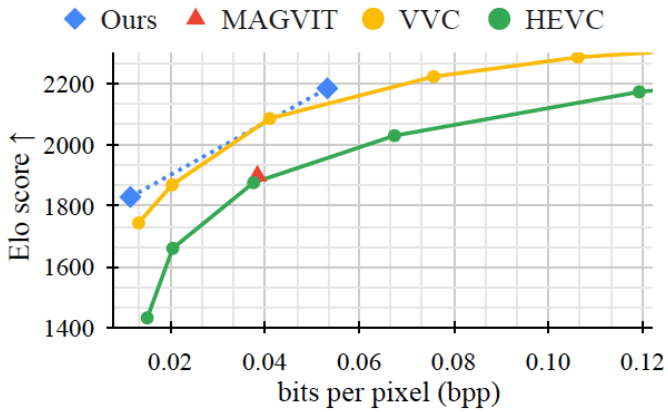
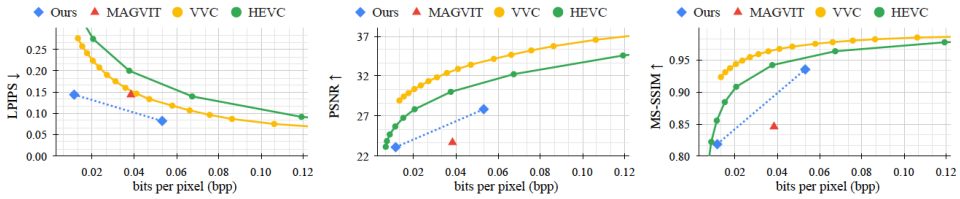


Figure 6: **Video compression rater study.**

- LPIPS, PSNR, MS-SSIM



(a) LPIPS↓ (b) PSNR↑ (c) MS-SSIM↑

Figure 9: **Video compression metrics**, supplementary to Tab. 3.

- Video Understanding
  - 行動認識のために動画の表現を獲得できているか評価
  - 設定
    - using tokens as prediction targets for the transformer's output
    - using tokens as the input to the transformer

Table 4: **Video action recognition performance** (classification accuracy↑ ×100).

Token as transformer's:	Output	Input		
Tokenizer	SSv2	SSv2	K400	K600
3D VQ-VAE	64.13	41.27	44.44	45.67
MAGVIT (Yu et al., 2023a)	67.22	57.34	72.29	74.65
MAGVIT-v2 (this paper)	**67.38**	**62.40**	**75.34**	**77.93**
Raw pixel	64.83	63.08	76.13	78.92
HoG descriptor (Wei et al., 2022)	65.86	n/a	n/a	n/a