LANGUAGE MODEL BEATS DIFFUSION — TOKENIZER IS KEY TO VISUAL GENERATION

Lijun Yu[‡†*] **José Lezama**[†] **Nitesh B. Gundavarapu**[†] **Luca Versari**[†] **Kihyuk Sohn**[†]
David Minnen[†] **Yong Cheng**[†] **Agrim Gupta**[†] **Xiuye Gu**[†] **Alexander G. Hauptmann**[‡]
Boqing Gong[†] **Ming-Hsuan Yang**[†] **Irfan Essa**[†] **David A. Ross**[†] **Lu Jiang**[†‡]
[†]Google, [‡]Carnegie Mellon University

Present by: Junqi Qu

What is a Token

- Smallest unit
  - ChatGPT is a language model -> “Chat”, ”GPT”, “is” ….
  - Image can be segmented into 16x16xD

Problem and Motivation

- **The Generative AI Landscape**
  - **For language tasks,** Large Language Models (LLMs) are the dominant, de facto models.
  - **For image and video generation,** diffusion models are widely considered state-of-the-art.
- **A Puzzling Performance Gap**
  - Despite advances, LLMs have historically underperformed diffusion models in visual synthesis.
  - For example, on the ImageNet 256x256 benchmark, the best language model had an FID score of 3.41, while a top diffusion model achieved 1.79—a substantial 48% performance gap

Central Hypothesis

- The paper hypothesizes that the primary reason for this gap is **not the LLM architecture itself, but the lack of a good visual representation**.
- LLMs depend on a **visual tokenizer** to convert pixel-space inputs into discrete tokens.
- This process creates a "visual vocabulary." If this vocabulary is poor or inefficient, the LLM's generative capabilities are severely limited, regardless of its power.
- The authors propose that with a better tokenizer, LLMs can surpass diffusion models.

Background: how LLMs perform visual generation

- It's a two-stage process built on the VQ-VAE (Vector Quantized-Variational Autoencoder) framework.
- **Tokenization (Encoding):**
  - A video V is fed into a CNN Encoder, producing latent embeddings Z.
  - A **Vector Quantizer** then maps each feature vector in Z to the closest vector in a learned codebook C.
  - The *index* of that codebook vector becomes the discrete token. The result is a 2D or 3D grid of tokens.
- **Generation (Modeling):**
  - The token grid is flattened into a 1D sequence.
  - An LLM (e.g., a Transformer) models this sequence, either by predicting the next token (Autoregressive) or by filling in masked-out tokens (Masked LM).
  - A decoder then converts the LLM's generated token sequence back into pixels.

The Solution: MAGVIT-v2

- A new video tokenizer designed to generate concise and expressive tokens for both images and videos using a common vocabulary.
- It features two key innovations:
  - **A Novel Lookup-Free Quantization (LFQ):** This enables the learning of a very large vocabulary, which is crucial for improving the generation quality of the language model.
  - **Architectural Enhancements:** The model uses a **causal 3D CNN** and other modifications to tokenizing both images and videos seamlessly, a challenge for previous models.

Innovation 1: Lookup-Free Quantization (LFQ)

- **The VQ Bottleneck:**
  - Improving a VQ-VAE's reconstruction quality by increasing its vocabulary size does not necessarily improve the LLM's generation quality.
  - In fact, very large vocabularies can *hurt* the LLM's performance, which is why most visual tokenizers use small codebooks (e.g., 8,192 tokens).
- **The LFQ Breakthrough:**
  - LFQ replaces the standard high-dimensional codebook lookup. It decomposes the latent space and uses a simple sign function for quantization, making each dimension an independent binary decision.
  - **Crucially, with LFQ, both reconstruction and generation quality consistently improve as vocabulary size increases**.
  - This allows them to successfully train with a vocabulary of 2^18 (262,144) tokens, which is far larger than what was previously feasible

Innovation 2: Causal 3D CNN Architecture

- State-of-the-art video tokenizers like the original MAGVIT use 3D CNNs, which makes it difficult to tokenize single images due to their temporal receptive field.
- MAGVIT-v2 uses **temporally causal 3D convolutions**.
  - This means the output for any given frame only depends on that frame and *previous* frames—never future ones.
  - As a result, the first frame is always processed independently, allowing the model to tokenize a single image naturally.
  - This design was empirically shown to be the most effective architecture.

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		307M	12
		4.3±**0.1**	**58**±3		24

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	192.4			307M	12
		3.07	213.1	**1.91**	**324.3**		64

Applications

- The power of the MAGVIT-v2 tokenizer extends beyond just image generation.
- **Video Generation:**
  - On the Kinetics-600 frame prediction benchmark, MAGVIT-v2 achieves an FVD score of **5.2**, significantly outperforming the previous MAGVIT (9.9) and a strong video diffusion model (16.2).
- **Video Compression:**
  - Human raters preferred MAGVIT-v2's compression quality over the **HEVC (H.265) standard** and found it comparable to the **next-generation VVC (H.266) codec** at similar bitrates. This is a first for a generative tokenizer.
- **Video Understanding:**
  - The learned tokens serve as a strong representation for downstream tasks like action recognition, outperforming the previous tokenizer and approaching the performance of models trained on raw pixels.

Conclusion

- **A Paradigm Shift:** This work shows the bottleneck for LLMs in vision was the **tokenizer**, not the generative model architecture itself.
- **Representation is Key:** A high-quality, discrete visual representation is crucial.
- **Lookup-Free Quantization (LFQ)** is a breakthrough that enables large, effective visual vocabularies.
- **SOTA Performance:** For the first time, an LLM-based approach has surpassed diffusion models on a major image generation benchmark, with better quality and higher efficiency.
- **Future Implications:** This research strongly advocates for more focus on visual tokenization methods and paves the way for more powerful and unified multimodal LLMs.