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

ABSTRACT

While Large Language Models (LLMs) are the dominant models for generative tasks in language, they do not perform as well as diffusion models on image and video generation. To effectively use LLMs for visual generation, one crucial component is the visual tokenizer that maps pixel-space inputs to discrete tokens appropriate for LLM learning. In this paper, we introduce MAGVIT-v2, a video tokenizer designed to generate concise and expressive tokens for both videos and images using a common token vocabulary. Equipped with this new tokenizer, we show that LLMs outperform diffusion models on standard image and video generation benchmarks including ImageNet and Kinetics. In addition, we demonstrate that our tokenizer surpasses the previously top-performing video tokenizer on two more tasks: (1) video compression comparable to the next-generation video codec (VCC) according to human evaluations, and (2) learning effective representations for action recognition tasks.

1 INTRODUCTION

Large transformer-based language models, commonly referred to as LMs or LLMs, are the de facto models for natural language generation (OpenAI, 2023; Google, 2023; Touvron et al., 2023). Over time, LMs have expanded their capabilities to generate content in various modalities, asserting their dominance in other domains like audio (Agostinelli et al., 2023), speech (Rubenstein et al., 2023), code generation (Li et al., 2023), medical applications (Singhal et al., 2023) and robotics (Zitkovich et al., 2023).

LMs are capable of generating images and videos. To do so, the image pixels are mapped into a sequence of discrete tokens by a visual tokenizer (*c.f.* Section 2). These tokens are then fed into the LM transformer, as if they were lexical words, for generative modeling. Despite notable advancements in employing LMs for visual generation (Esser et al., 2021; Chang et al., 2022), LMs still do not perform as well as diffusion models (Rombach et al., 2022). For instance, when evaluating on the ImageNet dataset, a gold standard benchmark for image generation, the best language model (Lee et al., 2022) underperforms the diffusion model (Gao et al., 2023) by a substantial 48% margin (FID 3.41 *vs.* 1.79 when generating images at the 256×256 resolution).

Why do language models lag behind diffusion models in visual generation? This paper suggests that a primary reason is the lack of a good visual representation, resembling our natural language system, for effectively modeling the visual world. To substantiate this hypothesis, this paper shows that, when utilizing a good visual tokenizer, the masked language model (Devlin et al., 2019; Chang et al., 2022; Yu et al., 2023a) surpasses the state-of-the-art diffusion models in terms of both generation fidelity and efficiency across image and video benchmarks, given the same training data, comparable model size, and training budget. To the best of our knowledge, this provides the first evidence that language models beat diffusion models on the hallmark ImageNet benchmark.

It is worth emphasizing that our intention is not to assert whether the language model is superior to others, but to promote the exploration of visual tokenization methods for LLMs. A fundamental difference of LLMs from other models, such as diffusion models, is that LLMs utilize a discrete latent format: tokens obtained from a visual tokenizer. We show that the values of these discrete visual tokens should not be overlooked considering their distinct advantages as follows. **(1) Compatibility with LLMs.** The main advantage of a token representation is that it shares the same form

[*]Work done during a research internship at Google Research.

as language tokens, making it straightforward to leverage the optimizations our community has developed over many years for LLMs. This includes faster training and inference speeds (Shazeer, 2019; Lester et al., 2021), advancements in model infrastructure (Dao et al., 2022; Du et al., 2022), learning recipes for model scaling (Brown et al., 2020; Chowdhery et al., 2022), and GPU/TPU optimization, among other innovations. Unifying vision and language by the same token space could set the stage for a true multimodal LLM that can understand, generate, and reason within our visual environment. **(2) Compressed representation.** The discrete token may offer a fresh perspective on video compression. The visual tokens can serve as a new video compression format to reduce disk storage and bandwidth during internet transfers. Unlike compressed RGB pixels, these tokens can be fed directly into generative models, bypassing the conventional decompression and latent encoding steps. This allows for faster processing in generative video applications, especially beneficial in edge computing cases. **(3) Visual understanding benefits**. Prior research has shown that the discrete tokens are valuable as a pre-training target in self-supervised representation learning, as discussed in BEiT (Bao et al., 2021) and BEVT (Wang et al., 2022). Additionally, research finds that using tokens as the model inputs improves the robustness and generalization (Mao et al., 2021).

In this paper, we introduce MAGVIT-v2, a video tokenizer designed to map videos (and images) into compact discrete tokens. Our model is built on the state-of-the-art video tokenizer, MAGVIT (Yu et al., 2023a), within the VQ-VAE framework (Van Den Oord et al., 2017). We propose two new techniques. First, a novel lookup-free quantization method enables the learning of a large vocabulary that is able to improve generation quality of the language model. Second, through extensive empirical analyses, we have identified modifications to the tokenizer that not only enhance generation quality but also enable the tokenization of both images and videos using a shared vocabulary.

We empirically demonstrate that our model outperforms the previously top-performing video tokenizer, MAGVIT, in three key areas. First, our model significantly improves the generation quality of MAGVIT, establishing the state of the art on the common image and video benchmarks. Second, user studies indicate that its compression quality exceeds that of MAGVIT and the current video compression standard, HEVC (Sullivan et al., 2012). Moreover, it is on par with the next-generation video codec, VVC (Bross et al., 2021). Finally, we show that, compared to MAGVIT, our new tokens are stronger for video understanding tasks across two setups and three datasets. The main contributions of this work are:

- A new video tokenizer that outperforms the previously best-performing video tokenizer in three areas: visual generation, video compression, and action recognition.
- A novel lookup-free quantization approach that enables improving the visual generation quality of language models by learning a large vocabulary.
- To the best of our knowledge, the first evidence suggesting that a language model can outperform diffusion models on ImageNet when provided with the same training data, an equivalent model size, and a similar training budget.
- A video compressor with better quality than HEVC and VVC, at similar bit rates, according to user studies. To our knowledge, this is the first successful attempt of a visual tokenizer designed for video generation to achieve comparable results to standard codecs.

2 Background

Language Model (LM) for visual generation. LMs have been extended to generate images and videos. A visual tokenizer f is used to first map visual inputs into a sequence of discrete tokens. A video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$ (or image when $T = 1$) is tokenized into a discrete representation $\mathbf{X} = f(\mathbf{V}) \in \{1, 2, \cdots, K\}^{T' \times H' \times W'}$, where K is the codebook (vocabulary) size of the visual tokenizer. $\mathbf{X}$ is flattened into a 1D token sequence obtained using raster scan ordering and then fed into an LM transformer for generative modeling.

Two types of LMs are commonly used for visual generation. The *Autoregressive LM (AR-LM)* includes ImageGPT (Chen et al., 2020), DALL-E (Ramesh et al., 2021), Parti (Yu et al., 2022b), *etc*. An AR-LM predicts the next token given the previous tokens along with additional conditioning information $\mathbf{c}$ using a categorical distribution for $p_\theta(\mathrm{x}_i \mid \mathbf{x}_{<i}; \mathbf{c})$. During inference, AR-LMs use the standard autoregressive decoding over the tokens. Finally, the tokens are converted back to pixels by a decoder associated with the visual tokenizer.

The *Masked LM (MLM)* is another type of language model for visual generation, such as: MaskGIT (Chang et al., 2022), MAGVIT (Yu et al., 2023a), Phenaki (Villegas et al., 2022), and MUSE (Chang et al., 2023), among others. An MLM is trained using a masked token objective (De-

vlin et al., 2019), where some tokens in the sequence are randomly masked and need to be predicted given the observed tokens. Let $\mathbf{m} \in \{0,1\}^n$ be a random binary sequence where $\mathbf{m}^\top \mathbf{1} \in [0, n-1]$. The MLM learns $p_\theta(\mathrm{x}_i \mid \{\mathrm{x}_j : \mathrm{m}_j = 1, \forall j\}; \mathbf{c})$ for all i where $\mathbf{m}_i = 0$. To generate a video or image during inference, the MLM uses the non-autoregressive decoding algorithms for images and videos (Chang et al., 2022; Yu et al., 2023a). The decoding starts with a fully masked sequence, which is iteratively filled by repeating two steps: (1) sample the whole sequence $\hat{\mathbf{x}}^{(t)}$ from p_θ given the non-masked tokens from the previous step, (2) re-mask the $\lfloor \lambda(t) \cdot n \rfloor$ tokens in $\hat{\mathbf{x}}^{(t)}$ with the lowest probability, following a decreasing masking ratio schedule $\lambda(t)$, according to timestamp t.

Denoising Diffusion Models (DDM). DDMs (Sohl-Dickstein et al., 2015; Song & Ermon, 2019) are regarded as the state-of-the-art in visual generation due to their high-quality image (Dhariwal & Nichol, 2021; Ho et al., 2022a) and video generation (Ho et al., 2022c). For instance, DDPM (Ho et al., 2020) learns a denoising process parameterized as conditional Gaussian distributions over image pixels. Recently, diffusion models and language models have displayed a significant overlap. Recent DDMs diffuse over latents rather than raw pixels. These latents are obtained using models similar to the visual tokenizer used by LMs. In fact, the very first latent in diffusion, proposed by Rombach et al. (2022), is derived from a visual tokenizer. Additionally, the diffusion model's architecture has been shifting from the U-Net to the transformer architecture (Peebles & Xie, 2022). Consequently, the boundaries between diffusion and language models in visual generation have become less distinct. Yet, a fundamental difference between DDMs and LMs lies in the latent format, *i.e.*, continuous *vs.* discrete. We have discussed the benefits of having discrete tokens in Section 1 and will show that the proposed tokenizer improves in these aspects.

Visual tokenization. Visual tokenization plays an essential role in mapping pixels into a discrete representation suitable for generative modeling. VQ-VAE (Van Den Oord et al., 2017) is a cornerstone work in image tokenization. A VQ-VAE model consists of a convolutional neural network (CNN) encoder, a vector-quantization (VQ) bottleneck, and a CNN decoder. Given a video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, the VQ-VAE's encoder E produces latent embeddings $\mathbf{Z} = E(\mathbf{V}) \in \mathbb{R}^{T' \times H' \times W' \times d}$. Each embedding vector $\mathbf{z} \in \mathbb{R}^d$ in $\mathbf{Z}$ is then passed through the vector quantizer q, which assigns it to the closest entry $\mathbf{c} \in \mathbb{R}^d$ in the learned codebook embedding $\mathbf{C} \in \mathbb{R}^{K \times d}$:

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

To get discrete tokens, we drop the embedding dimension and represent $\mathbf{Z}$ by its indices $\mathbf{X} \in \{1, 2, \cdots, K\}^{T' \times H' \times W'}$. For decoding, embeddings of all image tokens are given as input to the decoder D to reconstruct the input $\hat{\mathbf{V}} = D(\mathbf{Z})$. Following VQ-VAE, VQGAN (Esser et al., 2021) introduces an adversarial loss and feature-level perceptual losses to enhance the image quality.

Video tokenization is more challenging and VQGAN has been adapted to meet this purpose (Ge et al., 2022; Villegas et al., 2022; Yu et al., 2023a). The state of the art in video tokenization is MAGVIT (Yu et al., 2023a), which introduces a better 3D architecture, an inflation technique for initialization using image pre-training, and robust training losses. With MAGVIT, the LMs achieve leading generation quality across multiple video benchmarks. However, MAGVIT struggles to tokenize images and often results in noticeable flickering in longer videos.

3 Method

We introduce a new **video tokenizer** designed to map the spatial-temporal dynamics from a visual scene into compact discrete tokens suitable for language models. Compared with image generation, video generation still faces substantial challenges in generating consistent and realistic motion. We are interested in exploring the capabilities of language models in tackling this unsolved challenge. Therefore, this paper focuses on a video tokenizer that can effectively represent video for generative modeling. Our approach builds upon the state-of-the-art video tokenizer, MAGVIT, as detailed in Yu et al. (2023a). This section highlights two new designs: a lookup-free quantizer and a collection of enhancements to the tokenizer model.

3.1 Lookup-Free Quantizer

Although the community has made great progress in developing VQ-VAEs, the relationship between improvements in the reconstruction quality and subsequent generation quality is still not well understood. A common misconception is that improving reconstruction equates to improving the generation of the language model. For example, enlarging the vocabulary can improve reconstruction quality. However, such improvement only extends to generation when the vocabulary size is small, and a very large vocabulary can actually hurt the performance of the language model.

As illustrated by the dashed curves in Fig. 1, the reconstruction FID, indicated by the right y-axis (where a lower value is better), improves as the vocabulary size (the x-axis) increases. The orange solid curve in Fig. 1 represents the LM's generation quality (the left y-axis). The generation FID initially improves but deteriorates for larger vocabulary. This may shed light on why the vocabulary size of most language models for visual generation is around 1-8k (Esser et al., 2021; Villegas et al., 2022), which is significantly smaller than the size of natural language vocabulary, *i.e.* over 200k.

A simple trick for training a larger codebook involves decreasing the code embedding dimension when increasing the vocabulary size (Yu et al., 2022a). This trick captures the intuition of limiting the representational capacity of individual tokens, which in turn facilitates learning over the distribution of a large vocabulary.

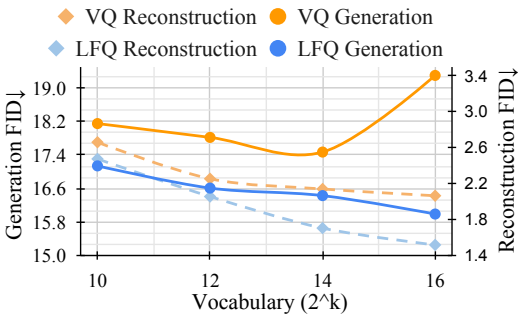


Figure 1: **Reconstruction and generation quality curves** in FID on ImageNet when scaling the tokenizer's vocabulary size with Vector Quantization (VQ) and Lookup-Free Quantization (LFQ). Comparison is done at 128×128 resolution using an MLM with 306-372M parameters.

Lookup-Free Quantization (LFQ). Motivated by the above observation, we reduce the VQ-VAE codebook's embedding dimension to zero. Formally, the codebook $\mathbf{C} \in \mathbb{R}^{K \times d}$ is replaced with an integer set $\mathbb{C}$ where $|\mathbb{C}| = K$. Recall that in VQ-VAE models, the quantizer must look up all K d-dimensional embeddings in the codebook, where d is typically 256, when computing the closest codebook entry to the encoder output. This new design eliminates the need for such embedding lookup entirely hence we call it *lookup-free quantization (LFQ)*. We found that LFQ can grow the vocabulary size in a way benefiting the generation quality of language models. As shown by the blue curves in Fig. 1, both reconstruction and generation consistently improves as the vocabulary size increases – a property not observed in current VQ-VAE methods.

While various LFQ methods are available, this paper discusses a straightforward variant that assumes independent codebook dimensions and binary latents. Specifically, the latent space of LFQ is decomposed as the Cartesian product of single-dimensional variables, as $\mathbb{C} = \times_{i=1}^{\log_2 K} C_i$. Given a feature vector $\mathbf{z} \in \mathbb{R}^{\log_2 K}$, each dimension of the quantized representation $q(\mathbf{z})$ is obtained from:

$$q(\mathrm{z}_i) = C_{i,j}, \text{ where } j = \arg\min_k \|\mathrm{z}_i - C_{i,k}\|, \tag{2}$$

where $C_{i,j}$ is the j-th value in C_i. With $C_i = \{-1, 1\}$, the $\arg\min$ can be computed by the sign function as

$$q(\mathrm{z}_i) = \mathrm{sign}(\mathrm{z}_i) = -\mathbb{1}\{\mathrm{z}_i \leqslant 0\} + \mathbb{1}\{\mathrm{z}_i > 0\}. \tag{3}$$

With LFQ, the token index for $q(\mathbf{z})$ is given by:

$$Index(\mathbf{z}) = \sum_{i=1}^{\log_2 K} \arg\min_k \|\mathrm{z}_i - C_{i,k}\| \prod_{b=0}^{i-1} |C_b| = \sum_{i=1}^{\log_2 K} 2^{i-1} \mathbb{1}\{\mathrm{z}_i > 0\}, \tag{4}$$

where $|C_0| = 1$ sets the virtual basis.

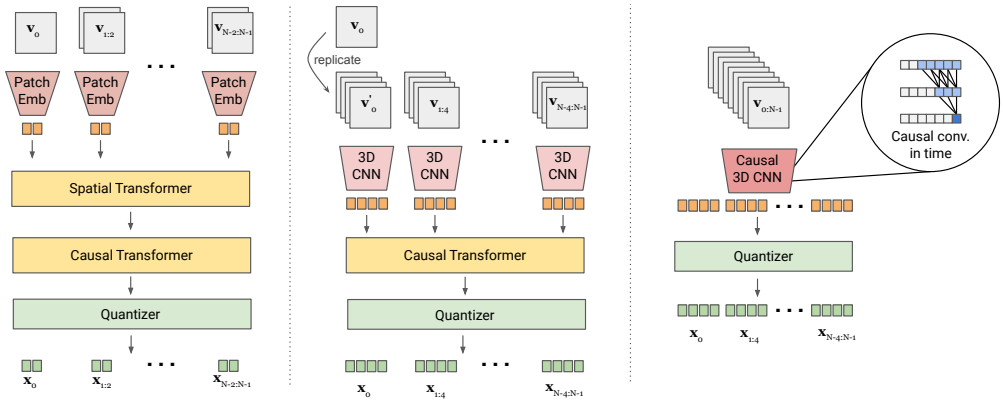
We add an entropy penalty during training to encourage codebook utilization:

$$\mathcal{L}_{entropy} = \mathbb{E}[H(q(\mathbf{z}))] - H[\mathbb{E}(q(\mathbf{z}))]. \tag{5}$$

This penalty is inspired by a similar loss used in image VQGAN model (Chang et al., 2022), which is also found in entropy-based clustering (Jansen et al., 2020). In LFQ, given the independence among dimensions, we rewrite $H(q(\mathbf{z})) = \sum_{i=1}^{\log_2 K} H(q(\mathrm{z}_i))$. The $H[\mathbb{E}(q(\mathbf{z}))]$ term can be approximated with sub-groups of dimensions for $K > 2^{18}$ where direct estimation is memory bound.

We note that there are various other variants of LFQ, *e.g.*, opting for the multivariant over the binary codebook C_i or employing other quantization techniques such as Agustsson et al. (2019). As the first paper to introduce this concept, we focus on the simplest form with independent binary dimensions, which shows promising improvements. Other LFQ methods merit further research.

In addition to the entropy penalty (Eq. (5)), an LFQ-based tokenizer is trained using the standard combination of *reconstruction*, *GAN*, *perceptual*, and *commitment* losses (Esser et al., 2021), excluding the inapplicable codebook loss. Following Yu et al. (2023a), we use LeCAM regularization (Tseng et al., 2021) for improved stability.



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

Figure 2: **Causal tokenizer architecture comparison**. The decoders, which are omitted from the figure, employ an architecture that is symmetric to the encoder. See detailed architecture diagram in the Appendix.

3.2 Visual Tokenizer Model Improvement

Joint image-video tokenization. A desirable feature of visual tokenization is the capability to tokenize images and videos using a shared codebook. However, the MAGVIT tokenizer, which utilizes the 3D CNN, faces challenges in tokenizing images due to the temporal receptive field.

To build a joint image-video tokenizer, a new design is needed. We begin our discussion by revisiting an existing method C-ViViT (Villegas et al., 2022). As depicted in Fig. 2a, C-ViViT employs full spatial transformer blocks combined with causal temporal transformer blocks. This approach performs reasonably well but has two drawbacks. First, unlike CNNs, the positional embeddings makes it difficult to tokenize spatial resolutions that were not seen during training. Second, empirically we found that 3D CNNs perform better than spatial transformer and produce tokens with better spatial causality of the corresponding patch.

To tackle these drawbacks, we explore two plausible designs. Fig. 2b combines C-ViViT and MAGVIT. Assuming a temporal compression ratio of 4, a 3D CNN processes blocks of 4 frames followed by a causal transformer. In Fig. 2c, we use the temporally causal 3D convolution to replace the regular 3D CNN. Specifically, the temporal padding scheme for a regular 3D convolution layer with kernel size (k_t, k_h, k_w) includes $\lfloor \frac{k_t - 1}{2} \rfloor$ frames before and $\lfloor \frac{k_t}{2} \rfloor$ frames after the input frames. In contrast, a causal 3D convolution layer pads with $k_t - 1$ frames before the input and nothing after, so that the output for each frame only depends on the previous frames. In consequence, the first frame is always independent of other frames, allowing the model to tokenize single images.

Temporal convolutional subsampling with stride s is sufficient for $s\times$ down-sampling by mapping $1 + s \times t$ frames into $1 + t$. After a regular $s\times$ up-sampling, we drop the first $s - 1$ resulting frames, which maps $1 + t$ frames into $1 + s \times t$ and allows for the tokenization of a single image. Tab. 5a empirically compares the designs in Fig. 2, and we find that the causal 3D CNN performs the best.

Architecture modifications. In addition to using causal 3D CNN layers, we made several other architectural modifications to improve upon the MAGVIT model. First, we change the encoder downsamplers from average pooling into strided convolutions to leverage learned kernels, and replace the decoder upsamplers from nearest resizing followed by convolution with a depth-to-space operator. Second, we defer the temporal downsampling from the first few encoder blocks to the last ones. In addition, the downsampling layer in the discriminator now utilizes 3D blur pooling (Zhang, 2019) to encourage shift invariance. Finally, we add one adaptive group normalization layer before the residual blocks at each resolution in the decoder to pass in the quantized latents as the control signal following StyleGAN (Karras et al., 2019). Tabs. 5b and 5c empirically verify these designs.

Token factorization for efficient prediction. The output tokens can be fed into language models to generate videos. To assist smaller transformers predicting in a large vocabulary, we can factorize the LFQ token's latent space into equal subspaces. For instance, rather than predicting using a codebook of size 2^{18}, we can predict in two concatenated codebooks, each of size 2^9. We embed each subspace token separately and use their embedding summation as the token embedding for the transformer input. We find it beneficial to use weight tying (Press & Wolf, 2017), a common technique in language modeling, which involves sharing the weights between the embedding and softmax layers. For the output layer with a factorized vocabulary, we use the embedding matrix for each subspace to obtain logits with seperate prediction heads.

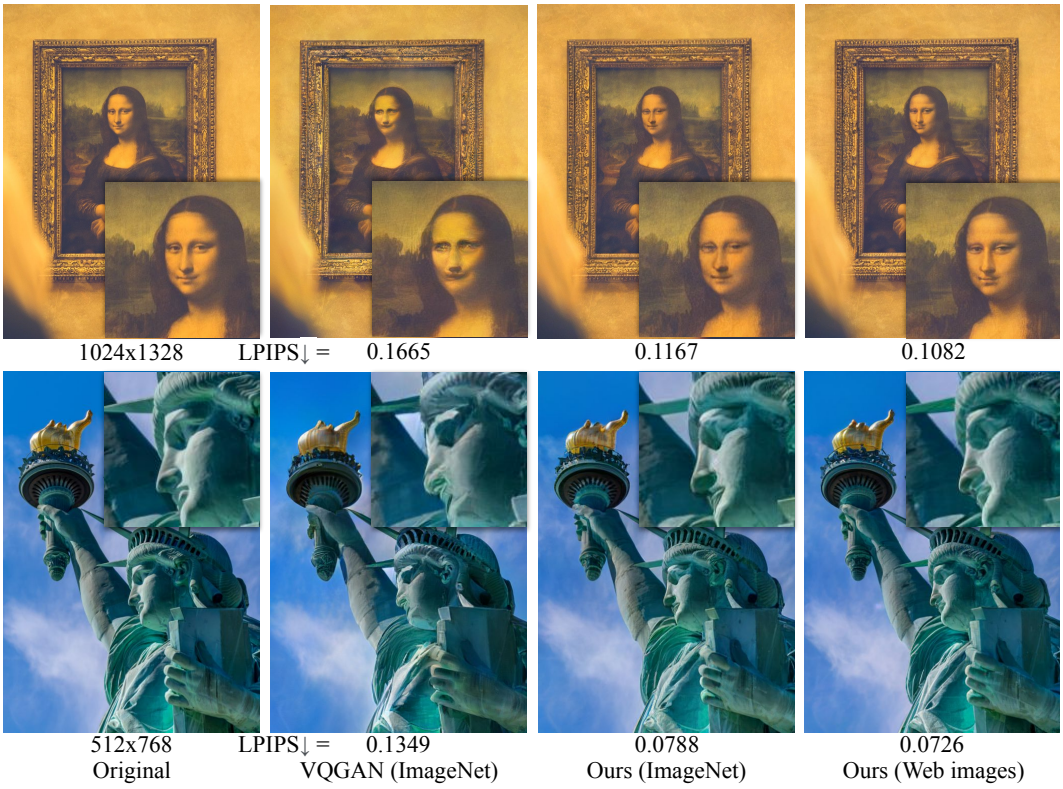


Figure 3: **Image reconstruction samples with different tokenizers**. We compare the VQGAN used in MaskGIT (Chang et al., 2022) with two of our models trained on ImageNet and web images (Chen et al., 2022). Original images are by Eric TERRADE and Barth Bailey on Unsplash.

4 Experiments

This section empirically verifies the proposed tokenizer across three distinct tasks: video and image generation, video compression, and action recognition. Fig. 3 visually compares the reconstruction quality of our tokenizer with prior works. More qualitative samples are shown at https://magvit.cs.cmu.edu/v2.

4.1 Experimental Setups

Datasets. We use Kinetics-600 (K600) (Carreira et al., 2018) and UCF-101 (Soomro et al., 2012) for video generation experiments, along with ImageNet (Deng et al., 2009) for image generaton. In addition, MCL-JCV (Wang et al., 2016) is used as the testbed for video compression, with Kinetics-400 (K400) (Kay et al., 2017) and SSv2 (Goyal et al., 2017) for video understanding.

Implementation details We follow the tokenizer training setting and hyperparameters in (Yu et al., 2023a), unless stated otherwise. LFQ is used, which eliminates the codebook embedding, to increase the default codebook size to $K = 2^{18}$. The weight of $\mathcal{L}_{entropy}$ follows an annealing schedule with a $3\times$ higher starting point and linearly decays to a fixed value of 0.1 within 2k steps. We defer details regarding the evaluation setup of each subsection to the Appendix.

4.2 Visual Generation

The masked language model (MLM) (Devlin et al., 2019) is used in image and video generation. To verify the tokenizer, we employ the same MLM transformers in MAGVIT (Yu et al., 2023a). We select the MLM due to its competitive performance on benchmark datasets (Yu et al., 2023a; Lezama et al., 2023). In the Appendix, we also show that an autoregressive language model (AR-LM) coupled with the proposed tokenizer outperforms the prior work MAGVIT. As we use a smaller MLM ($\sim$300M parameters) with a large codebook (2^{18} $\approx$262K), the token factorization as discussed in Section 3.2 is applied using two heads with each predicting from a codebook of size 2^9.

Video generation. We consider two standard video benchmarks, UCF-101 for class-conditional generation and K600 for frame prediction with 5-frame condition. FVD (Unterthiner et al., 2018) is used as our primary evaluation metric. Tab. 1 shows that our model surpasses all prior arts in both

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		w/ guidance		#Params	#Steps
		FID↓	IS↑	FID↓	IS↑		
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

benchmarks. Specifically, it outperforms the previous best model MAGVIT by a large margin, while using the same MLM transformer backbone. In addition, it significantly outperforms the non-causal baseline on frame prediction, highlighting the contribution of the causal tokenizer. These results demonstrate the essential role of a good visual tokenizer in enabling LMs to generate high-quality videos. Fig. 4 shows qualitative samples from the model.

Image generation on ImageNet. We evaluate MAGVIT-v2 on image generation under the standard ImageNet class-conditional setting. We present results for resolution 512×512 in Tab. 2 and refer to the Appendix for 256×256 results. FID (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) are used as evaluation metrics. Our model surpasses the best performing diffusion models both in sampling quality (FID and IS) and inference-time efficiency (sampling steps).

It is worth noting that all the models compared are trained using the same ImageNet training data, with a comparable model size and training budget. Therefore, the performance primarily evaluates the model's capabilities. The masked language model, equipped with our tokenizer, exhibits a notable improvement in FID over the best diffusion model baseline at 512×512 (FID=1.91 *vs*. 2.65, 28%↓). While this margin narrows at 256×256 resolution, the MLM uses a 50% reduced model size and needs much fewer decoding steps (*e.g*., 64 *vs*. 250) to get the image generation quality. Qualitative samples in comparison with other models are shown in Fig. 5.

4.3 Video Compression

We conduct a subjective rater study to assess the compression quality of MAGVIT-v2. The study is conducted on the 30 videos of the MCL-JCV dataset, resized to a resolution of 640×360. Sixteen raters are engaged, each providing responses to an average of roughly 800 pairwise-preference questions.

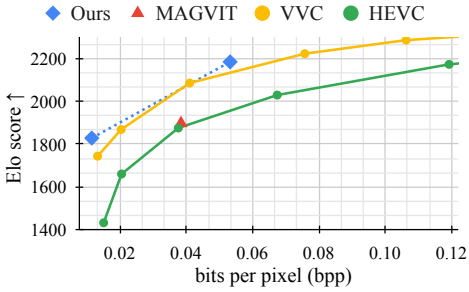


Figure 6: **Video compression rater study**.

Condition → Generation

Figure 4: **Frame prediction samples on Kinetics-600**.

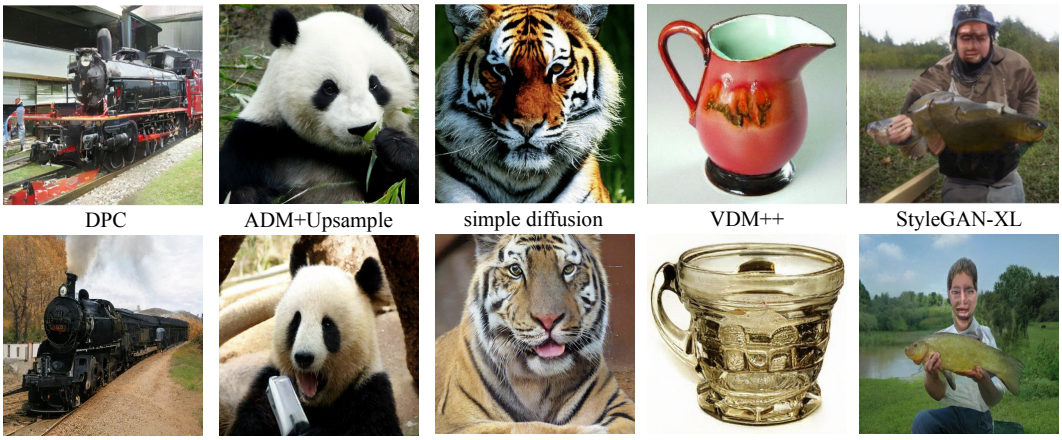


Figure 5: **Class-conditional generation samples on ImageNet 512×512**. We compare with each of the previous works with a random sample from the same image class.

We calculate Elo scores (Elo & Sloan, 2008) based on pairwise preferences to quantify the relative visual quality between the models. The study compares our model with MAGVIT as well as the current video compression standard HEVC (H.265) video codec (Sullivan et al., 2012) and the next-generation codec VVC (H.266) (Bross et al., 2021). As shown in Fig. 6, raters prefer our model to the compared methods at multiple bit rates.

Table 3: **Video compression metrics**.

Method	LPIPS↓	PSNR↑	MS-SSIM↑
HEVC (Sullivan et al., 2012)	0.199	30.10	0.943
VVC (Bross et al., 2021)	0.153	**32.65**	**0.966**
MAGVIT (Yu et al., 2023a)	0.144	23.70	0.846
MAGVIT-v2 (this paper)	**0.104**	26.18	0.894

We also compare the compression quality using common distortion metrics (LPIPS, PSNR, and MS-SSIM). Tab. 3 compares at 0.0384 bpp, the bit rate of MAGVIT, with full curves in the Appendix. The results show that our model outperforms MAGVIT on all metrics, and it outperforms all methods on LPIPS, a metric which correlates more closely with subjective quality assessments. At equal bit rates, standard codecs may render local details more accurately than neural models but also introduce block artifacts, detrimental to perceptual quality yet not captured by PSNR and MS-SSIM (Agustsson et al., 2019). Despite promising results with TPUs, further research is needed to adapt our model to run efficiently on CPUs like standard codecs.

4.4 Video Understanding

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

In this subsection, we assess the tokenizer's capability to learn a video understanding model for action recognition. Two setups are examined: (1) using tokens as prediction targets for the transformer's output, and (2) using tokens as the input to the transformer. For the former setup, we use a similar architecture following the BEVT (Wang et al., 2022) pre-training. For the tokens as inputs, to work with the ViViT backbone (Arnab et al., 2021), we detokenize the tokens to pixels before feeding them to frozen ViViT transformers trained on raw pixels.

Tab. 4 shows that MAGVIT-v2 outperforms the previous best MAGVIT in these evaluations. Specifically, when using the decoded tokens as input, the performance approaches that of the model trained

Table 5: **Ablation study verifying key design choices**.

(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**

(b) Image tokenization on ImageNet 128×128.

	FID↓	LPIPS↓
MAGVIT	2.65	0.1292
+ LFQ	2.48	0.1182
+ large vocabulary	1.34	0.0821
+ up/downsampler	1.21	0.0790
+ deeper model	1.20	0.0686
+ adaptive normalization	**1.15**	**0.0685**

(c) Video tokenization on UCF-101.

	FVD↓	LPIPS↓
MAGVIT	24.55	0.0988
+ LFQ & large vocabulary	16.12	0.0694
+ up/downsampler	15.37	0.0678
+ late temporal downsample	11.11	0.0653
+ deeper model	8.90	0.0542
+ 3D blur pooling	**8.62**	**0.0537**

with ground-truth pixels using the same ViViT backbone. While these numbers are still worse than the state-of-the-art in action recognition, they represent solid improvements credited to the new tokenizer.

4.5 Ablation Study

In Fig. 1, we have ablated LFQ *vs*. VQ and the vocabulary size. In Tab. 5, we validate the key designs proposed in Section 3.2. Specifically, Tab. 5a compares the architecture illustrated in Fig. 2; Tab. 5b and Tab. 5c verify the LFQ and other improvements on ImageNet and UCF-101, respectively.

5 Related Work

Visual tokenization. Beyond the VQ-VAE models discussed in Section 2, additional models have been proposed. ViT-VQGAN (Yu et al., 2022a) introduces transformer blocks as a substitute for CNNs for image tokenization. C-ViViT (Villegas et al., 2022) further extends this idea for video tokenization. Early studies on video tokenization treat frames as independent images with no temporal compression (Wu et al., 2022; Gupta et al., 2022). Later research (Yan et al., 2021; Ge et al., 2022; Yu et al., 2023a) integrates 3D CNNs to tokenize spatial-temporal volumes. Despite these advances in vector quantization (VQ), the codebook learned by previous VQ models is relatively small (*e.g.*, 8k) due to the difficulty in improving the generation quality with larger vocabularies. In contrast, our tokenizer can induce a large vocabulary (*e.g.*, 262k) that can be effectively modeled by an LM, leading to enhanced image and video generation quality.

Text-to-{image, video}. Text-to-image and text-to-video generation has garnered significant rapid advancements using both language models (Yu et al., 2023b; Chang et al., 2023) and diffusion models (Ho et al., 2022a; Blattmann et al., 2023; Singer et al., 2022; Ge et al., 2023; Ramesh et al., 2022). Although diffusion models, such as Midjourney, are considered the top performers in these tasks, it is unclear whether their advantage stems from the model, data, or some other unidentified factors. Indeed, it is challenging to scientifically compare these text-to-image models as they are trained on varied datasets, with some even being proprietary data, under inconsistent training conditions. To facilitate a fairer comparison, this paper prioritizes using the ImageNet and Kinetics benchmarks.

Diffusion models. Exhibiting high quality sampling, pixel-space diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) raised to the top of the generative modeling space for both image (Ho et al., 2020; Dhariwal & Nichol, 2021; Saharia et al., 2022) and video (Ho et al., 2022c;a; Singer et al., 2022) synthesis. The pixel-space denoising diffusion models (DDMs) are later refined by the latent-space DDM (Rombach et al., 2022), which conducts diffusion over the *continuous* latent embeddings derived from a pre-trained variational autoencoder (VAE). Binary latents for image modeling were used in Wang et al. (2023), where the diffusion process is parameterized with Bernoulli distributions. Recent studies have identified advantages in substituting the U-Net (Ronneberger et al., 2015) denoising backbone with a Transformer (Peebles & Xie, 2022; Jabri et al., 2023) or a hybrid of both (Hoogeboom et al., 2023), making the distinctions between diffusion and language models in visual generation more blurred, with a key distinction being their latent format — continuous for diffusion and discrete for language models.

6 Conclusion and Future Work

We introduce MAGVIT-v2, a novel video tokenizer that exploits lookup-free quantization along with architectural advancements to tokenize images and video with a shared vocabulary. The experiments show that our tokenizer outperforms the previously leading video tokenizer across three areas: visual generation, video compression, and action recognition in videos. Our results suggest that a good visual tokenizer is key for enabling language models to excel in image and video generation. These results demonstrate the great capabilities of LMs in visual generation, and advocate for further exploration of advanced visual tokenization methods designed for LLMs.

ACKNOWLEDGMENTS

We would like to express our gratitude to Yu-Chuan Su and Sergey Ioffe for their valuable comments on our work, to Josh Dillon for discussions, to Vighnesh Birodkar for help in baseline implementation, and to Eirikur Agustsson for help in compression evaluation.

REFERENCES

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating music from text. *arXiv:2301.11325*, 2023. 1

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *ICCV*, 2019. 4, 8

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, 2021. 8, 17

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2021. 2, 17

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 9

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018. 19

Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 2, 8

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about Kinetics-600. *arXiv:1808.01340*, 2018. 6

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7, 19

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 2, 9

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 2

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. 6

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022. 2

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 2, 6

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 3, 7, 9, 19

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. GLaMs: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022. 2

Arpad E. Elo and Sam Sloan. *The rating of chessplayers : past and present*. Ishi Press International, 2008. 8, 17

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1, 3, 4, 19

Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860. 17

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *arXiv:2303.14389*, 2023. 1, 15, 19

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *ECCV*, 2022. 3, 7, 9

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv preprint arXiv:2305.10474*, 2023. 9

Google. PaLM 2 technical report. *arXiv:2305.10403*, 2023. 1

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 6

Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked visual pre-training for video prediction. In *ICLR*, 2022. 9

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 7, 15, 19

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 9

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022a. 3, 9

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(1):2249–2281, 2022b. 19

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *ICLR Workshops*, 2022c. 3, 7, 9

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 2023. 7, 9, 19

Allan Jabri, David J Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *ICML*, 2023. 7, 9, 19

Aren Jansen, Daniel PW Ellis, Shawn Hershey, R Channing Moore, Manoj Plakal, Ashok C Popat, and Rif A Saurous. Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision. In *ICASSP*, 2020. 4

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 6

Diederik P Kingma and Ruiqi Gao. Understanding the diffusion objective as a weighted integral of elbos. *arXiv:2303.00848*, 2023. 7, 19

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and WOOK SHIN HAN. Draft-and-revise: Effective image generation with contextual rq-transformer. In *NeurIPS*, 2022. 1, 19

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 2

José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with Token-Critic. In *ECCV*, 2022. 19

José Lezama, Tim Salimans, Lu Jiang, Huiwen Chang, Jonathan Ho, and Irfan Essa. Discrete predictor-corrector diffusion models for image synthesis. In *ICLR*, 2023. 6, 7, 15, 19

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. StarCoder: may the source be with you! *arXiv:2305.06161*, 2023. 1

Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv:2003.04035*, 2020. 7

Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. In *ICLR*, 2021. 2

OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023. 1

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv:2212.09748*, 2022. 3, 7, 9, 19

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *EACL*, 2017. 5

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 9

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 9, 19

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 9

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. AudioPaLM: A large language model that can speak and listen. *arXiv:2306.12925*, 2023. 1

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 9

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 7

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022. 7, 19

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv:1911.02150*, 2019. 2

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022. 9

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv:2305.09617*, 2023. 1

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3, 9

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3, 9

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 6

Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. 2, 8

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1

Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *CVPR*, 2021. 4

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv:1812.01717*, 2018. 6

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 3

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv:2210.02399*, 2022. 2, 3, 4, 5, 7, 9

Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset. In *ICIP*, 2016. 6, 17

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: BERT pretraining of video transformers. In *CVPR*, 2022. 2, 8, 17, 18

Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *CVPR*, 2023. 9, 19

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 8

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜWA: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 9

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv:2104.10157*, 2021. 9

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQ-GAN. In *ICLR*, 2022a. 4, 9

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv:2206.10789*, 2022b. 2

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. MAGVIT: Masked generative video transformer. In *CVPR*, 2023a. 1, 2, 3, 4, 6, 7, 8, 9, 15, 19

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv:2309.02591*, 2023b. 9

Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019. 5

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv:2306.09305*, 2023. 19

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 1