A Implementation Details

A.1 Image and Video Generation

We set up two image tokenizers to downsample by 16× and 32×, where they are used for generation at 256×256 and 512×512, respectively. In both cases, an image is represented as 16×16 tokens. We train them on the ImageNet training set for 270 epochs using a batch size of 256, both with 256×256 images.

With this tokenizer we train a Masked Language Model following Yu et al. (2023a), using the token factorization described in Section 3.2. We train for 1080 epochs in accordance with the prior best model MDT (Gao et al., 2023), with batch size 1024 for better efficiency. For preprocessing and data augmentation, we randomly crop 80-100% of an image while keeping the aspect ratio, followed by random horizontal flipping. The class label is dropped for 10% of the training batches to enable classifier-free guidance (Ho & Salimans, 2021). For unguided generation, we use temperature 30 for 512×512 and 15 for 256×256 in the non-autoregressive decoding. For guided generation, we adopt the guidance schedule from Gao et al. (2023) with temperature scaling (Lezama et al., 2023), where we use guidance scale 25 with temperature 15.

We inflate an image tokenizer trained at 128×128 for video modeling. Different from the inflation in Yu et al. (2023a), we fill in the temporally last slice to correspond to the causal padding scheme. In addition, we disable the inflation for the discriminator and train it from scratch for better stability. We train the causal video tokenizer on Kinetics-600 training set for 190 epochs with batch size 256. This tokenizer is also used in subsequent evaluations of video compression and action recognition.

With the causal tokenizer producing 5×16×16 tokens for a 17×128×128 clip, the first 2×16×16 tokens are provided as the condition of the first 5 frames, per the standard setup of Kinetics-600 frame prediction benchmark. We train the MLM transformer following Yu et al. (2023a) with token factorization for 360 epochs with batch size 256. The model is sampled with a cosine schedule using temperature 32.

A.2 Model Setup and Hyperparameters

Fig. 7 illustrates the architecture of our proposed MAGVIT-v2. We provide detailed training hyper-parameters for our models as listed below:

- Video input: 17 frames, frame stride 1, 128×128 resolution.
- Base channels: 128.
- VQVAE channel multipliers: $1, 2, 2, 4$.
- Discriminator channel multipliers: $2, 4, 4, 4, 4$.
- Number of residual blocks: 4.
- Latent shape: $5 \times 16 \times 16$.
- Vocabulary size: 2^{18}.
- Initialization: central inflation from a 2D model trained on ImageNet with this setup.
- Entropy loss weight: 0.1.
- Entropy loss annealing steps: 2000.
- Entropy loss annealing factor: 3.
- Reconstruction loss weight: 5.0.
- Generator loss type: Non-saturating.
- Generator adversarial loss weight: 0.1.
- Discriminator gradient penalty: r1 with cost 10.
- Perceptual loss weight: 0.1.
- Commitment loss weight: 0.25.
- LeCAM weight: 0.001.
- Peak learning rate: 10^{-4}.
- Learning rate schedule: linear warm up and cosine decay.
- Optimizer: Adam with $\beta_1 = 0$ and $\beta_2 = 0.99$.
- EMA model decay rate: 0.999.
- Batch size: 256.

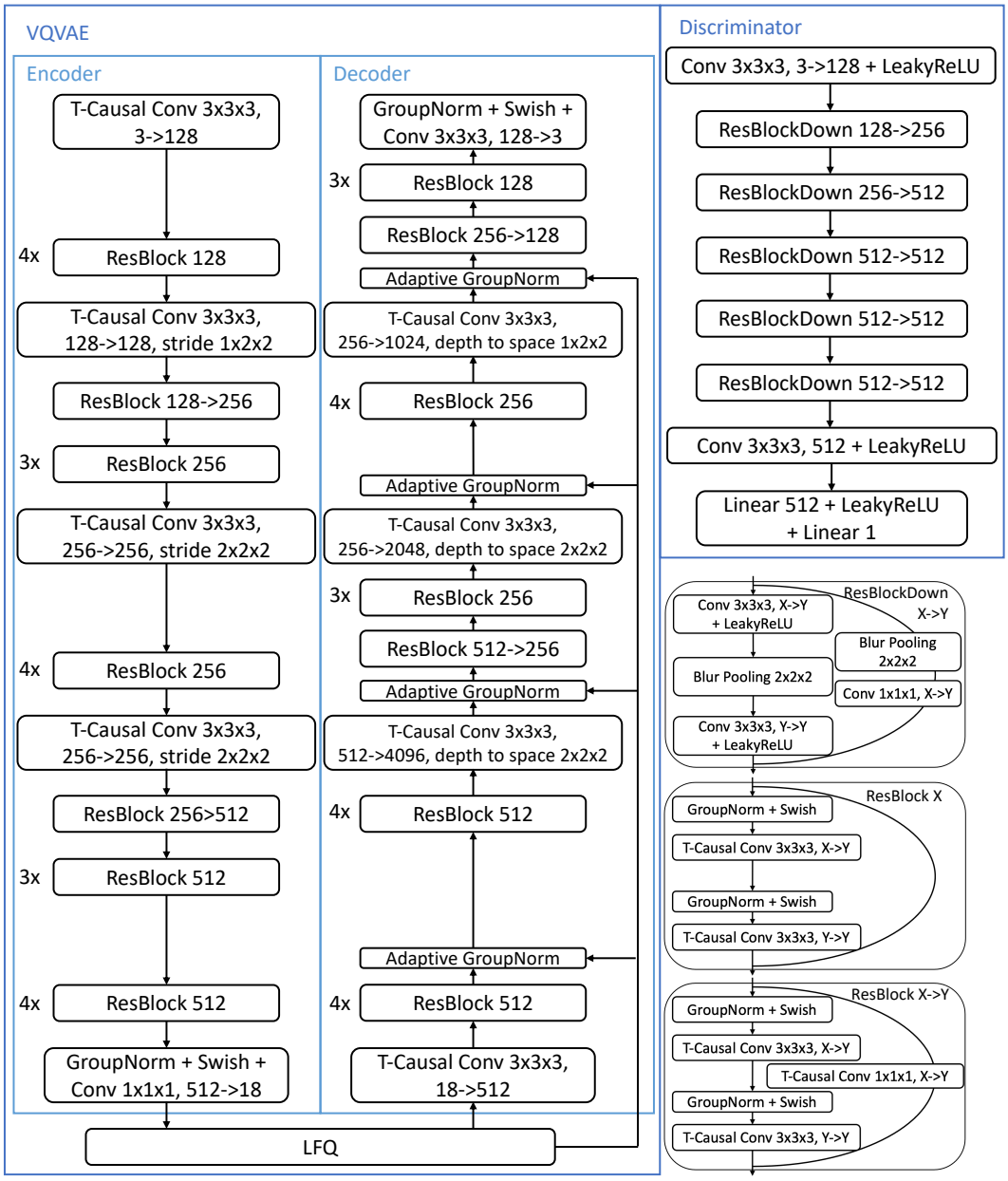


Figure 7: **MAGVIT-v2 tokenizer architecture**. T-Causal Conv refers to temporally causal convolution.

A.3 VIDEO COMPRESSION EVALUATION

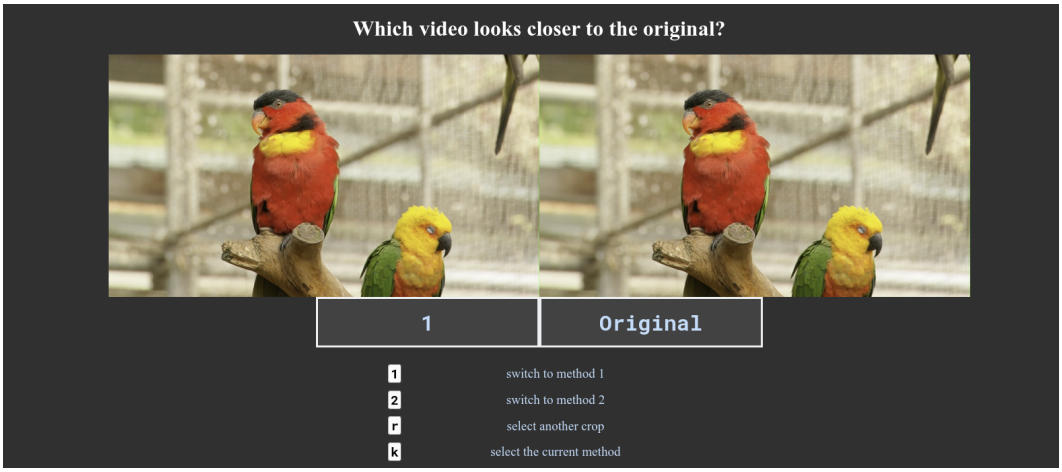


Figure 8: **Rating interface for subjective compression evaluation**.

To rate the quality of the different methods, we use a two-alternative forced choice rating methodology (Fechner, 1860). As this methodology produces a sequence of binary decisions, we calculate Elo scores (Elo & Sloan, 2008) based on pairwise preferences to quantify the relative visual quality between the models. The study was conducted on the 30 videos of the MCL-JCV dataset (Wang et al., 2016), scaled down to a resolution of 640×360 pixels. Sixteen raters are engaged, each providing responses to an average of roughly 800 pairwise-preference questions. The questions are presented with an interface that parallels the one used for the Challenge on Learned Image Compression (`http://compression.cc/`), extended to comparing videos, as shown in Fig. 8. Raters are instructed to compare the two videos and are not allowed to pause the videos.

A.4 VIDEO UNDERSTANDING EXPERIMENTS

Tokens as prediction targets. BEiT (Bao et al., 2021) and BEVT (Wang et al., 2022) class of models pretrain visual encoders on pixel inputs by predicting tokens as targets in a masked-modeling framework, and demonstrate state-of-the-art downstream results. We use a simplified BEVT pre-training setup to test the effectiveness of our video tokens as targets for masked modeling. The main difference is that we drop the image-stream from pre-training and only use the video stream and for this reason, we also drop the multiple decoders completely and adopt an encoder-only architecture similar to BEiT. Detailed pre-training and fine-tuning setup is presented in Tab. 6. In Tab. 4 of the main paper, we show that our video tokens are effective targets for masked modeling based video understanding.

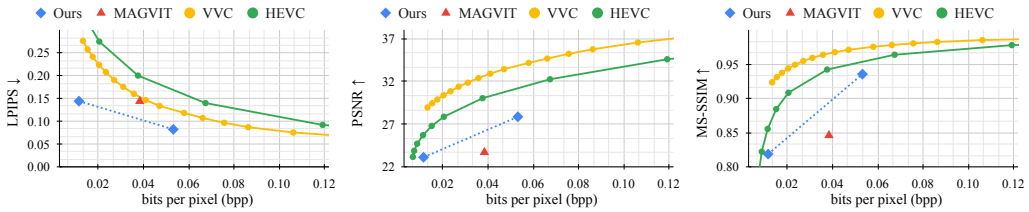
Tokens as inputs. In Tab. 4, we show that we can re-use video understanding models trained on pixels using our video tokens as input, with very minimal performance drop. For this experiment, we train a factorized variant of the ViViT model (Arnab et al., 2021) on pixels, and evaluate it on de-tokenized pixels from our model. We use the same hyper-parameters as used in Arnab et al. (2021) with a Base sized model operating on 32 frames of inputs at 224p resolution. For the Kinetics-600 experiment, we use the same hyper-parameters as the Kinetics-400 experiments.

B ADDITIONAL RESULTS

For better visualization, the generated video samples can be viewed at `https://magvit.cs.cmu.edu/v2`.

Table 6: **Experimental configurations with tokens as targets**.

Config	SSv2 Pre-Training	SSv2 Fine-tuning
inputs	pixels	pixels
input size	$16 \times 224 \times 224 \times 3$	$16 \times 224 \times 224 \times 3$
targets	tokens	classes
encoder	ViT-B	ViT-B
decoder	linear	linear
masking	block-tube (Wang et al., 2022)	none
masking ratio	0.75	0.0
mask temporal length	16	0
batch size	1024	512
training epochs	800	50
ViT sequence length	$8 \times 16 \times 16$	$8 \times 16 \times 16$
optimization		
optimizer	AdamW	AdamW
optimizer momentum	0.9	0.9
layer decay	0.75	0.75
weight decay	0.05	0.05
learning rate schedule	cosine decay	cosine decay
warmup epochs	40	5
data augmentations		
random horizontal flip	true	false
label smoothing	0.1	0.1
mixup	none	0.8
cutmix	none	1.0
droppath	0.0	0.1
dropout	0.1	0.0
random color augmentation	false	false



(a) LPIPS↓ (b) PSNR↑ (c) MS-SSIM↑

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

Where are the text-to-image results? We want to emphasize that our goal is to develop a video tokenizer, and many of the proposed techniques are designed specifically for videos. Text-to-image may be out of the scope of our paper. We are currently training text-to-video models that require considerable computational resources. Due to time constraints, these results are not available at the moment. We intend to add the generated videos in the next revision. However, it is important to note that comparing these text-to-image or text-to-video models scientifically is challenging. These models were trained on different datasets, and some were even based on proprietary or non-public data, all under varying training conditions.

Table 7: **Class-conditional image generation on ImageNet 256×256**. 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	BigGAN-deep (Brock et al., 2018)	6.95	171.4			160M	1
GAN	StyleGAN-XL (Sauer et al., 2022)			2.30	265.1	166M	1
Diff. + VAE*	LDM-4 (Rombach et al., 2022)	10.56	103.5	3.60	247.7	400M	250
Diff. + VAE*	DiT-XL/2 (Peebles & Xie, 2022)	9.62	121.5	2.27	278.2	675M	250
Diff. + BAE	Binary latent diffusion (Wang et al., 2023)	8.21	162.3			172M	64
Diffusion	ADM+Upsample (Dhariwal & Nichol, 2021)	7.49	127.5	3.94	215.8	608M	2000
Diff. + VAE*	MDT (Gao et al., 2023)	6.23	143.0	1.79	283.0	676M	250
Diff. + VAE*	MaskDiT (Zheng et al., 2023)	5.69	178.0	2.28	276.6	736M	40
Diffusion	CDM (Ho et al., 2022b)	4.88	158.7				8100
Diffusion	RIN (Jabri et al., 2023)	3.42	182.0			410M	1000
Diffusion	simple diffusion (Hoogeboom et al., 2023)	2.77	211.8	2.44	256.3	2B	512
Diffusion	VDM++ (Kingma & Gao, 2023)	2.40	225.3	2.12	267.7	2B	512
AR-LM + VQ	VQGAN (Esser et al., 2021)	15.78	78.3			1.4B	256
MLM + VQ	MaskGIT (Chang et al., 2022)	6.18	182.1			227M	8
MLM + VQ	Token-Critic (Lezama et al., 2022)	4.69	174.5			368M	36
MLM + VQ	Contextual RQ-Transformer (Lee et al., 2022)	3.41	224.6			1.4B	72
MLM + VQ	DPC (Lezama et al., 2023)	4.45	244.8			454M	180
MLM + LFQ	*MAGVIT-v2 (this paper)*	3.65	200.5	**1.78**	**319.4**	307M	64

Table 8: **Video generation results**: class-conditional generation on UCF-101 with AR-LM models. We use the same transformer configuration as MLM experiments but without vocabulary factorization and weight tying. As a result, the AR-LM with MAGVIT-v2 uses more parameters in the embedding table and the softmax layer.

Tokenizer	FVD↓	#Params	#Steps
MAGVIT (Yu et al., 2023a)	265	306M	1024
MAGVIT-v2 (this paper)	**109**	840M	1280