A Architecture Details and Hyperparameters

A.1 VQ-VAE Encoder and Decoder

Table 1: Hyperparameters of VQ-VAE encoder and decoder models for Moving MNIST, ViZDoom (HGS = Health Gathering Supreme), and BAIR

	Moving MNIST	ViZDoom (HGS)	BAIR / ViZDoom (Battle2)
Input size	$16 \times 64 \times 64$	$16 \times 64 \times 64$	$16 \times 64 \times 64$
Latent size	$4 \times 16 \times 16$	$4 \times 16 \times 16$	$8 \times 16 \times 16$
β (commitment loss coefficient)	0.25	0.25	0.25
Batch size	64	64	64
Learning rate	7×10^{-4}	7×10^{-4}	7×10^{-4}
Hidden units	240	240	240
Residual units	128	128	128
Residual layers	2	4	4
Uses attention	No	Yes	Yes
Codebook size	512	2048	2048
Codebook dimension	64	256	256
Encoder filter size	3	3	3
Upsampling conv filter size	4	4	4
Training steps	20k	75K	75k

A.2 Prior Networks

Table 2: Hyperparameters of prior networks for Moving MNIST, ViZDoom (HGS), BAIR and ViZDoom (Battle2).

	Moving MNIST	ViZDoom (HGS)	BAIR	ViZDoom (Battle2)
Input size	$4 \times 16 \times 16$	$4 \times 16 \times 16$	$8 \times 16 \times 16$	$8 \times 16 \times 16$
Conditional sizes	$1 \times 64 \times 64$	60	$3 \times 64 \times 64$, 64	315
Batch size	32	32	32	32
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Vocabulary size	512	2048	2048	2048
Attention heads	4	4	4	4
Attention layers	8	22	22	22
Resnet depth	18	None	34	None
Resnet units	512	None	512	None
Dropout	0.1	0.1	0.1	0.1
Training steps 40k	80k	80k	80k	80k