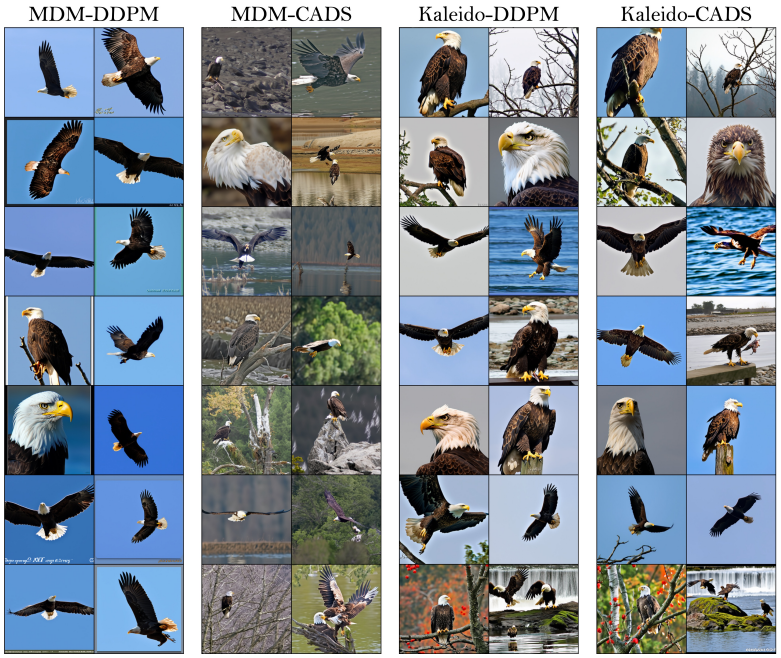

Fig.R.1: Comparison with CADS on ImageNet (class: *Bald Eagle*).

Fig.R.2: Comparison with CADS on T2I (prompt: *a cat sleeping on the bed*).

Fig.R.3: Comparison with MDM directly using synthetic captions as condition (prompt: *a raccoon playing piano*)

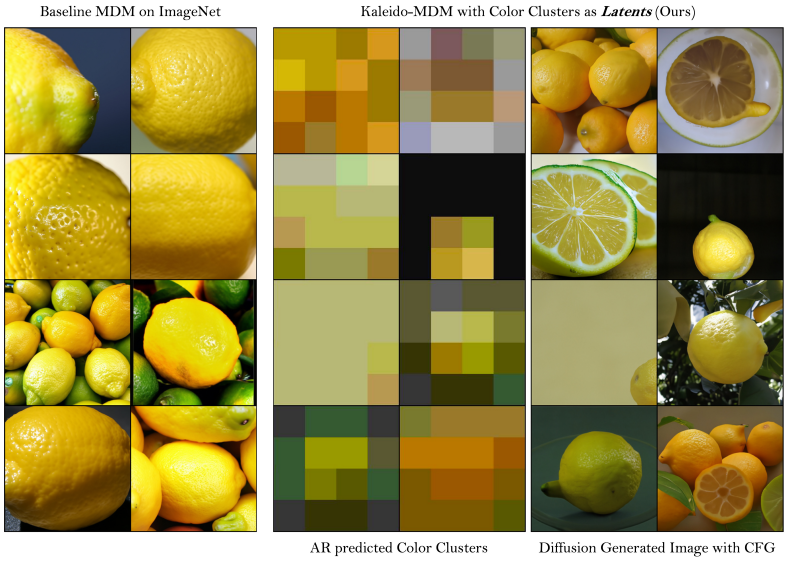

Fig.R.4: Kaleido-MDM with color clusters as latents on ImageNet (class: *lemon*)

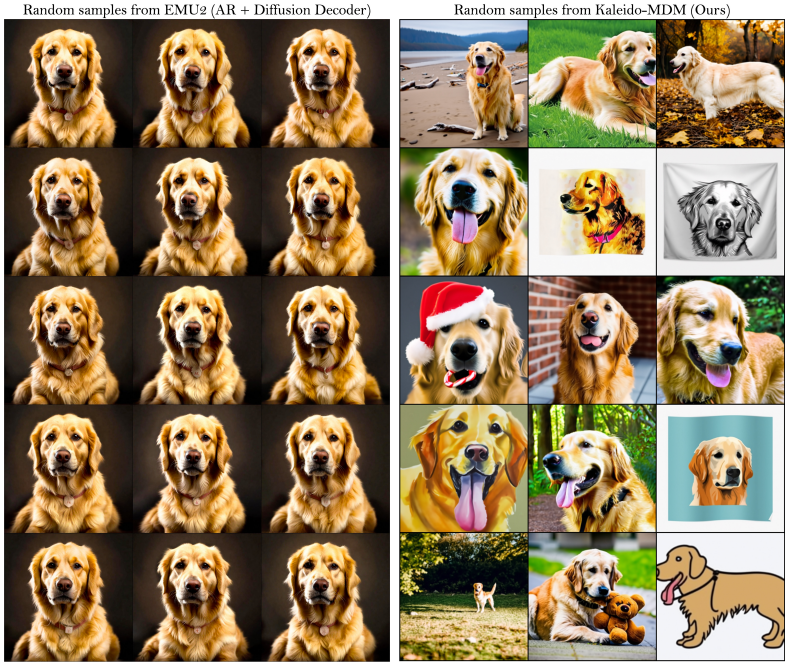

Fig.R.5: Comparison with AR+Diffusion decoder (prompt:*a golden retriever*)

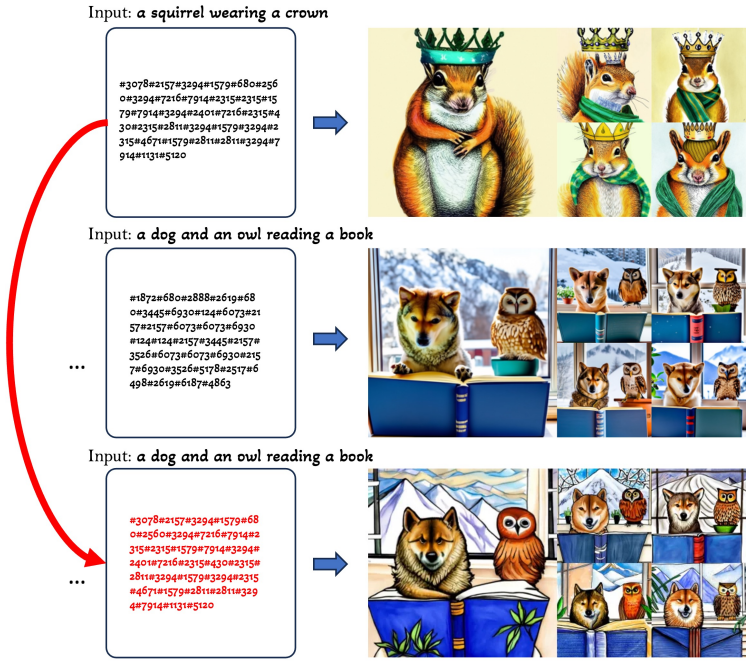

Fig.R.6: Image editing by altering the autoregressively predicted visual tokens.