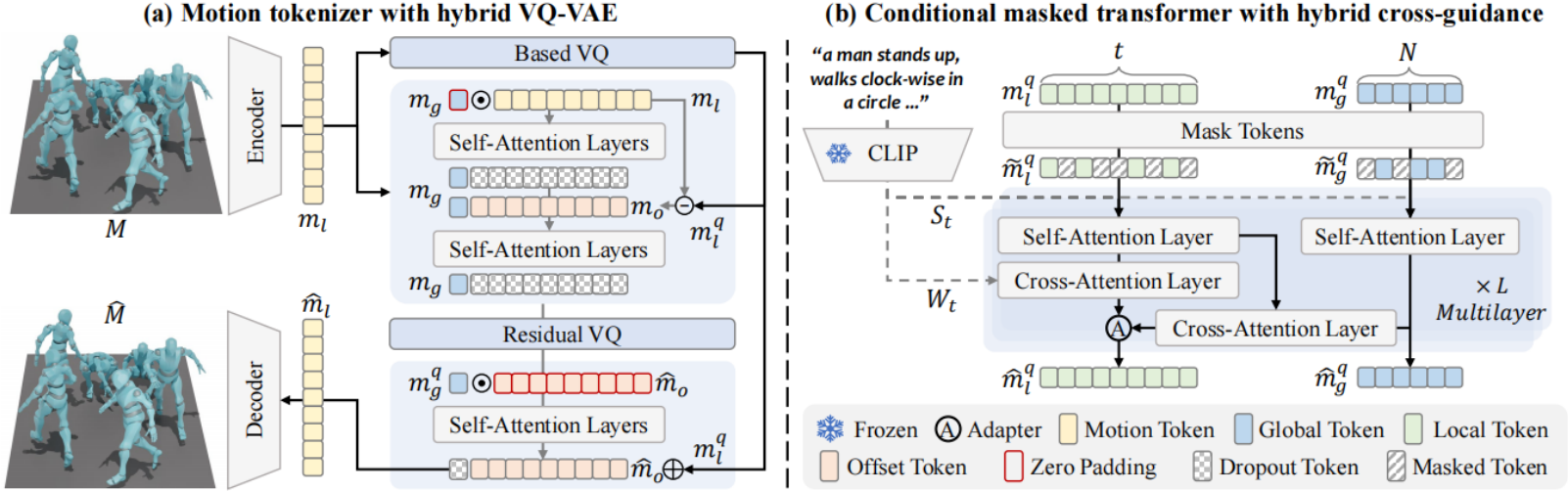
Generative Masked Text-to-Motion Model with Hybrid Vector Quantization

Jiaqi Zhang, Jiajun Wang, Fanglue Zhang, Miao Wang

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50904-0

Problems & Ideas

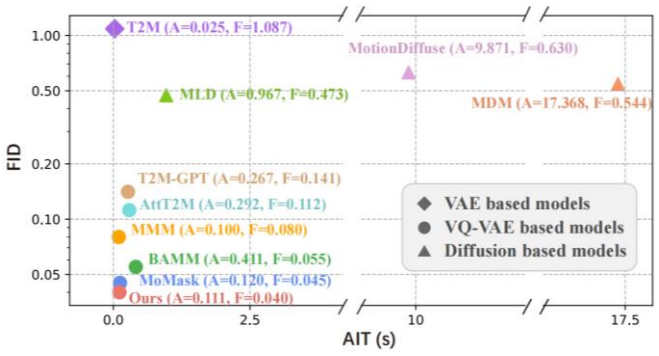
- Problems of conventional Text-Driven Motion Generation:
  - Methods based on diffusion models exhibit faster generation speeds, yet tend to compromise on generating quality.
  - Methods based VQ-VAE achieve higher generation quality but necessitate a substantial increase in the length of token sequences.
- Ideas: A hybrid VQ-VAE utilizes RVQ for global features to mitigate errors from the basic VQ-VAE, ensuring high-quality generation without substantially increasing token length.



Left Side: Displays the hybrid VQ-VAE, which decomposes motion into global and local components. The local motion is quantized using a based VQ to preserve details, while the global motion is reconstructed through a RVQ to compensate for the inaccuracies introduced by the VQ of the local components. Right Side: Featuring a conditional masked transformer guided by text and a global motion cross-attention module.

Main Contributions

- Contributions:
  - An advanced motion quantization approach (Hybrid VQ-VAE) that separats motion into global and local components, with Residual VQ applied to global motion and basic VQ-VAE to local motion;
  - The Hybrid VQ-VAE improves the quality of motion quantization while maintaining a minimal increase in code length;
  - A conditional masked transformer model that leverages high-quality reconstructions from a hybrid VQ-VAE and employs mixed cross-guidance to accurately predict global and local motion token labels.



Methods	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →	MModality ↑
		Top-1 ↑	Top-2 ↑	Top-3 ↑			
Real	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
TM2T [35]	$1.501^{\pm.017}$	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
T2M [5]	$1.087^{\pm.021}$	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	$2.219^{\pm.074}$
MDM [6]	$0.544^{\pm.044}$	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
MotionDiffuse [8]	$0.630^{\pm.001}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
MLD [13]	$0.473^{\pm.013}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
Fg-T2M [7]	$0.243^{\pm.019}$	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$	$1.614^{\pm.049}$
M2DM [34]	$0.352^{\pm.005}$	$0.497^{\pm.003}$	$0.682^{\pm.002}$	$0.763^{\pm.003}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$	$3.587^{\pm.072}$
T2M-GPT [14]	$0.116^{\pm.004}$	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
AttT2M [33]	$0.112^{\pm.006}$	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
MotionLCM [9]	$0.304^{\pm.012}$	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$3.012^{\pm.007}$	$9.607^{\pm.066}$	$2.259^{\pm.092}$
MMM [15]	$0.089^{\pm.005}$	$0.515^{\pm.002}$	$0.708^{\pm.002}$	$0.804^{\pm.002}$	$2.926^{\pm.007}$	$9.577^{\pm.050}$	$1.226^{\pm.035}$
MoMask [16]	$0.045^{\pm.002}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$2.958^{\pm.008}$	$9.636^{\pm.068}$	$1.241^{\pm.040}$
BAMM [19]	$0.055^{\pm.002}$	$0.522^{\pm.003}$	$0.715^{\pm.003}$	$0.808^{\pm.003}$	$2.936^{\pm.077}$	$9.636^{\pm.009}$	$1.732^{\pm.055}$
HyT2M	$0.040^{\pm.002}$	$0.556^{\pm.003}$	$0.749^{\pm.002}$	$0.839^{\pm.002}$	$2.723^{\pm.008}$	$9.607^{\pm.057}$	$1.042^{\pm.038}$

The left side shows a comparison of generation quality and inference speed of various methods on the HumanML3D test set, while the right side presents multiple quantitative metrics for these methods on the same set.