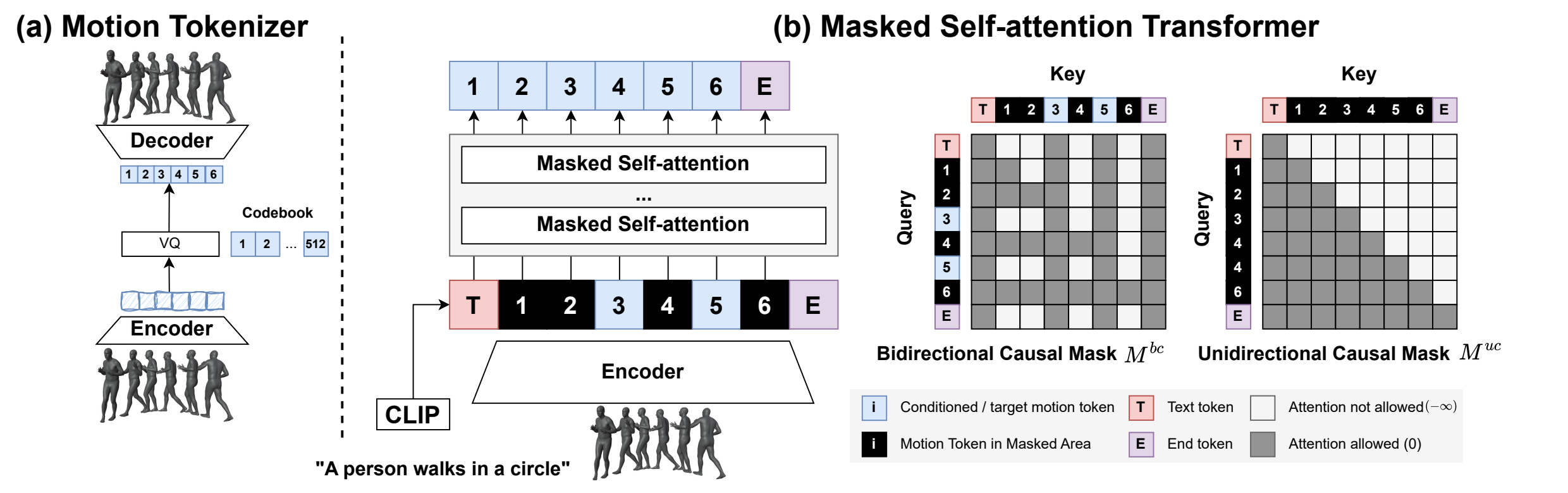
(a) Motion Tokenizer
Decoder
1 2 3 4 5 6
Codebook
VQ
1 2 ... 512
Encoder
1 2 3 4 5 6 E
Masked Self-attention
...
Masked Self-attention
T 1 2 3 4 5 6 E
Encoder
CLIP
"A person walks in a circle"
(b) Masked Self-attention Transformer
Key
Query
T 1 2 3 4 5 6 E
Bidirectional Causal Mask M^{bc}
Unidirectional Causal Mask M^{uc}
i Conditioned / target motion token
i Motion Token in Masked Area
T Text token
E End token
Attention not allowed $(-\infty)$
Attention allowed (0)