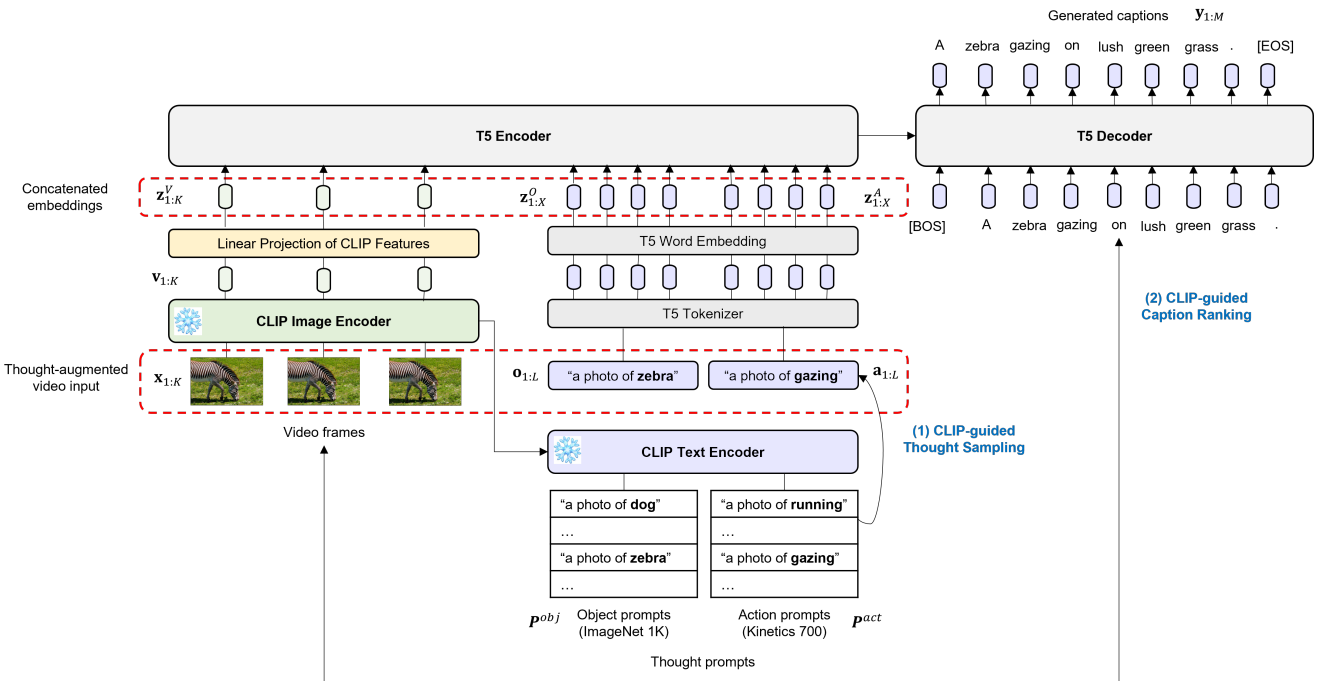

Generated captions $\mathbf{y}_{1:M}$
A zebra gazing on lush green grass . [EOS]
T5 Encoder
T5 Decoder
Concatenated embeddings
$\mathbf{z}^V_{1:K}$
$\mathbf{z}^O_{1:X}$
$\mathbf{z}^A_{1:X}$
[BOS] A zebra gazing on lush green grass .
Linear Projection of CLIP Features
T5 Word Embedding
$\mathbf{v}_{1:K}$
CLIP Image Encoder
T5 Tokenizer
(2) CLIP-guided Caption Ranking
Thought-augmented video input
$\mathbf{x}_{1:K}$
$\mathbf{o}_{1:L}$
"a photo of zebra"
"a photo of gazing"
$\mathbf{a}_{1:L}$
Video frames
(1) CLIP-guided Thought Sampling
CLIP Text Encoder
"a photo of dog"
…
"a photo of zebra"
…
"a photo of running"
…
"a photo of gazing"
…
$\boldsymbol{P}^{obj}$ Object prompts (ImageNet 1K)
Action prompts (Kinetics 700) $\boldsymbol{P}^{act}$
Thought prompts