

This file contains examples with textual descriptions that include unseen words. Each description is displayed in the corresponding video filename. The analysis is provided below, and unseen words are highlighted in red.

1. A person is playing tug-of-war .	
MoMask	The person bends over with some hand gestures, but the motion does not resemble tug-of-war.
BAMM	The person is sliding around and clearly not engaging in a tug-of-war motion.
MATE	The person bends over and appears to pull something with their hands, which is the closest to a tug-of-war motion among the three.
2. A person is pulling a rope backward, as if in a tug-of-war .	
MoMask	With richer contextual information, MATE generates the motion that best aligns with the input text, depicting a person pulling a rope forcefully while slightly moving backward in place.
BAMM	
MATE	
3. A person is closing the curtains .	
MoMask	All three motions fail to resemble the action of closing the curtains.
BAMM	
MATE	
4. A person is pulling the curtains closed from left to right.	
MoMask	With more detailed descriptions, MATE generates a motion that best aligns with the input text, portraying a person pulling the curtains with a clear left-to-right arm movement.
BAMM	
MATE	