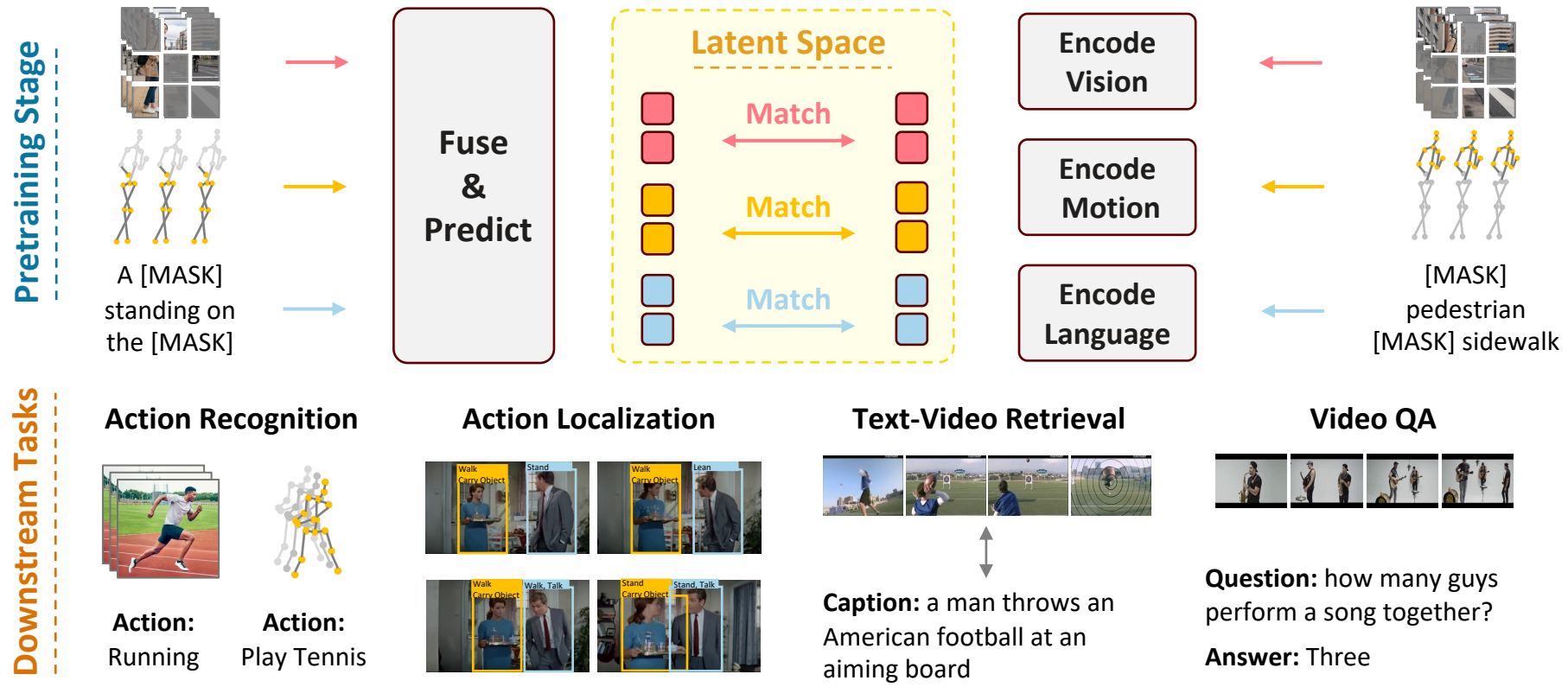
Pretraining Stage
A [MASK] standing on the [MASK]
Fuse & Predict
Latent Space
Match
Match
Match
Encode Vision
Encode Motion
Encode Language
[MASK] pedestrian [MASK] sidewalk
Downstream Tasks
Action Recognition
Action: Running
Action: Play Tennis
Action Localization
Walk
Carry Object
Stand
Walk
Carry Object
Lean
Walk
Carry Object
Walk, Talk
Stand
Carry Object
Stand, Talk
Text-Video Retrieval
Caption: a man throws an American football at an aiming board
Video QA
Question: how many guys perform a song together?
Answer: Three