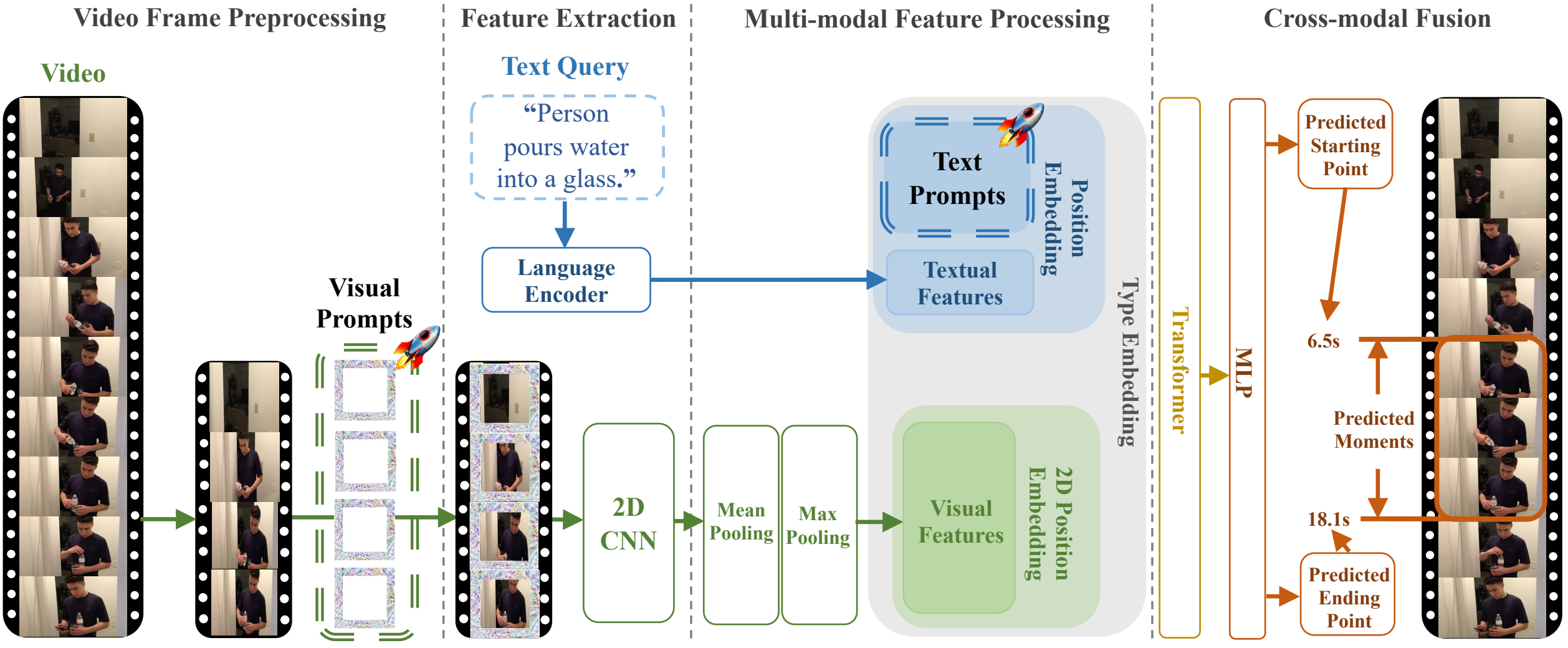
Video Frame Preprocessing
Feature Extraction
Multi-modal Feature Processing
Cross-modal Fusion
Video
Text Query
"Person pours water into a glass."
Language Encoder
Visual Prompts
Text Prompts
Position Embedding
Textual Features
Type Embedding
2D CNN
Mean Pooling
Max Pooling
Visual Features
2D Position Embedding
Transformer
MLP
Predicted Starting Point
6.5s
Predicted Moments
18.1s
Predicted Ending Point