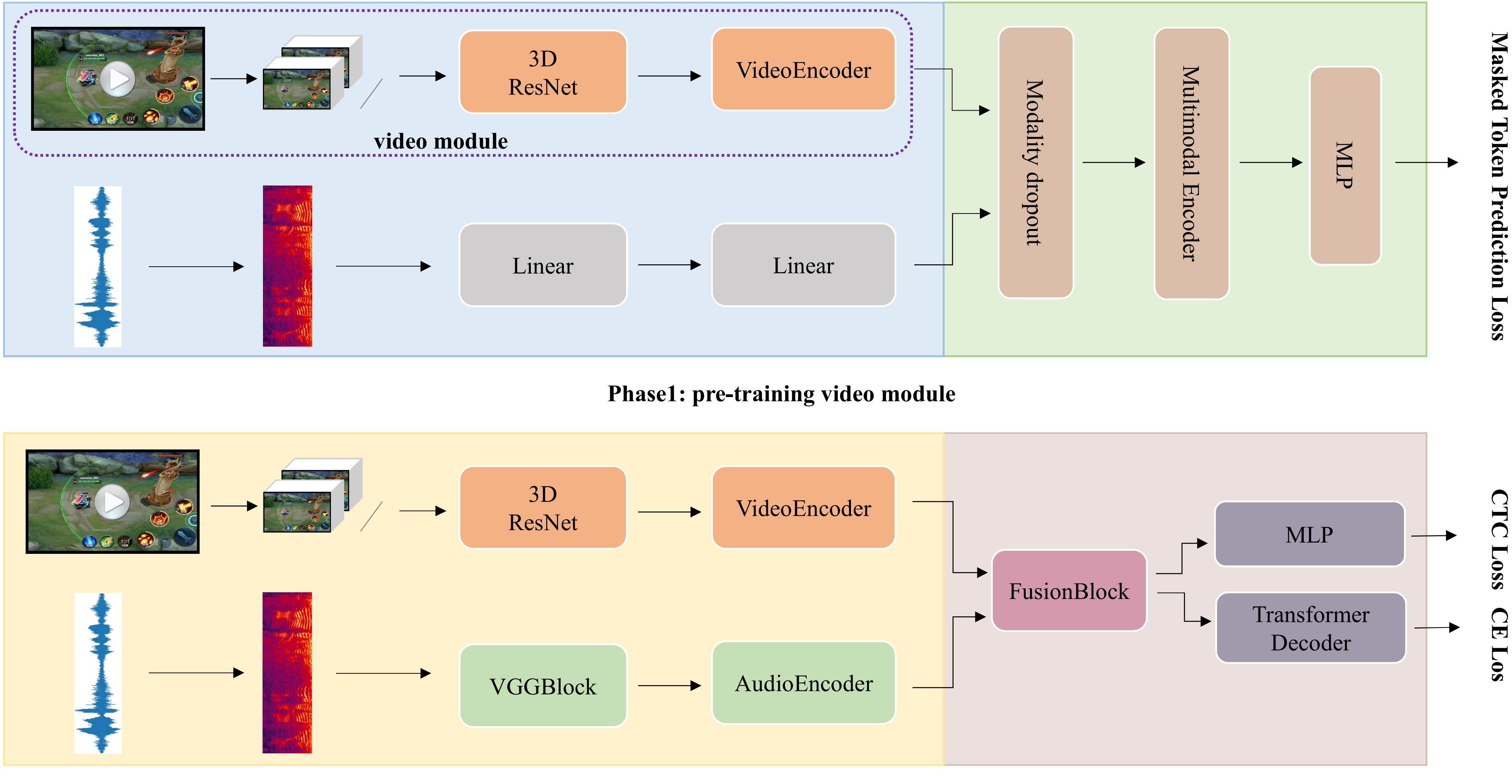

Phase1: pre-training video module

Phase2: fine-tuning speech recognition model