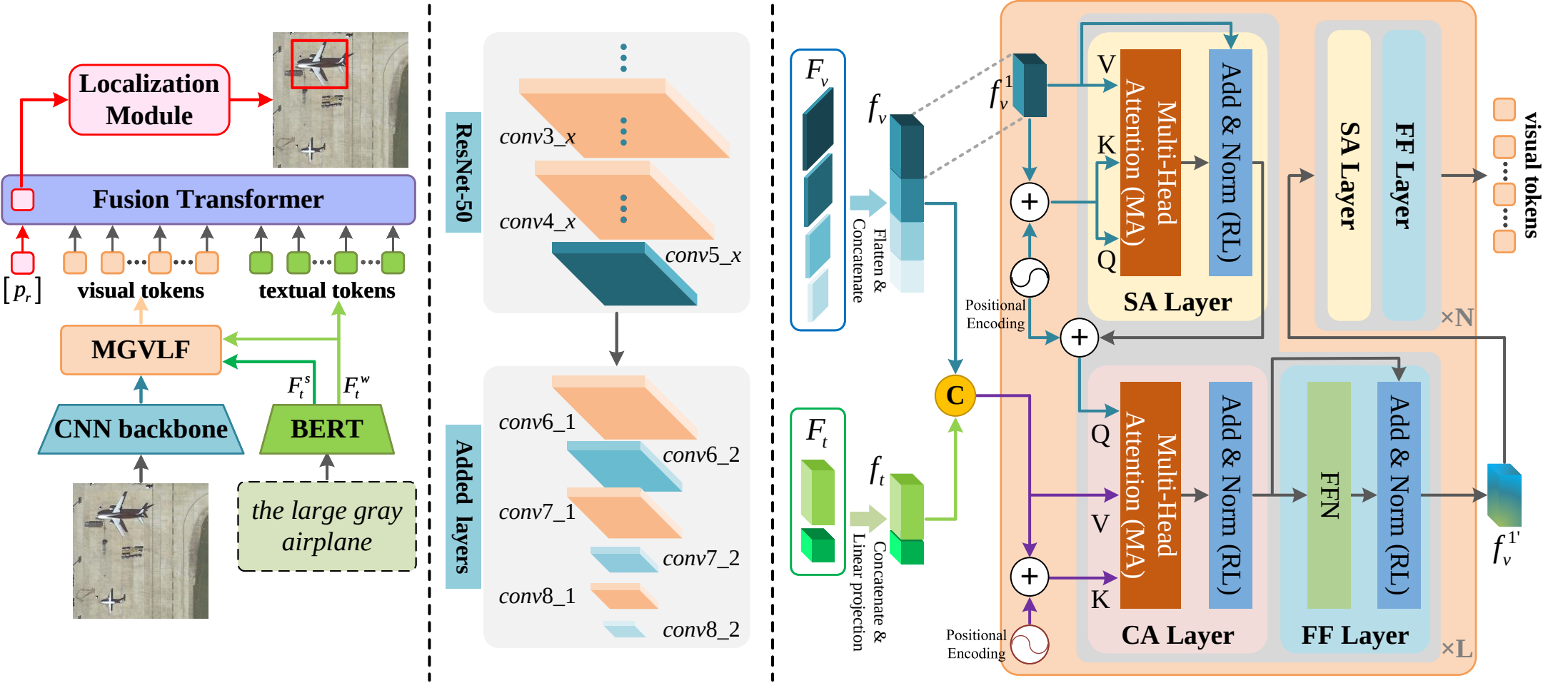

(a) Our overall architecture (b) CNN backbone (c) Multi-Granularity Visual Language Fusion (MGVLF) module