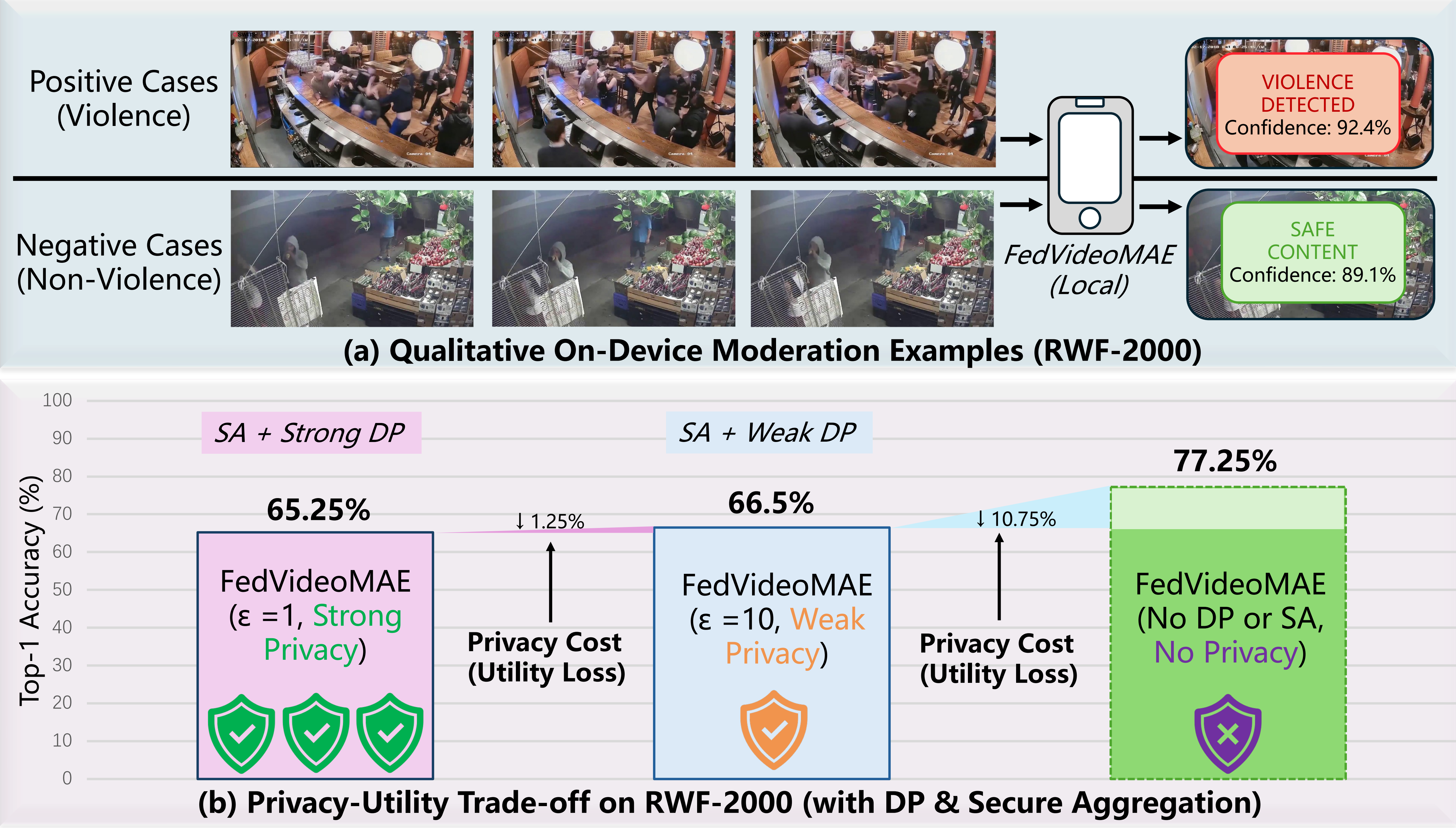

(a) Qualitative On-Device Moderation Examples (RWF-2000)

(b) Privacy-Utility Trade-off on RWF-2000 (with DP & Secure Aggregation)