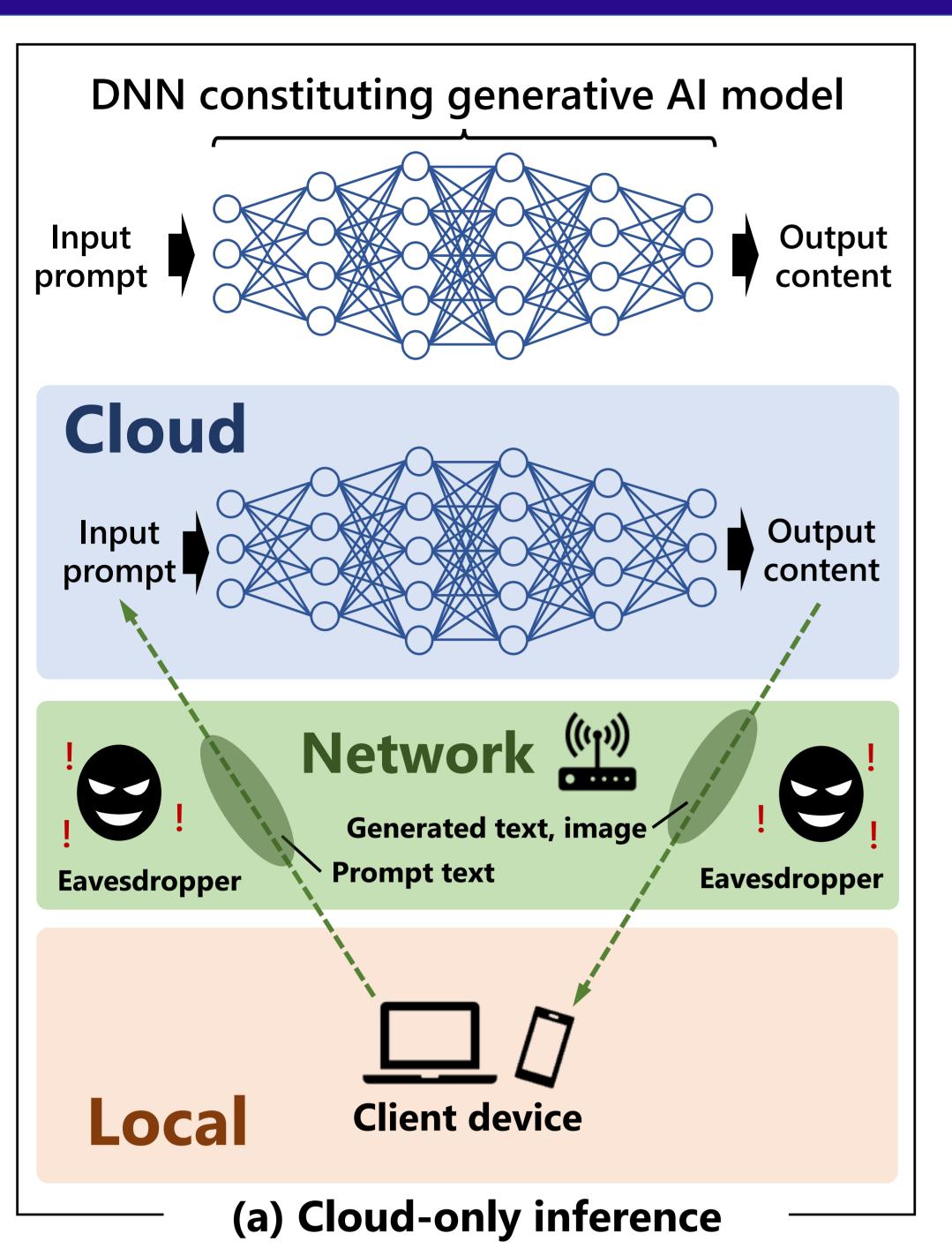
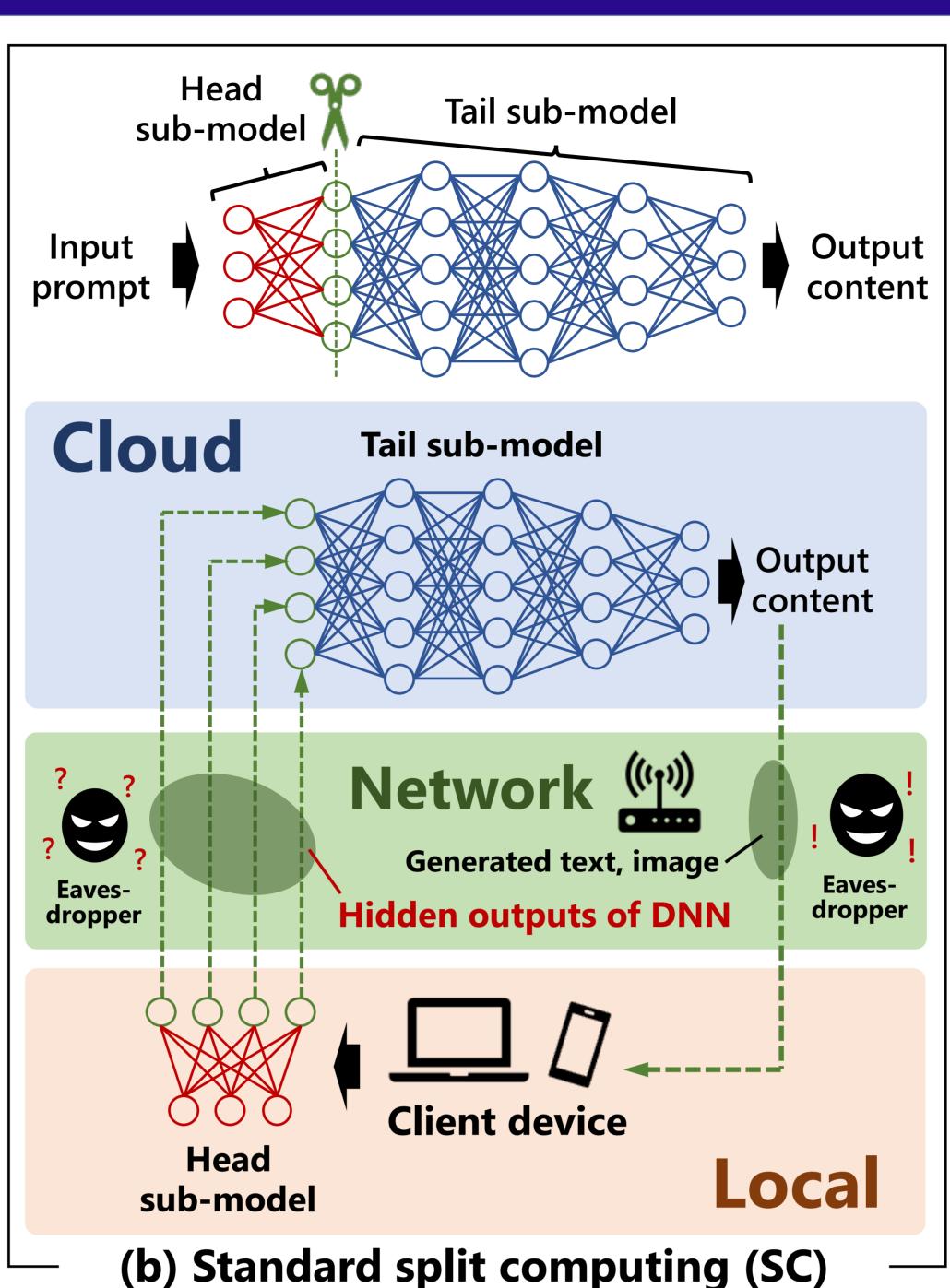
プライバシに配慮した生成AIサービスに向けた 三分割 Split Computing フレームワーク

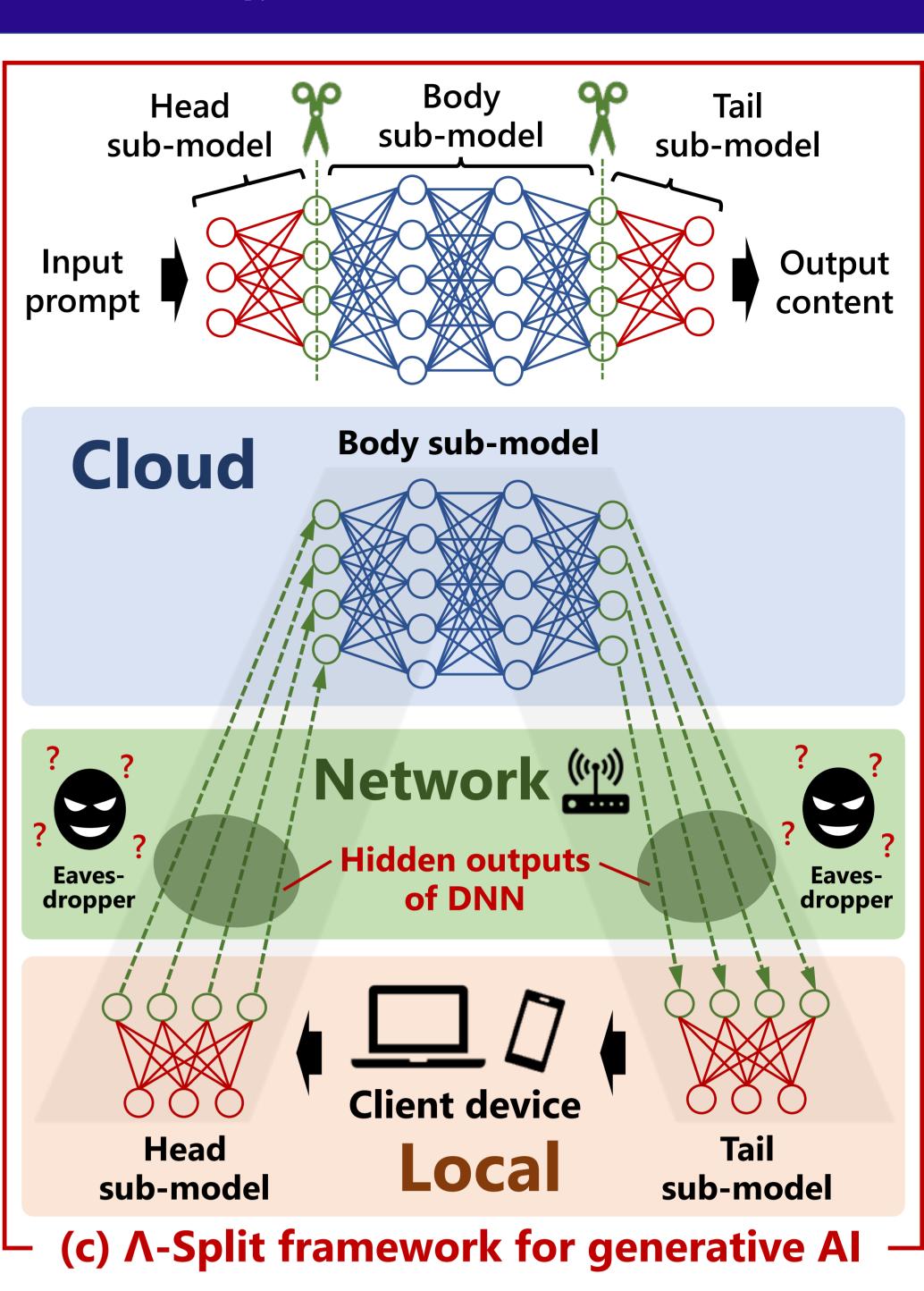


東京工業大学 西尾研究室 修士課程2年 太田翔己

三分割 Split Computing フレームワーク「Λ-Split」・既存手法との比較



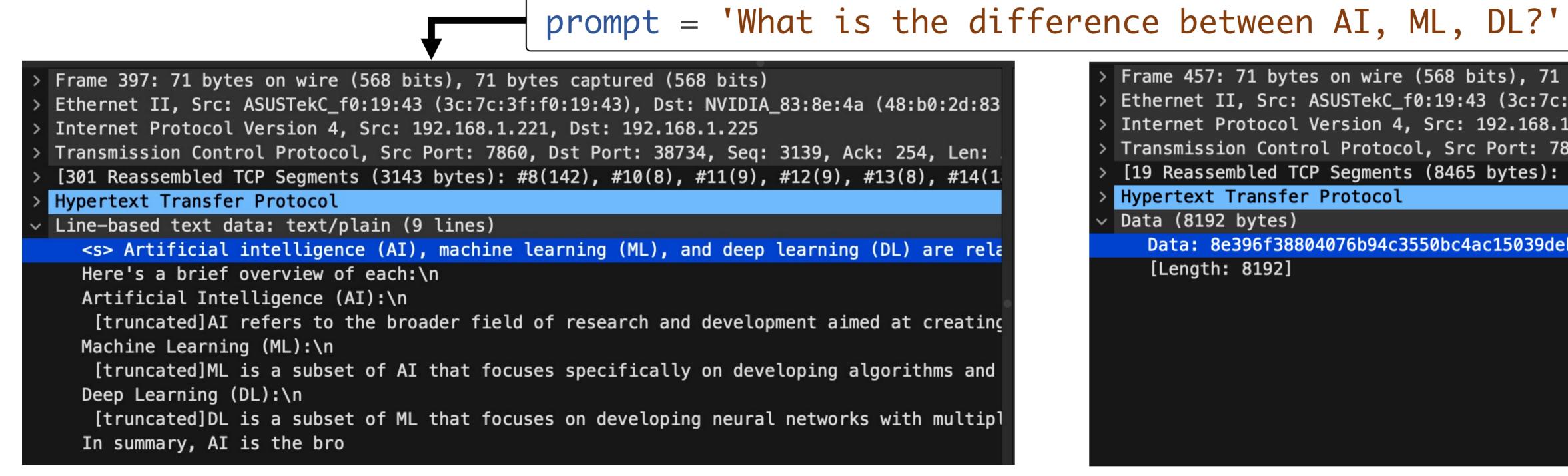




Λ-Split では、送受信されるデータが 全てDNNの中間層出力(hidden outputs)であるため、 入力プロンプトや生成されたテキスト・画像・音声などの出力コンテンツ自体は送信されない また、中間層出力は高次元のベクトルであり、単独で意味を復元することは難しい

テキスト生成AIへの適用:Llama 2 (Metaの大規模言語モデル:LLM)

LLMを三分割してLocalとCloudに配置・HTTPで端末間通信を実装し、送受信パケットを可視化



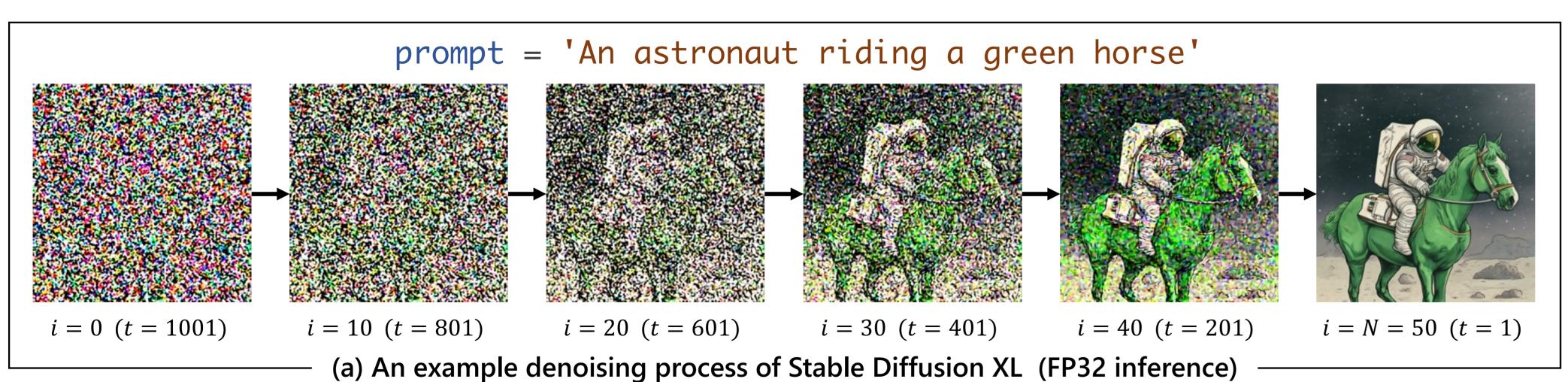
(a) Cloud-only LLM inference:送受信データはテキスト

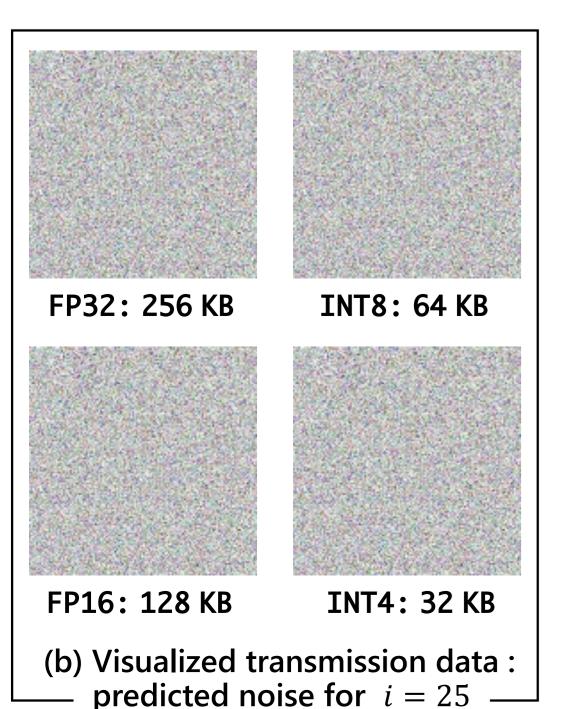
> Frame 457: 71 bytes on wire (568 bits), 71 bytes captured (568 bits) Ethernet II, Src: ASUSTekC_f0:19:43 (3c:7c:3f:f0:19:43), Dst: NVIDIA_83:8e:4a (48:b0:2d:83 Internet Protocol Version 4, Src: 192.168.1.221, Dst: 192.168.1.225 Transmission Control Protocol, Src Port: 7860, Dst Port: 58716, Seq: 8461, Ack: 8377, Len: [19 Reassembled TCP Segments (8465 bytes): #422(141), #423(1588), #424(1271), #425(254), # Hypertext Transfer Protocol Data (8192 bytes) Data: 8e396f38804076b94c3550bc4ac15039deba5e3b08bfdcb9383e4ebd2a4010c04d3f1435... [Length: 8192]

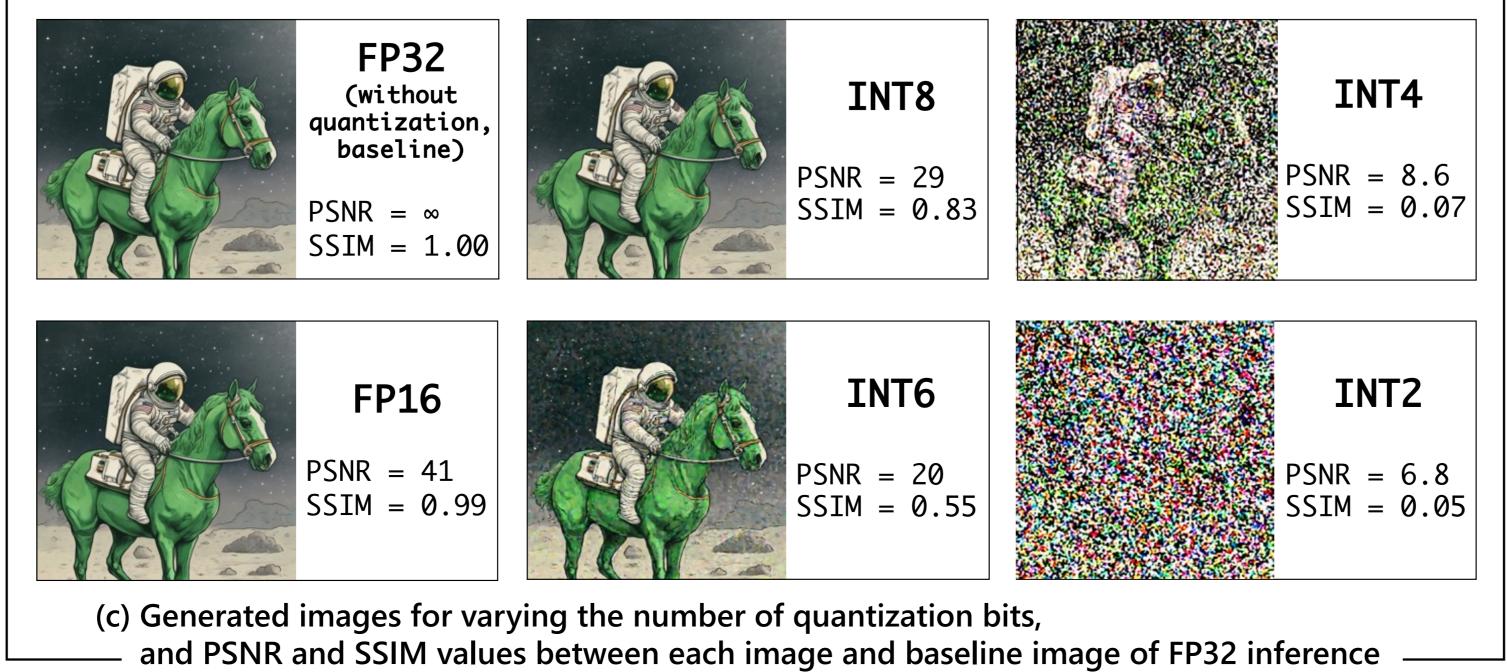
(b) Λ-Split framework:送受信データは中間層出力ベクトル

画像生成AIへの適用:Stable Diffusion XL (拡散モデル:Diffusion Model)

拡散モデルの分割点の工夫・量子化による通信データ量削減









GitHubにて デモ動画と実装を公開中

https://github.com/nishio-laboratory/lambda_split