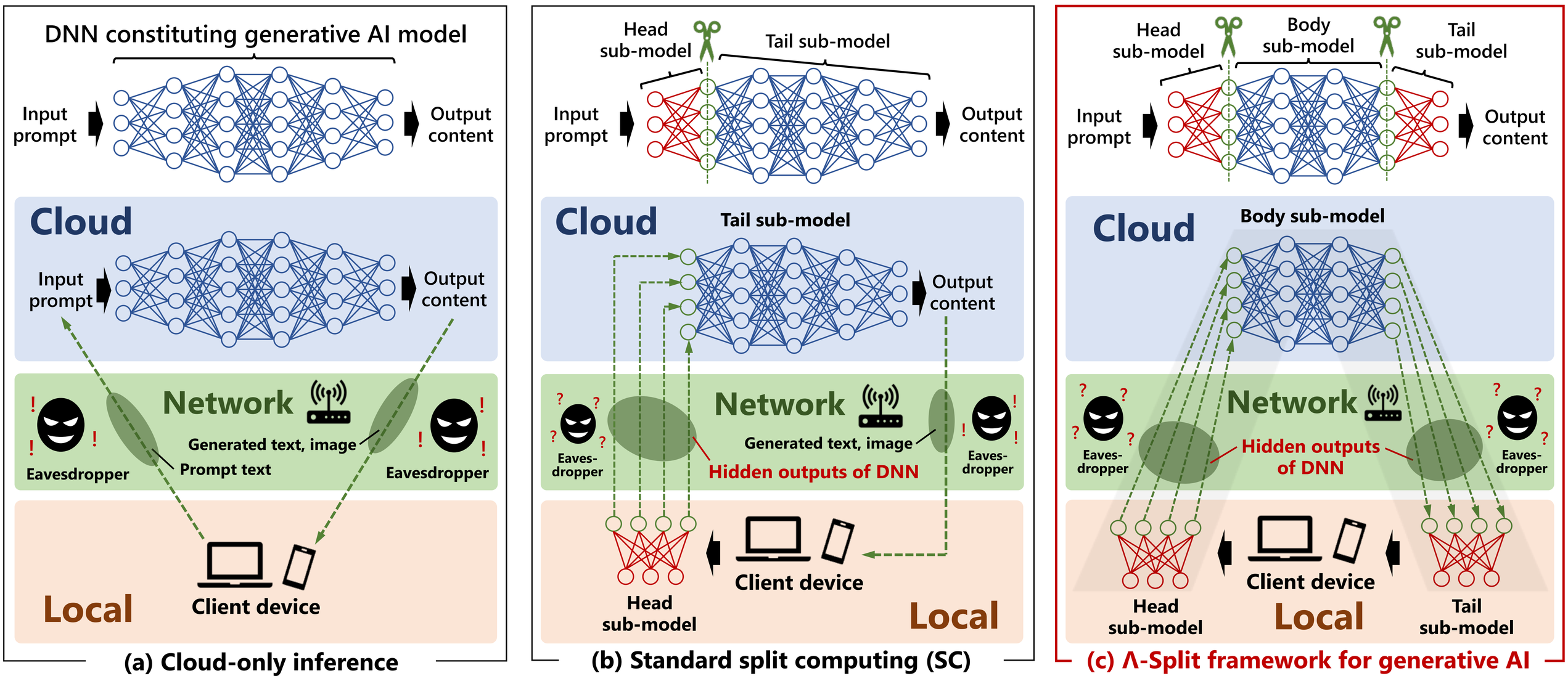
プライバシに配慮した生成AIサービスに向けた三分割 Split Computing フレームワーク

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

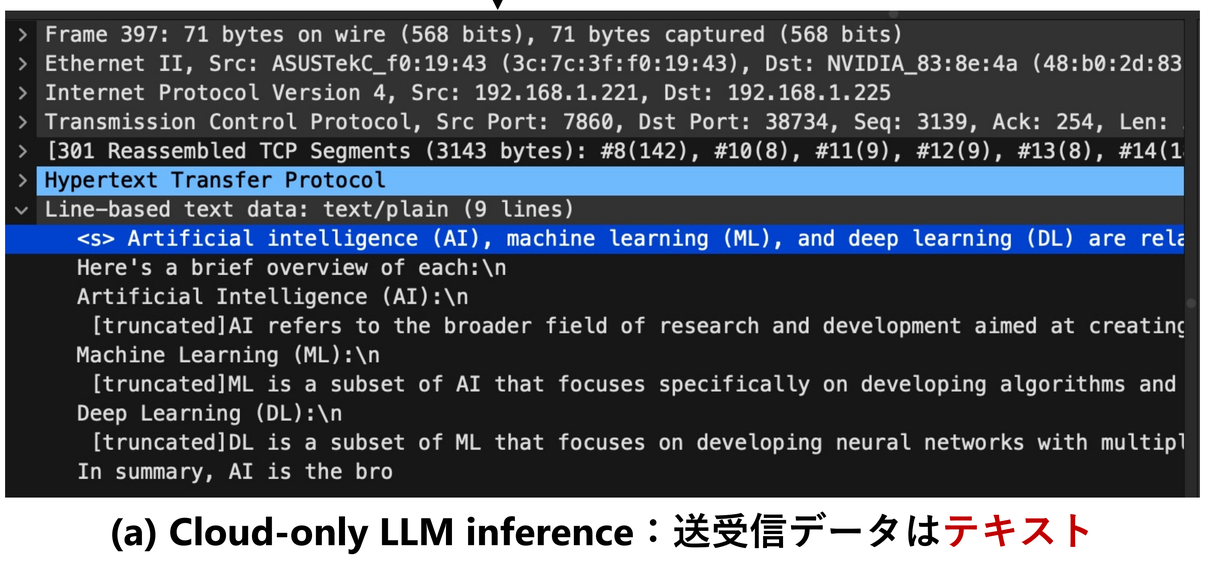


Λ-Split では、送受信されるデータが 全てDNNの中間層出力（hidden outputs）であるため、
入力プロンプトや生成されたテキスト・画像・音声などの出力コンテンツ自体は送信されない
また、中間層出力は高次元のベクトルであり、単独で意味を復元することは難しい

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

```
prompt = 'What is the difference between AI, ML, DL?'
```

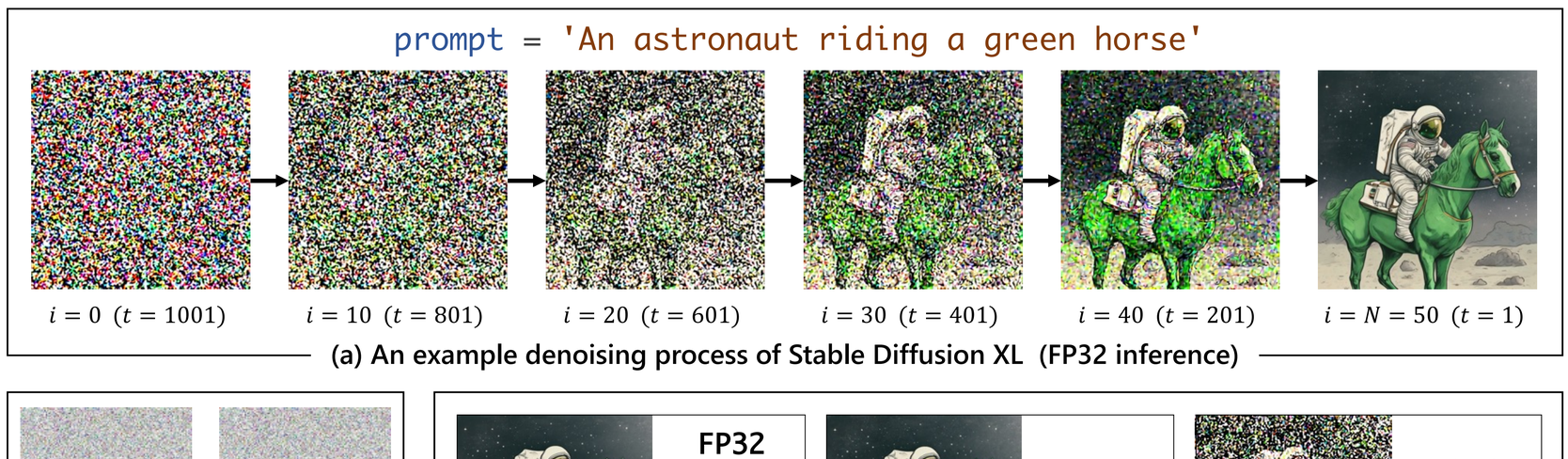


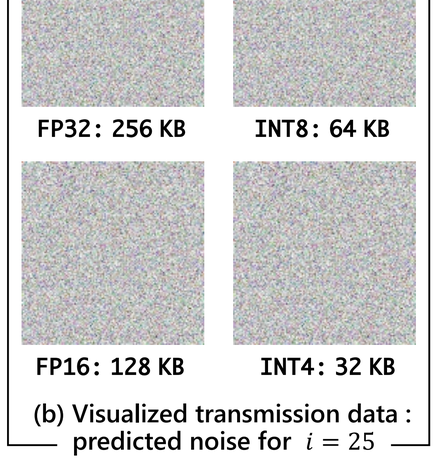
(a) Cloud-only LLM inference：送受信データはテキスト

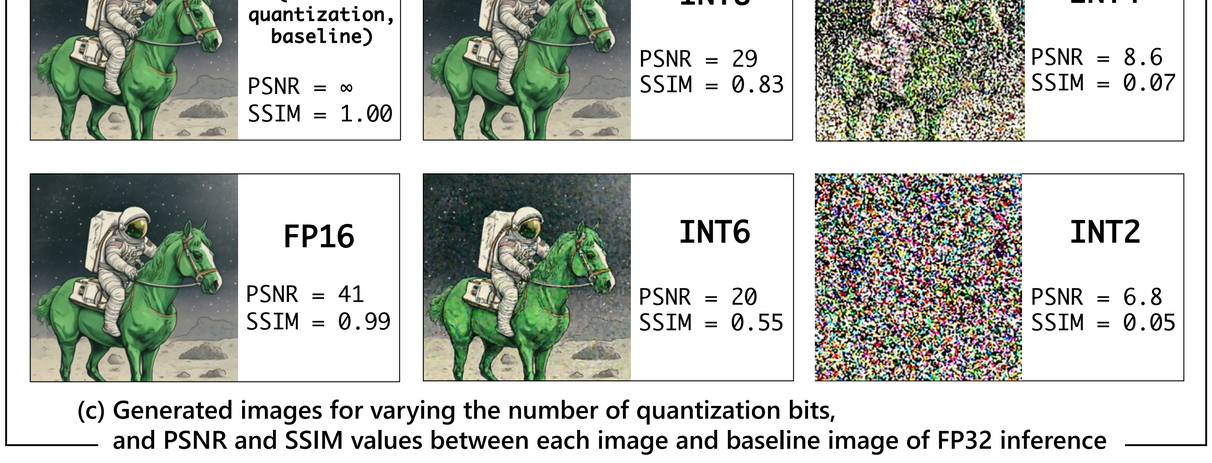
(b) Λ-Split framework：送受信データは中間層出力ベクトル

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