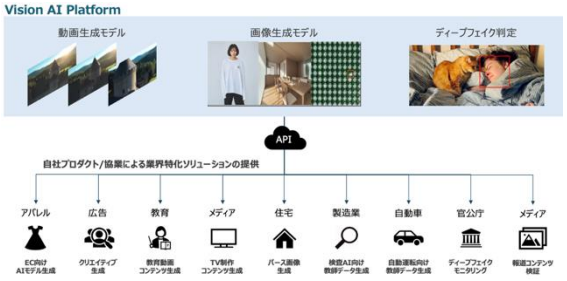
ユーザー意図を反映する選択的編集能力を備えたVision系基盤モデルの開発・事業成果概要

実施者	株式会社データグリッド
事業概要	本事業では、高い選択的編集能力を持つ動画・画像生成基盤モデルを開発する。同時に、生成AIコンテンツの悪用対策としてディープフェイク検知基盤モデルも開発する。

本事業では、高品質な大規模データセットを整備し、ユーザー意図に沿った動画・画像が生成可能な選択的編集能力の高い動画・画像生成基盤モデルを構築する。画像生成基盤モデルに関してはさらに製造業ドメイン特化基盤モデルを開発する。並行して、こうした動画・画像生成AIで作成されたメディアを識別するために、既存の基盤モデルベースのアプローチに加え、本事業で構築する動画像生成の基盤モデルをフェイクメディアの情報抽出器として活用することで、偽情報の検出性能を高めた強固なディープフェイク検知基盤モデルを構築する。

社会実装イメージ



本研究開発で実施する動画像生成基盤モデルとディープフェイク検知基盤モデルを含むVision系基盤モデルの開発成果を利用することで、各種モデルを搭載したAPIプラットフォーム”Vision AI Platform”を構築し、自社サービスへの組み込みや各業界のキープレイヤーとの協業により業界特化のAIソフトウェアを展開する。

事業成果

＜開発成果＞

動画生成AI基盤モデル：ロイヤリティフリーの動画共有サイトから約16万の動画データを収集し、Open-Sora-Planv1.3を基盤としたtext2video型の動画生成基盤モデル（DATAGRID-Open-Sora-Plan-v1.3.0-0.16M）と選択的編集モジュールを開発した。最終目標に掲げたFVD及びLPIPSで、LPIPSは目標値を達成したが、FVD（37.78）では目標値（32）に及ばなかった。しかし、学習可能なオープンな動画生成モデルの中でFVD=37.78は2025年5月時点で世界最高精度（SOTA）であり、国際競争力のある動画生成基盤モデルを構築できたといえる。

画像生成AI基盤モデル：PD12MやCommonCatalogなどのオープンなデータセットから約3600万枚のデータセットを構築し、Local Attention DiTという独自のアーキテクチャの画像生成基盤モデル（DATAGRID-Local-Attention-DiT-v1.0.0-0.52B）をスクラッチから構築するとともに、SDベースの基盤モデルを製造業特化データで追加学習することで最終目標に掲げたCLIP, DISTSなど主要6項目のうち選択的編集で特に重要な前景の3項目では目標値（SOTA）を達成することができ、編集技術の観点で国際的に優位な編集性能の高い画像生成基盤モデルを構築したといえる。

ディープフェイク検知モデル：DINOv2をベースにしたディープフェイク検知モデルと、今回構築した動画生成AI基盤モデルの一部を活用するという国際的にも独自のアプローチでディープフェイク検知基盤モデルでACC、EER、HTERなどの評価指標で、全てにおいて最終目標値を達成した。

＜成果の公開＞

公開項目	公開範囲・方法
開発成果・ノウハウ	Zennにて開発成果と開発ノウハウを共有済み
学習済みモデル	Hugging faceにて動画生成・画像生成モデルを公開済み
開発コード	GitHubにて推論コードと、学習コードの一部を公開済み