(報道発表資料)

2026.11.11
NTT 株式会社

ロボットが「動く前に未来を予測」、失敗を回避するフィジカル AI 技術を開発

～物理法則に基づき荷崩れや落下を予測、
空港手荷物など多様な物品の積込み・積降ろし自動化へ～

発表のポイント:

- ◆ NTT は、ロボットが動く前に、複数の行動候補について荷崩れや落下の可能性を予測・評価し、成功につながる行動を選択可能にする技術を実現しました。具体的には、フィジカル AI の基盤として、物体の重さや向き、動く速さなどの情報(物理量)を明示的に扱い、運動方程式に基づいて状態変化を予測するワールドモデル※1 を開発しました。
- ◆ 独自開発した「物理制御アダプタ」を映像生成モデル※2 に統合することで、荷崩れや落下など、物体の相互作用を高精度に予測します。公開ベンチマークで従来手法を上回る精度を確認するとともに、荷物の積付け作業を模擬した小型ロボットによる棚積みタスクでは、成功率が 73.1%から 92.3%に向上しました。
- ◆ 本技術の最初の適用領域として、形状や重量の異なる多様な空港手荷物の積込み・積降ろしを想定しています。今後は株式会社 Mujin と連携をしながら社会実証を進め、将来は物流・倉庫、製造、建設、小売など、幅広い分野のフィジカル AI へ展開します。

NTT 株式会社(本社:東京都千代田区、代表取締役社長:島田　明、以下「NTT」)は、ロボットが実際に行動する前に、複数の行動候補について荷崩れや落下などの結果を予測・評価し、成功につながる行動を選択できる技術を実現しました。その基盤となるワールドモデルには、物体の物理量を明示的に扱い、運動方程式に基づいて状態変化を予測する独自開発の「物理制御アダプタ」を映像生成モデル※2 に統合しています。これにより、物体の相互作用を高精度に予測し、学習データに含まれない場面にも対応します。本技術は、形状や重量の異なる多様な空港手荷物の積込み・積降ろしを最初の適用領域として想定し、株式会社 Mujin と連携をしながら社会実証を進めて参ります。将来は物流・倉庫、製造、建設、小売など、幅広い分野のフィジカル AI※3 へ展開し、現場の省人化と生産性向上への貢献をめざします。本研究成果は、2026 年 11 月 18 日～25 日に開催される NTT R&D FORUM 2026 ※4 に展示予定です。

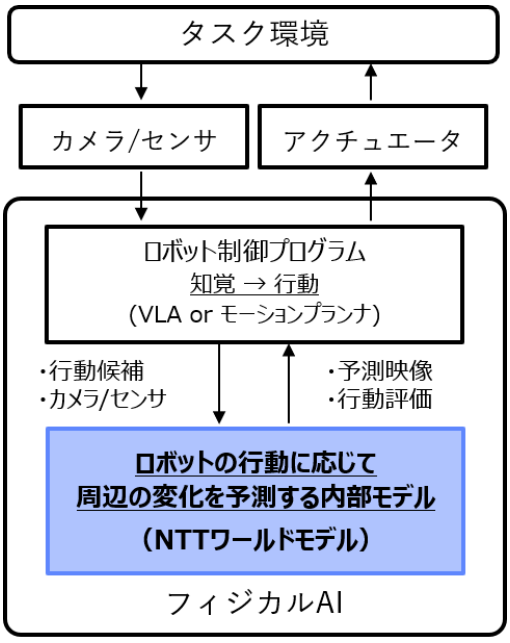


図 1：フィジカル AI における本技術の位置づけ。
既存のロボット制御プログラムと連携し、行動前に周辺環境の変化を予測・評価します。

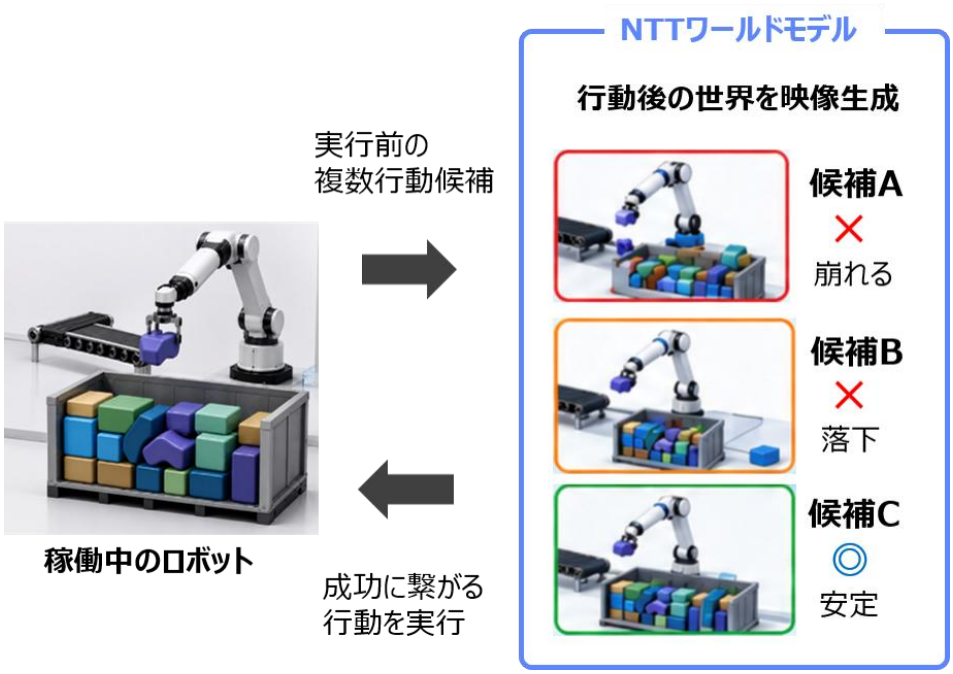


図2：本技術による行動選択の概要。
複数の行動候補について行動後の結果を予測・評価し、成功につながる行動を選択します。

1. 背景

ロボットなどの機械が実世界で自律的に判断・行動するフィジカル AI において、ロボットの行動に応じた未来の状態変化を予測するワールドモデルが注目されています。しかし、従来のワールドモデルは物体の物理量を内部で明示的に扱わないため、生成映像のなかで物体が消失・変形する、衝突などの物体の相互作用を正しく予測できない場合があるという課題がありました。また、大量の学習データに依存するため、学習データに含まれない場面では予測が破綻しやすいという課題もありました。

こうした課題が顕著に現れる例の１つが、空港手荷物や宅配便などの小口荷物を代表とする、形状や重量、配置が多様な物品の積込み・積降ろしです。対象物の形状・重量が一定でないため、従来のロボット制御だけでは、行動後に荷崩れや落下などが生じるかをあらかじめ見極めることが難しく、あらゆる物品を網羅した学習データを収集することも現実的ではありません。

NTT では、物体の物理量を明示的に扱い、物理法則に基づいて状態変化を予測することで、未知の場面にも対応可能な独自のワールドモデルを開発しました。

2. NTT ワールドモデルの特徴

特長 1　既存のロボット制御と連携し、ロボットの行動前の判断に利用可能

本技術は既存のロボット制御ソフトウェアと連携し、その判断を補強する形で適用できます。また、ロボット制御フローに直接組み込めるよう、精度を維持しつつ高速に動作する構成としています。具体的には、VLA※5 やモーションプランナ※6 などが生成する行動候補について、それぞれを実行した場合の未来を予測し、荷崩れや落下などの結果を行動計画の評価として出力します（図3）。

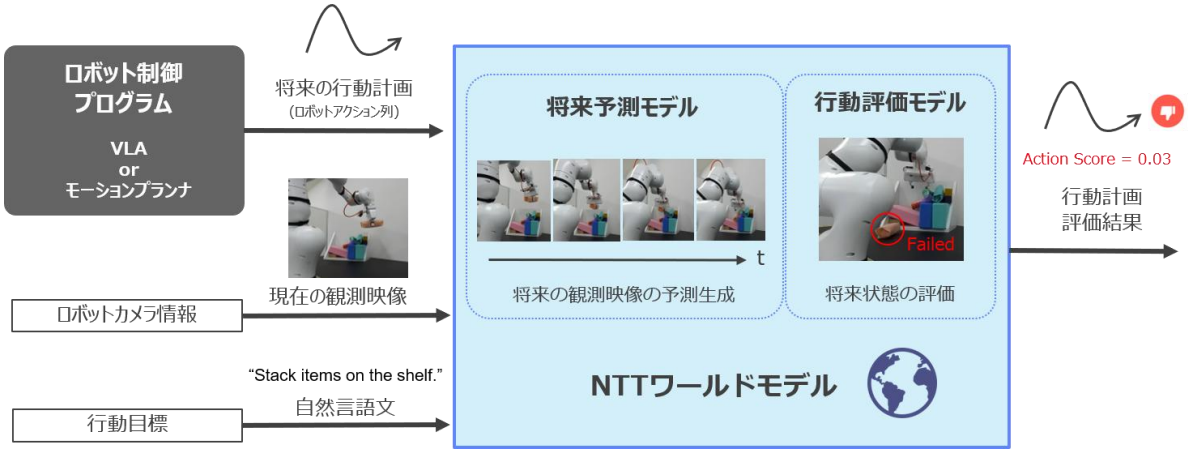


図3：NTT ワールドモデルの概要。
ロボットの行動計画ごとに将来を予測し、実行前に成否を評価します。

特長 2　複数の未来を比較・評価し、成功につながる行動を選択可能に

ロボット制御プログラムが生成した複数の行動候補について、それぞれを実行した場合の未来を映像として生成します。生成した映像から、物体の落下や衝突の有無などの結果を予測・評価することで、成功につながる行動を選択可能にし、ロボットの行動失敗率を低減します（図4）。

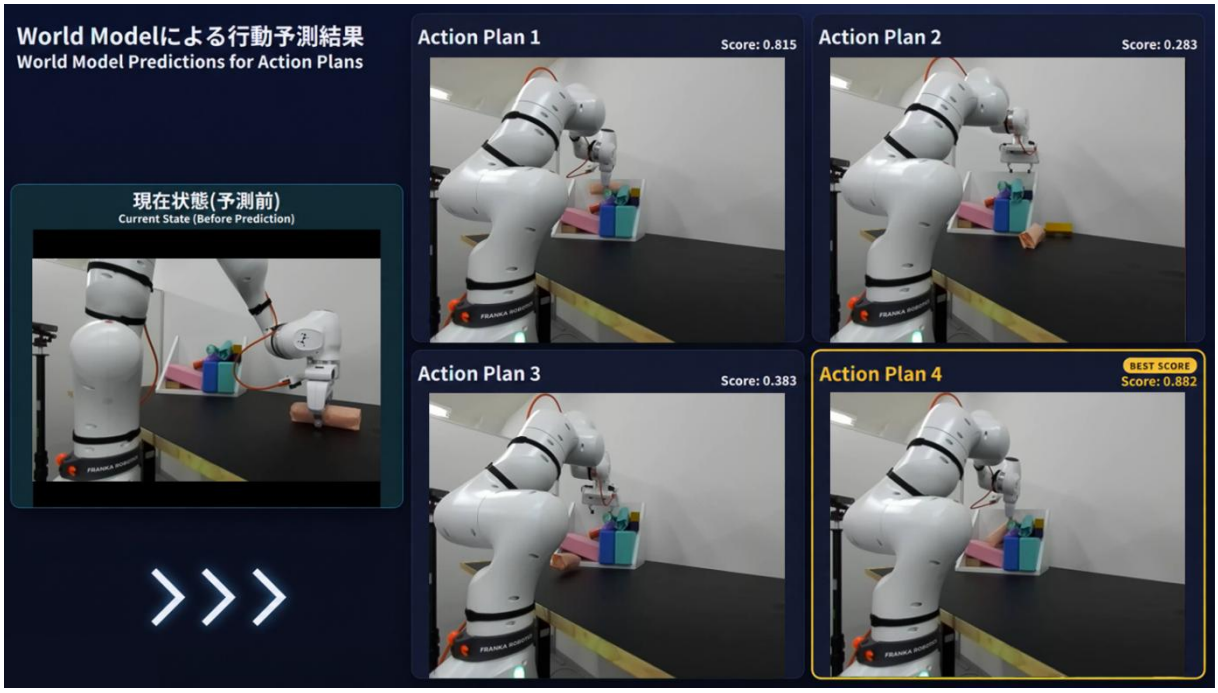


図4：NTT ワールドモデルによる複数の行動候補の予測・評価。
各候補を実行した場合の将来映像を生成し、成功への繋がりやすさをスコアとして評価します。

特長 3　物理法則に基づく高精度な未来予測

物体ごとの位置・姿勢・速度・質量などの物理量を明示的に扱い、運動方程式に従う動きとして未来を予測します。これにより、物体の消失や変形を抑え、衝突や崩落などの物体どうしの相互作用をより正確に予測できます（図5）。

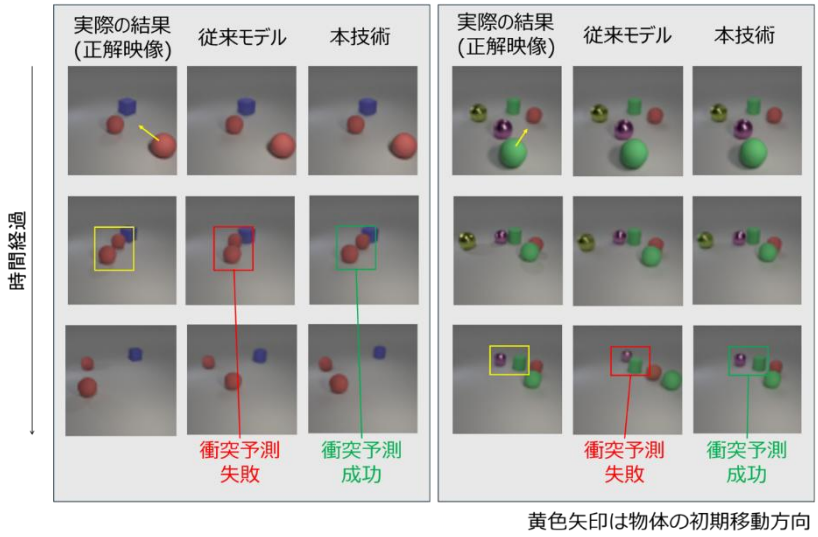


図5：公開ベンチマーク（OBJ3D[※7]）における、将来予測結果の比較。物体の衝突を含む場面で、従来モデルは衝突の予測に失敗し（赤枠）、本技術は成功しています（緑枠）。

特長 4　学習データに含まれない未知の場面への対応

物体の物理量を明示的に扱い、物理法則に基づいて状態変化を予測することで、学習データに含まれない場面や動きにも対応できます（図6）。空港手荷物や宅配便の小口荷物など、形状・重量・配置が多様な物品を扱い、学習データの網羅が難しい現場に適しています。

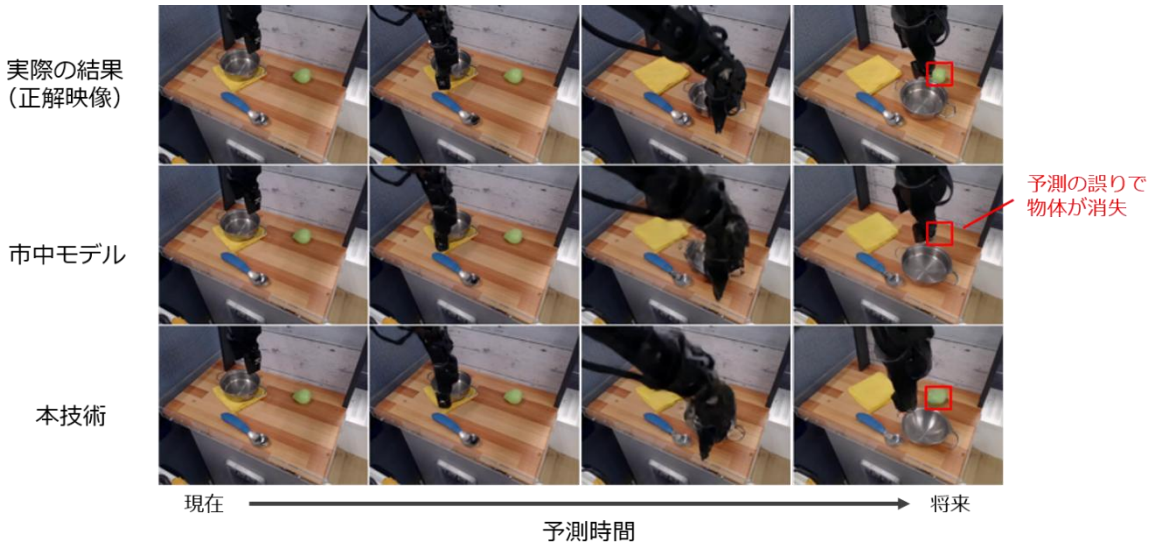


図6：ロボットアームで物体が隠れる、学習データに含まれない場面での予測結果

3. 技術のポイント

物理法則に基づく「物理制御アダプタ」の開発と、その映像生成モデルへの統合

NTT は、映像から将来を予測する仕組みに、物理法則に基づく計算を組み合わせる「物理制御アダプタ」を開発しました。具体的には、映像から物体ごとの位置や向き、動く速さ、重さなどを推定し、物理法則に基づいて次にどのように動くかを計算します。さらに、この物理制御アダプタを、ワールドモデルの基盤となる映像生成モデルに世界で初めて統合しました。物理制御アダプタの計算結果を映像生成に反映することで、物体の消失や変形を抑え、衝突や崩落などの物体どうしの相互作用をより正確に予測できます（図7）。

本技術は、次の 4 つの優位性を備えています。

（1）物体ごとの物理量を推定し、物理法則に基づいて将来の挙動を高精度に予測できます。
（2）物理法則を明示的に扱うことで、学習データに含まれない場面や動きの予測にも対応できます。
（3）映像生成モデルとの統合により、予測結果を自然な映像として出力でき、ロボットの行動の成否を事前に評価できます。
（4）予測の根拠となる物理量を明示的に出力できるため、なぜそのような予測になったのか、生成結果の理由を把握しやすくなります。

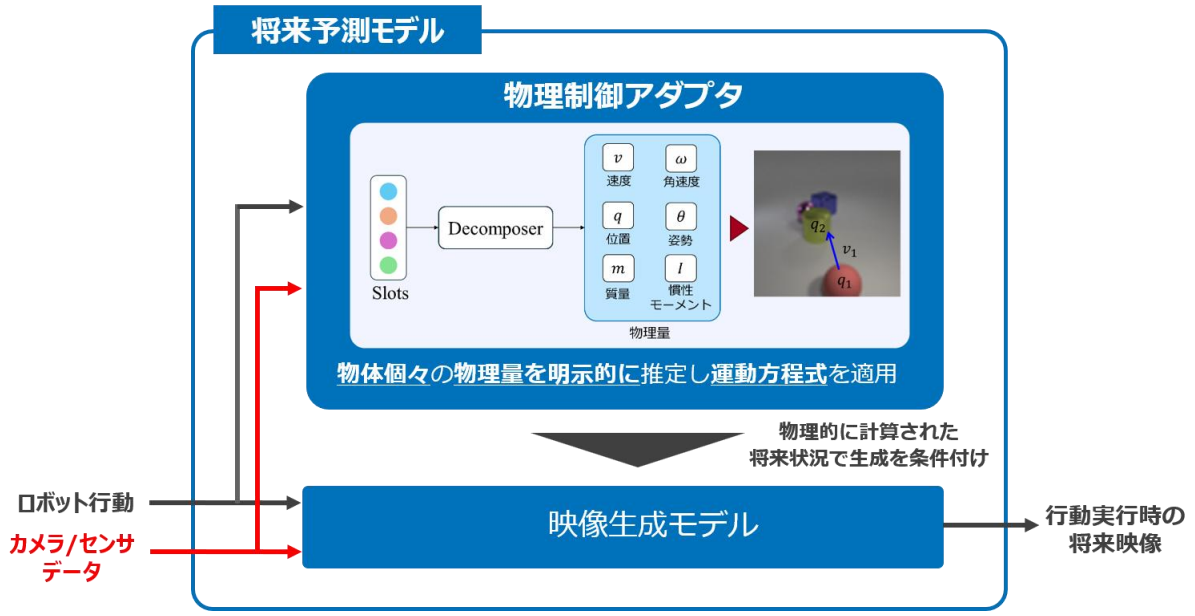


図7：本技術の中核となる将来予測モデルの構成。
映像生成モデルに物理制御アダプタを統合し、アダプタで推定した結果を
映像生成へ反映することで、高精度な将来映像の生成を実現します。

4. 実験の概要

小型ロボットを用いて荷物の積付け作業を模擬した棚積みタスクを設定し、本技術の適用有無によるタスク成功率を比較しました。その結果、本技術の適用により成功率が 73.1%から 92.3%に向上することを確認しました。

あわせて、将来予測そのものの精度を確認するため、複数物体の衝突を含む公開ベンチマーク（OBJ3D※7）を用いて予測映像を評価しました。人間の知覚に近い感覚で類似度を測る評価指標 LPIPS※8 において、従来手法を上回る予測精度を確認しました（表 1）。

手法	LPIPS（×100、低いほど高精度）
市中の映像生成モデル（NVIDIA Cosmos Predict 2.5）	21.90
既存の最高精度手法(OCK※)	7.92
NTTワールドモデル	**7.81**

※Song, YeonJi, et al. "OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics." , *ICCV*, 2025.

表 1：公開ベンチマーク（OBJ3D）における予測映像の精度評価結果。

5. 今後の展開

本技術は、形状や重量の異なる多様な空港手荷物の積込み・積降ろしを最初の適用領域として想定しています。今後は、株式会社 Mujin と連携をしながら社会実装を進めるとともに、ロボットサービスベンダ、物流・倉庫事業者などとの連携も検討していきます。

さらに荷物の積込み・積降ろしを起点として、製造、建設、小売など多様な物体を扱う幅広い分野へ展開し、人手不足が深刻な現場の省人化と生産性向上に貢献していきます。

なお、本技術は 2026 年 11 月に開催される「NTT R&D フォーラム 2026」で展示予定です。
https://www.rd.ntt/forum/2026/

【用語解説】

※1 ワールドモデル

外界の状態や変化をモデル化し、自身の行動によって生じる未来の状態変化を予測する仕組み。本技術では、ロボットの行動候補とカメラ・センサ情報から、行動後の周辺環境の変化を映像として生成し、実行前に行動計画の成否をスコアで評価します。

※2 映像生成モデル

大量のデータから学習し、映像を生成する深層学習モデル。一般的なワールドモデルの基盤として用いられる。

※3 フィジカル AI（Physical AI）

ロボットなどの機械が、カメラやセンサで実世界を知覚し、自律的に判断・行動するための AI。

※4 NTT R&D FORUM 2026 https://www.rd.ntt/forum/2026/

Mehr Licht!
More Photonics for Intelligence

※5 VLA（Vision-Language-Action model）

カメラ映像と言語による指示を入力として、ロボットの行動を直接出力する基盤モデル。

※6 モーションプランナ

ロボットの目標と周辺環境の情報から、タスク目標に到達するためのロボット制御を計算するプログラム。

※7 OBJ3D

複数の物体の衝突などの相互作用を含む映像で構成される、映像予測の公開ベンチマーク。

※8 LPIPS（Learned Perceptual Image Patch Similarity）

人間の知覚に近い感覚で、正解映像と予測映像の類似度を測定する評価指標。値が小さいほど類似度が高いことを示す。

■本件に関する報道機関からのお問い合わせ先
NTT 株式会社
サービスイノベーション総合研究所
広報担当
問い合わせフォームへ