視覚条件付き拡散ポリシーによるロボットマニピュレーションの実装と評価

forgottencow

概要

近年，視覚・言語・行動を統合してロボットを直接制御する Vision-Language-Action（VLA）モデルが注目されている.RT-2 は 大規模の視覚言語の事前学習をロボットの行動へ接続し，汎化能力の向上を報告した [1]. 一方で，現実のロボットで再現可能な検証を行うには，ハードウェア・データ収集・学習・評価までを一貫させた実験基盤が不可欠である. 本研究は Hugging Face の LeRobot[2] と低コストアーム（SO-101）を用い，視覚条件付き拡散ポリシー（Diffusion Policy）[3, 4] に基づくマニピュレーション学習の実装パイプラインを構築した. しかし実機データに対する学習では不安定化が観測され，画像表現（予測）段階で生成がノイズで崩壊する現象を確認した. そこで本稿では，データ収集・前処理・時系列整合・行動表現の設計が収束性に与える影響を整理し，再現可能な切り分け手順（ログ，定量指標，失敗の分類）を提示する. さらに比較として，VLA 基盤モデルである NVIDIA Isaac GR00T[5] のファインチューニングも同一タスク設定で試行し，適用要件と課題を報告する. 参考ベースラインとして，LeRobot が初手として推奨する模倣学習手法 ACT[6] との位置づけも併記する.

1　はじめに

近年, ロボティクス分野における技術革新は目覚ましく, オープンソースハードウェアと AI モデルの融合による「フィジカル AI」の潮流が世界中で加速している. 特に LeRobot[2] らが主導する「SO-101」等の 3D プリント可能なロボットアームの登場は, 開発の敷居を劇的に下げた. 東京で開催された AMD 主催の AI ハッカソンにおける熱狂が示すように, これまで専門機関に限られていたロボット開発は, 個人のエンジニアやコミュニティ主導へと移行しつつある. これらの安価で再現性の高いハードウェアが, 近い将来, 現実社会のあらゆる場面に実装されていくことは明白である.

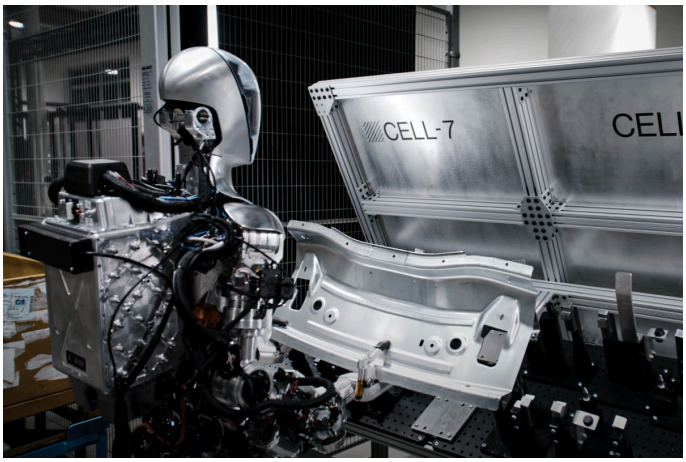

図 1: Figure AI 社のロボット（Figure 01）

こうしたハードウェアの急速な普及に対し, 制御モデルの分野では「汎用性」と「知見の共有」に課題が残されている. 現在, 最も注目されているのは LLM（大規模言語モデル）とロボット操作の融合であり,Figure 01（Figure AI 社）が BMW 工場で稼働する事例 [7] や, 複雑な家事タスクをこなすデモンストレーションは, 対話をベースとした自律的な操作が可能になる未来を示唆している. しかし, こうした高度な汎用モデルの多くは巨大企業によるクローズドソースとして独占されており, その詳細な知見が一般に広まることは少ない. また, 個人開発者の間で主流となっている ACT（Action Chunking with Transformers）[6] 等の模倣学習手法は, 少数のデータで動作を習得できる反面, 未知の環境やタスクに対する汎用性に欠けるという側面がある.

そこで本研究では, LeRobot と低コストアーム（SO-101）を用い，視覚条件付き拡散ポリシー（Diffusion Policy）[3, 4] によるマニピュレーション学習の実装を試みる.Diffusion Policy は，観測に条件付けされた行動分布を扱える点で有効性が報告されている.

一方で，本研究の実装では学習の不安定化が観測され，とくに（画像表現学習／予測の段階で）生成物がノイズで予測として成り立たない現象を確認した. そこで本稿では，データ収集・前処理・時系列整合・行動表現の設計が収束性に与える影響を整理し，再現可能な切り分け手順（ログ，定量指標，失敗の分類）を提示する. さらに比較として，VLA 基盤モデル（NVIDIA Isaac GR00T）のファインチューニングも同一タスク設定で試行し，適用要件と課題を報告する.

2　関連研究

本節では，本研究の基礎となる拡散モデルを用いたロボット制御手法，近年急速に発展している VLA（Vision-Language-Action）基盤モデル，およびそれらのオープンソースハードウェアへの適用に関する研究動向について述べる.

2.1　低コスト実機におけるデータ収集と模倣学習

低コストな実機ロボットによる精密操作では，機体剛性やセンサ品質，キャリブレーション誤差等の影響により，古典的

† 慶應義塾大学　環境情報学部

な状態推定や計画手法の前提が崩れやすい. この制約に対し，Zhao らは市販アームと 3D プリント部品から構成される低コスト両腕システム（ALOHA）を提示し,RGB 画像観測から行動を直接生成する e2e 模倣学習の有効性を示した [6]. 同研究は，テレオペレーションを前提としたデモ収集設計や，模倣学習における誤差蓄積・非定常性といった実運用上の課題を明示しており，低コスト環境での評価設計の基礎となっている. また，オープン実装基盤である LeRobot は，低計算資源でも学習しやすいベースラインとして ACT (Action Chunking with Transformers) を推奨している.ACT は行動系列をチャンク化して生成することで時間的な一貫性を保つ手法であり，個人環境での初期検証に適したベースラインとなる [6]. 筆者らの予備検討においても，SO-101 と ACT を用いたブロック積みタスクでは，固定配置で 90% の成功率を達成した. しかし，配置をランダム化した条件下では成功率が 10% 台まで低下しており，単純な模倣学習（ACT）では小規模データセットにおける空間的な汎化能力に限界があることが確認されている. 本研究で Diffusion Policy を採用するのは，この課題を行動分布の学習によって解決できるか検証するためでもある.

2.2 拡散モデルに基づく視覚運動ポリシー

模倣学習において，人間の操作分布を単一の回帰モデルで近似すると，平均化された行動が出力され，特に接触を含むタスクでの失敗要因となる. これに対し Chi らは，ポリシーを条件付き拡散過程（denoising diffusion）として定式化し，視覚条件付きで行動分布を生成する Diffusion Policy を提案した [3, 4]. 本手法は行動分布の表現能力に優れ, Receding Horizon 制御による頑健性も示されている. 一方で，拡散型ポリシーは小規模データセットでの学習の安定性に課題があり生成品質の劣化に繋がりやすい. 本研究では，この「低リソース実機」条件下で観測された生成の不安定化（ノイズで崩壊）に着目し，データ処理や表現設計の観点からその要因を整理する.

2.3 動画予測と行動生成の統合

視覚的な未来予測を行動生成に組み込むアプローチとして，Unified Video Action (UVA) 等が提案されている [8]. これらは動画生成モデルの強力な世界モデル能力を行動計画に利用するが，動画生成が要求する高い空間解像度と，行動生成が要求する高い時間分解能のミスマッチが，推論の遅延や誤差伝播が行動の生成に悪影響を及ぼす課題が指摘されている. 本研究の実装において観測された「画像表現（予測）段階の崩壊」は，まさにこれらの先行研究が指摘する生成品質と行動精度のトレードオフに関連する現象である. したがって本稿では，この失敗挙動を単なる実装不備ではなく，拡散モデルを用いた視覚運動ポリシーにおける構成的な課題として位置づけ，その回避策を議論する.

2.4 VLA 基盤モデルとファインチューニング

視覚と言語を統合し，ロボット行動を直接生成する VLA モデルは，タスク汎化の鍵として注目されている.RT-2 は 大規模の視覚言語の事前学習をロボットの行動へと接続する枠組みを示した [1]. さらに Project GR00T は，多様な身体性に対応可能な基盤モデルとして，マルチモーダル入力から Diffusion Transformer ヘッドを用いて行動を生成するアーキテクチャを採用している [5].GR00T は，ユーザが収集したデータを LeRobot 互換スキーマへ変換し，タスクに合わせてファインチューニングする運用を想定している. 本研究では，スクラッチ実装の Diffusion Policy との比較として，この「基盤モデル＋ファインチューニング」のアプローチを同一タスク設定で試行し，小データの環境におけるデータ効率の向上を検証する.

3 問題設定

本章では，本研究で取り扱うタスク，ハードウェア構成，および検証に用いるデータセットとモデルアーキテクチャについて定義する.

3.1 タスクとハードウェア構成

本研究では，先行研究と同様の「3 段ブロック積み上げタスク」を採用する. ロボットには，オープンソースの 3D プリント可能なロボットアーム SO-101 を使用した. 視覚センサとして，作業領域全体を俯瞰する RGB カメラを設置し，解像度 640×480，フレームレート 30 fps で画像を取得した. 取得された画像は，後述するモデルの入力仕様に基づきリサイズおよび正規化が行われる.

3.2 データセットと小規模性の制約

学習データセットは，筆者らの予備実験において ACT モデルが高い成功率（90%）を記録した「固定配置条件（C1）」に準拠して収集した. 具体的には，作業空間内の 20 通りの異なる初期配置に対し，各 10 エピソード，合計 200 エピソードの実演データを収集した. この「200 エピソード」という規模は，一般的な動画生成や VLA モデルの学習要件（数万〜数百万件）と比較して極めて小規模である. 本研究の狙いは，この「個人レベルで収集可能なデータ量」という制約下において，高度な生成モデルがどの程度機能するか，あるいはどのような様態で破綻するかを検証することにある.

3.3 採用モデル：Unified Video Action（UVA）

制御モデルとして，LeRobot フレームワーク上に実装された Unified Video Action（UVA）を採用した [8].UVA は，現在の観測画像から「数フレーム先の未来画像（動画）」と「ロボットの行動系列」を，共有された潜在空間を用いた拡散モデルによって同時生成するアーキテクチャである. 本実装における重要な仕様として，入力画像および生成される未来画像はす

べて 256×256 にリサイズされて処理される. 本来，この視覚的な未来予測は，環境の変化を予測することで行動のロバスト性を高めることを目的としている. しかし本研究では，この動画生成プロセスが小規模データ環境下において学習の不安定化要因（ノイズで崩壊）となり得る点に着目し，その挙動を評価対象とする.

4 実装

本章では，LeRobot フレームワーク上に構築したデータ収集パイプライン [2]，および実装した UVA モデルと GR00T の具体的な学習設定について述べる.

4.1 データ収集・前処理パイプライン

データの収集には LeRobot の標準記録スクリプト [2] を用い，SO-101 アームの関節状態・エンドエフェクタ位置およびカメラ画像を同一の制御ループ内でサンプリングした. 各フレームは `precise_sleep` 機構により厳密な周期（30 fps）で維持され，取得されたタイムスタンプに基づいて同期保存される. 学習時のデータローダには `UVALeRobotDataset` アダプタを実装した. ここでは，動画と行動の同期ズレを防ぐため，`check_delta_timestamps` 機構により 1/fps の整数倍から逸脱するサンプルをエラーとして排除している. モデル入力として，現在時刻 t を基準に過去・未来を含むウィンドウを切り出した. 具体的には，

$$\text{画像系列: } t \in [-12, \ldots, 16] \quad \text{(step 4)}$$
$$\text{行動系列: } t \in [-3, \ldots, 16]$$

を抽出し，すべての画像を 256×256 にリサイズして入力テンソルを構築した. このリサイズ処理は UVA の事前学習済み VAE の仕様に準拠したものであるが，元の撮影解像度（640×480）からの情報の圧縮を伴う.

4.2 視覚・行動統合モデル（UVA）の構成

本研究では，動画生成と行動生成を統合的に扱うため，UVA（Unified Video Action）アーキテクチャを採用した. モデルは主に以下の 2 つのモジュールで構成される. 画像エンコーダ/デコーダ（KL-VAE）：観測画像を潜在空間へ圧縮・復元する役割を持つ. 本実験では，計算コスト削減と学習安定化のため，VAE の重みは凍結（Freeze）して使用した. 自己回帰モデル（MAR）：VAE によって圧縮された潜在トークン列を入力とし，Masked Autoencoder（ViT backbone）によって「未来の画像潜在表現」と「行動系列」を同時に予測する.

4.3 比較対象：VLA 基盤モデル（GR00T）の学習設定

比較実験として，NVIDIA の基盤モデル GR00T-N1.5-3B のファインチューニングを行った [5]. 実装には LeRobot の `GrootConfig` ラッパーを使用した.GR00T の学習では，入力画像は UVA と異なり 224×224 にリサイズされ，状態・行動ベクトルは最大次元に合わせてゼロパディングされる. 計算資源の制約および過学習抑制のため，LLM バックボーンと Vision Tower（画像エンコーダ）は凍結し，拡散モデルヘッド（Diffusion Head）とプロジェクタ層のみを学習対象とする選択的なファインチューニングを行った（`tune_diffusion_model=True, tune_llm=False`）. 最適化には AdamW オプティマイザ（学習率 1.0×10^{-4}）を用い，バッチサイズ 32，ウォームアップ比率 0.05 で 40,000 ステップの学習を行った. これにより，スクラッチ学習である UVA と，事前学習済み重みを活用する GR00T の挙動差を検証する.

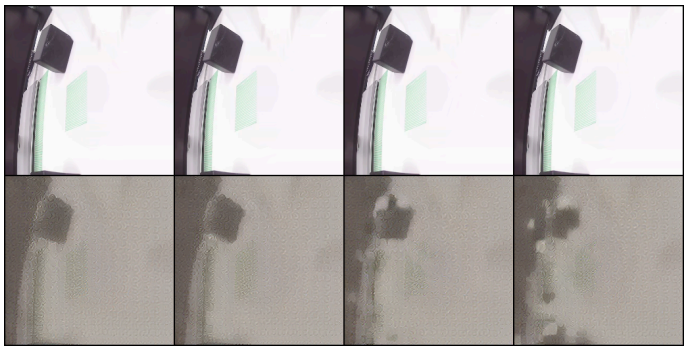

図 2: UVA 生成結果（epoch 2）

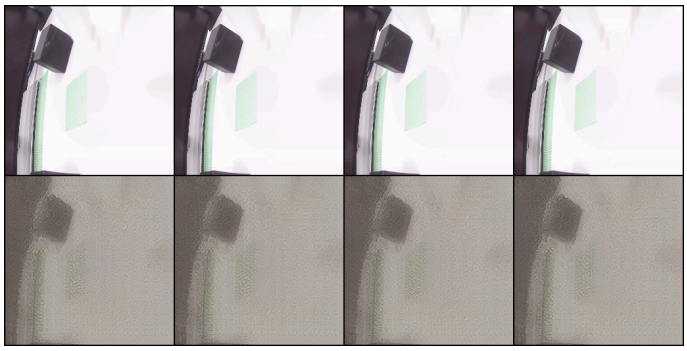

図 3: UVA 生成結果（epoch 10）

5 評価実験と考察

本章では，実装した UVA モデルおよび比較対象の GR00T モデルを用いた実機タスク評価の結果を述べ，観測された失敗事例と学習挙動（Loss 推移）の関係について分析する.

5.1 定量評価：タスク成功率の比較

各モデルに対し，学習データに含まれる初期配置（固定配置）および未学習のランダム配置にて各 10 エピソードの評価を行った. ベースラインとして，筆者らの先行実験における ACT のスコアも併記する. 結果を表 1 に示す.

モデル	条件	固定配置	ランダム配置
ACT (Baseline)	200 eps	90%	10%
UVA (Ours)	200 eps	0%	0%
GR00T (Fine-tune)	200 eps	0%	0%

表 1: タスク成功率の比較（%）

特筆すべきは，UVA と GR00T の双方が成功率 0% であった点である.ACT が少なくとも学習済み配置では高い成功率を示したのに対し，高度な生成モデルを用いた両手法は，本実験のデータ規模においてタスク遂行能力を獲得できなかった.

5.2　失敗の様態の分類

両モデルとも成功率は 0% であったが，その失敗の様態には明確な差異が確認された．

要因 A: 再構成の停滞（UVA）

UVA モデルでは，前述の通り生成画像が「灰色のノイズ」で停滞した（図 2，3）．これは視覚的な未来予測が機能せず，ロボットが視覚的なフィードバックを得られなかったことに起因する「感覚遮断」に近い状態である．

要因 B: 過剰適合による硬直（GR00T）

GR00T では，UVA とは異なり，鮮明な視覚認識能力を保持していたと推測されるが，アームがブロックの近辺に到達するものの，実際にはうまくブロックを掴めない挙動が見られた. 図 4 に GR00T の学習曲線（train/loss）を示す.Loss は 40,000 ステップにかけて極めて順調に減少しており，数値上は収束している. しかし，200 エピソードという小規模データに対し 40,000 ステップという長時間の学習を行ったことで，モデルが訓練データを過剰に記憶し，実機環境におけるわずかな変化（照明や配置の微差）に対して脆弱化した可能性が高い. 推奨される 5,000 ステップ程度で学習を止めるべきであったことが示唆される．

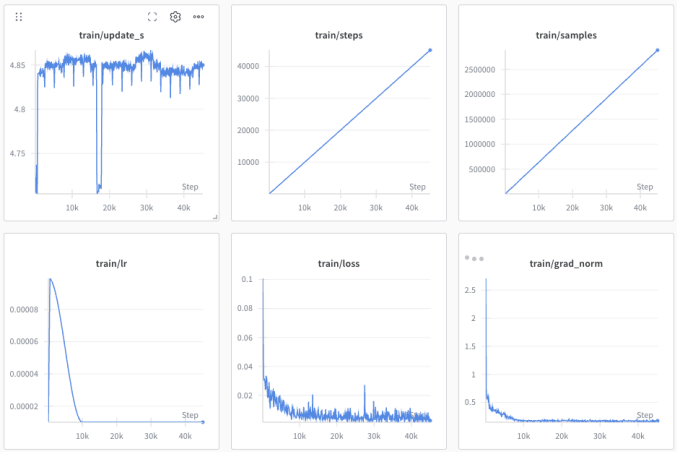

図 4: GR00T 学習曲線（train/loss）

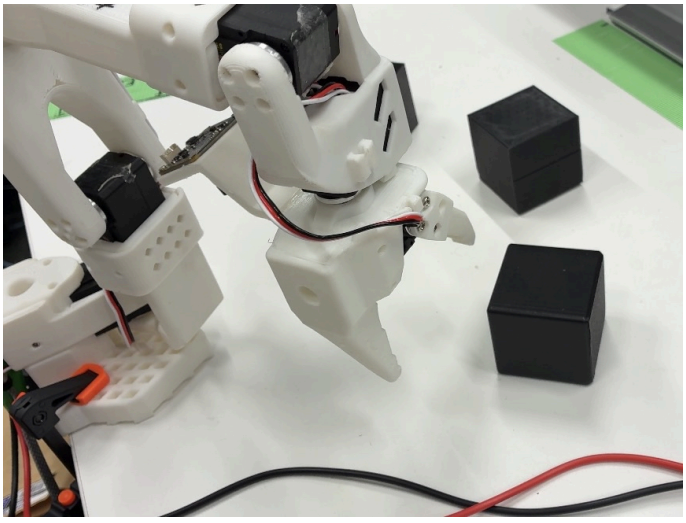

図 5: GR00T 評価時の様子

5.3　考察：なぜ「強いモデル」ほど動かなかったのか

ACT（単純な模倣）が動き，UVA/GR00T（高度な生成モデル）が動かなかった事実は，「モデルの表現力とデータ量の不均衡」を示している. 表現力の高いモデルほど，少数のデータを完璧に「丸暗記」できてしまうため，Loss は下がるが汎化しない（GR00T のケース）. あるいは，タスク設定が野心的すぎて（動画生成など），データ不足で学習が進まない（UVA のケース）. この結果は，小規模ラボにおけるロボット学習において，「むやみに SOTA モデルを導入するよりも，データ規模に見合った軽量なモデル（ACT 等）を選択するか，あるいはデータ収集自体を自動化して規模を拡大する戦略が先決である」という教訓を与えている．

6　考察

本章では，実験結果に基づき，小規模データセット環境下における生成モデルの挙動特性と，実機実装におけるボトルネックについて考察する．

6.1　表現能力とデータ規模の不均衡

本研究の最も重要な示唆は，「モデルの表現能力が高いほど，タスク遂行に必要なデータ規模の閾値が上がる」という逆説的な事実である. 先行実験における ACT（模倣学習）は，比較的単純なアーキテクチャであったため，200 エピソードという小規模データでも軌道の平均的特徴を捉え，固定配置において 90% の成功率を達成した. 対して，UVA や GR00T は動画生成や言語理解を含む巨大な表現能力（Capacity）を持つ. これらをわずか 200 件のデータで学習させた結果，UVA は「動画生成タスクの複雑性」に圧倒され（学習不足），GR00T は「データを丸暗記する余裕」がありすぎて過学習に陥った. つまり，個人開発レベルのデータ規模（数百件）においては，SOTA レベルの巨大モデルはむしろ「扱いきれない凶器」となり，単純な模倣学習の方が実用性が高いという，ある種の技術的ジレンマが浮き彫りとなった．

6.2　補助タスクとしての動画生成の代償

UVA における「再構成の停滞（要因 A）」は，行動生成の補助として導入したはずの「未来画像予測」が，逆に学習の阻害要因となったことを示している. 本来，視覚的な未来予測は，物理法則や環境ダイナミクスを理解させるための強力な信号となるはずである. しかし，それは「十分なデータ量」があって初めて成立する. 本実験のようにデータが疎（Sparse）な状態では，モデルは多様な未来画像を生成する術を持たず，損失関数を下げるために最も安全な「平均的な灰色の画像」を出力する局所解へと収束した. その結果，本来の目的である「行動生成」に必要な計算リソースや勾配情報までもが，機能しない動画生成タスクによって汚染され，共倒れを招いたと結論付けられる．

6.3 基盤モデルの微調整における落とし穴

GR00T の失敗は，基盤モデルといえども「データ不足を魔法のように解決するわけではない」という現実を示している. 学習曲線に示した通り，学習損失は極めて順調に減少したが，これは典型的な過学習の兆候である. 特に 40,000 ステップという学習量は，200 エピソードに対して過剰であり，事前学習によって獲得していた汎用的な特徴表現が，特定データの丸暗記によって上書きされた可能性が高い. 基盤モデルを実機適応させる際は，単に Loss を下げることを目指すのではなく，Validation Set による厳密な監視や，Early Stopping の導入，あるいは LoRA のランク数を下げるなどの正則化戦略が，スクラッチ学習以上に重要となると考えられる.

6.4 ハードウェア制約と空間解像度

視覚情報の圧縮（リサイズ）も，失敗の隠れた要因である.SO-101 による精密なブロック積み（許容誤差数ミリ）に対し，UVA の 256×256, GR00T の 224×224 という入力解像度は，空間情報を粗視化しすぎている恐れがある. 特に 640×480 からのリサイズ過程でアスペクト比や高周波成分が失われた結果，ロボットは「ブロックの正確なエッジ」を視覚的に捉えられなくなり，これが要因 B（行動不全）を助長したと推察される. 高精度タスクにおいては，画像全体をリサイズするのではなく，対象物体周辺を切り出す（Crop）などの前処理工夫が必須であったと言える.

7 総括と今後の展望

7.1 総括

本研究では，低コストアーム SO-101 と小規模データセット（200 エピソード）を用い，最先端の VLA 手法である UVA の実装および GR00T のファインチューニングを試みた. 実験の結果，単純な模倣学習（ACT）が特定の配置条件で成功を収めた一方で，高度な生成モデルを用いた両手法は成功率 0% に留まった. その要因として，UVA における動画生成タスクの負荷による学習の停滞，および GR00T における過学習による汎化性能の欠如を特定した. 本研究の貢献は，個人や小規模ラボにおけるロボット開発において，SOTA モデルの導入が必ずしも性能向上を約束しないことを実証し，「データ規模に見合ったモデル選択」と「失敗の様態」の重要性を定量・定性の両面から明らかにした点にある.

7.2 今後の展望

本研究で浮き彫りになった課題を解決するため，以下の方向性を展望とする.

- データ収集の自律化とスケールアップ：現在の 200 エピソードという規模が最大のボトルネックであることは明白である. テレオペレーションによる手動収集には限界があるため，ALOHA のような低コストなマスタースレーブシステムの導入や，シミュレーション（Sim2Real）によるデータ増強を行い，データ規模を数千件オーダーへ拡大する必要がある.
- 生成タスクの分離と軽量化：UVA の失敗は「動画生成」と「行動生成」を結合したことにあった. 小規模データ環境下では，未来画像の生成（画素レベルの予測）を放棄し，CLIP や R3M といった事前学習済み視覚エンコーダを用いて「潜在表現レベル」でのみ未来予測を行う，あるいは単純な Diffusion Policy（Action Only）に回帰することで，計算資源とデータ効率のバランスを最適化できると考えられる.
- 高解像度化と空間的注意機構：リサイズによる情報損失を防ぐため，高解像度画像をパッチ分割して入力する ViT アーキテクチャの導入や，把持対象周辺を動的に切り出す（Crop）Attention 機構の実装が求められる. これにより，低解像度の制約下でも精密な操作が可能になると期待される.

参考文献

[1] Yevgen Chebotar and Tianhe Yu. Rt-2: New model translates vision and language into action. Google DeepMind Blog, July 2023.

[2] Remi Cadene, Simon Alibert, Alexander Soare, et al. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. GitHub repository, 2024.

[3] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

[5] NVIDIA, Johan Bjorck, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[6] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[7] Figure. News. Company news page.

[8] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.