LLM の常識的知識を活用した行動認識のための物体中心特徴量の自律エージェント型学習

関井 大気 [1,*]　　佐藤 文彬 [1,*]
[1] サイバーエージェント
{sekii_taiki, sato_fumiaki}@cyberagent.co.jp

概要

動画基盤モデルに基づく近年の行動認識は，背景や物体の共起関係など行動と無関係な外観を過学習する一方，物体検出結果を入力として用いる手法は，入力の情報量の欠落が汎化を阻害する要因となる．本稿では，このようなトレードオフを緩和することを目的として，これら 2 種類の手法を組み合わせることで，互いの欠点を補い合う特徴量の学習を目指す．具体的には，人物の行動に関する LLM の常識的知識を活用し，自律エージェント型学習の枠組みの中で，LLM エージェントが 2 種類の手法を統合して，対象行動に合った動作特徴量を自動設計するフレームワークを実現する．実験では動作認識の公開データセットを用いて，提案手法の有効性を確認した．

1　はじめに

大規模言語モデル（Large Language Model，LLM）と視覚・言語事前学習の飛躍的な進歩は，マルチモーダル LLM とその多くの応用に結実した [1, 2, 3]．マルチモーダル LLM に対してテキストや画像に加えて動画像 [4] の入力が可能となり，さらには，音声認識モデル [5] を導入することにより，テキストなしで音声と動画を介してコンピュータと人がコミュニケーションできるアプリケーションが実現されている [1]．この潮流の中で，モデルが 2 次元の動画像中の人物の行動を認識する能力がより注目されるようになり，人と協調動作するロボット [6] やスポーツ [7]，エンターテインメント [8] といった数多くの応用において重要性が増している．

近年の行動認識技術は，視覚･言語事前学習 [9, 10] を通じて動画から比較的従来より汎化した表現を獲得できるようになった．これは，インターネット上に集積された動画とテキストのキャプションの大量のペアからなる大規模なデータセットの恩恵を受けている．獲得される表現は言語埋め込みの空間にアライメントされているため，多様なシーンにおいてプロンプトを通じてゼロショットでの行動認識が可能となった．このゼロショット推論によって幅広い用途で行動認識を活用できるようになった一方，事後学習時に獲得される表現の汎化性が，情報量がスパースなキャプションの範囲に限定される．この問題は偽相関やショートカット学習として顕在化し，認識対象の行動の表現に無関係の前景や静止した背景の外観，物体間の共起関係をモデルが過学習する結果になる [11]．特にこの問題は，外観でなく動作の理解が必要とされるベンチマーク [12, 13] において精度が低下する現象として確認されてきた．

一方，動作理解を目的としたアプローチが近年研究されており，動画ベース手法 [14, 15, 16] や物体中心手法 [17, 18, 19] が提案されている．この 2 つのアプローチはそれぞれ異なる方法で偽相関の影響の少ない特徴量を抽出する．前者は，キャプションを用いず動画内の動きを予測する自己教師あり学習により，動画基盤モデルを事前学習する．後者は，動画から物体領域 [17] や注目領域 [20] を特定した上で，それらを入力として教師あり学習 [17, 18] などの方法で特徴量を抽出する．

両者とも動画内の動きや物体に着目することで，動作理解に有益な特徴量をそれぞれ抽出する．しかしながら動画ベース手法は，視覚・言語事前学習と同様，動画の時空間ボリュームが入力であることから，入力の冗長性に起因する偽相関を完全には排除できない．これに対して物体中心手法は，偽相関の原因となる物体以外の領域の情報をあらかじめ入力から除き，モデルが着目する領域を制限することで，検出対象の物体の行動を比較的頑健に認識できる．ただし，入力の情報量が欠落し，検出対象以外

＊　著者は同等に貢献．

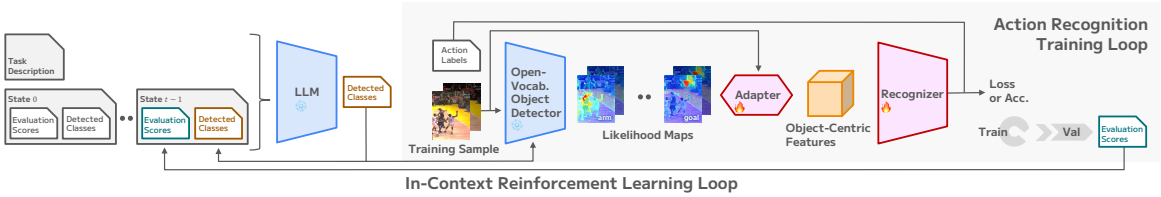

図 1: 提案手法の概要（2 章参照）.

の物体の認識が阻害されるため，動画ベース手法との間でトレードオフの関係にある．本研究では，このような情報量の欠落を避けることで，両者のトレードオフを緩和することを目指す．

一方ここまでの内容とは別の文脈において，近年では LLM のもつ常識的知識を行動認識に活用することが盛んである．例えば従来研究 [19] では，自己教師あり学習を通じて獲得された特徴量を，テキストとともに視覚トークンとして LLM に入力する．LLM は，言語のモダリティで学習された人物の行動に関する知識を活用することにより，視覚トークンから人物の行動を識別する．これらの従来研究では，LLM は認識器や識別器と呼ばれるヘッドの役割でシステムの後段で活用されるが，この場合，バックボーンの特徴抽出器の学習に LLM の知識は活用されない．これに対し AutoML 領域において，LLM は特徴抽出器の実装に活用されるが，認識対象の行動に依存しない実装を学習しているため，対象行動に関する知識を特徴抽出の学習に反映できない [21].

以上を踏まえ本研究では，Something-Something v2（SSv2） [12] などの教師あり動作認識のベンチマークを対象として，より頑健な特徴量を学習する方法の実現に取り組む．本稿では，トレードオフの関係にある動画ベース手法と物体中心手法を組み合わせることで，互いの欠点を補い合う特徴量の学習を目指す．特に，先述の人物の行動に関する LLM のもつ常識的知識を活用し，自律エージェント型学習の枠組みの中で LLM エージェントが動作理解の 2 つのアプローチを統合して，対象行動に合った動作特徴量を自動設計するフレームワークを実現する．

具体的には，自己教師あり動画基盤モデルを対象行動の認識に必要な動画中の物体に着目させるため，物体検出技術を用いて動画中の対象物体とそれ以外の物体をモデルに教示する．ここで，検出対象の物体の内訳は，LLM エージェントのもつ常識的知識によって対象行動から導かれるとともに，文脈内強化学習（In-Context Reinforcement Learning, ICRL） [22] の枠組みを通じて汎化するよう最適化される．図 1 に提案手法の ICRL の概観を示す．検出対象の物体クラスは，オープンボキャブラリ物体検出器（Open-Vocabulary Object Detector, OVD）を介して各フレームで検出される．検出結果を入力の動画と合わせて特徴量に変換し動画基盤モデルに入力するアダプターを新たに導入することで，事後学習において動画基盤モデルの特徴抽出層を対象物体にのみ着目させることが可能となる．実験では，動作認識の公開データセットを用いて提案手法で学習される特徴量の頑健性を検証する．

本稿の技術的な貢献は，(1) LLM の常識的知識が認識対象の行動に関係する物体に着目した動作特徴量の設計に有効であり，(2) ICRL を用いることで，OVD の検出対象とすべき物体のクラスを自動的に最適化できることを示すことの 2 点である．

2 提案手法

提案手法のパイプラインを図 1 に示す．本研究が着目するタスクは教師あり行動認識であり，学習にはラベル付き動画を用いる．OVD は，対象行動の認識に必要な物体を入力の動画像の各フレームで検出する．次に，物体検出結果の尤度マップをアダプター（線形層）により 3 チャンネルの特徴量（以降，物体中心特徴量）に変換し，入力の動画と加算した上で認識器に入力する．この物体クラスは LLM エージェントによって最大 $C_{\max}$ 通り設定される．

本研究では，認識器として動画基盤モデル（例えば V-JEPA2 [16]）を採用し，対象の学習データで認識器を事後学習することにより，大規模なデータセットを用いた事前学習で得られる特徴量の汎化性を転用する．アダプターからの物体中心特徴量を認識器の入力として各行動を分類するタスクを学習した後，行動認識精度を評価する．ここで，LLM エージェントによる物体クラスの選択は，単なる前処理ではなく認識器の特徴抽出層の学習に直接作用する．

ここまでが行動認識モデルの学習と評価であり，本研究ではさらに認識器に入力する物体中心特徴量を認識対象の行動に汎化させるために，LLM エー

ジェントを用いて学習と評価のプロセスを自動的に繰り返して検出対象の物体クラスを最適化する．LLM エージェントへの入力は，生成タスクの指示文（認識対象の行動の説明や出力形式）と，それまでの行動認識モデルの学習の履歴（前回までの検出対象の物体クラスや精度評価の結果）であり，LLM エージェントは履歴に応じて振る舞い，次の物体クラスを決定する．このプロセスは，LLM エージェントが物体クラスを決定するという制御行動を最適化する ICRL と捉えられ，また，最適な物体クラスは ICRL の副産物として獲得される．ただし，強化学習としての厳密な収束判定は設けず，ハイパーパラメータ探索に準じてあらかじめ定めた最大反復回数 T まで反復を行う．

2.1 ICRL に基づく学習フレームワーク

2.1.1 問題設定と記法

行動認識用のラベル付き動画を $\mathcal{D} = \{(X_n, y_n)\}_{n=1}^{N}$ とし，学習用の集合 $\mathcal{D}_{\text{train}}$ と，報酬の算出・方策選択に用いる検証用の集合 $\mathcal{D}_{\text{val}}$ を準備する．$\mathcal{D}_{\text{val}}$ は ICRL の各反復の最後に認識器の評価に使用される．

LLM エージェントが行う生成タスクの記述を c とする．c は学習対象のデータセット，認識対象の行動，出力形式などを説明するテキストから構成され，各反復で不変である．エージェントが各反復 t で観測する状態 s_t は，エージェントの過去の制御行動 a_{t-1} と報酬 r_{t-1} の履歴 H_{t-1} のみからなり，次式のように定義される．

$$s_t = H_{t-1} = (a_1, r_1), \ldots, (a_{t-1}, r_{t-1}), \tag{1}$$

ただし，a_t は OVD に与える物体クラスのテキストの集合であり，最大要素数を $C_{\max}$ とする．報酬 r_t は，a_t に基づく行動認識モデルの学習の後に $\mathcal{D}_{\text{val}}$ に対して得られる行動認識精度である（例えば分類精度）．

2.1.2 方策と環境

LLM エージェントの方策を $\pi_\phi(a \mid s, c)$ とする．ICRL の各反復 t において，エージェントは生成タスク c と履歴 s_t を連結したテキストを入力として制御行動 a_t を生成する．環境は a_t をシステムに入力し，OVD を用いた物体検出，アダプターによる特徴抽出，および認識器と識別器から構成される行動認識の学習と評価を行い，報酬 r_t を得る．状態は $s_{t+1} = H_{t-1} \cup (a_t, r_t)$ として更新される．したがって方策最適化は，パラメータ更新を伴わず履歴 H を手がかりに LLM の文脈内学習を通じてのみ逐次的に改善される．このとき，方策の変化は各反復での新たな行動認識の学習を通じて認識器の重みに反映されるため，LLM のもつ常識的知識が ICRL を介して間接的に特徴抽出器へ蒸留される．

3 評価実験

3.1 実験設定

本研究では，動作理解のベンチマークとして SSv2 [12] と MultiSports [13]（MS）を用いる．また，評価指標として行動分類精度を用いる．LLM エージェントとして GPT-5 ベースの ChatGPT [1] を利用し，ICRL の反復を 5 回行い検出対象の物体クラスを最適化した．OVD として Grounded SAM2 の `small` モデル [23] を用い，フレームごとに候補物体の尤度マップを算出した．行動認識モデルの学習には V-JEPA2 の公開実装（モデルは `ViT-L` [24]）を用いた．

3.2 従来研究に対する比較実験

3.2.1 Few-Shot 学習設定での評価

少数のサンプルのみ学習する Few-Shot の設定において，従来手法と提案手法の行動認識精度を比較した結果を表 1 に示す．同表より，単純に少数の学習サンプルで事後学習した V-JEPA2 でさえ，その他の State-of-The-Art（SoTA）手法よりも高い精度を達成できることがわかる．これは，`ViT-L` ベースで実装された V-JEPA2 が，SoTA 手法より表現力の高い DNN アーキテクチャを用いていること（CPR-CLIP は `ViT-B` を利用）に加え，大規模な動画データセットを用いて動作理解に適した自己教師あり学習を行ったことに起因する．しかしながら，提案手法を用いて V-JEPA2 の事前学習済みモデルを事後学習することでさらに高い精度を達成できている．これは，提案手法が LLM のもつ常識的知識を OVD を通じて行動認識モデルに転移させて，特徴量の頑健化を促した結果である．

3.3 Ablation Study

本節では，精度評価に MultiSports データセットを採用し，実験の簡略化のためにサブセットの学習サ

表 1: Few-Shot 設定における従来手法と提案手法の行動認識精度（%）の比較結果.

Method	Param.	SSv2 2-shot	SSv2 4-shot	SSv2 8-shot	MS 2-shot	MS 4-shot	MS 8-shot	Avg.	Δ (Avg.)
Vision-Language Pretraining									
TC-CLIP [25]	0.15B	7.3	8.6	9.3	-	-	-	-	-
CPR-CLIP [26]		8.0	9.1	10.2	-	-	-	-	-
Self-Supervised Pretraining									
V-JEPA2 [16]	0.3B	10.0	20.5	33.3	22.9	31.9	41.1	26.6	0.0
V-JEPA2 + Ours		**11.7**	**22.6**	**36.1**	**23.7**	**33.4**	**42.7**	**28.4**	**+6.8**

表 2: ICRL の有効性の検証結果．†：3 度 ICRL を実行し精度の平均を報告.

Method	Acc. (%)
PoseConv3D++	56.1
Ours w/o ICRL†	56.9
Ours†	**59.0**

表 3: ICRL の全反復で生成された物体クラスリストの言語埋め込みの分散.

Ours	Var.
w/o ICRL	0.10
w/ ICRL	**0.07**

ンプル（約 5,000）で行動認識モデルを学習した.

ICRL の有効性の検証　はじめに，提案手法の ICRL の有効性を検証するために，2 種類のベースライン手法を導入する．一方は，物体中心のアプローチ（PoseConv3D [17]）を最新の動画基盤モデルを用いて精度が向上するよう再実装したもの（PoseConv3D++）である．もう一方は，状態の履歴を用いず生成タスクの記述のみを用いて物体クラスを生成することで，ICRL の学習の効果を除いた方式（Ours w/o ICRL）である．ただし，提案手法と同じ反復回数学習を行う.

これらのベースライン手法と提案手法の精度を比較した結果を表 2 に示す．同表より，提案手法の精度はベースライン手法を上回っていることから，提案手法は LLM の常識的知識を活用することにより，物体中心アプローチで盛んに用いられる人物骨格や，提案する ICRL なしで LLM により予測された物体クラスよりも優れた物体クラスを獲得できることがわかる．また表 3 に示すように，ICRL を用いないベースライン手法と提案手法それぞれで生成される全反復における物体クラスリストを ChatGPT により言語埋め込みに変換し分散を比較すると，提案手法の埋め込みが一定の値に収束し分散が小さくなっている.

ICRL の各反復において検証用集合に対して得られた精度の推移を図 2 に示す．反復を重ねるごとに物体クラスが最適化され（Listing 1 参照），精度が改善していることがわかる．追加されたクラスの一部は背景に対応しており，対象行動の認識に不要な情報を捉えていると考えると，これは図 3 の特徴量の背景が視覚的に暗くなっている現象と整合する.

表 4: 計算量に関する Ablation Study.

Method	Acc. (%)	Time (ms)
V-JEPA2	58.8	776
Ours	**59.8**	764

表 5: LLM の選択に関する Ablation Study．3 度 ICRL を実行し精度の平均を報告.

Method	Acc. (%)
ChatGPT-4o	56.2
ChatGPT-5	**59.0**

```
# Before ICRL
object_classes_before = [
  "person", "ball",
  "net", "hoop", "goalpost",
  "arm", "leg", "foot"
]
```

```
# After ICRL
object_classes_after = [
  "person", "ball",
  "net", "hoop", "goalpost",
  "arm", "leg", "foot",
  "torso", "forearm", "head",
  "floor", "court", "grass", "rim"
]
```

Listing 1: 最適化される前後の物体クラスの例.

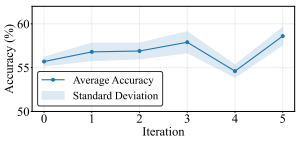

図 2: ICRL における反復中の行動認識精度（%）の遷移．3 度 ICRL を実行し精度の平均を報告.

図 3: 提案手法における入力の動画像の例と物体中心特徴量の可視化結果.

計算量に関する検証　提案手法は，OVD を前処理に利用する点において，ベースラインの V-JEPA2 よりも多くの計算量を必要とする．したがって，V-JEPA2 においてアンサンブルする認識器のモデル数を増加（6 倍）させて提案手法と推論時の計算量を完全に一致させた上で，精度を比較した結果を表 4 に示す．同表より，提案手法が V-JEPA2 を超える精度を達成していることから，精度と処理速度のトレードオフの観点からも，提案手法はより頑健な特徴量を抽出できることがわかる.

LLM の選択　提案手法で ICRL に用いる LLM を 2 パターン切り替えて精度を比較した結果を表 5 に示す．同表より，LLM の性能に合わせて学習される行動認識モデルの精度が改善していることから，LLM の表現力が行動認識モデルの学習の試行錯誤に直接的に影響していることがわかる.

4　まとめ

本稿では，人物の行動に関する LLM の常識的知識を活用し，自律エージェント型学習の枠組みの中で，LLM エージェントが近年の行動認識の手法を統合して，対象行動に合った動作特徴量を自動設計するフレームワークを提案した．実験では，動作認識の公開データセットを用いて提案手法で学習される特徴量の頑健性を確認した.

参考文献

[1] OpenAI. ChatGPT.

[2] Stability AI. Stable Diffusion.

[3] GitHub, Inc. GitHub Copilot.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2025.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. **arXiv preprint arXiv:2212.04356**, 2022.

[6] Lingyi Meng, Lin Yang, and Enhao Zheng. Hierarchical human motion intention prediction for increasing efficacy of human-robot collaboration. **IEEE Robotics and Automation Letters**, 9(9):7637–7644, 2024.

[7] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. A survey on video action recognition in sports: Datasets, methods and applications. **IEEE Transactions on Multimedia**, 25:7943–7966, 2023.

[8] Inc. Runway AI. Gen-4.5.

[9] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. **arXiv preprint arXiv:2508.18265**, 2025.

[10] Yi Wang, Xinhao Li, Ziang Yan, et al. Internvideo2.5: Empowering video mllms with long and rich context modeling. **arXiv preprint arXiv:2501.12386**, 2025.

[11] David Steinmann, Felix Divo, Maurice Kraus, Antonia Wüst, Lukas Struppek, Felix Friedrich, and Kristian Kersting. Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation, 2025.

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In **ICCV**, 2017.

[13] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In **ICCV**, 2021.

[14] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In **CVPR**, 2023.

[15] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. **arXiv preprint 2404.08471**, 2024.

[16] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. **arXiv preprint arXiv:2506.09985**, 2025.

[17] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting Skeleton-Based Action Recognition. In **CVPR**, 2022.

[18] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. SkeleTR: Towards Skeleton-based Action Recognition in the Wild. In **ICCV**, 2023.

[19] Haoxuan Qu, Yujun Cai, and Jun Liu. Llms are good action recognizers. In **CVPR**, 2024.

[20] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In **ECCV**, 2024.

[21] Alexander Tornede, Difan Deng, Theresa Eimer, Joseph Giovanelli, Aditya Mohan, Tim Ruhkopf, Sarah Segel, Daphne Theodorakopoulos, Tanja Tornede, Henning Wachsmuth, and Marius Lindauer. Automl in the age of large language models: Current challenges, future opportunities and risks. **TMLR**, 2024.

[22] Amir Moeini, Jiuqi Wang, Jacob Beck, Ethan Blaser, Shimon Whiteson, Rohan Chandra, and Shangtong Zhang. A survey of in-context reinforcement learning. **arXiv preprint arXiv:2502.07978**, 2025.

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. **arXiv preprint arXiv:2401.14159**, 2024.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **ICLR**, 2021.

[25] Minji Kim, Dongyoon Han, Taekyung Kim, and Bohyung Han. Leveraging temporal contextualization for video action recognition. In **ECCV**, 2024.

[26] Hao Song, Yangjun Ou, and Chen Wang. Cpr-clip: Cross-modal consistent and prompt-diverse regularized clip for action recognition. In **MMAsia**, 2025.

[27] Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. AutoML-agent: A multi-agent LLM framework for full-pipeline autoML. In **ICML**, 2025.

Algorithm 1 提案手法の ICRL フレームワーク.

Require: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, task description c
Require: number of iterations T, maximum set size $C_{\max}$, policy $\pi_\phi(a \mid s, c)$
Ensure: best object-class set a_{best} and best DNN weights W_{best},

$H \leftarrow \emptyset$ (H_0, hence $s_1 = H_0$)
$a_{\text{best}} \leftarrow \emptyset, r_{\text{best}} \leftarrow -\infty, W_{\text{best}} \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  $s_t \leftarrow H$ ($s_t = H_{t-1}$)
  $a_t \sim \pi_\phi(\cdot \mid s_t, c)$
  $a_t \leftarrow \text{Truncate}(a_t, C_{\max})$
  $W_t \leftarrow \text{TrainDNNs}(\mathcal{D}_{\text{train}}; a_t)$
  $r_t \leftarrow \text{Accuracy}(\mathcal{D}_{\text{val}}; a_t, W_t)$
  $H \leftarrow H \cup \{(a_t, r_t)\}$ ($H \leftarrow H_t$)
  if $r_t > r_{\text{best}}$ **then**
    $a_{\text{best}} \leftarrow a_t, r_{\text{best}} \leftarrow r_t, W_{\text{best}} \leftarrow W_t$
  end if
end for

return a_{best} and W_{best}

表 6: MultiSports データセットにおける LLM ベース AutoML 手法と提案手法の行動認識精度の比較結果. 3 度 ICRL を実行し精度の平均を報告.

Method	Acc. (%)
AutoML-based	26.3
Ours	**59.0**

A LLM ベース AutoML 手法との比較

LLM を用いる AutoML ベースの学習手法と提案手法の認識精度を比較することによって，提案手法による LLM の活用方法の有効性を検証する．提案手法の実装のうち，行動認識モデルの部分だけ従来研究 [27] と同様に LLM エージェントに実装させる手法をベースラインとする．ただしこれは，動画変換や評価プロトコルといった行動認識モデルの学習能力に関係のない部分の実装を手法間で共通化することで，公正な精度の比較を実現するためである．ベースライン手法は，提案手法のように，実装したモデルと評価結果を状態として ICRL により実装を反復的に改善する．提案手法とベースライン手法の認識精度の比較結果を表 6 に示す．同表より，提案手法はベースライン手法を超える精度を達成している．これは，ベースライン手法の LLM エージェントによる実装が既存知識の組み合わせに留まっている一方，提案手法が LLM の常識的知識を物体中心特徴量の発見に活用できていることを表している．

Listing 2: 提案手法において ICRL 時に LLM に入力するプロンプトの例.

```
You are an adaptive reasoning agent collaborating on the framework. Using the
 class label set and the iteration history, propose improved prompts to
enhance classification performance.
Your primary goal is to analyze the iteration history to identify weaknesses
and propose optimized prompt sets that directly improve overall and class-
wise classification accuracy. Focus on data-driven adjustments that lead to
measurable accuracy gains.
In each new round, adjust gradually with a few edits, avoiding large jumps to
 prevent training complexity.

## Output Format Instructions
CRITICAL: Output prompts as a plain text list, one per line, following these
rules STRICTLY:
- Each line = one single, atomic visual concept (short noun/noun phrase)
- NO headers, titles, numbering, or explanations
- Do not include any other text before or after the list

Note: open vocabulary object detector uses Grounded-SAM2. Keep prompts simple
 and atomic (single concept; avoid commas, slashes, logic words, or compound
phrases), as complex prompts are not reliably recognized.

## Instructions
* Diagnose likely causes of low accuracy (e.g., missing contextual cues,
insufficient body-part/object coverage, ambiguous wording).
* Design prompts that better capture discriminative evidence for
underperforming classes while keeping the vocabulary concise and
generalizable.
* Prefer compact, single-concept phrases suitable for open vocabulary object
detector (Grounded-SAM2-friendly).
* Prompt Count Policy (soft guidance): expand cautiously and incrementally;
prefer swapping out weak cues over adding many new ones at once.

## Framework
We propose an agent that turns commonsense into adaptive feature redesign for
 action recognition. For each action, a language model proposes a set of
evidence-bearing entities (objects, targets, body parts). These cues are
localized in video via text-conditioned, open vocabulary object detector (
Grounded-SAM2), yielding spatio-temporal regions that guide representation.
The agent then rewires the extractor on the fly-via region-conditioned
routing/adapters—to emphasize actor-object interactions and suppress
background bias. Learning proceeds in a closed loop: class-level rewards from
 recognition performance and attribution consistency drive in-context
reinforcement learning over the textual cues-implemented as iterative prompt
edits without updating the language model' s weights—followed by renewed
localization and feature updates. Thus, prompts, detections, and features co-
evolve as a single trainable pipeline.

## Classification Target Set
["aerobic push up", "aerobic helicopter", "volleyball serve", "volleyball
block", "football shoot", "football dribble", "basketball pass", "basketball
block"]

## Iteration History
### Iteration 1
Prompts: ["person", "ball", "net", "hoop", "goalpost", "arm", "leg", "foot"]

Overall Accuracy: 57.900%

Class-wise Accuracy:
aerobic push up: 0.00%
aerobic helicopter: 100.00%
volleyball serve: 82.03%
volleyball block: 96.73%
football shoot: 1.96%
football dribble: 88.57%
basketball pass: 93.91%
basketball block: 0.00%

### Iteration 2
..
```