動画に基づく教本参照型コーチングエージェントの構築

川田　拓朗†　　藤若　雅也††　　ジショウトン††　　劉　　健全††

† 法政大学大学院理工学研究科　〒 184–8584 東京都小金井市梶野町
†† NEC ビジュアルインテリジェンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部
E-mail: †takuro.kawada.3g@stu.hosei.ac.jp, ††{fujiwaka, xiaotong-ji, jqliu}@nec.com

あらまし　コーチングとは, 観測された行動を参照基準と比較し, その差分に基づいて改善を促す教育的支援である. 動画に基づくコーチングの既存研究の多くは, 学習者の動画に対してお手本動画を参照基準としてきたが, 高品質なお手本動画を大規模に用意することは困難である. 本研究では, 教本などのドキュメントを参照基準とし, 学習者の行動が, いつ・どの規範から・どの程度, 基準から逸脱したかを推定するコーチングエージェントを提案する. 提案手法では, ドキュメントをルーブリックとして構造化し, 動画中の行動と整列させることで, 各時刻の逸脱度を定量化する. これにより, 参照動画や追加学習を必要としない解釈可能なコーチング支援を実現する.

キーワード　マルチモーダル検索, 動画ベースコーチング, 映像解析

1　はじめに

コーチングとは, 観測された行動を何らかの参照基準と比較し, その差分に基づいて学習者の理解や行動の改善を促す教育的支援である [2]. この考え方は, スポーツ指導や医療教育, 専門職教育など多様な分野において, 効果的なフィードバックや指導の在り方として広く議論されてきた [3], [9], [7]. 近年, 大規模言語モデル (Large Language Models; LLM) や視覚言語モデル (Vision Language Models; VLM) の進展に伴い, 人間の指導プロセスを自動化あるいは支援するコーチングモデルへの関心が高まっている. これらの研究では, 学習者の行動を入力として解析し, 適切な助言や改善点を提示することで, 人手による指導を補完または代替することが目指されている.

スポーツや技能学習の分野では, 学習者の動作を動画として入力し, 熟練者のお手本動画との比較に基づいてフィードバックを生成する参照動画ベースのコーチング手法が提案されている [1], [10]. このような参照動画に基づくアプローチでは, 学習者と熟練者の動作を時間的に対応付けた上で, 姿勢や運動の時間変化に基づく表現を用いて両者の動作差分を明示的に捉えることができるため, 改善すべき箇所を直接的に特定できるという利点がある. しかし, 多くの既存手法では技能や運動種目ごとに収集された熟練者動画を用いたモデルの学習を前提としているため, 新たな技能や指導内容に適用する際には追加のデータ収集や再学習が必要となる. 特に, 専門性の高い技能や限定的な指導領域において, その内容を十分に反映した高品質なお手本動画を大規模に収集することは容易ではなく, これら手法の適用範囲や拡張性は限られている.

一方, 実際の技能学習や指導の現場では, 書籍や教本といった文書資料が参照基準として広く用いられている. このようなドキュメントには, 技能の手順や注意点, 評価観点が文章や図解として体系的に整理されており, 技能に関する規範的知識を明示的に提供している. しかし, これらの記述は人間による理解を前提として構成されており, 実世界の学習者の動画中に含まれる連続的な行動に対して, ドキュメント内のどの記述を対応付けるべきかはあらかじめ明示されていない. また, 技能の記述方法や粒度, 図解の有無といった文書の構成やスタイルはドキュメントごとに大きく異なるため, 熟練者動画との比較のように時系列的な対応関係や動作差分を直接導出することは容易ではない. したがって, ドキュメントを参照基準として動画中の行動を評価するためには, 記述された技能知識を動画から判定可能な構造へと整理した上で, それらと動画中の行動との対応関係を体系的に推定する枠組みが求められる.

本研究では, ドキュメントを参照基準とするコーチングエージェントの構築を目的とし, ドキュメントに記述された技能知識を手順構造および評価規範からなるルーブリックに変換し,

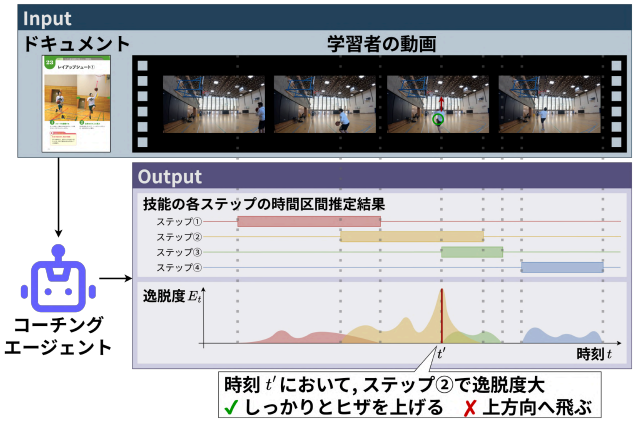

図 1: 学習者動画と教材ドキュメントを入力とし, 技能遂行を時間構造と品質を可視化する提案手法の出力例. 参照動画や追加学習を必要とせず, 入力動画における行動が「いつ・どの行動が・どの行動規範から・どの程度」ドキュメントに記述された規範を参照基準から逸脱しているかの特定を可能とする.[1]

1：ドキュメントのスクリーンショットは文献 [11] より引用.

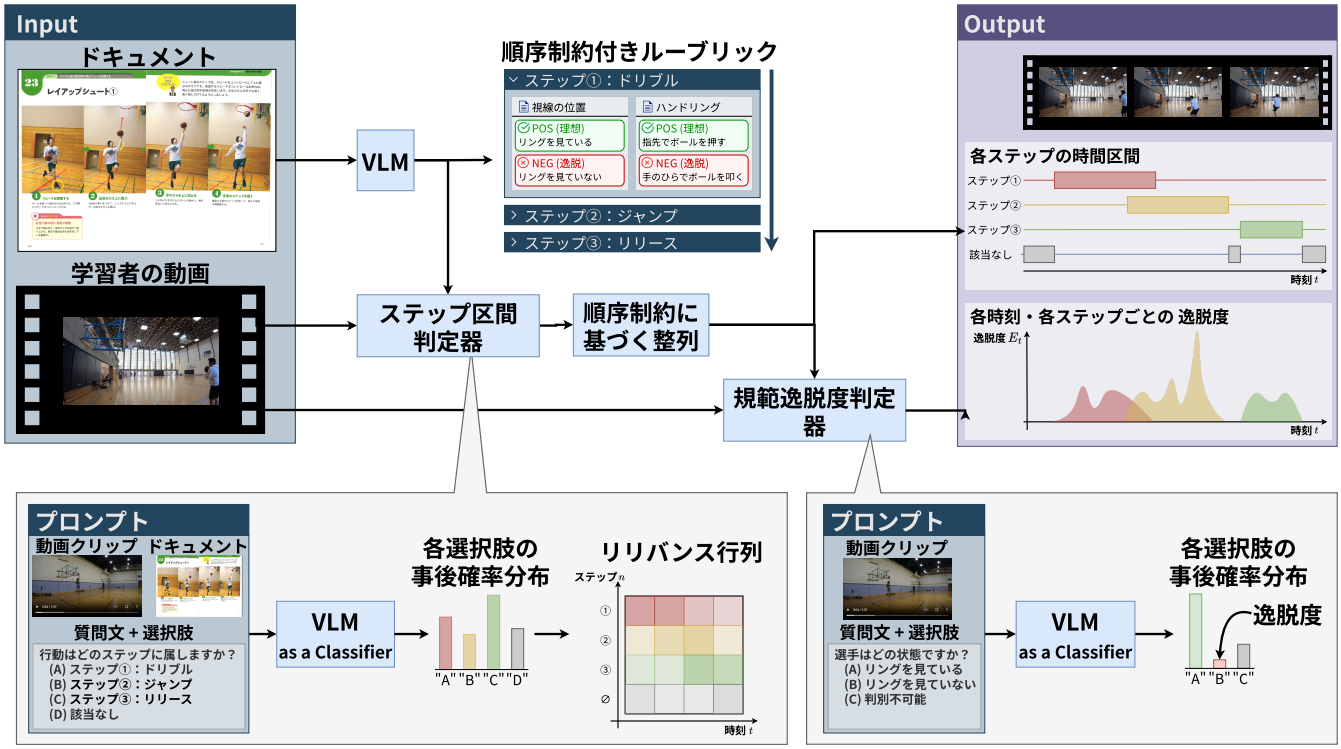

図 2: 提案するコーチングエージェントの概要図．ドキュメントと学習者の動画を入力とし，(1) 文書記述を手順ステップおよび評価規範からなる順序制約付きルーブリックへ変換し，(2) 各ステップが動画中のどの区間に対応するかを順序制約の下で推定・整列し，(3) 各ステップに定義された規範に基づいて行動の逸脱度を時系列的に算出する．ドキュメント内の知識と動画中の行動を一貫した枠組みで対応付けることで，技能遂行の時間構造と品質を同時に把握可能な解析基盤を実現する．[2]

学習者の動画中の行動がそれらの規範からどの程度逸脱しているかを定量的に推定する枠組みを提案する．我々が提案する枠組みは以下の 3 段階の処理から構成される: (1) まず，ドキュメント内の記述を解析することで当該技能の一連の動作をステップに分解し，各ステップに対応する評価規範を抽出・構造化することで動画中の行動を評価可能なルーブリックを構築する．(2) 次に，各ステップが動画内のどの区間に該当するかを順序制約を考慮しながら推定することで，動画全体に対するステップ区間を整列する．(3) 最後に，各ステップごとに，学習者がドキュメント内に記載された評価規範を満たしているか否かを判定し，その結果を参照基準からの逸脱度として定量化する．この枠組みにより，入力された動画に対して，「いつ・どの行動が・どの行動規範から・どの程度」逸脱しているのかを推定し，フィードバックとしてユーザに提供することが可能となる (図 1)．我々は収集が困難なお手本動画を用いることなく，ドキュメントに基づく学習不要なコーチングエージェントの構築した．

本研究の貢献は以下の通りである:

- ドキュメントに記述された技能知識を構造化された順序付きのルーブリックへ変換し，動画内の行動を評価するために利用可能とする枠組みを示した．
- ドキュメントと動画の間に時系列的な対応関係が明示されていない状況において，技能の手順ごとの時間区間推定を導入し，後段の行動評価と整合的な形で動画中の行動と文書知識を対応付ける方法を示した．
- ドキュメントを参照基準として用いることで，参照動画や追加学習を必要とせず，動画中の各時間区間における行動の逸脱度を定量的かつ解釈可能に推定できることを示した．

2 提 案 手 法

本研究では，ドキュメントを参照基準として用い，学習者の動画中の行動をステップ単位で評価する手法を提案する．入力として学習者の動画およびドキュメントの該当ページを与え，出力として動画中の各時間区間に対応するステップと，各ステップに定義された規範に対する逸脱度を推定する．これにより，学習者は入力した動画の各時刻において，どのような規範をどの程度逸脱したかを明示的に把握することが可能となる．提案手法は，図 2 に示すように，以下の 3 つの処理から構成される: (1) ルーブリック構築: ドキュメント内の記述をステップおよび評価規範からなるルーブリックへ変換する; (2) ステップ区間推定: 動画中の各クリップと各ステップとの対応区間を推定する; (3) 逸脱度推定: 各ステップに定義された規範に基づき，動画中の行動の逸脱度を算出する．

2.1 ルーブリック構築

我々はドキュメントに記述された技能知識を，動画中の行動を評価可能な参照基準として利用するため，ドキュメントの内容をルーブリックとして構造化する． ルーブリックとは，技能の遂行過程を構造化された手順と評価観点として明

2：ドキュメントのスクリーンショットは文献 [11] より引用．

示的に定義した評価基準を指す．本研究においては，技能の遂行過程を表す順序付きのステップ集合 $S = \{s_i \mid i \in \mathcal{I}\}$, $\mathcal{I} = \{1, 2, \ldots, N\}$ と，各ステップに対応する規範項目の集合 $R^{(i)} = \{r_j^{(i)} \mid j \in \{1, 2, \ldots, M^{(i)}\}\}$ から構成されるチェックリスト形式の評価基準としてルーブリックを定義する．ここで s_i は技能手順における第 i ステップに関するドキュメント内の記述を表し，N はステップ数である．また，$r_j^{(i)}$ は第 i ステップにおいて，ある行動がドキュメントの記述に沿っているかを判定するための評価項目を表し，$M^{(i)}$ は規範項目の数である．

ステップ抽出． ステップ S は，情報抽出器として VLM を用いてドキュメント内に記述された技能手順を解析することで構築される．この処理は，ドキュメント内の記述を変更することなく，技能の遂行過程を複数の手順段階に分割し，各段階に対応する説明文を対応付けることを目的とする．VLM への入力には，該当技能について記載された該当ページ群のスクリーンショット画像および PDF テキストを用いる．これらの入力を基に，各ステップを簡潔に表すタイトルと関連する全てのドキュメント内の文の集合を抽出する．このように構築された各ステップ s_i は，後続の動画解析において単一の時間区間に割り当て可能な技能遂行上の動作単位として構成される．

規範項目への極性付与および補完． 各ステップ s_i に対応する規範項目集合 $R^{(i)}$ は，ステップ抽出段階で得られた各ステップに関連する全ての説明文を VLM に入力し，評価可能な命題へと整形することで構築する．各規範項目 $r_j^{(i)}$ は，行動がドキュメントに記述された規範を満たしている状態を表す命題文 $r_{j,\mathrm{POS}}^{(i)}$ と，満たしていない状態を表す命題文 $r_{j,\mathrm{NEG}}^{(i)}$ のペアからなる二値的な表現として定義される．これら命題は第 i ステップを実行している学習者の行動が，どちらの状態に該当するかを動画から判定可能な形式で記述される．VLM は，ドキュメント内の記述を基に，これら 2 つの状態を区別するための命題文を生成・整形し，両者が意味的に補集合となるよう規範表現を構成する．ここで，ドキュメントの記述形態によっては，これら 2 つの状態のいずれか一方のみが明示されている場合がある．すべての規範項目を一貫して二値的に扱うため，記載されている命題を基に，対応するもう一方の状態を VLM を用いて補完する．この補完処理は，ドキュメントの記述内容を逸脱しない範囲で行われ，新たな評価基準や知識は付加されない．

以上の処理により，ドキュメントは各ステップの順序構造と，各ステップに対応する二値的な評価規範集合からなるルーブリックへと変換される．各ステップおよび規範項目は，ドキュメント内の記述に基づいて構成されており，評価の根拠を原文に対応付けて保持している．本ルーブリックは，ドキュメントに記述された技能知識を手順構造と評価観点の両面から構造化したものであり，後続のステップ区間推定および逸脱度推定における参照基準として用いられる．

2.2 ステップ区間推定

続いて，学習者動画中において，各ステップが実行されている時間区間を推定する．この問題は，クエリとなる各ステップがあらかじめ順序を持つという制約の下で，対応する動画中の時間区間を検索する Video Moment Retrieval タスクとして捉えられる．本研究では，以下の 2 つの処理でステップ区間推定を行う: (1) リリバンス行列の構築: 動画中の各時間区間と各ステップとの関連度を推定する; (2) 順序制約に基づく整列: 各ステップの順序制約に基づき，リリバンス行列を用いて動画中の区間を一貫した形で整列する．

リリバンス行列の構築． まず，学習者動画 V を固定フレーム長 τ の短いクリップ列 $V = \{c_t \mid t \in \{1, 2, \ldots, T_{\mathrm{clip}}\}\}$ に分割する．ここで c_t は動画中の時刻 t に対応するクリップ，T_{clip} はクリップ数を表す．動画の総フレーム数を T_{frame} とすると，$T_{\mathrm{clip}} = \lceil T_{\mathrm{frame}}/\tau \rceil$ となる．

次に，各動画クリップ c_t と，各ステップ s_i の関連度 $p_{t,i}$ あるいは，いずれのステップにも属さない状態との関連度 $p_{t,\emptyset}$ を推定し，これらを要素として持つリリバンス行列 $P \in \mathbb{R}^{T_{\mathrm{clip}} \times (N+1)}$ を構築する．ここで，$N+1$ 列目はいずれのステップにも属さない状態に対応する．我々は動画クリップとドキュメント中の技能記述という異なるモダリティ間の対応関係を VLM を用いて推定する．動画クリップ c_t，該当ページのスクリーンショット画像，「このクリップがどのステップを実行しているか」を問う質問文，選択肢となるステップ集合 S および，いずれのステップにも属さない状態における選択肢を VLM に入力し，$N+1$ 択の QA タスクとして推論させ，各選択肢に対応する識別子トークンを 1 文字生成する．逐次的なトークン生成に基づいて推論を行う VLM が最初に出力する各選択肢識別子トークンの対数尤度に softmax 関数を適用することで各クリップと各ステップの関連度 $p_{t,n}$ を定義する．

このように VLM を確率推定可能な分類モデルとして用いることで，ドキュメントに含まれるテキスト，図表などの複雑な視覚的情報，VLM が持つ行動理解能力を活用し，異なるモダリティ間の対応関係を連続値として柔軟に推定できる．

順序制約に基づく整列． リリバンス行列 P は各クリップごとに独立に推定されるため，局所的な誤りやノイズを含む可能性がある．また，各ステップはドキュメントに定義された順序に従って単調に進行するという構造的制約を持つ．そこで，ステップ順序の制約を明示的に考慮するため，動画全体に対するステップ区間を推定する．

まず，各時刻 t における動画中の行動に対応するステップを，ステップ番号の集合列 $\mathcal{A} = \{A_t \subseteq \mathcal{I}\}$ として表す．$i \in A_t$ であるとき，時刻 t において，第 i ステップ s_i がアクティブであることを意味する．また，$A_t = \{\emptyset\}$ であるとき，時刻 t において，いずれのステップにも該当しない区間であることを表す．ステップ番号が時間と共に単調に進行するという制約の下，$A_t \neq \{\emptyset\}$ ならば，開始ステップ $\alpha_t \in \mathcal{I}$ と同時にアクティブなステップ数 $L_t \coloneqq |A_t|$ を用いて，$A_t = \{\alpha_t, \alpha_t + 1, \ldots, \alpha_t + (L_t - 1)\}$ となる．ただし，$\alpha_t + (L_t - 1) \leqq N$ である．

次に，各時刻 t におけるスコア ϕ_t を，割り当てられたステップ集合に含まれる確率の最大値として次のように定義する:

$$\phi_t = \max_{n \in A_t} p_{t,n} \tag{1}$$

そして，全時刻におけるスコアの和が最大となる区間系列を推

定するため, 以下の最適化問題を解く:

$$\mathcal{A}^* = \underset{\mathcal{A}}{\operatorname{argmax}} \sum_{t=1}^{T} \phi_t. \tag{2}$$

ここで, ステップ境界付近における曖昧性を表現するため, 連続する時刻間で共有されるステップ数が高々 d 個となるよう, $|A_t \cap A_{t+1}| \leqq d$ という制約を課す. ここで, $d \geqq 1$ である. これにより, 各ステップの時間的独立性を考慮するとともに, 全ての時刻において $A_t = I$ となるような自明解を防ぐ. また, 各ステップは動画中で 1 つの連続区間として出現するものとし, $\max A_t \leqq \min A_{t+1}$ という制約を課す. これにより, 各ステップの進行の単調性を担保し, 一度終了したステップが再度出現するような遷移を防ぐ. 本研究では, この最適化問題を動的計画法で解き, 各ステップの時間区間を推定する. A^* は後続の逸脱度推定において参照される.

2.3 逸脱度推定

最後に, 2.1 節で構築したルーブリックと 2.2 節で推定したステップ区間に基づき, 学習者動画中の行動が各規範をどの程度逸脱しているかを定量的に評価する. 各時刻 t においてアクティブなステップ群 A_t において, 各ステップで定義された規範項目集合 $R^{(i)}$ を用いて行動評価を行い, どの規範がどの程度満たされていないかを定量化する.

我々は 2.1 節のリリバンス行列の推定と同様に, VLM を分類モデルとして用い, 各クリップが規範項目を満たしているか否かを推定する. 動画クリップ c_t, 「このクリップにおける行動はどの状態に該当するか」を問う質問文, 選択肢となる命題 $r^{(i)}_{j,\mathrm{POS}}$, $r^{(i)}_{j,\mathrm{NEG}}$ および, いずれの状態にも該当しない場合の選択肢 r_{UNK} を VLM に入力し, 3 択の QA タスクとして推論させ, 各選択肢に対応する識別子トークンを 1 文字生成する. VLM が最初に出力する各選択肢識別子トークンの対数尤度を取得し, softmax 関数を適用することで, 動画クリップ c_t が各状態に属する確率を取得する. ここで, 規範を満たしていない状態 $r^{(i)}_{j,\mathrm{NEG}}$ が選択される確率を, 時刻 t における規範項目 $r^{(i)}_j$ に対する 逸脱度 $e^{(i,j)}_t$ として定義する. この定義により, 動画の品質や撮影アングル, 遮蔽などの要因によって規範の判定が困難な場合には, r_{UNK} に対応する確率が大きくなり, 結果として逸脱度 $e^{(i,j)}_t$ は小さく抑えられる. これにより, 観測情報が不十分な状況において誤って逸脱を検出することを避け, 評価の不確実性を反映させる. 最終的に, 時刻 t における行動全体の逸脱度 E_t は, アクティブなステップ群 A_t に含まれる全ての規範項目に対する逸脱度の総和として定義する:

$$E_t = \sum_{s_i \in A_t} \sum_{r^{(i)}_j \in R^{(i)}} e^{(i,j)}_t. \tag{3}$$

このように定義された逸脱度は, 動画中の各時間区間において, どの程度ドキュメントに記述された規範から逸脱しているかを連続値として表現するものであり, 後続のコーチング支援において定量的かつ解釈可能な指標として利用可能である.

3 評価実験

3.1 実験設定

本研究では, スポーツの練習動画とそれに対応する指導書を用い, 提案手法により動画中の行動が指導書に記述された規範からどの程度逸脱しているかを推定する実験を行った. 我々は動画データとして, Ego-Exo4D データセット [4] に含まれるバスケットボール練習動画 280 件を使用した. 本データセットには, 学習者が実際にバスケットボールの基礎的な技能練習を行っている様子が三人称視点で収録されており, 撮影環境や被写体のばらつきを含む実環境に近い条件が含まれている. また, 参照基準となるドキュメントとして, バスケットボールの基礎技能を解説した指導書 [11] を用いた. 本教本は文章および図解を用いて技能の手順や注意点を説明しており, Ego-Exo4D データセットの動画が対象とするジャンプシュートおよびレイアップに関する全てのページを対象とした.

本研究において用いる VLM は, ルーブリック構築, ステップ区間推定, 逸脱度推定のすべての段階において GPT-5.2 [6] を使用した. また, 各動画は固定フレーム長 $\tau = 4$ の短いクリップ列に分割し, 提案手法の各段階においてこれらのクリップを基本単位として処理を行った. ステップ区間推定の順序制約に基づく整列において, 隣接する時刻間で共有されるステップ数を $d = 1$ とし, ステップの切り替わりに伴う短時間の曖昧さを表現しつつ, 不自然な長時間の重なりを抑制した.

ステップ区間推定におけるリリバンス行列の構築については, 提案手法である VLM に QA を解かせて対数尤度を取得する手法に加えて, 既存の Video Moment Retrieval 手法である InternVideo2 [8] および R^2-Tuning [5] を比較手法として用いた. これらの手法では, 各ステップの説明文と動画クリップとの類似度に基づいてリリバンス行列を構築し, 提案手法と同一の動的計画法による整列処理を適用した.

3.2 評価指標

ルーブリック構築の妥当性評価. 提案手法により構築されたルーブリックが, ドキュメントに記述された内容を過不足なく構造化できているかを検証するため, 人手による妥当性評価を行った. 提案手法では, ルーブリック構築を以下の 2 段階に分けて行っている: (1) ステップ抽出; (2) 規範項目への極性付与および補完. そこで, 我々はそれぞれの段階に対応した観点を設定し, 定性的にその妥当性を評価した. まずステップ抽出について, 次の 3 つの観点から妥当性を評価した: (i) 原文忠実性: 各ステップがドキュメント原文から直接導出可能な記述に基づいて構成されているか; (ii) ステップ混在の不存在: 1 つのステップに, 異なる技能段階における動作が混在していないか; (iii) 手順的曖昧性の不存在: ステップが技能遂行のどの段階に対応するかが不明確でないか. 次に, 規範項目への極性付与および補完について, 次の 3 つの観点から妥当性を評価した: (iv) 意味逸脱の不存在: 規範項目がドキュメントに記載されていない新たな評価基準や解釈を含んでいないか; (v) 極性の妥当性:

命題 $r^{(i)}_{j,\mathrm{POS}}$ および $r^{(i)}_{j,\mathrm{NEG}}$ が意味的に自然な補集合として構成されているか; (vi) 動画判定可能性: 動画を観察することで, 当該規範を満たしているか否かを判断可能な形式になっているか. 各観点に関して, 構築されたルーブリックが条件を満たすか否かを手動で二値的に判断し, 全評価対象項目のうち条件を満たした割合を指標とする.

ステップ区間推定の妥当性評価. 提案手法におけるステップ区間推定の妥当性を検証するため, 推定された各ステップの時間区間が, 人間の直感的な理解と整合した大まかな区切りを捉えているかを評価した. 我々は動画全体を観察し, 各ステップが実行されていると考えられる時間区間を手動で指定した. そして, モデルによって推定されたステップ区間と手動でアノテーションされた区間との一致度を, 各ステップごとの Intersection over Union (IoU) により評価し, 提案手法および比較手法について結果を比較した. ここで, IoU は推定区間と正解区間の重なり長を, それらの和集合長で割った値として定義される.

熟練度に基づく逸脱度の比較. 提案手法により推定される逸脱度が技能の熟練度の違いを反映しているかを検証するため, 熟練者の動画と初心者の動画をそれぞれ入力した際の逸脱度の差の比較を行った. 初心者と熟練者との分割には, コーチングモデルに関する先行研究 ExpertAF [1] における拡張データセットのラベルを用いた. 我々は, 各手法について初心者群と熟練者群の時間方向の平均逸脱度の差 Δ(初心者 − 熟練者) を比較した. この差は, 提案する逸脱度が技能レベルの違いをどれだけ明確に反映しているかを表す量であり, 値が正の方向に大きいほど初心者の動画においてより高い逸脱度が付与されていることを意味し, 熟練度の違いを適切に捉えられていることを示す.

逸脱度の定性評価. 提案手法により推定される逸脱度が, 人間の直感的な良否判断と整合した挙動を示すかを検証するため, 定性的な評価を行った. 我々は少数の動画サンプルを対象とし, 動画全体の中から明らかに上手くできている区間, および明らかに不適切である区間を手動で選択した. その上で, 時系列方向における逸脱度 E_t の推移を可視化し, 不適切と判断されたクリップが出現する区間において, 逸脱度が相対的に高くなる傾向が見られるかを定性的に評価した.

4 結果と考察

ルーブリック構築の妥当性評価. 表 1 に, 各教材ページから構築されたルーブリックに対する妥当性評価結果を示す. ステップ抽出に関する観点である (i) 原文忠実性 および (iii) 手順的曖昧性の不存在 は, すべての対象において高い達成率を示した. また, 多くの場合において (ii) ステップ混在の不存在 も高い値を示し, 構築されたルーブリックにおいて, 技能手順が概ね明確な単位として分割されていることが確認された. 規範項目への極性付与および補完に関しては, (v) 極性の妥当性 が一貫して高い値を示し, POS/NEG が意味的に自然な補集合として構成されていることが確認された. さらに, (iv) 意味逸脱の不存在 も高い達成率を示しており, 構築された規範項目が原文に含まれない新たな評価基準を過剰に導入していないことが示された. 一方で, (vi) 動画判定可能性 は他の観点と比較して相対的に低い値となった. すなわち, 構築された規範項目の一部は, 動画から直接二値判定するには曖昧さを含む形式となっていることが確認された.

表 1: ルーブリック構築の妥当性評価結果. ステップ抽出および規範項目の極性付与・補完という二段階の処理に対して, 各評価観点を満たすと判断された項目の割合を示す.

評価する処理段階	評価観点	妥当な項目の割合
ステップ抽出	原文忠実性	1.000
	ステップ混在の不存在	1.000
	手順的曖昧性の不存在	1.000
極性付与・補完	意味逸脱の不存在	0.967
	極性の妥当性	1.000
	動画判定可能性	0.833

これらの結果は, いずれも提案手法の設計方針に起因するものである. 提案手法は, 技能手順を再解釈して再編成するのではなく, 原文中に現れる節構造や記述単位を保持したまま, ドキュメントの構造をルーブリック形式へ忠実に写像することを目的として設計されている. そのため, 原文においてステップが明示的かつ時間的に整合的に分割されている場合には, 高い (i) 原文忠実性 と (iii) 手順的曖昧性の不存在 が得られる. 一方で, 原文側の分割が曖昧であったり, 複数の動作段階が一つの節に混在している場合には, その構造的曖昧さや時間的非整合性もまたそのままルーブリックに反映される. また, 原文に含まれる抽象的・感覚的表現も忠実に写像されるため, 一部の規範項目は (vi) 動画判定可能性 の観点からは不十分な形式となる. このように, 本手法は原文忠実性を最大化する設計であるがゆえに, 評価基準としての可観測性や, 手順構造を再編成する柔軟性との間にトレードオフを内包している.

以上より, 本手法はドキュメントに記載された知識を過不足なく構造化するという目的に対して有効であり, とりわけ技能手順がステップとして明示的に記述された教材に対しては妥当なルーブリックを自動的に構築できることが確認された. 一方, 原文中の手順構造が曖昧であったり, 感覚的・抽象的な表現に依存する教材に対しては, その曖昧さ自体がルーブリックに反映されるという限界も明らかとなった. このようなドキュメントに対しても適用可能な枠組みとするためには, 原文構造に依存せずに潜在的な手順境界を推定する仕組みや, 抽象的な記述を動画上で観測可能な行動表現へと変換する補正処理を組み込むことが重要である.

ステップ区間推定の妥当性評価. 表 2 に, ステップ区間推定の妥当性評価結果を示す. 提案手法は 0.415 ± 0.230 の IoU を達成し, InternVideo2（0.199 ± 0.157）および R^2-Tuning（0.230 ± 0.095）を大きく上回った. この結果は, 提案手法が各ステップの位置と広がりを, 人手で指定した区間と整合する形で安定して推定できていることを示す. 従来の Video Moment Retrieval 手法では, 各ステップの説明文を固定的なテキスト表

表 2: ステップ区間推定手法ごとの妥当性評価および熟練度に基づく逸脱度の比較結果. 各手法が利用する入力情報 (説明テキスト／スクリーンショット) とともに, ステップ区間推定精度を IoU により評価する. 逸脱度は, 初心者群および熟練者群における平均値と標準偏差, 両者の差 Δ（初心者 − 熟練者）を報告する. Δ が正で大きいほど, 初心者に対して一貫して高い逸脱度が付与されており, 推定結果が熟練度の違いをより明確に反映していることを示す.

ステップ区間推定手法	入力情報		ステップ区間 IoU	逸脱度の平均		
	説明テキスト	スクリーンショット		初心者	熟練者	Δ(初心者 − 熟練者)
InternVideo2 [8]	✓	✗	0.199 ± 0.157	2.748 ± 1.007	2.704 ± 1.019	0.044
R^2-Tuning [5]	✓	✗	0.230 ± 0.095	6.590 ± 1.378	6.400 ± 1.370	0.190
ours (GPT-5.2 [6])	✓	✗	0.312 ± 0.154	0.848 ± 0.321	0.732 ± 0.400	0.116
	✓	✓	$\mathbf{0.415 \pm 0.230}$	0.928 ± 0.440	0.695 ± 0.388	**0.233**

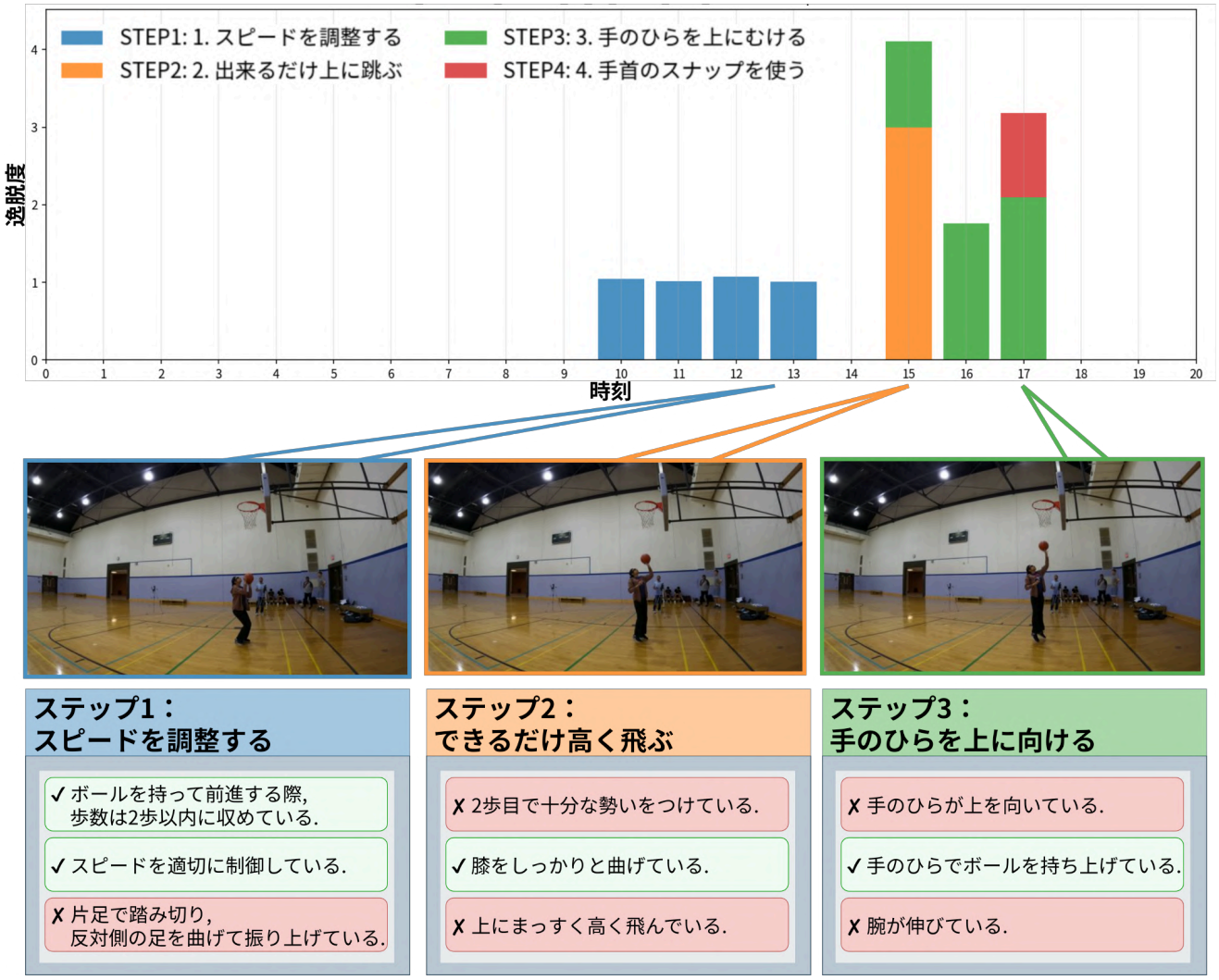

図 3: 逸脱度推定結果の可視化例. 上段は各時刻における各ステップごとの逸脱度を示し, 下段は対応する動画フレームと, 各ステップで定義された評価規範を満たしているか否かを手動で判定結果を示している. 逸脱度が大きく推定された区間では, 対応するステップにおいて複数の評価規範が満たされていないことが確認できる.

現へと埋め込み, 動画クリップとの類似度に基づいて対応付けを行う. 一方, 提案手法ではテキストと図解を含むドキュメント全体と動画クリップを VLM に与え, 「このクリップはどのステップに対応するか」という QA 形式の問題として定式化することで, VLM を確率出力可能な分類器として利用している. この枠組みにより, 本来は対話や記述理解を目的として学習された VLM を追加学習を行うことなく Video Moment Retrieval の文脈へと適用できる. 入力情報に関するアブレーション結果を見ると, 説明テキストのみの場合でも IoU は 0.312 ± 0.154 と既存手法を上回っており, この定式化自体が zero-shot かつ

専門的な技能に対して有効に機能していることが分かる．さらに，スクリーンショットを併用すると IoU は 0.415 ± 0.230 まで向上し, 図やイラストに含まれる視覚的文脈をそのままクエリに取り込むことで曖昧な記述や技能特有の表現と動画中の状態との対応付けがより安定することが確認できる．このように, 提案手法は教本に含まれるテキストだけでなく, 図やイラストといった視覚情報を含む文脈全体をクエリとして扱うことができる．これは VLM を基盤とすることで初めて可能となる性質であり, 固定的なテキスト表現に基づく従来の Video Moment Retrieval 手法では原理的に扱えない情報である．教本が本来備えているマルチモーダルな表現を歪めることなく検索過程に持ち込める点は, 専門的かつ曖昧な技能記述に対しても頑健に対応できるという, 本手法の重要な特徴である．

熟練度に基づく逸脱度の比較． 表 2 に, 熟練者動画群および初心者動画群における逸脱度の比較結果を示す．いずれの手法においても, 初心者群の逸脱度平均は熟練者群より大きい傾向を示すが, その差の大きさには手法間で顕著な違いが見られた．InternVideo2 の差は 0.044 と小さく，熟練者と初心者の分布がほぼ重なっていることが示唆された．R^2-Tuning では差が 0.190 と一定の分離が見られ, 提案手法では差が 0.233 と最も大きく, 熟練度差をより明確に反映できていることが確認できた．特筆すべき点は, 本手法が熟練度ラベルを直接用いることなく, 手順に基づく規範への逸脱という観点から算出された指標のみで, 結果として初心者と熟練者を分離できていることである．これは, 提案手法によって構築されたルーブリックおよびステップ区間推定により, 技能遂行に内在する質的な差異を適切に捉え, それを逸脱度として定量化できていることを示唆している．この傾向は, 提案手法が推定する逸脱度がステップ依存の規範集合に基づき計算される点と整合的である．すなわち, ステップ区間推定が妥当であるほど各クリップは適切なステップ文脈の下で評価され, 初心者に特有の違反や規範未達が逸脱度として顕在化する．一方, ステップ割当てが不安定な場合, クリップが無関係なステップ規範と照合されることで逸脱度が過度に増減し, 熟練度差がノイズに埋もれやすい．したがって, ステップ区間推定の精度が逸脱度を熟練度指標として安定に機能させる上で重要であることが示された．

逸脱度の定性評価． 図 3 は，提案手法によって推定された各時刻における逸脱度と，対応するステップの評価規範に基づく逸脱判定の例を示している．ステップ 1 およびステップ 4 に対応する区間では逸脱度が相対的に小さい一方で，15 クリップ目付近ではステップ 3，17 クリップ目付近ではステップ 4 に対応する逸脱度が大きくなっていることが確認できる．各規範項目に対する人手による判定結果と比較すると, 逸脱度が大きく推定された区間では対応するステップにおいて満たされていない規範項目の数が相対的に多いことが確認できる．すなわち, 提案手法によって推定された逸脱度の増減は, 人手判断と整合した挙動を示すことを示している．

5 おわりに

本研究では, 書籍や教本に記述された技能知識を参照基準として用い, 学習者動画中の行動を評価するコーチングエージェントの構築手法を提案した．提案手法では, ドキュメント内の記述を手順ステップと評価規範からなるルーブリックとして構造化し, ステップの順序制約を考慮した区間推定を行った上で, 各時刻における行動の逸脱度を定量的に算出する．評価実験の結果, 提案手法は参照動画や追加学習を用いることなく, 動画中の行動をドキュメントに基づく規範と対応付けられることが確認された．また, 推定されたステップ区間は人手による大まかな区切りと整合する傾向を示し, さらに算出された逸脱度は熟練度の違いや, 人手で確認した規範未達の区間と対応した挙動を示した．これらの結果は, 提案手法が技能遂行の時間構造と品質を同時に捉えるための基盤として有効であることを示す．

一方, 動作が極めて短時間で生じるステップや, 空間的に局所的な運動に依存する規範に対しては, 区間推定や逸脱度推定の精度に課題が残ることも確認された．今後の方向として, より細粒度な時間分解能での解析や, 文書記述の曖昧さを補正する仕組みを導入することで, より幅広い技能や教材に適用可能なコーチング支援へと拡張していくことが有望である．

文　献

[1] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. ExpertAF: Expert actionable feedback from video. In *CVPR*, 2025.

[2] Adelle Atkinson, Christopher J. Watling, and Paul L. P. Brand. Feedback and coaching. *European Journal of Pediatrics*, Vol. 181, pp. 441–446, 2022.

[3] Phillip Dawson, Michael Henderson, Patrick Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, Vol. 44, No. 1, pp. 25–36, 2019.

[4] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *CVPR*, 2024.

[5] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R^2-Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. 2024.

[6] OpenAI. GPT-5, 2025. `https://platform.openai.com/docs/models/gpt-5`.

[7] Stephanie J. Sohl, Deborah Lee, Heather Davidson, Blaire Morriss, Rebecca Weinand, Katherine Costa, Edward H. Ip, James Lovato, Russell L. Rothman, and Ruth Q. Wolever. Development and validation of the Health Coaching Index. *Patient Education and Counseling*, Vol. 104, No. 3, pp. 642–648, 2022.

[8] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. 2024.

[9] Benedikt Wisniewski, Klaus Zierer, and John Hattie. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, Vol. 10, , 2020.

[10] Wei-Hsin Yeh, Yu-An Su, Chih-Ning Chen, Yi-Hsueh Lin, Calvin Ku, Wenhsin Chiu, Min-Chun Hu, and Lun-Wei Ku. CoachMe: Decoding Sport Elements with a Reference-Based Coaching Instruction Generation Model. In *ACL*, 2025.

[11] 森圭司. 目で学ぶシリーズ 3 見るだけでうまくなる！バスケットボールの基礎. ベースボール・マガジン社, 東京, 2020. 第 1 版 第 1 刷.