対話型ビデオ理解モデルにおける動作特徴量の活用

中溝 雄斗[†1,a)]　柳井 啓司[†1,b)]

概要：近年，ビデオ理解分野では大規模言語モデルを活用し，対話的なビデオ理解を可能にしたモデルが登場している．しかし，既存のモデルではビデオの各区間に含まれる動作については注目されていない．そこで本研究では，動作特徴を用いた対話型ビデオ理解モデル Act-ChatGPT を提案する．Act-ChatGPT は定量的な比較においてベースモデルを上回り，定性的な比較においても動作の認識などで応答を改善する例が確認された．

1. はじめに

近年，自然言語処理分野における大規模言語モデルの発展を受け，ビデオ理解分野でも視覚エンコーダーにより抽出された視覚特徴量を大規模言語モデルのトークン空間に射影することで視覚エンコーダーと大規模言語モデルを統合し，対話的なビデオ理解を可能にした対話型ビデオ理解モデルが提案されている．しかしながら，既存の対話型ビデオ理解モデルは，視覚エンコーダーとして画像言語モデルやビデオ全体のモデリングを意識したビデオ言語モデルを使用することが一般的であり，ビデオの各区間に含まれる動作に関しては注目されていない．一方で，近年ではビデオ領域においても Transformer [21] が普及し，自己教師あり学習の有効性が証明されたことにより，動作認識分野でも大量のビデオデータで事前学習された Transformer ベースのモデルが成功を収めている．これらのモデルは高い動作認識能力を有し，特にビデオセグメント単位で動作するモデルを用いることにより，ビデオの各区間から優れた動作特徴を抽出することが可能である．

そこで，本研究ではビデオの各区間に含まれる動作特徴を活用した対話型ビデオ理解モデルである Act-ChatGPT を提案する．Act-ChatGPT では視覚エンコーダーとして画像言語モデルを用いた Video-ChatGPT [18] に対して，ビデオセグメント単位で特徴を抽出する動作認識モデルを追加の視覚エンコーダーとして導入することにより，ビデオの各区間に含まれる動作特徴を大規模言語モデルの入力に追加している．また，Act-ChatGPT では従来のシングルエンコーダー方式とは異なり，デュアルエンコーダー方式を採用することにより，画像言語モデルの持つ物体認識能力と動作認識モデルの人間動作認識能力の両方を活用している．

[†1] 現在，電気通信大学
Presently with The University of Electro-Communications
[a)] nakamizo-y@mm.inf.uec.ac.jp
[b)] yanai@uec.ac.jp

2. 関連研究

2.1 大規模言語モデル

大規模なコーパスによる自己教師あり学習で事前学習された言語モデルを事前学習済み言語モデルという．近年，この事前学習済み言語モデルのモデルパラメータや学習データのスケーリングが下流タスクの性能向上に繋がる [9] という知見に基づき，非常に多くのパラメータを持ち，特に膨大なデータで学習された大規模な事前学習済み言語モデルが構築されている．それらのモデルは創発的能力 [27] と呼ばれる小規模な事前学習済み言語モデルでは見られなかった能力を有しており，一連の複雑なタスクを解く際に驚異的な能力を発揮することから，小規模な事前学習済み言語モデルと区別して，大規模言語モデル（Large Language Model: LLM）と呼ばれている [25]．大規模言語モデルは言語生成や常識的な推論を行う能力に秀でており，自然言語処理分野のみならず，他分野でもその活用が多方面から研究されている．例を挙げると，特に優れた指示応答性能が報告されている OpenAI の GPT-4 はモデル構造や重みが非公開であるものの，API を介した活用が可能であるため，データセットの作成やフィルタリング，入力テキストの拡張等に活用されている．また，LLaMA [7] はモデルや重みが公開されている大規模言語モデルであるため，Alpaca [19] や Vicuna [3] を始めとする多くの派生

モデルの基盤となり，特に大規模言語モデルをモデルの一部に組み込み，End-to-End で学習を行う研究において，広く活用されている．

本研究は視覚領域における大規模言語モデルの活用例の一つである対話型ビデオ理解モデルに関する研究であり，特にビデオの各区間に含まれる動作に着目することによる応答精度の向上を目的とした研究である．

3. マルチモーダル大規模言語モデル

現在の視覚領域におけるマルチモーダル大規模言語モデルは大きく分けて，大規模言語モデルを用いて様々な視覚タスクのエキスパートモデルを接続させる手法と，視覚エンコーダーにより抽出された視覚特徴量を大規模言語モデルのトークン空間に射影することで視覚モデルと大規模言語モデルを統合し，End-to-End で学習可能な一つのモデルを構築する手法に分けられる．前者の手法である Visual ChatGPT [2] は大規模言語モデルがユーザーからの指示と画像を理解し，指示の遂行に必要な外部の視覚基盤モデルを呼び出すことを可能にすることで，大規模言語モデルを介して多数のエキスパートモデルをつなぎ合わせたシステムである．Visual ChatGPT は緻密に設計されたプロンプトマネージャーにより視覚情報を言語に落とし込み，各視覚基盤モデルの機能や使い方等とともに大規模言語モデルに提供することで，大規模言語モデルがユーザーの指示に従い適切な視覚基盤モデルを活用することを可能にしている．一方で，後者の手法である Li ら [11] の手法では，視覚特徴量を大規模言語モデルのトークン空間に射影する際のモダリティギャップに着目し，そのギャップを埋めるための Q-Former と呼ばれる軽量なモジュールを導入したマルチモーダル大規模言語モデル BLIP-2 を提案した．BLIP-2 では，Q-Former を画像とテキストの対比，画像とテキストのマッチング，画像に基づくテキスト生成により学習することでモダリティギャップを埋め，より少ない学習パラメータで画像質問応答や画像テキストマッチングなどの視覚言語タスクにおける性能向上を実現した．また，同じく後者の手法である Liu ら [5] の手法では自然言語処理分野における Instruction Tuning [23] の成功に着目し，それを視覚領域に拡張した Visual Instruction Tuning 及びそれに従い fine-tuning されたマルチモーダル大規模言語モデル LLaVA を提案した．Instruction Tuning は，指示応答テキストで構成されるデータセットを用いて大規模言語モデルを fine-tuning することにより，指示応答性能を向上させる学習フレームワークである．なお，指示応答テキストとは，モデルに対する指示文とそれに対する適切な応答文のペアからなるテキストである．この学習フレームワークに従い Fine-Tuning されたモデルは，未知のタスクに対しても指示内容に従いタスクを実行する能力を獲得するため，未知のタスクに対するゼロショット性能が改善されることが示されている [23]．Liu らはこの学習フレームワークを視覚領域におけるマルチモーダル大規模言語モデルの学習に初めて導入し，優れたマルチモーダルな対話能力を実現した．

なお，本研究では後者の手法に着目しており，以降は後者を Vision-LLM と定義し，特にビデオ領域に焦点を置いた Vision-LLM を Video-LLM と定義する．

3.1 Video-LLM

ビデオ理解分野では，近年の大規模言語モデルの発展を受け，対話的なビデオ理解を可能にした Video-LLM が多数提案されている．既存の Video-LLM はビデオのエンコード手法によって，画像言語モデルを用いてフレーム単位でエンコードするモデルとビデオ言語モデルを用いてビデオ全体を一度にエンコードするモデルの 2 種類に分けられる．まず, 画像言語モデルを用いてフレーム単位でエンコードするモデルとしては，VideoChat [12] や Video-LLaMA [4], Video-ChatGPT [18]，LLaMA-VID [16] などが挙げられる．これらのモデルはビデオからサンプリングしたフレームから画像言語モデル CLIP [1] を用いて画像単位の特徴量を抽出した後に，プーリングや追加のモジュールにより，情報の圧縮やビデオ全体の時間軸モデリングを行い，得られた特徴量を全結合層を用いて大規模言語モデルのトークン空間に射影することで Video-LLM を構築している．

一方で，ビデオ言語モデルを用いてビデオ全体を一度にエンコードするモデルとしては VideoChat2 [13] や Video-LLaVA [17] などが挙げられる．これらのモデルは UMT [15] や LanguageBind [26] などのビデオ言語モデルを用いてビデオ単位の特徴量を抽出した後に，全結合層により大規模言語モデルのトークン空間に射影することで Video-LLM を構築している．しかしながら，これらに用いられているビデオ言語モデルでは効率化のため，ビデオ全体から 4〜16 フレームのみをサンプリングしており，ビデオ全体を通してのモデリングに焦点が置かれている．

そのため，これらに代表される既存の Video-LLM はビデオの時間的な特徴に対して，明示的なモデリングを行わない，もしくはビデオ全体を通してのモデリングに焦点を置いており，ビデオの各区間における動作については着目していない．したがって，本研究は Video-LLM に対して，ビデオの各区間に含まれる動作特徴を導入した点で既存手

法とは異なる．

3.2 動作認識モデル

近年では自己教師あり学習の有効性が証明されたことで，大量のビデオデータで事前学習されたモデルを fine-tuning することにより構築された VideoMAE v2 [20] や UMT [15] などの Transformer [21] ベースのモデルが動作認識モデルとしても優れた性能を残している．他方で，現在の動作認識モデルはそのフレームサンプリング戦略から 2 種類に分けられる．一つ目は密なサンプリングと呼ばれるビデオから複数の既定フレーム長のビデオセグメントをサンプリングする手法を採用するモデルであり，VideoMAE v2 などがこれに属する．二つ目は疎なサンプリングと呼ばれるビデオの長さに関わらず，ビデオ全体から既定の数のフレームをサンプリングする手法を採用するモデルであり，UMT などがこれに属する．それぞれ，前者はビデオの各区間における特徴を，後者はビデオ全体の特徴をモデリングすることが可能である．

本研究はビデオの各区間に含まれる動作特徴の活用を目的として，密なサンプリングを採用した動作認識モデルを Video-LLM に導入した初の研究である．

4. Act-ChatGPT

4.1 概要

Act-ChatGPT では，Video-ChatGPT [18] に動作特徴量を追加で導入することにより，新しい Video-LLM を実現する．Act-ChatGPT の概要図を**図 1** に示す．Act-ChatGPT は視覚エンコーダーとして，フレームから画像特徴を抽出する画像言語モデルとビデオセグメントから動作特徴を抽出する動作認識モデルを併用するデュアルエンコーダー方式を採用している．Act-ChatGPT では，まず入力されたビデオから T 個のフレーム $F \in \mathbb{R}^{T\times W\times H\times C}$ 及び T 個の 16 フレームのビデオセグメント $S \in \mathbb{R}^{T\times 16\times W\times H\times C}$ をサンプリングし，画像言語モデルにより前者から各フレームの画像特徴量 $v_f \in \mathbb{R}^{T\times N\times D_f}$ を，動作認識モデルにより後者から各セグメントの動作特徴量 $v_s \in \mathbb{R}^{T\times D_s}$ をそれぞれ抽出する．このとき，D_f，D_s はそれぞれ，画像言語モデルと動作認識モデルの埋め込み次元数であり，N は画像言語モデルのパッチサイズ p を用いて，$N = W/p \times H/p$ と表される．次にモデル間アダプターを用いて，各特徴量の大規模言語モデルのトークン空間への射影及び特徴量の融合を行い，画像特徴量 v_f と動作特徴量 v_s を視覚トークン $Q_v \in \mathbb{R}^{(2T+N)\times D_h}$ に変換する．このとき，D_h は大規模言語モデルのトークン空間の次元数である．なお，モデル間アダプターにおける変換については 4.3 節にて詳細に述べる．最後に変換された視覚トークン Q_v と入力されたテキストを変換して得られた言語トークン Q_t に基づき，Next token prediction により大規模言語モデルで応答となるテキストを生成する．また，Act-ChatGPT では学習コスト削減のため，視覚エンコーダーと大規模言語モデルに学習済みのモデルを活用し，モデル間アダプターのみを学習する．

4.2 使用モデル

Act-ChatGPT では画像言語モデル，動作認識モデル，大規模言語モデルにそれぞれ学習済みのモデルを用いる．まず，画像言語モデルには OpenAI CLIP [1] ViT-L/14 モデルを採用し，後ろから 2 層目の出力を画像特徴量として扱う．続いて，動作認識モデルには Kinetics-710 [14] で fine-tuning された VideoMAEv2 [22] ViT-g/14 モデルを採用し，最終層の出力の平均に Layer Normalization を適用した値を動作特徴量として扱う．最後に，大規模言語モデルには LLaVA [5] のために fine-tuning された Vicuna v1.1 [3] の 7B モデルを採用する．

4.3 モデル間アダプター

Act-ChatGPT のモデル間アダプターの概要図を**図 2** に示す．Act-ChatGPT のモデル間アダプターは画像特徴量変換モジュール，動作特徴量変換モジュール，特徴量融合モジュールの 3 種類のモジュールで構成される．以下では，各モジュールの構成要素を説明した後に，処理手順を説明する．

4.3.1 画像特徴量変換モジュール

このモジュールには，ベースである Video-ChatGPT [18] で用いられたモデル間アダプターを採用する．このモジュールにおける画像特徴量から画像特徴トークンへの変換では，まず画像言語モデルにより抽出された各フレームの画像特徴量 $v_f \in \mathbb{R}^{T\times N\times D_f}$ に対して，時間的及び空間的平均プーリングを行い，時間特徴量 $v_t \in \mathbb{R}^{T\times D_f}$ と空間特徴量 $v_n \in \mathbb{R}^{N\times D_f}$ を得る．その後，それらの特徴量を連結し，一層の全結合層 f_f を用いて大規模言語モデルのトークン空間への射影することで変換後の画像特徴トークン $Q_f = f_f([v_t, v_n]) \in \mathbb{R}^{(T+N)\times D_h}$ を得る．このとき，$[a, b]$ はベクトル a，b の連結を表す．

4.3.2 動作特徴量変換モジュール

このモジュールでは，各ビデオセグメントの動作特徴量間の関係のモデリングと大規模言語モデルのトークン空間への射影を行う．前者は各ビデオセグメント間の関係を考慮せず，連続する動作の一部を別の動作として誤認識する

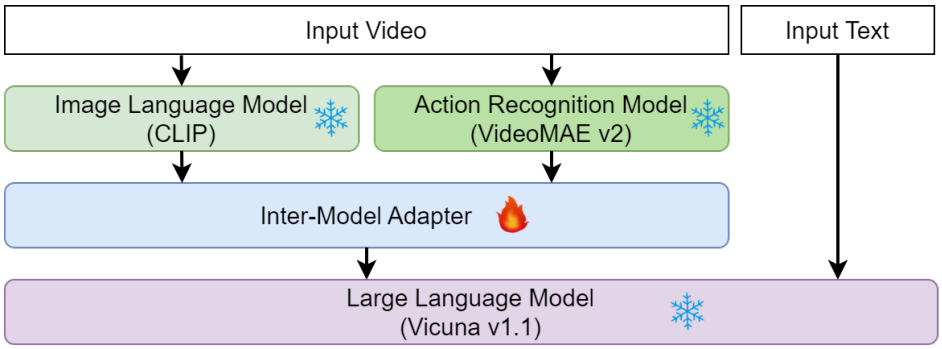

図 1 Act-ChatGPT の概要図

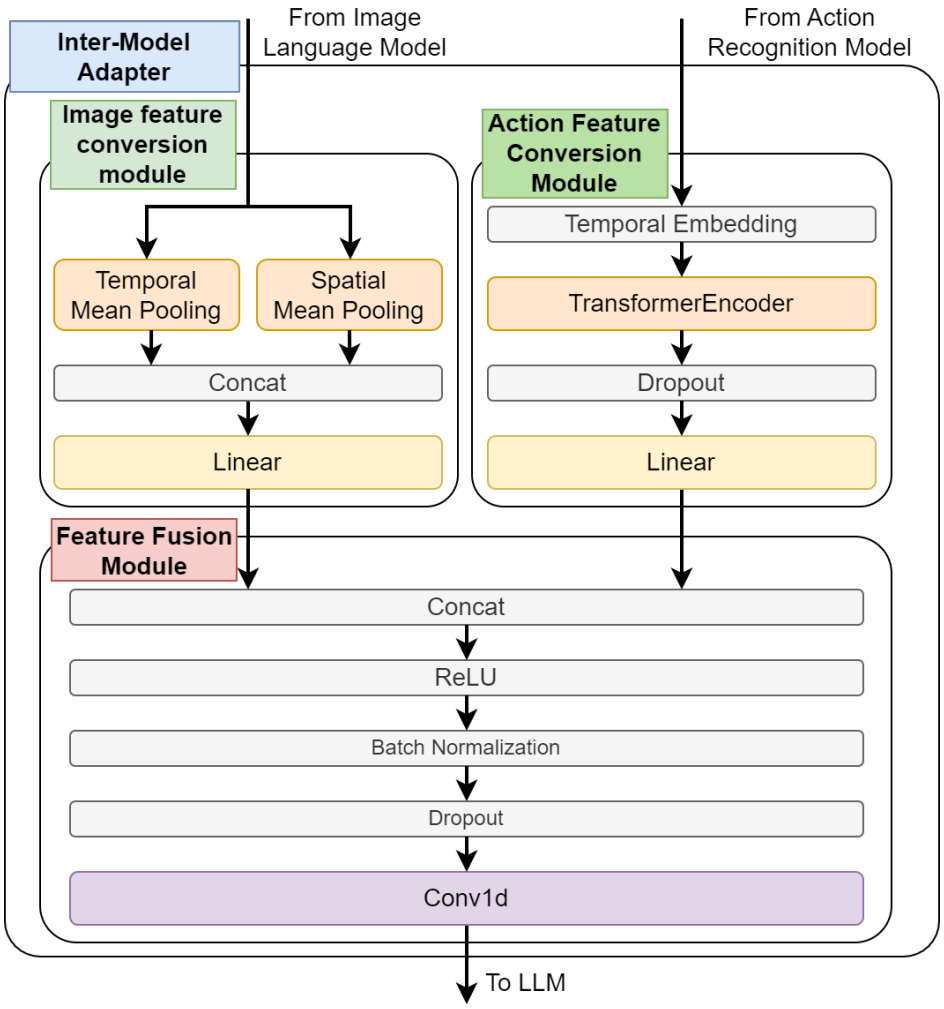

図 2 モデル間アダプターの概要図

ことを防止するためのものであり，時間埋め込みと TransformerEncoder を採用する．なお，TransformerEncoder のレイヤー数は 1，各レイヤーの Multi-Head Attention の heads 数は 2 である．一方で，後者には一層の全結合層を採用する．このモジュールにおける動作特徴量から動作特徴トークンへの変換では，まず動作認識モデルにより抽出された各ビデオセグメントの動作特徴量 $v_s \in \mathbb{R}^{T \times D_s}$ に時間埋め込みを施し，TransformerEncoder で変換することでビデオセグメント間の関係を考慮した動作特徴量 $v'_s = \mathrm{TransformerEncoder}(v_s + pos) \in \mathbb{R}^{T \times D_s}$ を得る．このとき，pos は時間埋め込みを表す．その後，Dropout 層及び一層の全結合層 f_s を用いて，動作特徴量 v'_s を大規模言語モデルのトークン空間への射影することで変換後の動作特徴トークン $Q_s = f_s(\mathrm{Dropout}(v'_s)) \in \mathbb{R}^{T \times D_h}$ を得る．

4.3.3 特徴量融合モジュール

このモジュールには，異なる性質を持つ二つの特徴量を明

示的に融合することを目的として，一層のカーネルサイズ 1 の 1 次元畳み込みを採用する．このモジュールにおける各特徴量の融合では，上記の特徴量変換モジュールにより得られた画像特徴トークン Q_f と動作特徴トークン Q_s を連結し，ReLU 層，Batch Normalization 層，Dropout 層，1 次元畳み込みで順次変換することで特徴量融合後の視覚トークン $Q_v = \mathrm{Conv1d}(\mathrm{Dropout}(\mathrm{BN}(\mathrm{ReLU}([Q_f, Q_s])))) \in \mathbb{R}^{(2T+N)\times D_h}$ を得る．

4.4 データ拡張

本研究では学習に用いるビデオ指示データセットに対して，データ拡張を適用する．このデータ拡張はモデルがテキストに過剰に反応することの抑制を目的としており，大規模言語モデルにより既存の指示応答テキストの指示内容を言い換えることにより行われる．詳細には，Vicuna v1.5 [3] 13B に対して，同義語や類義語を積極的に用いる，言い換えには外部情報を反映しない，与えられた指示と応答の関係を壊さない範囲で言い換えるという条件のもと，指示内容の言いかえを作成する．

4.5 学習

Act-ChatGPT では Vision Instruction Tuning [5] に従い，学習を行う．詳細には，Video-ChatGPT [18] に倣い，ビデオと指示応答テキストのペアで構成されるビデオ指示データセットを用いて，学習データとモデルの応答テキストの間のトークンごとのクロスエントロピー誤差を最小化することを目的として，学習を行う．また，Act-ChatGPT では学習を 2 段階に分けて行う．まず，1 段階目の学習では一方の視覚エンコーダーのみを用いて，それぞれに対応する特徴量変換モジュールを独立して学習する．なお，このときの画像特徴量変換モジュールを学習する場合のモデル構造は Video-ChatGPT と同様であり，Video-ChatGPT に倣い，画像特徴量変換モジュールに含まれる全結合層は LLaVA [5] の全結合層で初期化して学習する．続いて，2 段階目では各特徴量変換モジュールを 1 段階目の学習で得られた重みで初期化し，特徴量融合モジュールも含め，モデル間アダプター全体を学習する．

4.6 プロンプト

大規模言語モデルへ入力されるプロンプトは Video-ChatGPT [18] に従い，以下のテンプレートに基づいて作成する．

USER: ⟨Instruction⟩ ⟨Video-token⟩ ASSISTANT:

このとき，⟨Instruction⟩ はビデオに関連する質問などの大規模言語モデルに対する指示を表し，⟨Video-token⟩ は視覚特徴量を表す．また，USER: と ASSISTANT: はユーザーによる指示内容と大規模言語モデルによる応答内容を分け，特にマルチターンの対話において大規模言語モデルが対話の流れを適切に理解することを補助するために用いる．Act-ChatGPT では上記のテンプレートの ⟨Instruction⟩ を指示テキストで置換してトークンに変換した後に，⟨Video-token⟩ に対応するトークンをモデル間アダプターにより得られた視覚トークン Q_v に置換し，大規模言語モデルに入力する．

5. 実験

5.1 実験設定

本実験ではフレーム及びビデオセグメントのサンプリング数 T は Video-ChatGPT [18] に倣い，$T = 100$ とし，各 Dropout 層のパラメータは 1 段階目の学習では $p = 0.0$，2 段階目の学習では $p = 0.5$ とした．また，推論時に大規模言語モデルの生成トークンの確率分布を調整し，モデルの創造性を制御する temperature パラメータ τ は特に言及がない限り，$\tau = 0.2$ とした．学習では 2 段階の学習で共通のデータ及び学習設定を用いた．学習データセットには Video Instruction Dataset [18] を用いた．Video Instruction Dataset は ActivityNet [6] のサブセットに基づき構築された約 100,000 ペアのビデオとシングルターンの指示応答テキストからなるビデオ指示データセットであり，BLIP-2 [11] や GPT-3.5 などを用いてビデオの内容に関連する指示応答テキストを作成することで構築されている．オプティマイザには AdamW を採用し，2×10^{-5} をピークとしたウォームアップ率 0.03 の linear warmup と cosine decay による学習率スケジューリングを実施した．各学習の epoch 数は 3 とした．

評価指標には Video-based Generative Performance Benchmarking [18] と AutoEval-Video [24] を採用した．まず，Video-based Generative Performance Benchmarking では，ActivityNet に基づくテストセットに対するモデルの出力を正答との比較に基づき，GPT-3.5 により 0 から 5 のスコアで評価した．この評価では各データは Correctness of Information（CI），Detail Orientation（DO），Contextual Understanding（CU），Temporal Understanding（TU），Consistency（C）の 5 項目に分類されており，以下では項目ごとの GPT-3.5 によるスコアの平均を報告する．そして，AutoEval-Video では，複数の能力ドメインとトピックにわたって Youtube[*1] からベンチマーク用に

[*1] https://www.youtube.com/

表 1 Video-based Generative Performance Benchmarking による評価結果

	CI↑	DO↑	CU↑	TU↑	C↑
Video-ChatGPT	2.41	2.59	3.00	2.07	2.19
Act-ChatGPT（w/o Aug）	2.48	2.63	3.11	2.13	2.07
Act-ChatGPT	**2.57**	**2.69**	**3.17**	**2.24**	**2.32**

表 2 AutoEval-Video による評価結果（全体）

	All↑
Video-ChatGPT	**0.107**
Act-ChatGPT（w/o Aug）	0.067
Act-ChatGPT	0.058

表 3 AutoEval-Video による評価結果（項目別）

	DP ↑	STP ↑	ComR ↑
Video-ChatGPT	**0.089**	**0.125**	**0.211**
Act-ChatGPT（w/o Aug）	0.043	0.094	0.158
Act-ChatGPT	0.022	0.062	0.158
	REK ↑	ER ↑	PR ↑
Video-ChatGPT	**0.088**	**0.091**	**0.125**
Act-ChatGPT（w/o Aug）	0.062	0.061	0.062
Act-ChatGPT	0.050	0.061	0.062
	D ↑	CouR ↑	CMP ↑
Video-ChatGPT	0.033	**0.158**	0.000
Act-ChatGPT（w/o Aug）	**0.067**	0.053	0.000
Act-ChatGPT	0.033	0.105	0.000

独自に収集され，注釈が付けられたデータセットに対するモデルの出力を，各データに定義された特定の評価ルールに基づき，GPT-4 により正誤で評価した．この評価では，各データは Dynamic Perception（DP），State Transition Perception（STP），Comparison Reasoning（ComR），Reasoning with External Knowledge（REK），Explanatory Reasoning（ER），Predictive Reasoning（PR），Description（D），Counterfactual Reasoning（CouR），Camera Movement Perception（CMP）の 9 項目に分類されており，以下では各項目の正答率と全体の正答率を報告する．

なお，本研究では，既存のビデオ LLM 評価において最も一般的に使用されている方法である前者の評価を重視した．一方，後者の評価は，学習データとは全く異なる方法で収集され，注釈が付けられたデータセットに基づいているため，モデルの汎化性能を確認するために使用した．また，本実験における評価では，GPT3.5 には gpt-3.5-turbo-0613 のチェックポイントを，GPT-4 には gpt-4-1106-preview のチェックポイントを使用した．

5.2 ベースモデルとの比較

Act-ChatGPT と Video-ChatGPT [18] の Video-based Generative Performance Benchmarking [18] による定量評価の結果を**表 1** に示す．公平な比較を行うために，データ拡張を除いた場合の結果も（w/o Aug）として示している．Act-ChatGPT は Video-ChatGPT と比較した場合，全ての評価項目で優れた性能を示した．特に，データ補強を行わない場合でも，Act-ChatGPT が一貫性以外のすべての項目で Video-ChatGPT を上回っており，動作特徴の導入により Video-LLM の性能が向上することが示された．Act-ChatGPT と Video-ChatGPT の AutoEval-Video [24] による定量評価の結果を**表 4**，**表 3** に示す．この評価では，Act-ChatGPT はほぼすべての項目で Video-ChatGPT を下回った．したがって，Act-ChatGPT は Video-ChatGPT よりも汎化性能が低いといえる．この結果は，それぞれのモデルに使用された学習データ量に起因すると考えられる．Video-ChatGPT の学習可能なパラメータは全て LLaVA [5] のパラメータで初期化された後に学習されるため，学習に使用される画像またはビデオ指示データの総数は $753\text{k} + 100\text{k} = 853\text{k}$ である．一方で，Act-ChatGPT の学習に使用されるビデオ指示データの総数，特に行動特徴変換モジュールと特徴変換モジュールに使用されるビデオ指示データの総数はデータ増強を考慮しても 200k のみである．このような学習データの差により，Act-ChatGPT が学習可能なデータ分布はより制限されたものになり，結果として汎化性能が低くなったと考えられる． 続いて，**図 3** から 5 に Act-ChatGPT と Video-ChatGPT の定性的な比較の結果を示す．図 1 より，Act-ChatGPT は Video-ChatGPT とは対照的に，男性の転倒するという動作を適切にとらえ，応答することが出来ていることが示された．また，図 3 では，Act-ChatGPT は少年の動作に付随して，少年が用いている用具の認識も改善しており，動作の認識を改善するだけでなく，動作内容に関連している物体の認識も改善することが示された．この物体認識の改善は，大規模言語モデルが画像言語モデルによる空間的な認識とは異なる経路で動作を認識し，応答を生成する際に文としての物体要素と動作要素の一貫性を考慮することが可能になるためであると考えられる．加えて，**図 4** より，Act-ChatGPT は Video-ChatGPT で見られている自由の女神のような固有の物体を認識する能力を保持していることが示された．

5.3 アブレーション

アブレーション研究では，Video-based Generative Performance Benchmarking [18] を使用した． 表 4 は Act-ChatGPT の学習構成とエンコーダー構成を変更した場合の定量的な評価結果を示している．具体的には，（w/o S1）は 2 段階目のみで学習した場合の結果，（w/o Img）と（w/o

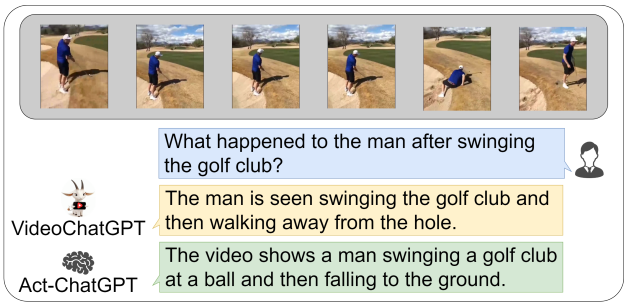

図 3 特に動作の認識が必要な応答例

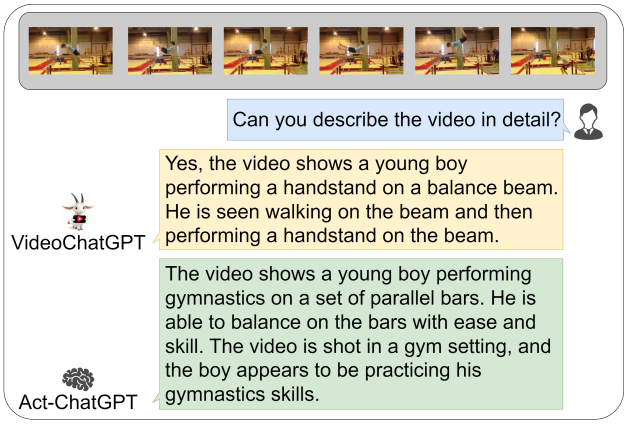

図 4 動作とそれに関連する物体の認識が必要な応答例

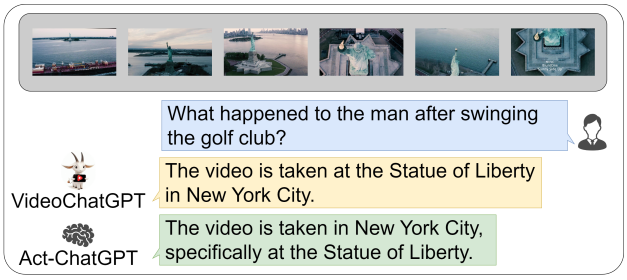

図 5 固有の物体の認識が必要な応答例

Act）は視覚エンコーダーとして動作認識モデルのみ，画像言語モデルのみを採用した場合の結果である．なお，視覚エンコーダーの一方のみを使用する場合でも，特徴量融合モジュールは次元数を調整し，適用した．図 4 の結果から，Act-ChatGPT は第 2 段階のみで学習した場合，大幅に性能が低下することが明らかになり，多段階学習の重要性が示された．また，一方の視覚エンコーダーのみを使用する場合には，どちらのエンコーダーを使用するかによらず，性能が大幅に悪化した．この結果から，ビデオ理解において，画像特徴と動作特徴が相補的な役割を果たしていることが示唆され，動作特徴がビデオ理解に有益であることが強調された．

6. 制限

本研究では，新たに導入された行動特徴を持つ新しい Act-ChatGPT を提案した．

しかし，Act-ChatGPT にはいくつかの限界が残っている．一つ目は学習データに関するものである．近年の Video-LLM の動向では，Peng Jin ら [8] が画像とビデオの共同学習の優位性を示したように，質が良く，より大規模な画像データセットも利用して様々な視覚表現を学習することが主流となっている．一方で，Act-ChatGPT では視覚エンコーダーの一部としてビデオセグメントベースで動作する行動認識モデルを採用している．そのため，画像データを学習に利用することが困難である．したがって，Act-ChatGPT で様々な視覚表現を学習することによる性能向上を目指すためには，ビデオのみで十分なデータ量を確保するための大規模で高品質のビデオデータセットを作成するか，画像を学習に利用する方法を模索するが必要がある．

二つ目は計算コストに関するものである．Act-ChatGPT は大規模な動作認識モデルを追加の視覚エンコーダーとして採用しているため，計算コストが高い．また，Act-ChatGPT で用いた動作認識モデルは，シーン情報のみでも比較的分類が容易とされる Kinetics データセット [10] で学習され，ある程度の物体認識能力を含んでいるといえる．一方で，Act-ChatGPT では画像言語モデルが既に物体認識を処理しているため，これらの能力は冗長である．そのため，この冗長性を解消することにより計算コスト上の制限を緩和することが可能であるといえ，より焦点を絞った小規模で高効率なモデルの必要性が示唆される．

表 4 学習構成とエンコーダー構成を変更した場合の Video-based Generative Performance Benchmarking による評価結果

	CI↑	DO↑	CU↑	TU↑	C↑
Video-ChatGPT	2.41	2.59	3.00	2.07	2.19
Act-ChatGPT（w/o S1）	2.08	2.41	2.78	2.01	1.98
Act-ChatGPT（w/o Img）	2.04	2.34	2.79	1.84	1.98
Act-ChatGPT（w/o Act）	2.26	2.48	2.92	2.19	2.00
Act-ChatGPT	**2.57**	**2.69**	**3.17**	**2.24**	**2.32**

7. 結論

本研究では，個々のビデオセグメントから得られる動作特徴を利用することで，ビデオに含まれる細かな動作要素を応答生成に反映させるように設計された Video-LLM である Act-ChatGPT を提案した．Act-ChatGPT は動作とそれに関する物体の認識能力を強化し，ベースモデルとして用いた Video-ChatGPT の応答性能の改善を達成した．また，Act-ChatGPT ビデオに含まれる固有の物体を識別する能力も維持しており，ベースモデルと比較してビデオ理解能力を全体的に引き上げることが確認された．

参考文献

[1] Alec, R., Jong, Wook, K., Chris, H., Aditya, R., Gabriel, G., Sandhini, A., Girish, S., Amanda, A., Pamela, M., Jack, C., Gretchen, K. and Ilya, S.: Learning transferable visual models from natural language supervision, *Proc. of International Conference on Machine Learning*, Vol. 139, pp. 8748–8763 (2021).

[2] Chenfei, W., Shengming, Y., Weizhen, Q., Xiaodong, W., Zecheng, T. and Nan, D.: Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models, *arXiv:2303.04671* (2023).

[3] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I. and Xing, E. P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, *Large Model Systems Organization. https://lmsys.org/blog/2023-03-30-vicuna/*, (online), available from ⟨https://lmsys.org/blog/2023-03-30-vicuna/⟩ (2023).

[4] Hang, Z., Xin, L. and Lidong, B.: Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding, *arXiv:2306.02858* (2023).

[5] Haotian, L., Chunyuan, L., Qingyang, W. and Yong, Jae, L.: Visual Instruction Tuning, *Proc. of Neural Information Processing Systems* (2023).

[6] Heilbron, F. C., Escorcia, V., Ghanem, B. and Niebles, J. C.: ActivityNet: A large-scale video benchmark for human activity understanding, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 961–970 (online), DOI: 10.1109/CVPR.2015.7298698 (2015).

[7] Hugo, T., Thibaut, L., Gautier, I., Xavier, M., Marie-Anne, L., Timothée, L., Baptiste, R., Naman, G., Eric, H., Faisal, A., Aurelien, R., Armand, J., Edouard, G. and Guillaume, L.: LLaMA: Open and efficient foundation language models, *arXiv:2302.13971* (2023).

[8] Jin, P., Takanobu, R., Zhang, C., Cao, X. and Yuan, L.: Chat-UniVi: Unified visual representation empowers large language models with image and video understanding, *arXiv:2311.08046* (2023).

[9] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D.: Scaling Laws for Neural Language Models, *arXiv:2001.08361* (2020).

[10] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M. and Zisserman, A.: The Kinetics Human Action Video Dataset, *arXiv:1705.06950* (2017).

[11] Li, J., Li, D., Savarese, S. and Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *Proc. of International Conference on Machine Learning*, pp. 19730–19742 (2023).

[12] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L. and Qiao, Y.: VideoChat: Chat-centric video understanding, *arXiv:2305.06355* (2023).

[13] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L. and Qiao, Y.: MVBench: A Comprehensive Multi-modal Video Understanding Benchmark, *arXiv:2311.17005* (2023).

[14] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L. and Qiao, Y.: UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding, *Proc. of IEEE International Conference on Computer Vision*, pp. 1632–1643 (2023).

[15] Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L. and Qiao, Y.: Unmasked Teacher: Towards Training-Efficient Video Foundation Models, *Proc. of IEEE International Conference on Computer Vision*, pp. 19948–19960 (2023).

[16] Li, Y., Wang, C. and Jia, J.: LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models, *arXiv:2311.17043* (2023).

[17] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P. and Yuan, L.: Video-LLaVA: Learning United Visual Representation by Alignment Before Projection, *arXiv:2311.10122* (2023).

[18] Muhammad, M., Hanoona, R., Salman, K. and Fahad, Shahbaz, K.: Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, *arXiv:2306.05424* (2023).

[19] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T. B.: Alpaca: A strong, replicable instruction-following model, *Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html*, Vol. 3, No. 6, p. 7 (2023).

[20] Tong, Z., Song, Y., Wang, J. and Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, *Proc. of Neural Information Processing Systems*, Vol. 35, pp. 10078–10093 (2022).

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proc. of Neural Information Processing Systems*, Vol. 30 (2017).

[22] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y. and Qiao, Y.: VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 14549–14560 (2023).

[23] Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. and Le, Q. V.: Finetuned Language Models are Zero-Shot Learners, *Proc. of International Conference on Learning Representations* (2022).

[24] Xiuyuan, C., Yuan, L., Yuchen, Z. and Weiran, H.: AutoEval-Video: An Automatic Benchmark for Assessing Large Vision Language Models in Open-Ended Video Question Answering, *arXiv:2311.14906* (2023).

[25] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y. and Wen, J.-R.: A Survey of Large Language Models, *arXiv:2303.18223* (2023).

[26] Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C. W., Li, Z., Liu, W. and Yuan, L.: LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment, *arXiv:2310.01852* (2023).

[27] Zoph, B., Raffel, C., Schuurmans, D., Yogatama, D., Zhou, D., Metzler, D., Chi, E. H., Wei, J., Dean, J., Fedus, L. B., Bosma, M. P., Vinyals, O., Liang, P., Borgeaud, S., Hashimoto, T. B. and Tay, Y.: Emergent Abilities of Large Language Models, *Proc. of Transactions on Machine Learning Research* (2022).