Act-ChatGPT: 動作特徴を用いた対話型ビデオ理解モデル

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，既存モデルではビデオに含まれる動作特徴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ラインを上回り，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（LLM）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，VQA のような一対の質問応答ではなく，対話形式での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。なお，本研究ではそのような形式の対話型ビデオ理解モデルを Video-LLM と定義する。しかしながら，既存の Video-LLM では，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 [15] の普及と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 Video-LLM 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ている Video-ChatGPT [13] 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] 電気通信大学
[a)] nakamizo-y@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

2. 関連研究

2.1 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 [7] や Video-LLaMA [3]，Video-ChatGPT [13]，LLaMA-VID [11] 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 [8] や Video-LLaVA [12] などが挙げられる。これらのモデルは UMT [10] や LanguageBind [18] 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。

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を行っており，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。

2.2 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 [14] や UMT [10] などの Transformer [15] 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サンプリング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。

3. 提案手法

3.1 概要

提案手法では，Video-ChatGPT [13] 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。提案手法の概要図を図 1 に示す。提案手法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。提案手法では，まず入力されたビデオから T 個のフレーム及び T 個の 16 フレームのビデオセグメントをサンプリングし、$F \in \mathbb{R}^{T\times W\times H\times C}$ と $S \in \mathbb{R}^{T\times 16\times W\times H\times C}$ を得る。その後，画像言語モデルにより前者から T 個の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 T 個の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 3.3 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，提案手法では計算コスト削減のため，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。

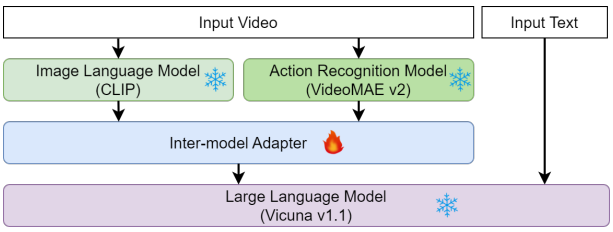

図 1 Act-ChatGPT の概要図

3.2 使用モデル

提案手法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 [9] で fine-tuning された VideoMAEv2 [16] 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 [4] のために fine-tuning された Vicuna v1.1 [2] の 7B モデルを採用する。

3.3 モデル間アダプター

提案手法のモデル間アダプターの概要図を図 2 に示す。提案手法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。

3.3.1 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 [13] 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 T 個の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。

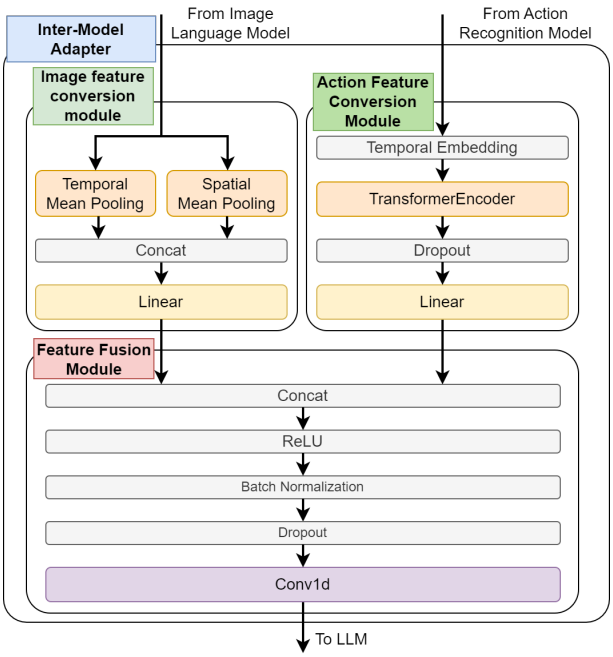

図 2 モデル間アダプターの概要図

3.3.2 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 T 個の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。

3.3.3 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。

3.4 学習

提案手法では Vision Instruction Tuning [4] に従い，学習を行う。詳細には，Video-ChatGPT [13] 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，提案手法では学習を 2 段階に分けて行う。まず，1 段階目の学習では片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 と同様であるため，画像特徴量変換モジュールに含まれる全結合層は Video-ChatGPT [13] 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. 実験

4.1 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 [13] 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 [13] を用いた。Video Instruction Dataset は ActivityNet [5] 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 [6] 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。なお，提案手法では学習データの不足に対処するため，指示内容を大規模言語モデルで言い換えることによるデータ拡張を適用している。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 [13] を採用した。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 5 のスコアで評価した。なお，Video-based Generative Performance Benchmarking 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

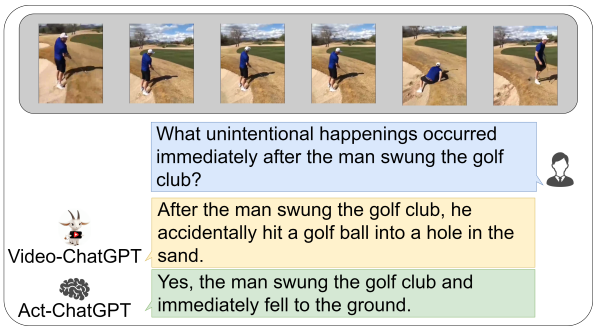

図 3 動作の認識が必要な応答例

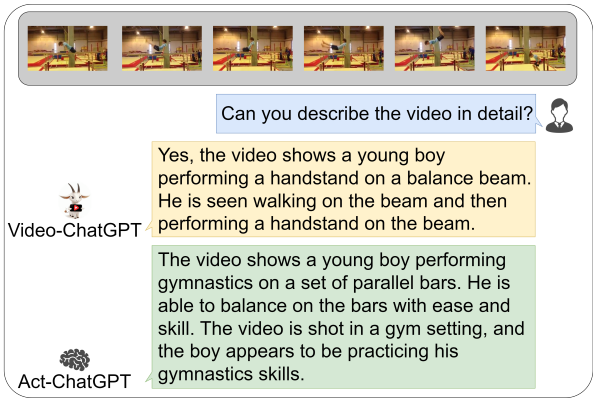

図 4 動作とそれに関連する物体の認識が必要な応答例

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。

4.2 定量・定性評価

提案手法とベースである Video-ChatGPT [13] の定量的な比較を表 1 に示す。なお，公正な比較のため，データ拡張を用いずに学習した結果も (w/o Expand) として示している。提案手法は全ての項目で Video-ChatGPT を上回る結果となった。また，提案手法はデータ拡張を用いていない場合であっても，Consistency を除く 4 項目で Video-ChatGPT を上回っており，動作特徴量の導入により正確性などの面で Video-LLM の応答精度が改善することが示された。 図 3，図 4，図 5 は Act-ChatGPT と

表 1 GPT-3.5 による評価の結果

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

Video-ChatGPT の定性的な比較を示している。 図 3，図 4 より，Act-ChatGPT 動作の認識だけでなく，動作やそれに直接的に関与する物体の認識においても Video-ChatGPT から応答を改善していることがわかる。このときの物体の認識における改善は，大規模言語モデルが空間特徴とは異なる認識経路で動作特徴を認識することにより，応答を生

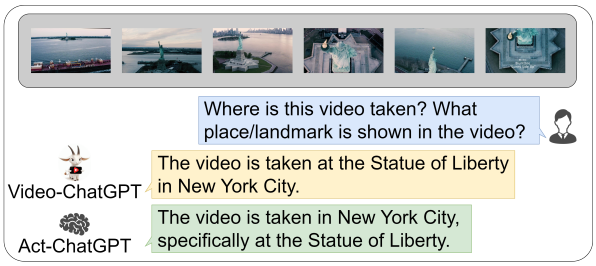

図 5 固有の物体の認識が必要な応答例

成する際に物体要素と動作要素の文章としての整合性を考慮することが可能になったことによるものと推測される。また，図 5 より，Act-ChatGPT は Video-ChatGPT で見られた固有の物体に対する認識能力を維持していることがわかる。

4.3 アブレーション研究

表 2 は Act-ChatGPT を 2 段階目の学習のみで学習した場合及び視覚エンコーダーの片方のみを用いた場合の定量的な比較である。なお，(w/o Stage1) は 2 段階目の学習のみで学習した場合を表し，(w/o Image) は動作認識モデルのみを用いた場合を，(w/o Action) は画像言語モデルのみを用いた場合をそれぞれ表す。Act-ChatGPT では 2 段階目の学習のみで学習した場合に大幅な指標の悪化が確認され，多段階の学習の重要性が示された。また，視覚エンコーダーの片方のみを用いた場合には，どちらも指標が大幅に悪化した。このことから，画像特徴量と動作特徴量は互いに補完する性質を持つといえ，ビデオ理解において動作特徴を活用することの有効性が示された。

表 2 2 段階目の学習のみで学習した場合及び視覚エンコーダーを片方のみ用いた場合の GPT3.5 による評価の結果

	CI↑	DO↑	CU↑	TU↑	C↑
Video-ChatGPT	2.41	2.59	3.00	2.07	2.19
Act-ChatGPT (w/o Stage1)	2.08	2.41	2.78	2.01	1.98
Act-ChatGPT (w/o Image)	2.04	2.34	2.79	1.84	1.98
Act-ChatGPT (w/o Action)	2.26	2.48	2.92	2.19	2.00
Act-ChatGPT	**2.57**	**2.69**	**3.17**	**2.24**	**2.32**

5. まとめ

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 Video-LLM である Act-ChatGPT を提案した。Act-ChatGPT では動作認識モデルを追加の視覚エンコーダーとして導入し，ビデオセグメントごとの動作特徴量を大規模言語モデルの入力に追加することにより，ビデオに含まれる動作も考慮した応答の生成を実現している。それにより，提案手法の Act-ChatGPT はベースである Video-ChatGPT と比較して，固有の物体に対する認識能力を維持したまま，動作だけでなく物体の面での認識も改善した。

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] 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/* (2023).

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

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

[5] 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 (2015).

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[10] 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).

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

[12] 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).

[13] 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).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).