深層状態空間モデルおよび双方向スキャンに基づく Multimodal LLM による動画像理解

八島 大地[1,a)]　栗田 修平[2,b)]　小田 悠介[3,c)]　鈴木 駿太郎[1,d)]　小槻 誠太郎[1,e)]　杉浦 孔明[1,f)]

概要

本研究ではマルチモーダル大規模言語モデル（MLLM）による動画像理解を扱う．動画像理解は時間的依存関係および系列長が長いため困難なタスクである．また，既存の Transformer ベース手法における注意機構は系列長に対して二乗の計算量を要し，計算量が大きくなるという問題がある．そこで本研究では，深層状態空間モデルを言語バックボーンとして使用し，従来の注意機構を置き換えることで計算量を線形に抑制する MLLM を提案する．提案手法は，階層的な時間解像度で動画像を処理する Aligned Hierarchical Bidirectional Scan モジュールを備え，大規模な動画像系列を効率的に扱う．本手法を VATEX および MSR-VTT といった標準的な動画キャプショニングベンチマーク上で評価した結果，既存の代表的な MLLM と同程度の精度を維持しつつ，約 3 倍のスループットを達成した．

1. はじめに

MLLM の発展 [22, 12] により，画像認識を超えて，複雑な時空間的な変化を含む動画像理解が可能となった．これにより，MLLM は動画像質問応答および動画キャプショニングなどの下流タスクに応用されている [24]．特に，動画キャプショニングは，動画要約および生活支援ロボットにおける視覚シーンの解釈など，幅広い分野において実用性を有する [24, 19]．一方，既存の動画像を扱う MLLM は，長時間の入力の処理やリアルタイム性において依然として限定的である．そこで本研究では，MLLM による動画キャプショニングに着目する．図 1 に，提案手法の代表的な使用例を示す．提案手法は動画像および自然言語プロンプトを入力とし，動画に関するキャプションを出力する．

1 慶應義塾大学
2 国立情報学研究所
3 国立情報学研究所 大規模言語モデル研究開発センター
a) ydaichi1207@keio.jp
b) skurita@nii.ac.jp
c) odashi@nii.ac.jp
d) shuntaro20021227@keio.jp
e) otsu8sei14@keio.jp
f) komei.sugiura@keio.jp

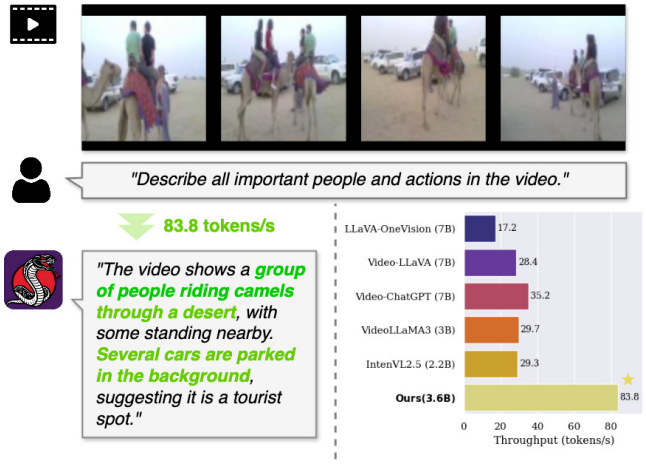

図 1　提案手法の代表的な使用例．

このタスクは，視覚入力が複雑な時間的依存関係を有し，系列長も大きいため困難である．従来の Transformer ベースの手法を直接適用すると，注意機構が系列長に対して二乗でスケールするため，計算量が非常に大きくなるという問題が存在する．この問題に対処するため，既存手法では入力の圧縮を行うことが一般的であるが，細かな時空間的な情報が捉えられないということが知られている．

そこで，本手法では深層状態空間モデルに基づく動画像理解が可能な MLLM を提案する．従来の言語バックボーンにおいて注意機構を深層状態空間モデルに置き換えることで，提案手法は長尺な動画系列を線形計算量でスケーラブルな処理を可能とする．また，複雑な時間的変化を捉えるために，新たに Aligned Hierarchical Bidirectional Scan（AHBS）モジュールを導入する．本モジュールは，複数の時間解像度にわたって情報を双方向に伝播させることで，単純なダウンサンプリングや射影に伴う時空間的な情報の損失を克服する．

本研究の主な貢献は以下の通りである．

- 長系列動画に対する効率的な時間モデリングを実現する，深層状態空間モデルベースの MLLM を提案する．本モデルは，系列長に対して線形の計算量でスケールする．
- 複数の時間解像度にわたって動画像を処理する AHBS

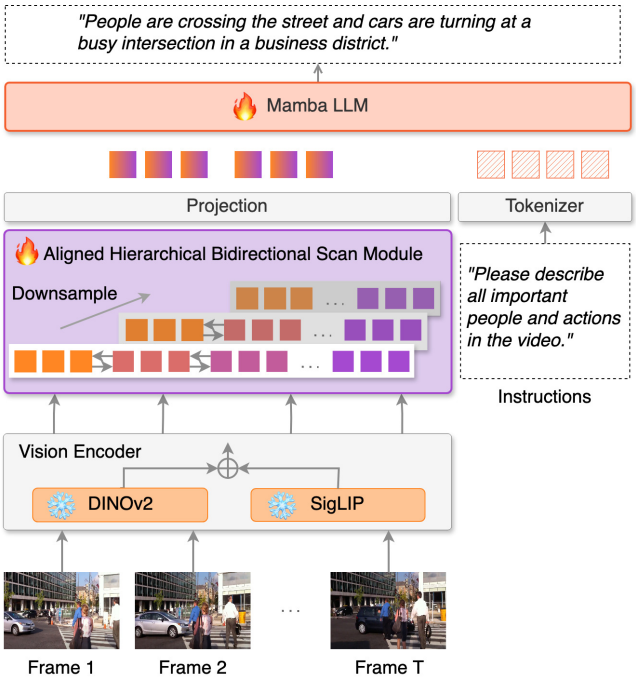

図 2 提案手法のモデル図

モジュールを導入する．このモジュールにより，動画に内在する複雑な時間的変化を効果的に捉えることが可能となる．

- 提案手法は，多くの評価指標において既存手法と同等の性能を示しつつ，ベースライン手法の約 3 倍の推論速度を実現する．

2. 提案手法

本研究では，動画に内在する複雑な時間的変化を捉えるための深層状態空間モデルに基づく，MLLM を提案する．既存の動画を扱う MLLM [16, 12, 31, 7] と比較して，提案手法は以下の点で異なる．本手法では，深層状態空間モデルの一種である Mamba [9] に基づく LLM バックボーンおよび，動画像を複数の時間解像度で処理する AHBS モジュールを導入する．AHBS モジュールは，階層的な双方向スキャンを用いることで動画像入力において MLLM の主要な課題である時間的依存関係の取得を可能とする．さらに，本提案手法は，長系列の時間的依存性を効率的に処理できることから，視覚言語行動モデル [2, 3, 11] や自動運転分野 [30, 6, 25] への応用も期待される．

図 2 に提案手法のモデル図を示す．提案手法は視覚エンコーダ，AHBS モジュール，および Mamba-LLM の 3 つのモジュールから構成される．

2.1 視覚エンコーダ

入力は $\boldsymbol{x} = (\boldsymbol{x}_{\text{vision}}, \boldsymbol{x}_{\text{txt}})$ と定義される．ここで，$\boldsymbol{x}_{\text{vision}} = (\boldsymbol{x}_{\text{vision}}^{(1)}, \boldsymbol{x}_{\text{vision}}^{(2)}, \ldots, \boldsymbol{x}_{\text{vision}}^{(T)}) \in \mathbb{R}^{T\times 3\times H\times W}$ は 1 枚の画像もしくは動画フレーム列を，$\boldsymbol{x}_{\text{txt}} \in \mathbb{R}^{V\times L}$ はトークナイズされたテキスト入力を表す．ここで，T，H，W，V，および L はそれぞれ，フレーム数，フレームの高さ，幅，語彙サイズ，および系列長を示す．ただし，$T = 1$ の場合は，単一の画像入力に相当する．先行研究 [34, 26, 8] に基づき，視覚エンコーダには，SigLIP [32] および DINOv2 [20] の 2 つを並列に使用した．SigLIP は画像と言語の対照学習に基づくため，意味的整合性に優れる．一方，DINOv2 は自己教師あり学習により微細な視覚特徴を捉えることができる．2 つのエンコーダを補完的に組み合わせることで，意味の抽出と細部の理解の両立が可能となると考えられる．

各フレーム $\boldsymbol{x}_{\text{vision}}^{(t)}$ は，以下の手順で独立に処理される．まず，$\boldsymbol{x}_{\text{vision}}^{(t)}$ をサイズ $p\times p$ の非重複パッチに分割し，総パッチ数 $N_{\text{p}} = (H \times W)/p^2$ とする．分割されたパッチ画像は，SigLIP および DINOv2 の 2 つの視覚エンコーダに入力され，それぞれ $\boldsymbol{V}_{\text{SigLIP}}^{(t)} \in \mathbb{R}^{N_{\text{p}}\times d_{\text{s}}}$ および $\boldsymbol{V}_{\text{DINOv2}}^{(t)} \in \mathbb{R}^{N_{\text{p}}\times d_{\text{d}}}$ を得る．ここで，d_{s} および d_{d} は，各エンコーダの出力次元を表す．次に，$\boldsymbol{V}_{\text{SigLIP}}^{(t)}$ と $\boldsymbol{V}_{\text{DINOv2}}^{(t)}$ を特徴次元方向に結合し，$\boldsymbol{V}^{(t)} = [\boldsymbol{V}_{\text{SigLIP}}^{(t)}, \boldsymbol{V}_{\text{DINOv2}}^{(t)}] \in \mathbb{R}^{d_{\text{v}}}$ を得る．ここで，$d_{\text{v}} = d_{\text{s}} + d_{\text{d}}$ である．この簡潔かつ包括的な特徴表現により，高次の意味情報と精緻な視覚情報の両方を効果的に捉えることができる．

2.2 Aligned Hierarchical Bidirectional Scan

動画特徴量を LLM に統合する際には，計算量を抑えつつ，時空間的依存関係を効果的に捉える構造が求められる．こうした依存関係は，複数の時間解像度にまたがる階層的な構造を伴うことが多い．そこで，AHBS モジュールは，動画像を異なる時間解像度で階層的な時間的経路を通じて視覚特徴を取得することで，粗・密両方の時間的変化を捉えることを可能とする．本モジュールは，従来の双方向スキャン手法 [35, 17] とは異なり，動画に内在する多様かつ複雑な時間的変化を明示的に扱う点に特徴がある．

図 3 に，AHBS モジュールのアーキテクチャを示す．本モジュールは，入力として $(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)}, \ldots, \boldsymbol{V}^{(T)}) \in \mathbb{R}^{T\times N_{\text{p}}\times d_{\text{v}}}$ を受け取る．はじめに，計算量を削減し，時間的依存関係を効率的に捉えるために，特徴空間方向にダウンサンプリングを適用し，$\boldsymbol{V}_{\text{d}} \in \mathbb{R}^{T\times N_{\text{d}}\times d_{\text{v}}}$ を得る．ここで，N_{d} はダウンサンプリング後の次元数である．続いて，$\boldsymbol{V}_{\text{d}}$ は，異なる時間解像度に対応した M 個の並列経路で処理される．各経路において，$\boldsymbol{V}_m \in \mathbb{R}^{T_{\text{m}}\times N_{\text{d}}\times d_{\text{v}}}$ は，$\boldsymbol{V}_{\text{d}}$ に対して時間的ダウンサンプリングを適用することで得られる．ここで，$T_m = \lfloor \frac{T}{2^{m-1}} \rfloor$ は経路 $m \in 1, \ldots, M$ におけるサンプリング率であり，T_1 は元の時間解像度に対応する．その後，各 $\boldsymbol{V}_m$ に対して双方向スキャンを適用し，時間的依存関係を考慮した特徴量を取得する．複数の時間解像度からの統合は，以下の式の通りに行われ，モジュールの出力 $\boldsymbol{H}_{\text{v}} \in \mathbb{R}^{T\times N_{\text{d}}\times d_{\text{v}}}$ が得られる．

$$\boldsymbol{H}_{\text{v}} = \underset{m}{\text{Aggregate}}(\text{SSM}(\boldsymbol{V}_m) + \text{SSM}(f_{\text{rev}}(\boldsymbol{V}_m)))$$

ここで，Aggregate$(\cdot)$ は統合（加算，連結など），SSM$(\cdot)$

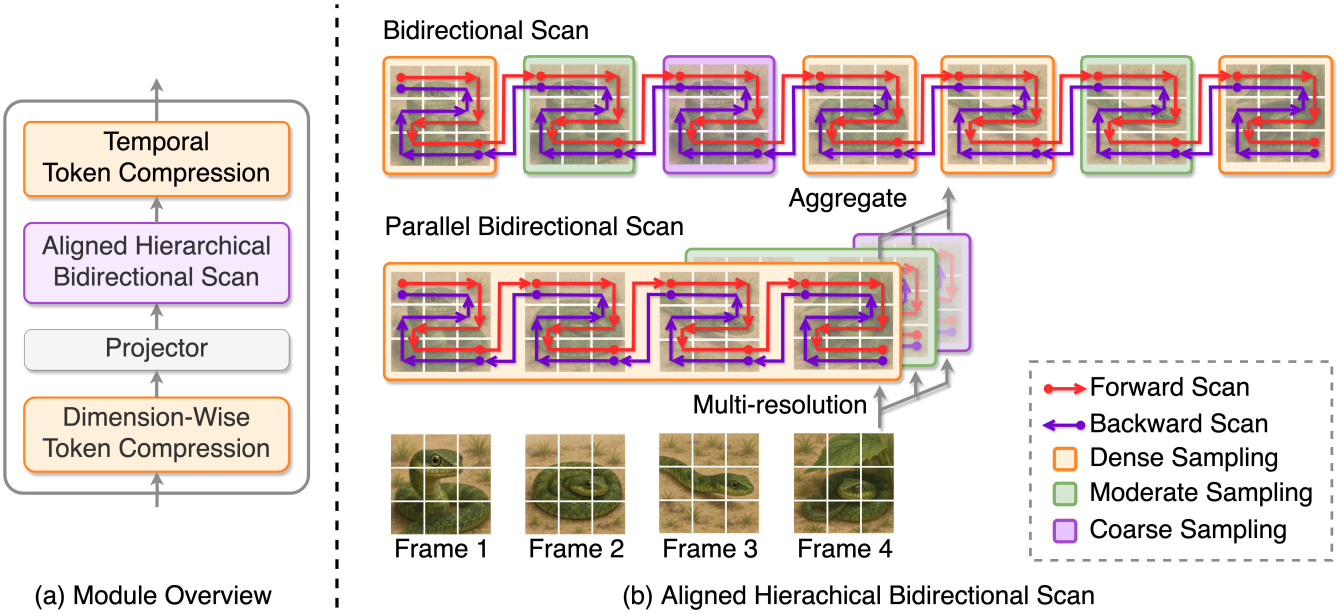

図 3 Aligned Hierarchical Bidirectional Scan モジュールのアーキテクチャ図

は Mamba[9]，$f_{\text{rev}}(\cdot)$ は入力系列の反転を行う関数を表す．本研究では，結合関数として加算を用いた．

2.3 Mamba-LLM

本手法では，結合された動画と言語の系列を効率的に処理できることから，バックボーン LLM として Mamba [9] を採用する．Mamba の中核である選択的スキャンにより，入力系列全体にわたる動的なモデル制御が可能となり，一部の言語モデリングタスクにおいて Transformer を上回る性能を示している．本手法では，Mamba-LLM の入力として，$\boldsymbol{H}_{\text{vision}}$ および $\boldsymbol{x}_{\text{txt}}$ をトークナイズして得られる埋め込みトークン列を連結した系列を用いる．Mamba バックボーンは，同一構造の基本ブロックを積層したものであり，各ブロックは畳み込み，Mamba ブロック，残差接続，および正規化層から構成される．

3. 実験

3.1 実験設定

本研究では，LLaVA の派生手法で一般的に用いられる事前学習フェーズ [15, 5, 34] を省略し，学習プロセスを簡略化した．この設計は，先行研究で指摘されている過学習不足の問題 [10] に対処するものである．そこで，本研究では，AHBS モジュールおよび Mamba-LLM を直接ファインチューニングする手法を採用した．ファインチューニングには，COCO，Visual Genome，GQA など多様なデータセットに由来する，LLaVA 1.5 [15] で導入された 665K の画像と言語のデータセットを使用した．さらに，Video-ChatGPT [18] データセットを追加し，10k の動画と言語のペアを学習に用いた．評価には，標準的な動画キャプショニングベンチマークである MSR-VTT [29] および VATEX [28] を使用した．各動画に対して，フレーム数 T を一様にサンプリングし，384 × 384 サイズにリサイズする前処理を行った．

ベースライン手法として，パラメータ数が 7B 程度の動画像を扱える MLLM である Video-ChatGPT [18]，Video-LLaVA [13]，および LLaVA-OneVision [12] を採用した．同様に，パラメータ数が 4B 以下の MLLM として InternVL2.5 [4] および VideoLLaMA3 [33] を使用した．また，評価指標には動画キャプショニングの標準的な評価指標である BLEU [21]，ROUGE-L [14]，METEOR [1]，CIDEr [27]，および PAC-S [23] を使用した．

3.2 定量的比較結果

表 1 に，VATEX および MSR-VTT ベンチマークにおける提案手法と各ベースライン手法の定量的比較結果を示す．両ベンチマークに対して，単一試行に基づく評価を実施した．表 1 より，提案手法は多くの評価指標においてベースライン手法と同等以上の性能を示している．特に，VATEX および MSR-VTT ベンチマークでは BLEU4 においてそれぞれ 28.6 および 23.6 であり，次点のモデルに対して 4.1 ポイントおよび 0.3 ポイント上回った．また，表 1 に，各手法のスループットについての結果も示す．MSR-VTT ベンチマークにおいて，提案手法のスループットの平均値は毎秒 95.4 トークンであった．一方，最も高速なベースラインである Video-ChatGPT は，同一評価条件下で毎秒 38.1 トークンであった．この結果から，提案手法はベースラインの約 3 倍の推論速度であることが分かる．加えて，提案手法は InternVL2.5 よりもパラメータ数が多いにもかかわらず，高いスループットを達成した．このスループットの向上は，長尺な動画像系列の処理において，提案手法が有効であることを示している．特に，深層状態空間モデルに基づく Mamba アーキテクチャは線形計算量で動作するた

表 1　ベースライン手法と提案手法の定量的比較結果

ベンチマーク	手法	パラメータ数	BLEU1↑	BLEU4↑	ROUGE↑	CIDEr↑	METEOR↑	PAC-S↑	Throughput↑ (tokens/s)
VATEX	Proprietary								
	Gemini-1.5-Pro	–	22.8	4.5	19.3	14.2	8.9	39.2	–
	Fully open MLLMs								
	Video-ChatGPT	7B	14.8	2.4	25.0	9.9	10.6	39.4	35.2
	Video-LLaVA	7B	67.3	24.5	44.9	37.7	19.6	40.8	28.4
	LLaVA-OneVision	7B	60.6	17.6	41.5	39.2	21.2	42.0	17.2
	Small MLLMs								
	InternVL2.5	2.2B	58.3	16.5	35.8	33.2	17.2	41.4	29.3
	VideoLLaMA3	2B	56.8	17.5	42.2	36.9	**24.0**	**43.0**	29.7
	Ours	3.6B	**73.4**	**28.6**	**47.7**	**44.4**	22.2	41.8	**83.8**
MSR-VTT	Proprietary								
	Gemini-1.5-Pro	–	51.6	12.0	36.8	19.4	20.8	40.8	–
	Fully open MLLMs								
	Video-ChatGPT	7B	56.3	14.4	42.4	17.8	24.8	39.4	38.1
	Video-LLaVA	7B	68.0	23.3	50.1	30.7	25.7	40.8	28.9
	LLaVA-OneVision	7B	52.5	12.4	37.4	10.8	22.8	42.5	24.8
	Small MLLMs								
	InternVL2.5	2.2B	**69.7**	19.1	43.0	**32.0**	22.2	41.0	31.5
	VideoLLaMA3	2B	59.4	15.7	41.5	17.9	25.3	**42.9**	33.5
	Ours	3.6B	68.1	**23.6**	**50.6**	27.3	**27.0**	40.1	**95.4**

(i) $\boldsymbol{x}_{\text{vision}}$	
(ii) $\boldsymbol{y}$	A man cleans a window with a squeegee while outside.
(iii) Ours	A man is cleaning a window with a squeegee.
(iv) Intern VL2.5	The man sprays the window with a spray bottle.
(v) LLaVA-OneVision	A man cleans a large glass window, focusing on the lower section.

図 4　定性的結果

め，通常の Transformer モデルにおける注意機構の二乗スケーリングと比較して，高速な処理が可能となる．

3.3　定性的結果

図 3.3 に，提案手法とベースライン手法である InternVL2.5 および LLaVA-OneVision の定性的結果を示す．図中の (a), (b), (c), (d), および (e) の各行は，それぞれ入力動画 $\boldsymbol{x}_{\text{vision}}$，正解キャプション $\boldsymbol{y}$，および提案手法，InternVL2.5，LLaVA-OneVision によって生成された動画キャプションを示す．

図 3.3 の (c) では提案手法が “A man is cleaning a window with a squeegee.” と出力し，動画内の主要な行動を正しくに認識できていることを示している．一方で，(d) の InternVL2.5 の出力では，“The man sprays the window with a spray bottle.” と男性が持っている物体を誤って認識している．さらに，(e) の LLaVA-OneVision によるキャプションでは，“A man cleans a large glass window, focusing on the lower section.” と記述されているが，実際には動画中で男は窓全体を掃除しており，動画の時系列情報が失われていることがわかる．

4.　おわりに

本研究では深層状態空間モデルに基づく MLLM による動画キャプショニングタスクを扱った．本研究の貢献は次の通りである．

- 長系列動画に対する効率的な時間モデリングを実現した，深層状態空間モデルベースの MLLM を提案した．
- 複数の時間解像度にわたって動画を処理する AHBS モジュールを導入した．このモジュールにより，動画に内在する複雑な時間的変化を効果的に捉えることが可能になった．
- 提案手法は，動画キャプショニングタスクにおいて多くの評価指標においてベースライン手法と同等の性能を示しつつ，それらの約 3 倍の推論速度を達成した．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものである．また，本研究の一部は，JSPS 科研費 23K28168，JST ムーンショットの助成を受けて実施されたものである．

参考文献

[1] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL*, pp. 65–72 (2005).

[2] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L., Tanner, J., Vuong, Q., Walling, A., Wang, H. and Zhilinsky, U.: π_0: A Vision-Language-Action Flow Model for General Robot Control, *arXiv preprint arXiv:2410.24164* (2024).

[3] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T. and Zitkovich, B.: RT-1: Robotics Transformer for Real-World Control at Scale, *arXiv preprint arXiv:2212.06817* (2022).

[4] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y. and Dai, J.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, *CVPR*, pp. 24185–24198 (2024).

[5] Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B. and Shen, C.: MobileVLM V2: Faster and Stronger Baseline for Vision Language Model, *arXiv preprint arXiv:2402.03766* (2024).

[6] Cui, C., Ma, Y., Cao, X., Ye, W. and Wang, Z.: Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles, *WACVW*, pp. 902–909 (2024).

[7] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.-H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N., Hajishirzi, H., Girshick, R., Farhadi, A. and Kembhavi, A.: Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, *arXiv preprint arXiv:2409.17146* (2024).

[8] Goko, M., Kambara, M., Saito, D. and Sugiura, K.: Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations, *CoRL* (2024).

[9] Gu, A. and Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces, *CoLM* (2024).

[10] Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T. and Sadigh, D.: Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models, *ICML* (2024).

[11] Kim, J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P. and Finn, C.: OpenVLA: An Open-Source Vision-Language-Action Model, *CoRL*, pp. 2679–2713 (2024).

[12] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z. and Li, C.: LLaVA-OneVision: Easy Visual Task Transfer, *arXiv preprint arXiv:2408.03326* (2024).

[13] 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, *EMNLP*, pp. 5971–5984 (2024).

[14] Lin, C.: ROUGE: A Package For Automatic Evaluation Of Summaries, *ACL*, pp. 74–81 (2004).

[15] Liu, H., Li, C., Li, Y. and Lee, Y. J.: Improved Baselines with Visual Instruction Tuning, *CVPR*, pp. 26296–26306 (2024).

[16] Liu, H., Li, C., Wu, Q. and Lee, Y. J.: Visual Instruction Tuning, *NeurIPS*, pp. 34892–34916 (2023).

[17] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J. and Liu, Y.: VMamba: Visual State Space Model, *NeurIPS*, pp. 103031–103063 (2024).

[18] Maaz, M., Rasheed, H., Khan, S. and Khan, F.: Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, *ACL*, pp. 12585–12602 (2024).

[19] Nguyen, T., Bin, Y., Xiao, J., Qu, L., Li, Y., Wu, Z., Nguyen, C.-D., Ng, S.-K. and Luu, T.: Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives, *ACL*, pp. 3636–3657 (2024).

[20] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba,

W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A. and Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision, *TMLR* (2024).

[21] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, *ACL*, pp. 311–318 (2002).

[22] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. et al.: Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context, *arXiv preprint arXiv:2403.05530* (2024).

[23] Sarto, S., Barraco, M., Cornia, M., Baraldi, L. and Cucchiara, R.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, *CVPR*, pp. 6914–6924 (2023).

[24] Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J. and Xu, C.: Video Understanding with Large Language Models: A Survey, *arXiv preprint arXiv:2312.17432* (2023).

[25] Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X. and Zhao, H.: DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models, *CoRL* (2024).

[26] Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, Z., Fergus, R., LeCun, Y. and Xie, S.: Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, *NeurIPS*, pp. 87310–87356 (2024).

[27] Vedantam, R., Zitnick, L. and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, *CVPR*, pp. 4566–4575 (2015).

[28] Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F. and Wang, Y.: VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research, *ICCV*, pp. 4580–4590 (2019).

[29] Xu, J., Mei, T., Yao, T. and Rui, Y.: MSR-VTT: A Large Video Description Dataset for Bridging Video and Language, *CVPR*, pp. 5288–5296 (2016).

[30] Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.-Y., Li, Z. and Zhao, H.: DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model, *IEEE RA-L*, Vol. 9, No. 10, pp. 8186–8193 (2024).

[31] Yue, X., Song, Y., Asai, A., Kim, S., Nyandwi, J., Khanuja, S., Kantharuban, A., Sutawika, L., Ramamoorthy, S. and Neubig, G.: Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages, *ICLR* (2024).

[32] Zhai, X., Mustafa, B., Kolesnikov, A. and Beyer, L.: Sigmoid Loss for Language Image Pre-Training, *ICCV*, pp. 11975–11986 (2023).

[33] Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L. and Zhao, D.: VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding, *arXiv preprint arXiv:2501.13106* (2025).

[34] Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S. and Wang, D.: Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference, *AAAI*, pp. 10421–10429 (2025).

[35] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W. and Wang, X.: Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model, *ICML* (2024).