GazeLLM: 視覚情報を利用するマルチモーダルLLM

暦本 純一[1,2,a)]

概要：大規模言語モデル（LLM）は，テキストだけでなく画像，音声，動画も処理できるマルチモーダルLLM（MLLM）へと発展している．MLLMが動画像と音声を通じて人間の活動を理解することで，人間の活動支援，実世界エージェント，ロボットや他者への技能移転など，多くのヒューマンコンピュータインタラクション応用が可能になる．しかし，高解像度で長時間の動画像の入力はLLMの記憶容量と処理時間を大きく消費してしまう．動画像の解像度を下げると，多くの場合必要とされる理解力が損なわれてしまう．本論文では，一人称視点映像のMLLMによる分析を最適化するために，視線情報に基づいて一人称視点映像を分割する方法を提案する．視線が集中する画像領域を選択的に処理することで，全画像と比較して約1/10の画素数で，全画像の場合と同等またはそれ以上の理解を達成できることを示し，人間の技能をAIが解釈・活用するためのMLLMの効率的な解決策を提供する．

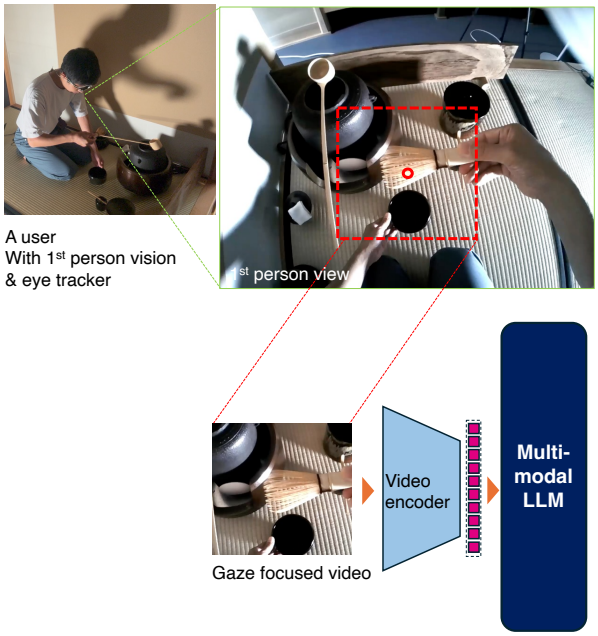


図 1 GazeLLM: 1人称視点ビデオ内の視線点を中心にトリミングされた矩形領域のみをMLLMの入力として使用する．原画像の1/10の画素数のみを入力としながら，全画像と同等またはそれ以上の理解を得られる．

1. はじめに

大規模言語モデル（LLM）は，自然言語処理において基盤的な存在となり，言語や文書を理解・生成する多くの応用を支えている．近年ではさらに，マルチモーダルデータを理解できるマルチモーダルLLM（MLLM）の研究開発が精力的に行われている [23]．これらのモデルは文章だけでなく，画像や動画，音声などの時系列情報も処理することができる．MLLMは，視覚，音声，その他の感覚モダリティを利用する人間の情報処理方法を模倣しながら，汎用人工知能（AGI）に向けた過程を表している．

ヒューマンコンピュータインタラクションの観点からも，MLLMには多くの可能性がある．このようなモデルが人間に似た方法で世界を認識できれば，様々な応用が可能となる．たとえば，熟練した人間の動作を記録して他者に伝達する技術，スキルの発展を評価する技術，実世界の行動を認識する支援する技術，ウェアラブルエージェント，実環境の感覚的認識を代替・補強することで障害者を支援する技術などが含まれる．

しかし，MLLMの計算負荷は依然として大きな課題である．LLMの中核となるtransformerモデル [2] は入力長の二乗に比例する記憶容量を必要とし，これは処理時間とメモリ要求の両方を増加させる．この問題は特に画像や動画像データの処理において顕著である．

MLLMで主に使用される画像エンコーダーであるVision Transformer（ViT） [4] では，画像は格子状の領域に分割され，各領域にトークンが割り当てられる．これにより画素数の増加に伴いトークン数も増加する．画像解像度を下げたりフレームレートをダウンサンプリングしたりしてトークン数を減らすことは可能だが，これはトレードオフをもたらす．解像度が環境を適切に認識するのに不十分になる可能性があるためである．例えば，利用者の目の前にある看板のテキストを読むなどのタスクでは，文字を理解

1 東京大学
2 ソニーCSL
a) rekimoto@acm.org

する OCR 処理に十分な画像解像度が必要となる．

一方，人間の視覚認知は網膜に投影されるすべての視覚情報を均等に処理するわけではない．人間は視線の移動を通じて特定の領域に選択的に注視し処理する．周辺視野の情報は低解像度で処理されるか，動きにのみ反応する．この視覚的注意メカニズムにより，人間は世界の理解の精度と視覚入力の処理効率のバランスを取っている．

本研究では，人間が有する視覚的注意メカニズムを活用して，MLLM における動的な視覚情報の処理負荷を軽減する方法，GazeLLM を提案する (図 1)．具体的には，MLLM への入力として一人称視点映像に焦点を当て，そこからタスクの内容を認識することを目指す．また，視線追跡機能を搭載した一人称視点カメラを使用し，一人称視点映像を視線が集中する領域に基づいて分割する機構を提案する．選択的に焦点を当てた入力を処理することで，全解像度の動画の 1/10 の画素数のみを入力としながら，全解像度の場合と同等またはそれ以上の理解を得られることを示す．

本研究の貢献は以下の通りである：

(1) 一人称視点映像を MLLM に適用することで，料理，修理，応急処置，スポーツを含む 6 種類の実世界活動を検証し，タスク手順の正確な理解が可能であることを示す．

(2) 利用者評価により，一人称視点映像から視線位置周辺のみを切り出した部分映像を利用した場合，画素数が 1/10 であるにも関わらず，全画像映像と同等またはそれ以上の品質のタスク記述が生成できることを実証する．

(3) さらに定量的評価（BLEU, ROUGE, Sentence-BERT, LLM ベースの手法）によっても，視線中心の切り出し映像がタスク記述生成を効果的に支援することを示す．

これらの知見は，提案する GazeLLM（視線ベースの LLM）が AI とウェアラブル技術を統合した実世界支援システムのための有望なフレームワークであることを示唆している．

2. 関連研究

マルチモーダル LLM

マルチモーダル大規模言語モデル（MLLM）は，近年急速に進歩している LLM 研究の分野である [22]．静止画像や音声に加えて，一部の MLLM は動画コンテンツも処理できるようになっている [6], [7]．特に，Gemini 1.5 Pro [18] は 1 時間以上の動画入力を処理することができ，一方 Qwen2-VL [19] はローカルで実行可能な MLLM システムとして動作することができる．

現在，これらの MLLM は動画処理用の専用エンコーダーを提供しており，多くの場合 Vision Transformer (ViT) [4] またはその変形を使用している．ViT のアプローチでは，画像は 2 次元パッチの格子に分割され，それぞれに潜在表現ベクトルが割り当てられる．このベクトル列をトークン列に変換して扱うことができ，他のモダリティと共に LLM の入力として統合することが可能となる．しかし，この方法にはトレードオフが存在する．画素数が増加するとトークン数も増加し，計算負荷が増大するか，画像解像度を下げる必要が生じ，視覚的理解の精度が損なわれる可能性がある．

たとえば，MLLM の一種である Qwen2-VL [19] では，画像を 28×28 のパッチに分割してそれぞれにトークンを割り当てている．これにより，たとえば 1000×1000 の画素数からなる静止画像一枚あたり 1296 個のトークンが生成され，動画像となるとさらに多くのトークンを必要とする．このようなトークン数の増加は，メモリ使用量と処理時間の増加をもたらす．

視線情報を利用した AI 支援

視線情報のユーザインタフェースへの活用は，Bolt の先駆的研究 [1] 以降，多くの試みがなされている．

GEAR は，視線情報を利用して AR ヘッドセット利用者のタスク種別を推定している [10]．認識に SVM, Random Forest 等が用いられていた．Multimodal Human-like Attention Network (MULAN) [17] は, 視覚的質問応答（VQA）において画像の顕著性に基づく人間らしい視覚的注意を使用している．GazeGPT は，画像と視線情報を入力とし，視線が集中する領域に焦点を当てることで，視覚的質問応答タスクの精度を向上させている **?**．本研究では，静止画に限定されず動画においても視線情報を利用し MLLM と連動させている．

Wang らは，人間の視線パターンと LLM の注意機構 (attention mechanism) の類似性について調査している [20]．テキストを解釈する際の人間の眼球運動と LLM 内の注意機構との間に類似点があることを確認している．我々の研究は主に処理効率を向上させるために利用者の視線情報を活用するものだが，将来の研究では人間の視線メカニズムとマルチモーダル LLM の注意機構の類似性を探究することも考えられる．

EgoScanning は，一人称視点映像の効率的なブラウジングのために，ジェスチャーや会話などの手がかりを用いてタイムラインの密度を調整する [10]．この技術は人間の動画像視聴効率を向上させるものだが，我々の研究では長時間の映像を解釈するための AI の支援としての有用性を探究している．

視線情報の取得手段，視線を利用した映像理解データセット

Ego-Exo4D プロジェクトは，一人称視点映像のための大規模データセットの構築を目指している [8]．一人称視点映像，視線，ジェスチャー認識が可能な Aria と呼ばれる眼鏡型デバイスが開発されている [5]．複数の研究機関が様々なタスクドメインにわたってデータを収集するため

に協力している．プロジェクトの主な目的は人間の行動を研究するための情報基盤を確立し提供することだが，本研究で提案するような視線と組み合わせたマルチモーダルLLMの具体的な試みは報告されていない．

VQA-MHUGは，視覚的質問応答（VQA）におけるマルチモーダルニューラル注意を研究するためのデータセットでもある [16]．

視線情報を得るための装置は，設置型や装着型 (眼鏡型) があるが，従来は使用前にキャリブレーションが必要など，日常での使用には課題があった．最近では，Pupillab社のNeonなど装着型でキャリブレーションも不要な装置も開発されており，外見も通常の眼鏡と同等程度で，視線情報の取得が容易になっている [5]．またMeta社が発表した眼鏡型デバイスOrionでも一人称映像，視線情報取得機能が搭載されており，次世代のAIスマートグラスではこれらの機能が一般化すると予想される．

3. GazeLLM

GazeLLMでは, 一人称視点動画映像を入力としてMLLMが利用者の実行しているタスクを理解し，映像ベースのガイダンスや質問応答を可能にする．利用者の一人称視点からの視線情報を利用して，その周辺領域のみを切り出して入力とする．実験では，全体の一人称視点映像と比較して，入力が約10分の1の画素数に削減された．

図2は，視線が集中する領域を使用してMLLM（Gemini1.5pro）で生成された映像記述を示している．

もう1つのGazeLLMの例として “stop and ask” がある（図3）．この応用例では，まず指導者が一人称視点動画映像を使用して料理手順のシーケンスをMLLMに与え，事前訓練する．学習者は同じ種類のレシピに従い，一人称視点を記録できるヘッドセットを装着して作業を進める．学習者は料理プロセスの途中で一時停止し，システムに質問する（例：「次は何をすべき？」や「今どの調理器具を使うべき？」）．システムは指導者が提供した手順のシーケンス内での学習者の進捗を正確に特定し，次に取るべき作業を指示する．

この応用例は，ウェアラブルシステムとAIを組み合わせた可能性を示している．利用者は一人称視点を利用してMLLMにタスクのコンテキスト（料理，自転車修理，応急処置など）を認識させ，現在の状況に応じた具体的なガイダンスを受けることができる．

このように，一人称視点をコンテキスト入力として使用するマルチモーダルLLMとウェアラブルシステムの統合により，実世界アプリケーションでAIをエージェントとして活用する可能性が開かれる．

4. 評価

GazeLLMアプローチの有効性を確認するため，MLLMにより生成された映像記述テキストを，数値評価指標とクラウドソーシングによる人間ベースの評価の両方で評価した．

4.1 データセット

評価には，Ego-Exo4Dデータセット [9] を使用した．Ego-Exo4Dデータセットには，Metaが開発した眼鏡型デバイスAria [5] で取得した視線情報，一人称視点動画像，および外部カメラから得られた三人称視点の映像が含まれている．このデータセットは料理，修理，スポーツなど，多様なシナリオを網羅しており，研究のための貴重なデータ基盤となっている．

本評価では，EgoExo4Dデータセットのうち，一人称視点映像と視線情報のみを使用した．タスクの多様性を確保するため，6つの異なるタスクカテゴリを選択した（図4）．表1は各タスクのデータ数，映像時間，標準偏差を示している．評価には合計137のタスクを使用した．

各タスクについて，以下の3種類の動画像データを準備した（図5）：

Full Ego-Exo4Dで提供される一人称視点映像をそのまま使用し，画像サイズは 1440×1440 画素，ただし映像フレームレートは利用するMLLMの仕様に合わせ1 fpsに減らしている．視線情報は含まれない．

Gaze Full映像から，視線位置を中心とした 448×448 の矩形領域で切り出したデータ．画素数はFull映像の約1/10になる．フレームレートはFullと同じく1 fpsにダウンサンプルされる．

Center Full映像の中心から 448×448 の矩形領域を切り出したデータ．画素数はGazeと同じになる．フレームレートもFull，Gazeと同じく1 fps．一人称映像では中心画像は頭部方向の映像であるため，Gazeに準じて作業情報を記録していることが期待される．

動画の切り出しのみを評価するために、いずれの映像も音声は含まれない．GazeとCenterの画素数（$448 \times 448 = 200,704$）はFullの画素数（$1440 \times 1440 = 2,073,600$）の10.3%である．評価指標は，動画をMLLMに与えて説明文を生成させた場合に，このように削減されたデータ量で生成される映像記述の品質の比較である．

評価対象のMLLMとして，Gemini 1.5 Pro [18] を選択した．Gemini 1.5 Proは約1時間の映像（フレームレートを1 fpsにダウンサンプリング）を処理できるマルチモーダルLLMである．データセットからの映像記述は，以下のプロンプトを用いて生成した：

```
Clear chat history and begin a new
    session.
Forget all the previous information and
    only use the information from this
    video.
```

Make the sauce: In a small bowl, combine the ketchup, Worcestershire sauce, salt, and pepper. Mix well. [00:01:42]

Make the omelet: In a small bowl, whisk together the eggs and milk. [00:04:43]

Cook the omelet: Heat a little oil in a clean frying pan over medium heat. Pour in the egg mixture and cook, swirling the pan to distribute the egg evenly. [00:06:13]

Fold the omelet: Fold the other half of the omelet over the filling. [00:07:02]

図 2 GazeLLM による動画像からのタスク記述: 1 人称視点ビデオ内の視線点を中心にトリミングされた矩形領域（赤の破線で示された部分）のみを MLLM の入力として使用し，タスクを言語化した例．LLM からの出力には，タスクの説明と対応するビデオのタイムスタンプが含まれており，生成された説明は該当時間のビデオ映像とともに表示される．

表 1 評価タスク: ビデオの長さの分布

ラベル	タスク	# ビデオ数	長さ（秒）
Bike	自転車修理	20	181.8（std = 162.5）
Sushi	寿司作り	14	627.7（std = 106.5）
Omelette	オムレツ作り	14	783.0（std = 269.9）
Soccer	サッカー活動	19	132.6（std = 81.0）
PCR	PCR 検査準備	53	253.8（std = 86.5）
CPR	心肺蘇生法トレーニング	17	95.4（std = 27.1）

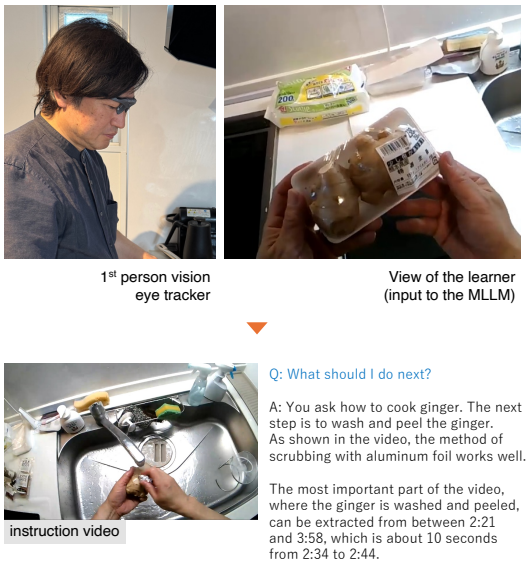


図 3 “stop and ask” アプリケーションの例: 1 人称視点ヘッドセットを装着した利用者が作業中に停止し，「次に何をすればいいか？」とシステムに尋ねる．システムは 1 人称視点ビデオに基づき，次に必要なステップを特定し，ガイダンスを提供する．また，事前に録画された教示用ビデオの対応する再生位置を再生する．

```
Please create a written procedure in
    English for the work in this video.
I would like to insert an image of the
    work in progress in the instructions
to make it easier to see, so please
    write the time of the frame of the
    video to be inserted.
```

4.2 利用者評価結果

人間による評価実験を実施した．評価実験参加者は，Full 映像を視聴した後，Full，Gaze，Center の各条件下で Gemini 1.5pro が生成した説明文を 10 点満点で評価した．

実験は Prolific クラウドソーシング基盤 [11] を使用して実施した．オンラインで 25 名の参加者を募集し，10 点満点での数値評価と（Full）映像記述に対する自由記述フィードバックを提供してもらった．予期効果を防ぐため，映像記述の提示順序はランダム化した．記述が英語で書かれているため，英語能力に基づいて参加者を限定した．参加者の平均年齢は 29.5 歳（$std = 8.4$）であった．

結果を図 6 に示す．Gaze 条件がすべてのタスクタイプにおいて Cener 条件を上回った．また 6 タスクのうち 5 タスクで Gaze 条件が Full 条件を上回る最も高い平均スコア

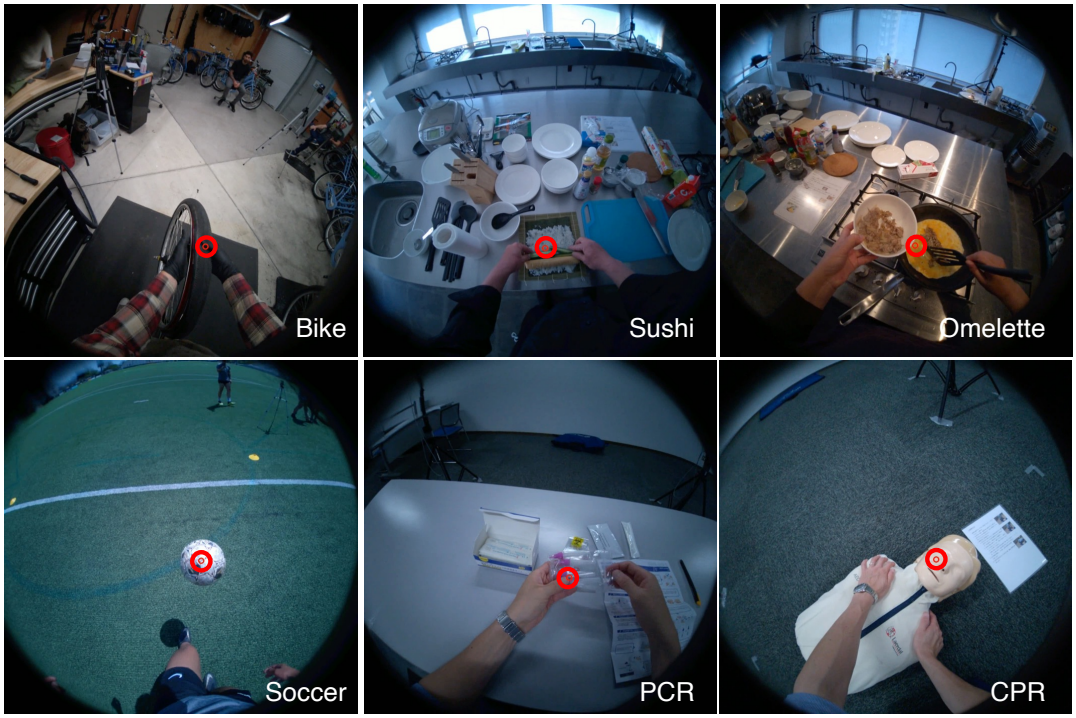


図 4 評価タスク: 「Bike（自転車修理）」「Sushi（寿司作り）」「Omelette（オムレツ作り）」「Soccer（サッカー）」「PCR（ポリメラーゼ連鎖反応検査）」「CPR（心肺蘇生法）」とラベル付けされたタスク．視点を示す赤い円が付加されている．

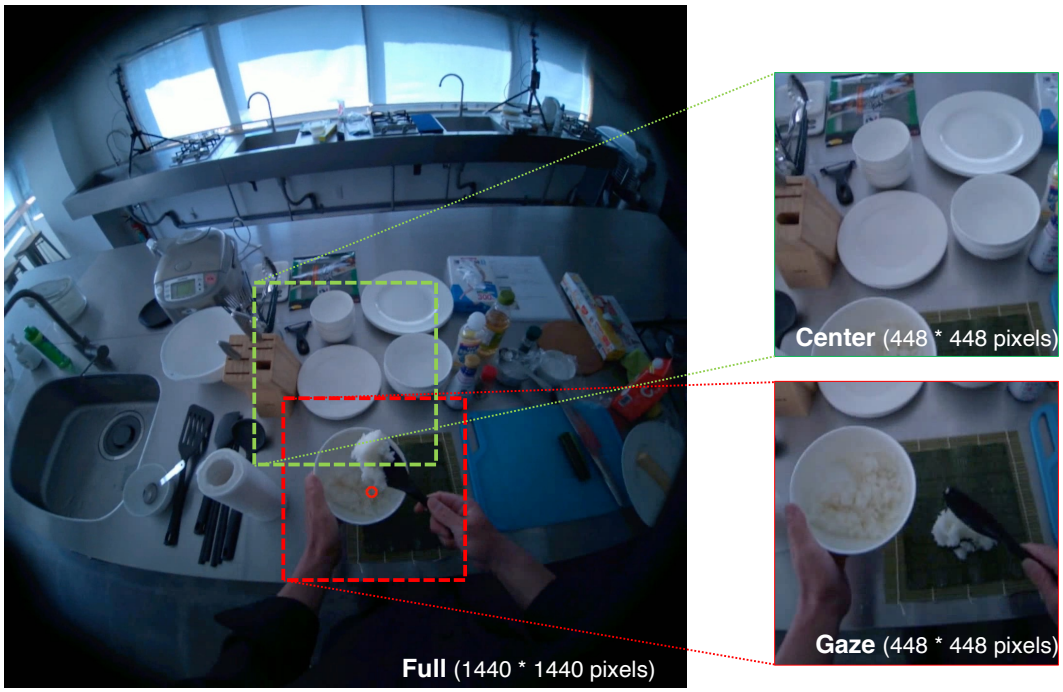


図 5 評価のためのビデオタイプ（Full: 全視野，Gaze: 視線点を中心にトリミングされた動画像，Center: 画像の中心からトリミングされたビデオ）

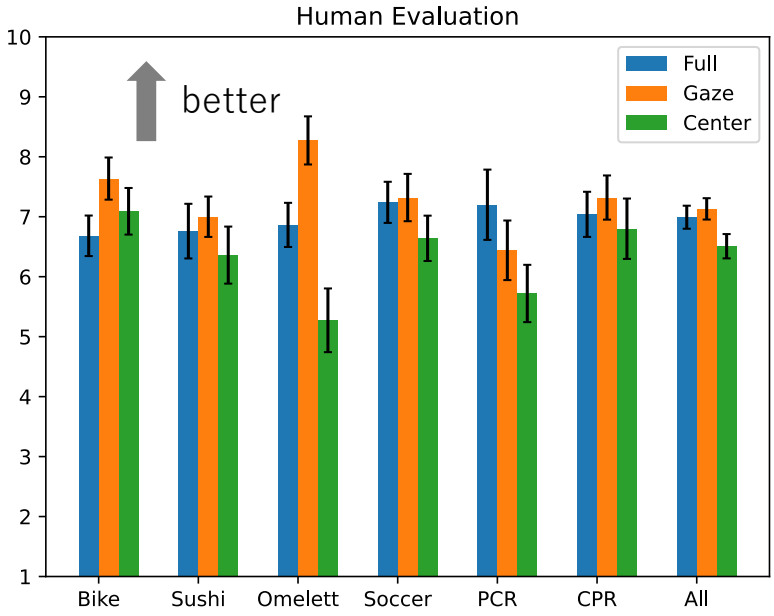


図 6　利用者によるビデオ記述の評価結果

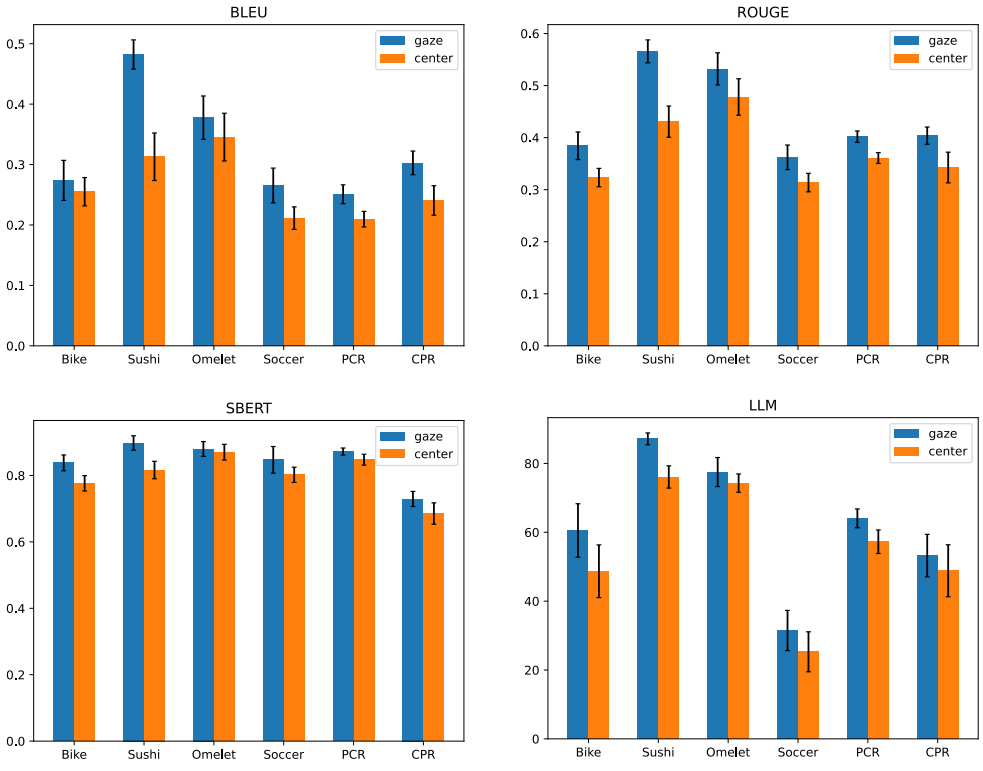
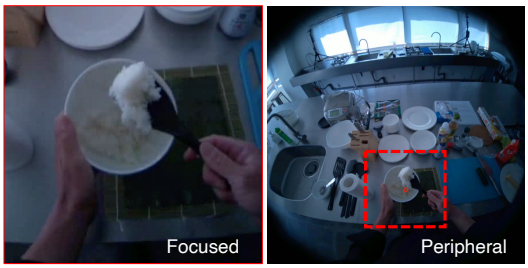
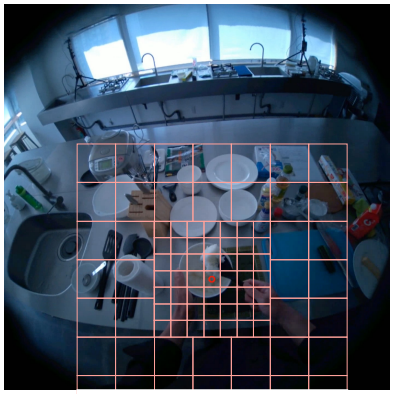


図 7　BLEU，ROUGE，SBERT（Sentence-BERT），および LLM によるビデオ記述の評価

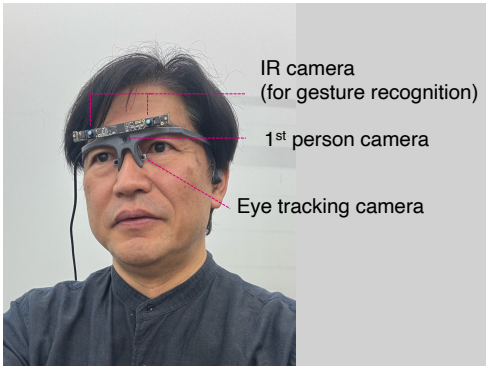


(a)



(b)

図 8　GazeLLM のための代替的なビデオエンコーディングスキーム：（a）注視領域 (高解像度) と周辺領域 (低解像度) のビデオを使用，（b）適応的な ViT グリッドサイズ



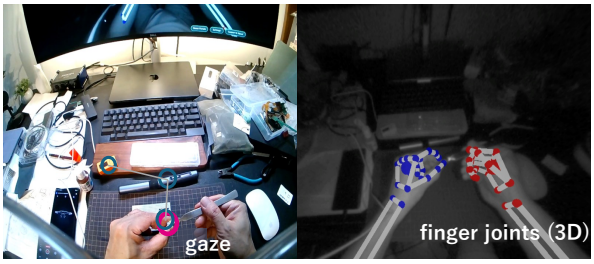


図 9　1 人称視点カメラ，アイ・トラッカー，ハンドジェスチャー認識機能を備えた将来の GazeLLM ウェアラブル構成

を獲得した．参加者は Full 条件の映像のみを視聴したにもかかわらず，Gaze 条件の映像から生成された記述をより高く評価したことになる．

映像の長さと記述の長さの関係についても調査を行った（表 2）．生成された映像記述の文字数において，条件間（Full，Gaze，Center）で有意な傾向は観察されなかった．また，映像の長さと記述の文字数の間にも弱い相関（相関係数 0.13）しか見られなかった．これは，Full 条件において，Gaze や Center 条件と比較してより広い撮影範囲に多くの物体が含まれ—その多くはタスクに直接関係しない可能性がある—モデルがこれらの追加要素を説明しようとすることで，記述の質が低下する可能性があることを示唆している．

4.3　数値的指標に基づく評価結果

次に，各種指標により記述の精度を自動評価した．この評価では，Gaze と Center 条件で生成された記述の精度を Full 条件で生成された記述と比較することとした．この目的のため，以下の 4 つの指標を採用した：

BLEU　BLEU スコア [14] は機械翻訳評価で最も広く使用される手法の 1 つである．機械生成テキストが人間の翻訳者の翻訳例に近いほど，精度が高いという思想に基づいている．この評価では，Gaze と Center 条件で生成された記述の精度を，Full 条件の記述を基準として，BLEU スコアで測定した．

ROUGE　ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[12] も自然言語処理において，人間が作成した参照と比較して機械が生成した要約や翻訳を評価するために広く使用されている．本評価では，最長共通部分列（LCS）に基づく ROUGE-L を使用して，Gaze と Center 条件の記述の精度を Full 条件の記述と比較して測定した．ROUGE-L は，対象と参照要約間の順序付きで共起する単語を数えることで，文レベルの類似性を評価する．

SBERT　Sentence-BERT (SBERT) [15] は，文の比較のために調整された BERT [3] の変形で，意味的類似性を評価する．SBERT は Siamese および triplet ネットワーク構造を利用して，コサイン類似度で比較可能な文章の潜在ベクトルを導出する．SBERT は特に複数文からなる記述間の類似性を測定するのに適している．ここでは，Full 条件で生成された記述と Gaze および Center 条件で生成された記述との類似性を評価した．類似性が高いほど，生成された記述が Full 条件の出力と密接に一致していることを示す．

LLM　LLM 自体を用いた評価も実施した．具体的には，ChatGPT-4o [13] を使用して，以下のプロンプトに基づき，Gaze と Center 条件の記述を Full 条件の記述と比較して 0 から 100 の範囲でスコア化した：

```
I will provide you with two texts: Text
    A and Text B. Please evaluate how
   well the content described in Text A
    is also covered in Text B. Assess
   the similarity between the two texts
    based on whether Text B includes
   the key information, ideas, and
   explanations found in Text A,
```

```
    regardless of wording or phrasing
    differences.

Please provide a score from 0 to 100,
    where:
- 100 means Text B completely covers
    all the key points and information
    from Text A,
- 0 means Text B does not cover any of
    the key points from Text A,
- A score between 0 and 100 represents
    partial coverage.

Show score as '** Score:50 **'
After assigning a score, explain the
    reasoning for your score in a few
    sentences.
### Text A:
{text_1}
### Text B:
{text_2}
```

BLEU と ROUGE は n-gram に基づく統計的な尺度で，同一の語彙やフレーズに反応する．対照的に，SBERT は文埋め込み空間内でのコサイン類似度を使用し，特定の語彙を超えてより広い範囲で文章の意味的な類似性を評価する．LLM 条件は全体的な文の理解に基づく評価を提供すると想定される．これら 4 つの異なる指標により，Full 条件と他の条件との間の包括的な比較が可能となる．

評価結果を図 7 に示す．すべての評価条件とタスク項目において，Gaze 条件のスコアは Center 条件を上回った．これは，視線ベースの切り出しが単純な中心ベースの切り出しよりも各種の作業をより正確に捉えていることを示唆している．

これらの評価において，定義上，Full 条件は完全なスコア（BLEU，ROUGE，SBERT: 1.0，LLM: 100 点）を達成する．しかし，Full 条件で生成された動画説明文が必ずしも最も正確に作業を表していると保証されているわけではない点は留意が必要である．一人称視点映像の広視野角の性質により，記述には作業に直接関係のない周辺視野の物体が含まれることがある．

二つの評価実験により，以下のことが確認された：

- 利用者評価において，Gaze 条件で生成された記述は，完全な一人称視点映像（Full 条件）や完全画像の中心から切り出した映像（Center 条件）からの記述と比較して，より高い評価を受けた．Gaze 映像の画素数が Full 映像の 10 分の 1 に過ぎないことを考えると，提案手法は MLLM への入力として効果的な情報削減を実現していることを示している．
- 4 つの定量的評価（BLEU, ROUGE, Sentence-BERT, LLM ベースの評価）においても，Gaze 条件は Center 条件より高いスコアを達成し，視線に基づく画像の切り出しがより効果的であることを示唆している．

5. 議論

本研究では，視線情報を活用して一人称視点映像からのデータ量を削減し，LLM による効率的な理解を可能にする手法を提案している．現在は視線位置を中心とした矩形領域を切り出しているが，他にも改善の可能性がある（図 8）．図 8（a）は，視線中心の映像とダウンサンプリングされた全体視野の両方を提供するアプローチを示しており，視線外の領域をより低解像度で理解することを可能にする．このアプローチは，人間の視覚における中心視野と周辺視野の関係に対応している．MLLM 研究では，複数の時間レイヤーを持つマルチストリーム入力に統一的な位置エンコーディングを提供する Multimodal Rotary Position Embedding（M-RoPE） [19] などの手法が探究されており，それを GazeLLM に適用することも可能かもしれない．

図 8（b）は，ViT における 2D パッチのグリッドに変形を導入する別のアプローチを示している．このアプローチでは，視線位置付近でパッチの密度が高く，周辺部に向かって疎になっていく．これもまた人間の視覚における中心視野と周辺視野の関係を反映している．この他にも，ViT の重み付けを適応的に変化させる研究 [21] の利用も考えられる．

本研究では 1fps にダウンサンプルされた動画と視線情報を利用したが，視線移動に鑑みて画像の時間間隔を変更することも考えられる．視線には作業に集中している際の fixation（注視）や，次の領域に移動する saccade（衝動性眼球運動）の情報が含まれており，これらの情報を活用する（たとえば fixation により高い重み付けを与える）ことで，より効果的な映像切り出しが可能になるかもしれない．

本研究では視線による画像切り出しの効果を検証するために映像の音声はあえて利用していない．実用的には，音声による説明を併用しながら作業を教示することが効果的であると予測できる．このような音声，一人称視点映像，視線情報を組み合わせたマルチモーダル LLM の研究も今後の課題である．

視線情報以外の，装着者に関する情報を取り入れることも可能である．例えば，視線情報と同様に，手の位置情報を LLM による認識の優先領域とすることができる（図 9）．手指の位置は現在の作業対象を示すのに対し，視線は次の作業対象を予測する傾向があるため，視線入力と手の活動情報を組み合わせることで，より正確な作業理解が可能になるかもしれない．

6. 結論

本研究では，一人称視点映像からタスク記述を生成する

表 2 ビデオ記述の文字数

タスク	Full	Gaze	Center
Bike	1312.7（std = 521.8）	1419.8（std = 563.4）	1561.7（std = 432.0）
Sushi	1968.1（std = 169.5）	1996.3（std = 192.3）	1711.9（std = 543.0）
Omelette	1600.2（std = 244.4）	1656.6（std = 316.5）	1468.2（std = 191.2）
Soccer	2107.2（std = 358.8）	1728.1（std = 336.0）	1768.1（std = 364.6）
PCR	1522.5（std = 403.9）	1386.8（std = 481.1）	1342.2（std = 513.3）
CPR	1845.6（std = 292.1）	1866.5（std = 394.8）	1593.7（std = 438.8）

マルチモーダル LLM（MLLM）の実験を行った．具体的には，視線に基づいて切り出された画像が元の 10 分の 1 の画素数しか含まない場合に，タスクがどの程度よく記述されるかを評価した．料理，自転車修理，医療，スポーツなどにわたる 6 種類の映像（合計 135 映像）を評価した．評価には BLEU，ROUGE，sentence-BERT，および LLM ベースの記述評価を使用した．すべての場合において，視線に基づいて切り出された映像は，画像中心から切り出された映像と比較して，より高い評価スコアを受けた．

利用者評価実験により，視線ベースで切り出された映像から生成された記述が，完全な映像や中心から切り出された映像からの記述より高く評価されることが確認された．

これらの知見は，MLLM で一人称視点映像の記述を生成する際に視線ベースの切り出しを使用することが効果的なアプローチであることを示している．この手法は処理される画素数を削減し，それによって LLM の計算負荷とメモリ使用量の削減に貢献する．このアプローチはまた，視線情報を活用した MLLM がより長いタスク映像を効率的に処理できる可能性も示唆している．

謝辞：本研究は JST ムーンショット型研究開発事業グラント番号 JPMJMS2012 の支援を受けた．

参考文献

[1] Richard A. Bolt. 1982. Eyes at the interface. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems (CHI '82)*. Association for Computing Machinery, New York, NY, USA, 360–362. DOI: http://dx.doi.org/10.1145/800049.801811

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020). https://arxiv.org/abs/2005.14165

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI:http://dx.doi.org/10.18653/v1/N19-1423

[4] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021). https://arxiv.org/abs/2010.11929

[5] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. (2023). https://arxiv.org/abs/2308.13561

[6] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and others. 2024a. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075* (2024).

[7] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024b. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv preprint arXiv:2408.05211* (2024).

[8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Ge-

breselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. 2024a. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. (2024). https://arxiv.org/abs/2311.18259

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. 2024b. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19383–19400.

[10] Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6536–6546. DOI: http://dx.doi.org/10.1145/3025453.3025821

[11] Prolific inc. 2014. Prolific. (2014). https://www.prolific.co

[12] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[13] OpenAI. 2023. GPT-4 Technical Report. (2023). https://openai.com/research/gpt-4

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. DOI:http://dx.doi.org/10.3115/1073083.1073135

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. (2019). https://arxiv.org/abs/1908.10084

[16] Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, Online, 27–43. DOI: http://dx.doi.org/10.18653/v1/2021.conll-1.3

[17] Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal Integration of Human-Like Attention in Visual Question Answering. In *Proc. Workshop on Gaze Estimation and Prediction in the Wild (GAZE), CVPRW*. 2647–2657.

[18] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. (2024). https://arxiv.org/abs/2403.05530

[19] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).

[20] Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024b. Probing Large Language Models from A Human Behavioral Perspective. (2024). https://arxiv.org/abs/2310.05216

[21] Hongxu Yin, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. AdaViT: Adaptive Tokens for Efficient Vision Transformer. (2022). https://arxiv.org/abs/2112.07658

[22] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).

[23] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. (2024). https://arxiv.org/abs/2306.13549