マルチモーダルAIの過去・現在・未来
―技術ブレークスルーと応用インパクトを一望する―

The Past, Present, and Future of Multimodal AI

国立研究開発法人産業技術総合研究所　情報・人間工学領域　人工知能研究センター　首席研究員
佐藤　雄隆

2001 年北大・院・博士後期課程修了。博士 (工学)。産総研知能システム研究部門コンピュータビジョン研究グループ長、人工知能研究センター副研究センター長等を経て、現在人工知能研究センター首席研究員、筑波大・理工情報生命学術院 教授 (連携大学院)。コンピュータビジョンおよびパターン認識に関する研究に従事。

yu.satou@aist.go.jp

1 モダリティと情報処理

人間は視覚・聴覚・触覚・嗅覚・味覚という五感を介して外界を把握する。これらの知覚チャネルは「モダリティ」の典型例であるが、情報工学の文脈ではさらに、温度・加速度・電磁波スペクトル・三次元形状などセンサ由来の物理信号、金融取引履歴・ネットワークログ・時系列 IoT データのようなサイバー空間で生成されるデータも広義のモダリティとして包括的に扱う。すなわち、意味を持つ独立した情報チャネルこそがモダリティであり、その範囲は技術の進歩とともに拡張し続けている（表 1）。

計算機は早くからこれらの多様なデータを「表層的に」記録・伝送・変換する能力を備えていたが、そこに潜む意味を読み取り、認識・理解することにつなげることは困難であった。転機は 2010 年代前半に訪れた。畳み込みニューラルネットワーク（CNN）が人間並みの視覚認識を実現し、音声・動画・三次元点群などにも応用が拡大した。さらに 2017 年に Transformer が提案され、大規模事前学習との組み合わせにより自然言語理解の性能も桁違いに向上した。これらの成果により、AI による単一モダリティ特化タスクは条件次第では人間を超える性能にも到達し、実社会で広く用いられるようになった。

しかし、人間が五感を駆使して生活している事実からも推察できるように、現実世界の課題の多くは図面を読み、説明書を理解し、音声指示に応じて行動するなど、複数モダリティが相互に絡み合う。単一モダリティ処理の成熟は、むしろ「複数の情報チャネルを統一的に扱う技術」の必要性を際立たせることになった。

表 1　モダリティの例

由来	モダリティの例	代表的データ
実世界パターン系 - 物理現象をセンサで"直接"計測して得るデータ	視覚（光学）	RGB画像・動画、深度マップ、LiDAR点群
	聴覚（音響）	生音波形、周波数スペクトル
	振動・運動	IMU 6軸データ、GPS軌跡
	触覚	圧力分布マップ、フォース時系列
	環境・化学	温湿度ログ、ガス濃度スペクトル
デジタル記号系 - 人・システムがデジタル空間で生成・整理し、一定の規則に従って表現した記号列データ	テキスト（自然言語）	ニュース記事、チャットログ、特許明細書
	構造化テーブル	金融取引CSV、センサ統合DB
	グラフ構造	ナレッジグラフ、ブロックチェーン取引グラフ
	プログラム・符号化言語	ソースコード、バイトコード
	専用記号列	化合物SMILES、DNA塩基配列

2 マルチモーダル処理

2.1　処理の概要

マルチモーダル処理とは、画像・音声・テキストといった異種モダリティを同時に解析し、単独では得られない知

見や、それに基づく新たな機能の創出を可能にする技術体系である。単に複数の入力チャネルを受け取るだけではなく、相互参照を前提とした表現の結合を通じて、高次の判断や生成を実現する点に特徴がある。今日では画像・言語などはもちろんのこと、触覚信号や深度データ、慣性計測ユニット（IMU）などのロボットが実世界で動作する際に不可欠なセンサ情報、さらには倉庫ロボット群の稼働ログのような大規模センサ時系列データまでもが統合対象となりつつあり、その応用範囲は急拡大している。

マルチモーダル処理では、まず各モダリティ固有の入力信号から意味的に有用な表現を抽出する表現学習を行う。初期には、エッジ検出や色ヒストグラムといった人手設計による低次特徴量（hand-crafted features）が主に用いられていたが、現在では CNN や Transformer に基づく end-to-end 型の深層ネットワークが主流であり、特に自己教師あり学習によって、大量の未ラベルデータから高次の抽象表現を自動獲得する手法が広く採用されている。

しかし、マルチモーダル処理における本質的な課題はその先にある。画像とテキスト、あるいは動画と音声といった異種モダリティを融合する際には、粒度や時間軸の違いに起因するアラインメント問題が不可避である。時間同期・空間位置合わせ・意味的対応付けのいずれか一つでも破綻すれば、後続の統合処理は著しく劣化する。

2.2　モダリティ融合の壁

こうしたアラインメント問題を受けて、さまざまなモダリティ融合戦略が設計されてきた。マルチモーダル処理の有効性は早くから広く認識されていたにもかかわらず、モダリティ間で粒度や時間軸の異なる信号を適切に整合させる決定的な手法は、近年まで存在しなかった。人間は自然に行っている処理を、機械で実現する困難さが、まさにこのアラインメント問題に集約されていたといえる。

モダリティ融合の方法は大きく三つに分類できる（図1）。Early Fusion は画素値や MFCC（メル周波数ケプストラム係数）のような低次特徴を連結し、単一ネットワークに流し込む方式で、1990 年代の音声・口形状認識から用いられてきた。Late Fusion はモダリティごとに別モデルで推論し、最終段でスコアを統合する決定レベル融合であり、2010 年代の CNN ベース医用画像・テキスト検索システムに広く採用された。これらの折衷として Intermediate（Hybrid）Fusion が提案され、中間層で段階的に結合することで、より積極的に融合のメリットを引き出そうとする試みが行われたが、当初は整合位置や重み付けを手設計せざるを得ず、決定的な解法には至っていなかった。

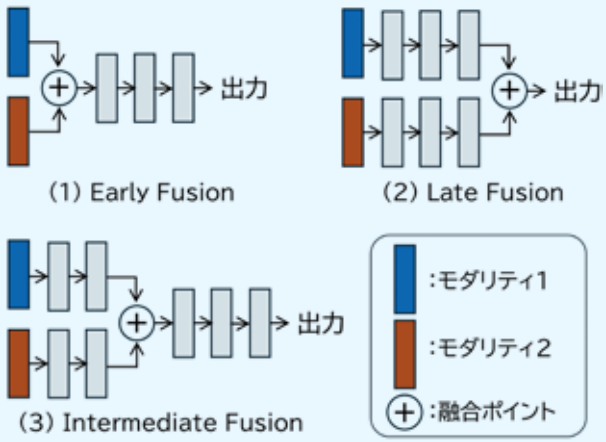

図1　モダリティ融合の基本概念

表2　各注意機構の概念

機構	対象	主な用途	テキストの例	画像の例
Self-Attention	同一系列内の1つの情報	要素間の関係把握	「それ」→「ソファー」	顔としっぽの関連を見る
Cross-Attention	異なる情報（例：テキストと画像）	対応づけ関連づけ	画像→「猫が座っている」	テキストの意味に合う画像部位を探す
Multi-Head Cross-Attention	複数の視点（ヘッド）でのCross-Attention	複数の関係を並列に把握し統合	背景/猫/ソファーを同時に見る	異なる特徴に並行注目して全体を解釈

2.3　Attention Is All You Need

大きな研究動向の転換点になったと評されるのが、Transformer に代表される自己注意アーキテクチャである。詳しい機構については参考文献[1]を参照いただきたいが、やや難解であるので、補足的な情報として表2に各注意機構の概念をまとめた。Self-Attention は系列内の近距離・長距離の両方の依存関係を一度に評価し、従来の CNN や RNN が苦手とした遠隔要素の関係までも損なわずに統合する。さらに Cross-Attention は、クエリとキー／バリューを異なる系列（異種モダリティ）から取得して重み付け可能であり、Intermediate Fusion の一形態として広範に適用されている。

こうした構造的革新により、視覚質問応答、動画字幕生成、ロボティクス操作支援など、多様なマルチモー

ダルタスクにおいて性能と汎用性の両立が進んだ。しかし一方で、これらのモデルは依然として、ラベル付きデータセットの拡充に依存しており、マルチモーダル共通空間の本格的な学習には大規模なペアデータ（例えば、画像とその説明文のペア）が必要とされていた。

この限界を大きく緩和したのが、次章で扱う対照学習の枠組みであり、その代表的実装である CLIP の登場が、マルチモーダル AI の進化に大きな転機をもたらすこととなる。

3 対照学習と大規模マルチモーダルモデルの飛躍

前章で述べたように、Transformer の導入によってアラインメントの柔軟性は飛躍的に高まった。しかし、高品質なペアデータの不足は、依然としてモデルのスケールアップを阻む主要因であった。本章では、このデータ障壁を突破し、画像と言語をはじめとする異種モダリティを実用水準の共通空間へ束ねた対照学習の原理と、その有効性を決定的に裏付けた CLIP について述べる。CLIP は、Cross-Attention を用いずとも、Web 由来の弱ラベルを対照学習で活用することでこのデータ障壁を突破し、画像とテキストの意味的な対応関係を問うタスクで飛躍的性能を示した。

対照学習（contrastive learning）[2] は、「意味的に対応する組を近づけ、無関係な組を遠ざける」という単一の損失関数を中核に据える。Web から収集した画像と周辺テキストを正例とし、ランダムに組み替えたペアを負例とすることで、モデルは自己教師あり学習によってクロスモーダル共通表現空間を獲得できる。大量データによりノイズを平均化しつつラベル作成コストを大幅に削減できる点が、従来の教師付き学習との決定的な違いである。端的に言えば、データ要求を質から量へと転換することを可能にする手法である。

この原理を桁違いのスケールで実装したのが、2021 年初頭に公開された CLIP[3] である。4 億組の画像・テキストペアによる対照学習により、ImageNet ゼロショット Top-1 精度が 76% を超えるなど、未学習の課題を説明文ひとつで解く性能を実現した。特筆すべきは、画像とテキストが同一の埋め込み空間に投影され、互いに共通の意味表現で接続されたことで、タスク固有のラベル付きデータを追加せずとも未知のタスクに対応可能な水準に到達した点である。

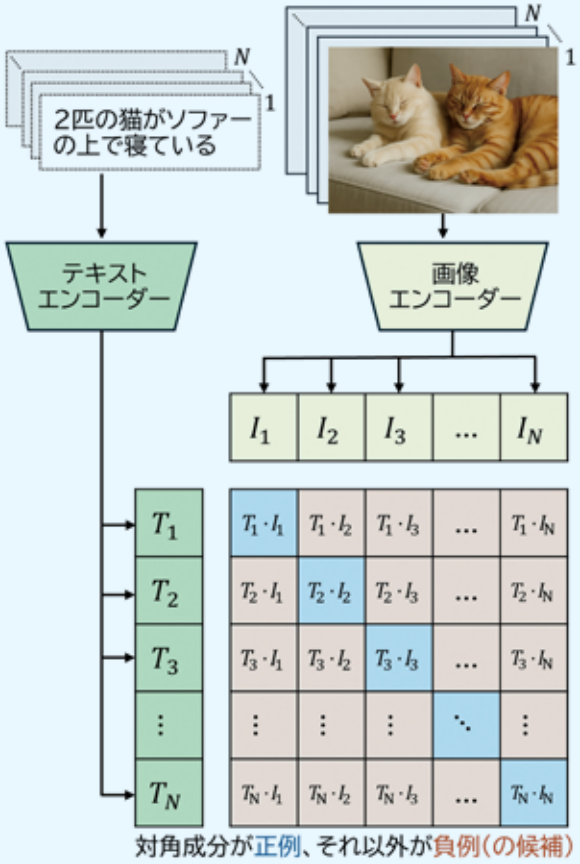

図２　対照学習

CLIP が構築したマルチモーダル埋め込みは、テキストと画像を単一の意味空間へ射影することで、両者のセマンティックな対応付けを高精度に実現した。同手法は統合位置の観点からは Late Fusion に分類されるが、画像・テキスト両エンコーダを共通の対照損失で同時最適化し、潜在空間でモダリティ整合を完遂しているため、積極的融合戦略に位置づけられる。発表からわずか一年余で、この汎用的な埋め込みベクトルが DALL·E 2 や Stable Diffusion などのテキスト条件付き拡散モデルへ流用され、検索や分類といった「理解」タスクから、生成タスクに至るまで、共通の言語表現で一貫して制御できる枠組みが実現された。さらに近年では Diffusion Policy[4] のように行動生成へも波及し、「コップをそっと持ち上げる」といった自然言語指示からロボットアームの三次元軌道を直接合成するプロトタイプが報告されている。

このように、対照学習は (i) 大規模データ獲得の壁、(ii) モダリティ間の表現統合、(iii) 多様な出力タスクへの展開という三重の壁を同時に突破した。その成果は拡散モデルと結び付き、マルチモーダル AI は「意味共有」を

核に、生成・推論・検索・編集・実世界操作へと展開可能な統合的基盤へと成熟しつつある。

4 マルチモーダル AI の社会実装と展開

前章では、対照学習がもたらした研究上のブレークスルーについて検証した。本章では、マルチモーダル AI の社会実装に向けた進展について、プラットフォーム整備、評価指標、説明性確保、応用事例の順に取り上げながら、全体像を俯瞰する。

4.1 プラットフォーム整備

まず注目すべきは、2024 年以降に急速に整備されたマルチモーダル API 群である。OpenAI の GPT-4o、Google の Gemini 1.5 Pro といったクラウド型モデルは、テキスト・画像・音声といった複数のモダリティを単一のインターフェースで扱えるだけでなく、マルチモーダル入力に対してマルチモーダル出力を返すストリーミング対話にも対応しつつある。たとえば、GPT-4o では音声で話しかけながら画像を提示すると、音声で応答しつつ画像の内容をほぼリアルタイムに解説するといったインタラクションが実現されている。GPT-4o などの最新マルチモーダルモデルは、従来は別個に扱われていた自動音声認識（ASR）と音声合成（TTS）の推論処理を単一のニューラルネットワーク内で統合的に実行する設計を採用している。これにより、開発者は OpenAI Realtime API（HTTP/JSON ベース）や、Gemini Live API が提供する gRPC ストリーミング SDK などを通じて、音声・画像を含むリアルタイム入出力をサブ秒レベルで扱えるようになりつつある。

一方、オープンソースの分野でも Meta の ImageBind、LLaVA-Next、Salesforce の BLIP-2 などが Hugging Face 上で公開されている。パラメータ数は数億から数十億まで幅広く、たとえば 7 ～ 13B 級のチェックポイントであれば、LoRA や QLoRA などのパラメータ効率化手法を用いて A100 80GB GPU 1 枚でもファインチューニングできるという報告がある。クラウド上で即時に利用可能な環境と、ローカルで機密性を確保できる環境の双方が整ったことで、企業は自社データの性質やコンプライアンス要件に応じて、柔軟かつ戦略的にデプロイ手法を選択できるようになった。

4.2 評価指標の整備

プラットフォームの整備に呼応して、汎用ベンチマークも急速に高度化している。画像＋テキスト領域では、従来広く用いられてきた VQA v2 や MS-COCO Caption に加え、大学レベルの専門知識と多段階推論を要求する MMMU、単一画像に対する多肢選択式 QA で多面的能力を測る MMBench などが公開され、単純なキャプション生成を超える複雑な推論能力を評価できるようになった。動画タスクでは、映像と字幕を同時に利用する TVQA ／ MovieQA などが広く参照され、時系列コンテキストとセリフ情報を統合した理解が評価対象に含まれている。

一方、正解率だけに依存する評価には限界が認識されつつあり、OpenAI の System Card などでは幻覚（ハルシネーション）発生率、毒性・有害表現スコア、回答の一貫性など安全性・品質にまたがる多軸指標が提示されている。こうした複合評価は、モデルのバージョン管理や A/B テストにおける品質比較の客観的根拠として活用され始めている。

4.3 説明性の確保

評価の厳格化と並行して、ユーザに対する説明責任も一層強く求められている。Transformer における Self-Attention および Cross-Attention は比較的可視化が容易であり、Grad-CAM、Attention Rollout といった手法を用いることで、モデルがどの領域やトークンに「重みを配分したか」をヒートマップとして提示できる。たとえば、視覚質問応答（VQA）では、回答根拠となる画像パッチが赤くハイライトされ [5]、医用画像 AI においては、病変候補の輪郭が自動的に描画されるなどの例がある [6]。

生成モデルの根拠提示手法としては、Retrieval-Augmented Generation（RAG）[7] が挙げられる。RAG は検索強化（外部知識の動的注入）に用いられる一方、引用付き回答を生成できるため説明支援の手法としても利用される。ただし出典が提示されるのは取得文書が実際に回答に反映された場合に限られ、モデル内部に残る暗黙知や幻覚までは説明し切れない点に留意が必要である。

近年では、出力文の各部分がどの根拠に基づいて生成されたかを明示可能にする Attributable

Generation[8] や、内部推論をテキスト形式で明示させる Chain-of-Thought[9] などの手法も提案されており、こうした複合的な説明支援の積み重ねを通じて、「根拠の提示が標準となる」文化の定着が期待されている。

4.4　応用に向けた実装・評価事例

実装・評価の取り組みは、すでに多様な産業領域において具体的な成果を生みつつある。医療分野では、胸部 X 線画像と電子健康記録（EHR）を同時に入力するマルチモーダル AI の有効性を示す報告が現れ始めている。たとえば Lin らの PrismICU [10] では、ICU 患者 3,798 例を対象に X 線画像と 74 項目の臨床指標を統合したモデルを構築し、30 日院内死亡予測の外部テストセットで EHR 単独 AUC　0.72 をマルチモーダル AUC 0.82 へ引き上げたと報告している。画像と構造化テキストの相補情報が、単独モダリティでは得られない診断支援の精度向上に寄与した好例であると考えられる。

モビリティ領域では、マルチモーダル学習を用いた自動運転の研究実装が相次いで報告されている。Waymo は 2024 年、カメラ画像と車両状態・経路指示を同時に処理して進路予測と物体検出を一体化する end-to-end モデル「EMMA」[11] を発表し、公開ベンチマークで従来法を上回る精度を示したとしている。Ghost も 2023 年、前方のレーダーと複数のステレオカメラ映像を融合し、複雑な走行環境において距離と速度を冗長的に推定する研究デモを公開している。

メディア分野では、映像と台本の両方を同時に解析して要約や内容抽出を行う研究が、Tencent や Korea University のプロジェクト [12] などで進められており、放送コンテンツのマルチモーダル理解に向けた技術基盤が整いつつある。

ロボティクス分野では、RT-2（Google DeepMind, 2023）や PaLM-SayCan（Google Brain & Everyday Robots, 2022）が実機ロボットに実装され、カメラ画像と自然言語指示に加え、ロボットの内部状態（関節角度・グリッパ開度など）を表す状態ベクトルを統合したモデルによって「コップをシンクに運ぶ」「本を棚に戻す」など高レベル指示を柔軟に遂行できることが報告されている。さらに Diffusion Policy は、映像とロボットの内部状態を条件に拡散過程で軌道を生成し、現実世界の把持や整列動作 18 タスクで最大 90% の成功率を達成した。これらの研究実装が相次いで公開されたことで、マルチモーダル AI を用いたロボットが、実用化に向けたステップへと着実に進みつつある。

4.5　社会実装に向けて

前節までで述べたように、プラットフォーム、評価法、説明性、応用事例の各側面において、マルチモーダル AI は実運用に向けた実装段階へ進展しつつある。とりわけ、視覚・言語・音声・力覚といった異なるモダリティを統合的に処理し、意思決定や身体動作へと接続するアーキテクチャが現実の産業応用に結び付き始めている点は、技術的にも社会的にも注目に値する。一方で、共通表現空間の安定性や推論過程の透明性、大規模データへの依存といった基盤的課題もなお存在するが、研究開発と社会実装が両輪となって加速しており、マルチモーダル AI は今まさに、新たな価値創出に向けて技術と現場が歩調を合わせて進む段階に入っていると考えられる。

5　まとめ

本稿では、単一モダリティ AI の成熟、アラインメントとフュージョンを核とするマルチモーダル基盤、対照学習による大規模モデルの躍進、そして GPT-4o や Gemini 1.5、RT-2 といったプラットフォームの登場を順に概観してきた。共通して浮かび上がったのは、異種情報を共通の潜在表現に束ね、その表現を起点に生成と実世界操作を循環させるサイクルが既に一部実務へ入り込みつつあるという事実である。医療では胸部 X 線 + EHR の統合診断が治験段階に達し、モビリティやロボティクスでもマルチモーダルモデルが実機運用を試行中だ。

このような情報処理サイクルが知財分野に応用されるとすれば、最も早く恩恵を受けるのは先行技術調査であろう。請求項、図面、化学構造式といった多様なモダリティを共通の潜在空間に統合できれば、たとえば「コイルばねで曲率半径 3mm 以上」といった自然言語のクエリに対して、関連図面を視覚的に表示し、対応するクレーム箇所を自動でハイライトするといった検索・可視化の統合的な支援が現実味を帯びる。調査、要約、翻訳などを単一のマルチモーダルモデルで連携処理できるようになれば、従来分業されていた役割の境界も緩やかに

なり、業務全体の効率向上が期待される。

さらに今後の展開として注目されるのが、出願文書の作成支援である。契約書や報告書の起草において生成モデルが活用されつつある中、特許出願でも、図面や構成要素の整合性を確認し、図番参照の抜けや請求項間の論理矛盾を検出する「補助的校閲エージェント」として、マルチモーダル AI が貢献する可能性がある。また、拡散モデルなどの生成技術と組み合わせることで、将来的には図面のプロトタイプ生成のように既存業務を支援する機能に加え、実験手順の動画化といった新たな参考用の補助表現の導入も検討可能であろう。

もっとも課題も大きい。

1. 幻覚・誤補完―AI が実在しない構造を図示・記述すると出願や審査が混乱する。

2. 権利帰属―特許図面・商標画像を学習に使う際は著作権・ライセンスを精査する必要がある。

3. 説明責任―係争で「何を根拠に類似判定したか」を示せなければ受け入れられない。

4. ベンチマーク不足―多モダリティ横断で公開データセットが乏しく客観比較が難しい。

解決には、

（1）権利整理済みの多言語データ整備

（2）判断根拠を追跡できるログ標準

（3）幻覚率や攻撃耐性を第三者が共通評価する枠組み

の三本柱が不可欠である。整備が進めば、調査からドラフティング、係争支援まで知財ワークフロー全体が一貫化し、専門家は創造的判断に集中できる。

さて、ここまでの議論を総括するとマルチモーダル AI は、人間が自然に行うような感覚統合を機械において実現し、〈知〉の構造を再編しつつあると言えるだろう。医療・モビリティ・ロボティクスなどの先行例が示すように、研究と社会実装は同時進行で加速し、サイバー空間の推論が物理世界へ直接作用する段階へ踏み出した。今後は技術と制度を両輪として連携させることで、新たな価値創出のプラットフォームとしてのマルチモーダル AI が、本格的に定着すると期待される。

参考文献

[1] A. Vaswani et al.: Attention Is All You Need. In Proc. NeurIPS 2017, pp. 5998–6008. (or arXiv:1706.03762)

[2] R. Hadsell, S. Chopra and Y. LeCun: Dimensionality Reduction by Learning an Invariant Mapping. In Proc. CVPR 2006, pp. 1735–1742.

[3] A. Radford et al.: Learning Transferable Visual Models from Natural Language Supervision. In Proc. ICML 2021, pp. 8748–8763.(or arXiv:2103.00020)

[4] S. Wang et al.: Diffusion Policy: Visuomotor Policy Learning via Diffusion Models. arXiv 2303.04103, 2023.

[5] H. Tan and M. Bansal: LXMERT: Learning Cross-Modality Encoder Representations from Vision and Language. In Proc. EMNLP 2019, pp. 5100–5111.

[6] H.-Y. Zhou et al.: Generalized Radiograph Representation Learning via Cross-supervision between Images and Free-text Radiology Reports. Nat. Mach. Intell., 4, pp. 32–40, 2022.

[7] P. Lewis et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401, 2020.

[8] T. Gao et al. “Enabling Large Language Models to Generate Text with Citations.” EMNLP 2023, pp. 6465-6488.

[9] J. Wei et al.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv 2201.11903, 2022.

[10] J. Lin et al.: Development and Validation of Multimodal Models to Predict the 30-Day Mortality of ICU Patients Based on Clinical Parameters and Chest X-Rays. J. Imaging Inform. Med., 37, pp. 1312–1322, 2024.

[11] Y. Ren et al.: EMMA: End-to-End Multimodal Model for Autonomous Driving. arXiv 2410.23262, 2024.

[12] Y.-J. Seo et al.: DramaQA: Character-Centered Video Story Understanding with Hierarchical Question-Answering. In Proc. AAAI 2021, pp. 2991–2999.