Image and Video Question Answering with Large Language Models: A Comprehensive Review

Alexander Davis, Justin Parker, Julian Perry
Universidad Autónoma de Santo Domingo

***Abstract*—Image Question Answering (IQA) and Video Question Answering (VQA) are pivotal tasks at the intersection of computer vision and natural language processing, aiming to enable machines to comprehend visual content and respond to human questions in natural language. Historically, these fields have advanced through specialized architectures and feature engineering, but their ability to handle complex reasoning, open-ended generation, and real-world ambiguity has been limited. The advent of Large Language Models (LLMs) has fundamentally transformed the landscape of AI, showcasing unprecedented capabilities in language understanding, generation, and intricate reasoning. This survey provides a comprehensive review of the state-of-the-art in IQA and VQA, specifically focusing on how LLMs are being integrated and leveraged to push the boundaries of visual-linguistic intelligence. We delineate the foundational concepts of VQA/IQA and LLMs, categorize prominent architectural paradigms for their integration, scrutinize existing datasets, benchmarks, and evaluation metrics, and critically analyze the current challenges and promising future directions. Our review highlights the transformative potential of LLM-enhanced visual QA systems in overcoming limitations of traditional models, while also shedding light on emergent issues such as hallucinations, computational costs, and the need for robust evaluation. This work aims to serve as a structured guide for researchers navigating this rapidly evolving domain, fostering further innovation at the confluence of vision, language, and artificial intelligence.**

I. Introduction

The ability of machines to perceive, understand, and interact with the visual world is a long-standing grand challenge in artificial intelligence. Image Question Answering (IQA) and Video Question Answering (VQA) represent critical steps towards this goal, requiring systems to interpret visual content (images or videos) and generate accurate, contextually relevant answers to natural language questions. Early approaches to IQA and VQA often relied on task-specific models, meticulously designed feature extractors, and rule-based reasoning, demonstrating progress on constrained datasets [1]–[3]. However, these traditional methods frequently struggled with the inherent complexities of real-world scenarios, such as disambiguating visual elements, performing multi-step logical inference, and generating free-form, human-like answers. The limitations often stemmed from their inability to capture the nuanced interplay between visual and linguistic modalities, as well as their lack of broad world knowledge.

The recent proliferation of Large Language Models (LLMs), exemplified by models like GPT-3 and its successors, has ushered in a new era for artificial intelligence. Trained on vast corpora of text data, these models have demonstrated remarkable emergent capabilities, including advanced language understanding, coherent text generation, and sophisticated reasoning across a multitude of natural language processing (NLP) tasks [4], [5]. Their success stems from the Transformer architecture and massive scale, enabling them to learn intricate linguistic patterns and store a significant amount of factual and commonsense knowledge. The challenge and opportunity lie in extending these powerful linguistic abilities to the visual domain, thereby enabling LLMs to understand and reason about images and videos.

Integrating LLMs into IQA and VQA systems holds immense promise. By leveraging LLMs as powerful reasoning engines, multimodal interfaces, or knowledge bases, researchers aim to equip visual QA systems with enhanced capabilities for: (1) Complex Reasoning: Moving beyond simple factoid answers to tackle multi-hop, compositional, and abstract reasoning tasks [6], [7]; (2) Open-ended Generation: Producing diverse and natural language answers, rather than selecting from a predefined set; (3) Knowledge Integration: Accessing and applying external world knowledge to answer questions that extend beyond the explicit content of the visual input [8], [9]; and (4) Adaptability: Rapidly adapting to new domains and tasks with few or no examples (zero-shot/few-shot learning) [10].

This survey aims to provide a comprehensive and structured review of the rapidly evolving field of LLM-enhanced IQA and VQA. We define the scope by focusing on approaches that explicitly integrate or leverage LLMs for visual question answering across both image and video modalities. We will cover the historical context, delve into modern architectures and methodologies, critically assess datasets, benchmarks, and evaluation metrics, and finally, highlight pressing challenges, open problems, and promising future research directions. By synthesizing existing knowledge and identifying key trends, this survey seeks to be a valuable resource for researchers and practitioners in developing more intelligent and versatile visual AI systems.

II. Background on Visual Question Answering and Large Language Models

This section provides foundational knowledge essential for understanding the convergence of visual question answering and large language models. We first trace the evolution of IQA and VQA, highlighting key milestones and inherent limitations of earlier paradigms. Subsequently, we delve into the core concepts and advancements of LLMs, particularly focusing on

their architectural underpinnings, pre-training objectives, and emergent reasoning capabilities.

A. Evolution of Visual Question Answering

Visual Question Answering (VQA) emerged as a challenging task requiring joint understanding of both visual content and natural language queries. Initial approaches often involved feature extraction from images using Convolutional Neural Networks (CNNs), combined with language models (e.g., LSTMs or GRUs) to process questions, followed by a fusion mechanism and an answer prediction layer [1], [3]. Attention mechanisms quickly became crucial, enabling models to focus on relevant image regions and question words, as exemplified by methods like Bottom-Up and Top-Down Attention for Image Captioning and VQA [2]. These advancements significantly improved performance on datasets like VQA v1 and v2.

However, traditional VQA models faced several limitations. One significant issue was their susceptibility to dataset biases and spurious correlations, often leading to models that could answer questions based primarily on linguistic priors rather than genuine visual understanding [11], [12]. For instance, if most questions asking "What color is the banana?" had "yellow" as the answer in the training set, a model might predict "yellow" even for a non-yellow banana. Addressing this required careful dataset design and bias mitigation strategies. Furthermore, complex reasoning beyond direct object recognition, such as multi-hop inference or understanding implicit relationships, remained a significant challenge [6].

The extension to Video Question Answering (VQA) introduced additional complexities, primarily temporal dynamics and higher dimensionality [13]. VQA systems need to track objects, understand actions, and reason about events unfolding over time. Early VQA models adapted IQA techniques by processing video frames independently or aggregating temporal features. More sophisticated methods began to incorporate recurrent neural networks or temporal attention to model the evolution of events. Despite these efforts, VQA often lagged behind IQA in performance due to the added challenge of understanding motion, temporal alignment, and the sheer volume of data.

Simultaneously, open-domain Question Answering (QA), initially text-based, laid groundwork for the need for external knowledge. Systems like those in [14]–[18] tackled questions requiring retrieval from large text corpora. Dense retrieval models, such as DPR [18], became prominent for efficiently finding relevant passages. However, even these systems faced challenges with entity-centric questions [19] or few-shot learning scenarios [20]. Approaches like SPARTA introduced sparse transformer matching retrieval for efficiency and performance [21]. The integration of knowledge graphs into QA further enhanced reasoning capabilities by providing structured knowledge [22]–[26]. These text-based advancements foreshadowed the necessity of integrating similar knowledge and reasoning capabilities into visual QA.

B. Large Language Models: Architectures and Capabilities

Large Language Models (LLMs) are deep neural networks, predominantly based on the Transformer architecture, that have achieved remarkable success in various NLP tasks. The Transformer, introduced by Vaswani et al., relies heavily on self-attention mechanisms, allowing the model to weigh the importance of different words in a sequence when processing any given word. This architecture allows for parallel processing, making it highly scalable and efficient for training on massive datasets.

1) Pre-training Objectives: LLMs undergo a two-stage paradigm: pre-training and fine-tuning. During pre-training, models are exposed to vast amounts of unlabeled text data, learning general linguistic patterns and world knowledge through self-supervised objectives. Common pre-training tasks include:

- **Masked Language Modeling (MLM):** As seen in BERT, portions of the input text are masked, and the model is trained to predict the original masked tokens [27]. This forces the model to learn bidirectional contextual representations.
- **Causal Language Modeling (CLM):** Used in models like GPT, the model predicts the next token in a sequence given the preceding tokens. This autoregressive objective is crucial for generative tasks.
- **Sequence-to-Sequence (Seq2Seq) Pre-training:** Models like T5 use both encoder-decoder architectures and objectives that unify various NLP tasks into a text-to-text format.

These objectives enable LLMs to develop a robust understanding of syntax, semantics, and pragmatics, which underpins their remarkable capabilities. The scale of these models (billions of parameters) and datasets is critical; for instance, [28] explored explicit alignment objectives for multilingual encoders. Work on synthetic pre-training tasks has also explored ways to mitigate issues with crawled corpora [29].

2) Instruction Tuning and Emergent Capabilities: After pre-training, LLMs are often fine-tuned on instruction datasets, a process known as instruction tuning, which aligns them with human instructions and preferences. This further enhances their ability to follow complex commands and perform diverse tasks without explicit task-specific training (zero-shot or few-shot learning) [30]. Prompt-based learning, where tasks are framed as natural language prompts, has become a standard approach to interact with LLMs [10], [31]. Techniques like pre-trained prompt tuning (PPT) and BitFit have been developed for efficient adaptation in low-resource settings [31], [32]. Knowledgeable prompt-tuning further incorporates external knowledge into the prompt verbalizer for improved text classification [33]. LLM-Adapters provide a framework for parameter-efficient fine-tuning, allowing smaller models to achieve competitive performance [34]. Compression techniques like LLMLingua also aim to optimize prompt length for accelerated inference [35].

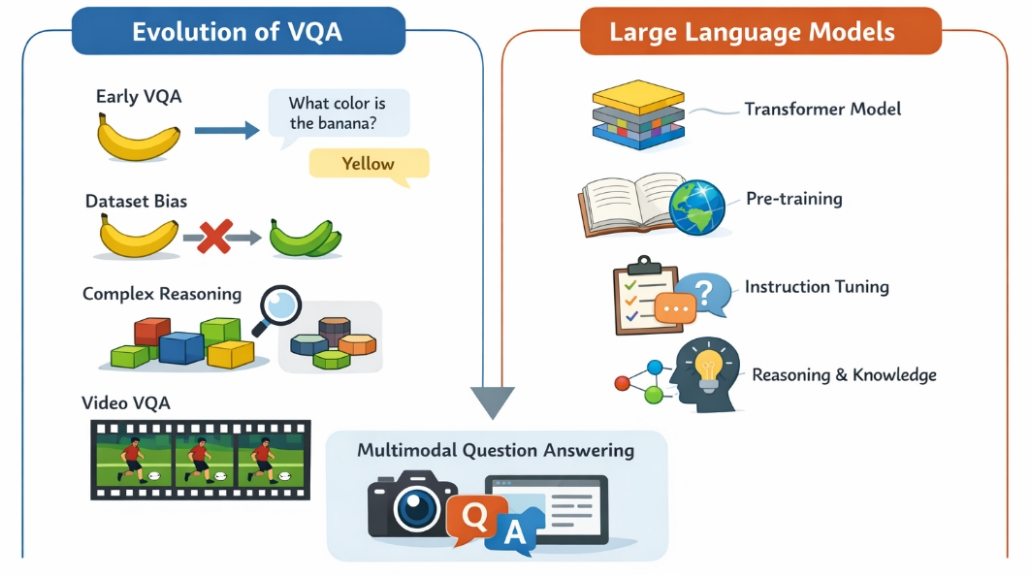

Fig. 1. An overview of how the evolution of Visual Question Answering and the capabilities of Large Language Models converge to enable modern multimodal question answering.

The emergent capabilities of LLMs relevant to visual QA include:

- **Reasoning:** LLMs can perform various forms of reasoning, including arithmetic, commonsense, and logical deduction. Chain-of-Thought (CoT) prompting has significantly improved their reasoning abilities by enabling them to articulate intermediate steps [6], [36], [37]. Self-correction and self-reflection mechanisms further enhance their reliability [5], [38]. However, their logical capabilities have limits, as shown by studies exploring when reasoning beats scale [39].
- **Knowledge Retrieval and Integration:** While not explicit knowledge bases, LLMs implicitly store vast amounts of information learned during pre-training. They can be augmented with external knowledge sources (e.g., knowledge graphs, web search) through Retrieval-Augmented Generation (RAG) to provide more accurate and up-to-date information [8], [9].
- **Generation and Summarization:** LLMs are highly proficient in generating coherent and fluent text, which is essential for producing answers in visual QA. They can also perform tasks like extractive summarization effectively [40].
- **Multilingualism:** Many LLMs are trained on multilingual datasets, enabling them to process and generate text in multiple languages, offering opportunities for cross-lingual visual QA [28], [41].

These foundational capabilities of LLMs serve as a powerful bedrock upon which modern multimodal systems are built, bridging the gap between perception and advanced linguistic intelligence. Furthermore, the development of sophisticated machine learning approaches, such as hybrid supervised-unsupervised learning pipelines, is crucial for addressing complex challenges in various other domains, including fraud detection in online transactions [42].

III. Architectures and Methodologies for LLM-Enhanced Image and Video QA

The integration of Large Language Models (LLMs) into image and video question answering has spawned a diverse array of architectural patterns and methodologies. These approaches largely aim to enable LLMs to ”see” and ”understand” visual information, extending their powerful language-centric reasoning to multimodal contexts. This section categorizes and analyzes these paradigms, highlighting their underlying mechanisms, strengths, and application scenarios.

A. Multimodal Large Language Models (MLLMs) and Vision-Language Pre-training (VLP)

A central theme in LLM-enhanced visual QA is the development of Multimodal Large Language Models (MLLMs) through Vision-Language Pre-training (VLP). These models are designed to learn joint representations across visual and textual modalities, typically by co-training on large datasets of image-text or video-text pairs [43].

1) Unified-Modal Pre-training: Early VLP efforts focused on learning cross-modal alignments. UNIMO [44] proposed a unified-modal pre-training architecture capable of adapting to both single-modal and multi-modal understanding and generation tasks. It leverages cross-modal contrastive learning (CMCL) to align textual and visual information, demonstrating the utility of non-paired single-modal data for generalization. Similarly, BriVL (part of the Chinese WenLan project) also employs cross-modal contrastive learning with a large queue-based dictionary to incorporate more negative samples, outperforming models like UNITER and OpenAI CLIP on various downstream tasks [45]. Models like mPLUG introduce

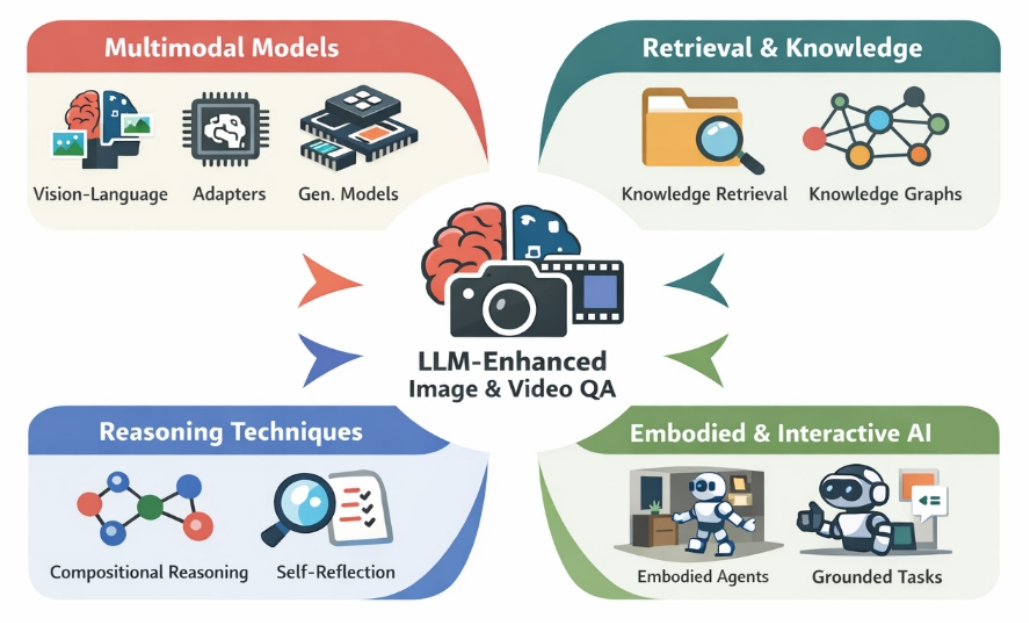

Fig. 2. Overview of architectures and methodologies for LLM-enhanced image and video question answering, highlighting multimodal modeling, knowledge retrieval, reasoning, and embodied interaction.

cross-modal skip-connections for effective and efficient vision-language learning [46].

2) Adapter-based Multimodal Augmentation: To leverage the encyclopedic knowledge and in-context learning abilities of pre-trained LLMs without modifying their immense weights, adapter-based finetuning has become popular. MAGMA [47] augments generative language models with additional modalities using adapter-based finetuning, allowing for end-to-end pre-training with a single language modeling objective. This preserves the LLM's language capabilities while enabling multimodal input. E2E-VLP [48] focuses on end-to-end vision-language pre-training enhanced by visual learning, further emphasizing the holistic learning of multimodal interactions.

3) Vision-Guided Generative Models: Some approaches explicitly inject visual information into generative pre-trained language models (GPLMs). For multimodal abstractive summarization, [49] presented a method to construct vision-guided GPLMs by adding attention-based add-on layers. This allows GPLMs to incorporate visual information without detrimental effects on their text generation abilities, leading to significant improvements in MAS tasks. Video-LLaMA and Video-ChatGPT represent further steps in this direction, fine-tuning audio-visual language models for comprehensive video understanding and detailed video discussions [50], [51]. This approach highlights the potential for richer, more interactive multimodal dialogues.

4) Token Efficiency and Visual Text Inputs: An interesting direction explores the efficiency of visual inputs for text. [52] investigates whether textual inputs can be compressed by feeding them as images to reduce token usage in multimodal LLMs. They show that rendering long text inputs as a single image can yield substantial token savings without degrading task performance, suggesting a new form of input compression.

B. Knowledge-Augmented and Retrieval-Augmented Generation (RAG)

While MLLMs learn implicit world knowledge, they often struggle with specialized, up-to-date, or fine-grained factual information. Retrieval-Augmented Generation (RAG) addresses this by dynamically retrieving relevant external knowledge (from text or images) to inform the generation process.

1) Multimodal RAG Architectures: MuRAG [53] proposes a multimodal retrieval-augmented generator for open question answering over images and text. This framework aims to retrieve relevant information from both modalities to formulate a comprehensive answer. A broader survey on multimodal RAG [54] systematically analyzes datasets, metrics, methodologies, and challenges in this evolving field.

2) Knowledge Graph Integration: Explicit knowledge graphs (KGs) can provide structured, factual knowledge to MLLMs. KAT [55] introduces a Knowledge Augmented Transformer for Vision-and-Language, integrating implicit and explicit knowledge in an end-to-end encoder-decoder architecture. This allows for joint reasoning over both knowledge sources during answer generation, showing strong performance on tasks like OK-VQA. Unifying LLMs and KGs is a growing research area, with roadmaps outlining frameworks like KG-enhanced LLMs and LLM-augmented KGs [9].

3) Dynamic and Hierarchical RAG: For complex knowledge-intensive VQA, dynamic and hierarchical retrieval strategies are emerging. QA-Dragon [56] proposes a Query-Aware Dynamic RAG System, featuring a domain

router and a search router that dynamically select optimal retrieval strategies, orchestrating both text and image search agents. This supports multimodal, multi-turn, and multi-hop reasoning. Wiki-LLaVA [57] integrates an external knowledge source of multimodal documents through a hierarchical retrieval pipeline, demonstrating improved effectiveness and precision for visual question answering requiring external data.

C. Reasoning Paradigms in LLM-Enhanced Visual QA

The most significant contribution of LLMs to visual QA is their advanced reasoning capabilities. Researchers are exploring how to leverage these for more sophisticated visual understanding.

1) Compositional and Multi-Hop Reasoning: Traditional VQA models often struggle with compositional reasoning, where the solution requires composing answers to multiple sub-problems. [6] investigated this "compositionality gap" in LLMs, showing that while larger models improve factual recall, their compositional reasoning doesn't necessarily scale at the same rate. Elicitive prompting methods like chain-of-thought (CoT) and self-ask can narrow this gap by explicitly asking and answering follow-up questions. Recent surveys highlight the importance of compositional visual reasoning, tracing its paradigm shifts from prompt-enhanced pipelines to unified agentic VLMs [7]. Visual Reasoning Tracer (VRT) introduces a task requiring models to explicitly predict intermediate objects in a reasoning path, aiming to make visual reasoning more transparent and human-like [58]. LogicVista [59] provides a benchmark to assess MLLMs' integrated logical reasoning in visual contexts.

2) Self-Reflection and Verification: Inspired by human cognitive processes, self-reflection and self-verification mechanisms are being introduced. [36] demonstrates that LLMs can self-verify their answers by performing a backward verification of deduced conclusions, leading to improved reasoning performance. Logic-LM integrates LLMs with symbolic solvers, using LLMs to translate problems into symbolic formulations and then leveraging the solver's error messages for self-refinement [37]. This combination allows for more faithful logical reasoning. Research also delves into whether LLMs can "lie," investigating the internal state and neural mechanisms underlying deception to enhance trustworthiness [60], [61].

3) Causal and Analogical Reasoning: Beyond logical reasoning, causal and analogical reasoning are critical for deeper understanding. CausalVLR [62] provides a toolbox and benchmark for visual-linguistic causal reasoning, encompassing tasks like VQA and image captioning. Analogical reasoning, the ability to map structural relationships between different domains, is fundamental to human cognition. Studies are exploring if LLMs can match human performance in such tasks, suggesting that LLMs might offer a "how-possibly" explanation for human analogical reasoning in contexts not well modeled by existing theories [63].

D. End-to-End Multimodal Learning and Embodied AI

The ultimate goal for many AI systems is end-to-end functionality, where raw multimodal inputs are directly mapped to outputs, often in interactive or embodied settings.

1) End-to-End Multimodal Architectures: The concept of end-to-end learning in multimodal contexts has been explored in various domains, from speech-to-text translation [64] and conversation modeling [65] to keyword spotting and voice activity detection [66]. In visual QA, end-to-end training allows for direct optimization of multimodal understanding. Recent work includes architectures for end-to-end autonomous driving using Vision Language Models (VLMs), demonstrating impressive performance with single camera inputs [67]. Similarly, generalized trajectory scoring and multi-target distillation are used in end-to-end multimodal planning for autonomous driving, showcasing improvements in generalization [68], [69].

2) Embodied and Grounded Visual Reasoning: Embodied AI, where agents interact with and perceive their environment, naturally extends to visual QA. Embodied 3D grounding aims to localize target objects from ego-centric viewpoints based on human instructions. VLM-Grounder [70] proposes a novel framework using VLMs for zero-shot 3D visual grounding based solely on 2D images, outperforming previous zero-shot methods without relying on 3D geometry. DEGround [71] enhances contextual understanding in embodied 3D grounding by sharing DETR queries for both detection and grounding, highlighting instruction-related regions and incorporating sentence-level semantics. A survey on visual grounding [72] provides a comprehensive overview of this field, including grounded pre-training and grounding multimodal LLMs. The notion of embodied and social grounding for LLMs is gaining traction, suggesting that active bodily systems, temporally structured experience, and social skills are crucial for true understanding [73]. The "chain-of-sketch" method also addresses global visual reasoning by breaking down complex tasks into intermediate visual steps [74].

E. Multimodal Sentiment Analysis and Emotion Recognition

Beyond factual QA, LLMs are increasingly being adapted for subjective and affective understanding in multimodal contexts. Multimodal Sentiment Analysis (MSA) and Emotion Recognition in Conversation (ERC) are key applications where the fusion of language, visual (facial expressions, gestures), and acoustic (tone, prosody) cues is critical.

MMGCN [75] proposes a multimodal fused graph convolutional network for ERC, leveraging speaker information and multimodal dependencies. CTFN [76] introduces hierarchical learning with a Coupled-Translation Fusion Network for MSA. ConFEDE [77] uses contrastive feature decomposition for MSA, while multimodal phased transformers also contribute to sentiment analysis [78]. UniMSE [79] aims for unified multimodal sentiment analysis and emotion recognition. Learning language-guided adaptive hyper-modality representation (ALMT) for MSA helps suppress sentiment-irrelevant and conflicting information across modalities [80]. Furthermore,

representations of facial affect have been introduced for automated multimodal deception detection, improving performance in high-stakes situations [81]. These works demonstrate the versatility of multimodal LLMs in understanding complex human social signals.

F. Continual Learning and Adaptation

As models are deployed in dynamic environments, the ability to continually learn from new data without forgetting previously acquired knowledge (catastrophic forgetting) is crucial. This is particularly relevant for MLLMs that need to adapt to evolving visual-linguistic domains.

Continual learning in practice involves managing ML models in production, adapting to shifting data distributions, and retraining when necessary [82]. Energy-Based Models (EBMs) have been proposed as a promising class for continual learning, reducing interference with previously learned information [83]. Routing networks with co-training enable sparse activation of experts for different tasks, minimizing interference while allowing positive transfer [84]. StackNet offers a method to learn additional tasks by stacking parameters, ensuring no degradation in previous task performance [85]. For MLLMs, continual learning can mitigate the performance decrease observed on natural language tasks when integrating vision models. [86] investigates continual learning methods to enhance visual understanding in MLLMs while minimizing linguistic performance loss. Dense Knowledge Distillation (DKD) uses a task pool to track model capabilities and distill cumulative knowledge from all previous tasks, outperforming state-of-the-art baselines [87]. Task Agnostic Continual Learning using Multiple Experts (TAME) detects shifts in data distributions and switches between expert networks online, even outperforming methods that assume task identities [88]. These advancements are vital for ensuring the robustness and longevity of LLM-enhanced visual QA systems in real-world applications.

IV. Datasets, Benchmarks, and Evaluation Metrics

The rapid progress in LLM-enhanced Image and Video Question Answering necessitates robust datasets, comprehensive benchmarks, and refined evaluation metrics. This section reviews the current landscape, highlighting the evolution from simple factoid QA to complex reasoning tasks, and the challenges in assessing multimodal understanding.

A. Image Question Answering Datasets

Traditional IQA datasets often focused on direct visual content. With the advent of LLMs, datasets that require more complex reasoning, external knowledge, or conversational context have become crucial.

- **OK-VQA**: This dataset is designed for Knowledge-based VQA, requiring models to answer questions by combining visual information with external world knowledge [55]. It serves as a benchmark for approaches leveraging knowledge graphs [55] or visual retriever-reader pipelines [89]. MAGMA [47] also reports state-of-the-art results on OKVQA.
- **EntityQuestions**: This dataset highlights a challenge for dense retrievers with simple, entity-rich questions based on Wikidata facts, revealing that dense models can struggle to generalize to uncommon entities [19].
- **ChartQA**: A benchmark specifically for question answering about charts, requiring both visual and logical reasoning over data presented in graphical forms [90].
- **Visual News**: A large-scale benchmark for news image captioning, emphasizing the importance of events and entities in news imagery. While for captioning, it represents a rich multimodal source for VQA about real-world events [91].

B. Video Question Answering Datasets

VQA introduces temporal reasoning and dynamic content, leading to specialized datasets.

- **Perception Test**: A diagnostic benchmark that evaluates the perception and reasoning skills of pre-trained multimodal video models across various skills (Memory, Abstraction, Physics, Semantics) and reasoning types (descriptive, explanatory, predictive, counterfactual) [92].
- **MM-Ego**: A large-scale egocentric QA dataset (7M samples) for egocentric video understanding, evaluating models' ability to recognize and memorize visual details across varying video lengths [93].
- **STRIDE-QA**: Designed for spatiotemporal reasoning in urban driving scenes, this dataset contains 16 million QA pairs over driving data, supporting object-centric and egocentric reasoning tasks [94].
- **MAQA**: A multimodal QA benchmark specifically designed to evaluate negation reasoning in MLLMs, adapting labeled music videos from AudioSet [95].
- **HeurVidQA**: A framework leveraging domain-specific entity-action heuristics to refine video-language foundation models for improved domain-specific VideoQA [96].
- **General Video QA Surveys**: Reviews like [13] provide comprehensive overviews of various methods and datasets for VQA.

C. Reasoning and Conversational QA Datasets

To evaluate LLMs' advanced capabilities, datasets demanding complex reasoning or interactive elements are vital.

- **QASPER**: A dataset of information-seeking questions and answers anchored in research papers, requiring complex reasoning about claims across multiple parts of a paper [97].
- **NuMGLUE**: A suite of fundamental yet challenging mathematical reasoning tasks that probe numerical understanding and calculation [98].
- **SQuAD, TriviaQA, Natural Questions**: Standard open-domain QA datasets (primarily text-based) used to benchmark retriever-reader systems [18], [20]. Variations like AdversarialQA [99] focus on model robustness.

Fig. 3. Overview of datasets and evaluation dimensions for LLM-enhanced image and video question answering, spanning visual perception, multimodal reasoning, and trustworthy assessment.

- **Conversational QA (CQA) datasets**: QReCC [15] is a dataset for Question Rewriting in Conversational Context, aiming to find answers to conversational questions across web pages. SIMMC 2.0 [100] provides task-oriented dialogs grounded in immersive multimodal scenes for situated and interactive conversations. Q^2 [101] evaluates factual consistency in knowledge-grounded dialogues via question generation and question answering.
- **Logical Reasoning Benchmarks**: FOLIO, LogicalDeduction, AR-LSAT, and ProofWriter are used to evaluate logical problem-solving abilities of LLMs, often in conjunction with symbolic solvers [37].

D. Evaluation Metrics and Frameworks

Evaluating LLM-enhanced visual QA is multifaceted, extending beyond traditional accuracy scores to encompass aspects like factual consistency, reasoning fidelity, and human preference.

- **Traditional Metrics**: Accuracy, F1-score, and Exact Match (EM) remain standard for extractive QA tasks [20], [102]. For generative tasks, linguistic metrics like BLEU, ROUGE, and METEOR are common, but their correlation with human judgment can be limited [102], [103]. CLIPScore [104] is a reference-free metric for image captioning that leverages vision-language embeddings.
- **Factuality and Hallucination Evaluation**: Given LLMs' propensity for hallucination, specific metrics and frameworks are crucial. GO FIGURE [103] is a meta-evaluation framework for factuality metrics in summarization, proposing conditions to evaluate metrics on diagnostic data. FactScore [105] offers fine-grained atomic evaluation of factual precision in long-form text generation. Studies like [106] systematically investigate object hallucination in LVLMs and propose improved evaluation methods like POPE. Specialized datasets and methods are being developed for hallucination detection in medical text summarization [107].
- **LLM-as-a-Judge**: Leveraging LLMs themselves to evaluate generated text has shown promising results. [108] demonstrates that LLM evaluation is consistent with human expert evaluation for tasks like story generation. [102] reassesses the performance of extractive QA models using LLM-as-a-judge, finding higher correlation with human judgments than traditional EM/F1. For conversational recommendation, iEvaLM [109] uses LLM-based user simulators for interactive evaluation.
- **Multi-Dimensional Evaluation**: UniEval [110] proposes a unified multi-dimensional evaluator for text generation by re-framing evaluation as a Boolean Question Answering task, correlating better with human judgments across various NLG tasks.
- **Robustness and Bias Evaluation**: Evaluating robustness involves assessing model performance under perturbed inputs or out-of-distribution data [111]. Measures of social biases are extended to grounded vision and language embeddings, identifying that biases can be equally or more significant than for ungrounded embeddings [112], [113].
- **Domain-Specific Evaluation**: For fields like biomedicine, specific QA benchmarks like BioASQ [114] are critical, along with specialized LLMs and evaluation methods for tasks like vulnerability reasoning in source code [115] or energy modeling [116]. Human-centric benchmarks like AGIEval [117] and the MULTIMODAL UNIVERSE [118] for astronomical data are also emerging.

The ongoing development of these diverse datasets and metrics is crucial for accurately measuring the progress and identifying the shortcomings of LLM-enhanced visual QA systems, pushing research towards more capable and trustworthy AI.

V. Challenges, Open Problems, and Future Directions

Despite the remarkable advancements driven by Large Language Models (LLMs) in visual question answering, several significant challenges and open problems remain. Addressing these issues will be crucial for the continued progress and practical deployment of LLM-enhanced IQA and VQA systems.

A. Hallucinations and Factual Inconsistencies

One of the most pressing challenges for LLMs, particularly in multimodal contexts, is the phenomenon of hallucination, where models generate factually incorrect or ungrounded information [8], [106]. In visual QA, this manifests as generated descriptions or answers that are inconsistent with the visual content [106].

- **Mitigation Strategies**: Current research explores various mitigation techniques. Self-reflection and consistency-based approaches, as discussed in [38], [119], aim to detect and correct hallucinations. ViHallu [120] proposes a vision-centric framework using visual variation image generation and visual instruction construction to enhance visual-semantic alignment and reduce hallucinations. Knowledge Distillation (KD) with smoothed knowledge can also mitigate hallucinations by reducing overconfidence [121]. However, studies like [61] delve deeper into whether LLMs can "lie" intentionally, a concern beyond mere hallucination.
- **Rethinking Reasoning**: [122] proposes the Visual Inference Chain (VIC) framework, where reasoning chains are constructed using textual context alone *before* introducing visual input, to mitigate hallucinations caused by misleading images. This suggests a re-evaluation of the "thinking while looking" paradigm.
- **Domain-Specific Hallucinations**: Hallucinations can be particularly critical in sensitive domains like medicine. [107] conducts an evaluation of hallucination detection methods in medical text summarization, highlighting that general-domain detectors struggle with clinical hallucinations and proposing fact-based approaches for diagnosis.

B. Computational Expense and Efficiency

Large-scale MLLMs are computationally intensive, requiring significant resources for training and inference.

- **Efficient Architectures and Retrieval**: Techniques like Binary Passage Retriever (BPR) [18] reduce memory costs by representing passage indexes with compact binary codes. SPARTA [21] learns sparse representations for efficient neural retrieval.
- **Parameter-Efficient Fine-Tuning (PEFT)**: Methods such as prompt tuning, BitFit, and LLM-Adapters [31], [32], [34] allow for adapting LLMs to downstream tasks with minimal trainable parameters, making them more accessible. However, their effectiveness for embedding specific facts is still being reassessed [123].
- **Input Compression**: [52] explores rendering text as images to achieve token savings, demonstrating a novel way to compress textual inputs for MLLMs. LLMLingua [35] compresses prompts to accelerate inference.
- **Scaling Reasoning**: While reasoning models are powerful, their optimal role might be as discriminators rather than generators, as they can achieve higher F1 and discrimination accuracy with fewer parameters compared to larger non-reasoning LLMs in planning frameworks [39].

C. Data Scarcity and Low-Resource Scenarios

Despite the abundance of web data, high-quality, diverse, and well-annotated multimodal datasets remain scarce for many specific tasks and low-resource languages.

- **Data Augmentation and Generation**: Synthetic data generation can improve model robustness to adversarial attacks [99] or generate answer candidates for quizzes [124]. Automatically deriving VQA examples from image-caption annotations can also generate high-quality data at volume [125].
- **Low-Resource NLP**: Surveys on low-resource NLP [126] highlight transfer learning and data augmentation as promising strategies. Few-shot learning, often combined with pre-training span selection [20] or prompt tuning [31], helps in scenarios with limited training examples.
- **Multilingual and Cross-Lingual QA**: Addressing information scarcity and asymmetry in non-English languages requires cross-lingual open-retrieval QA (XOR QA) [17]. Research on multilingual encoders with explicit alignment objectives also facilitates zero-shot cross-lingual transfer [28]. The WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages focuses on MT and QA for low-resource languages [127].
- **Scientific Corpus Distillation**: For biomedical LLMs, knowledge-driven agentic scientific corpus distillation frameworks like m-KAILIN [128] address insufficient quantity and quality in open-source annotated scientific corpora by generating and refining domain-specific QA pairs.

D. Real-time Processing and Dynamic Knowledge

Many real-world applications require instantaneous responses and access to up-to-date information, posing challenges for static LLMs.

- **Real-time QA Platforms**: REALTIME QA [129] is a platform that evaluates systems on current world events, highlighting the need for up-to-date retrieval and the ability to identify unanswerable cases. Real-time state monitoring systems for industrial applications also demonstrate the need for immediate data processing [130].

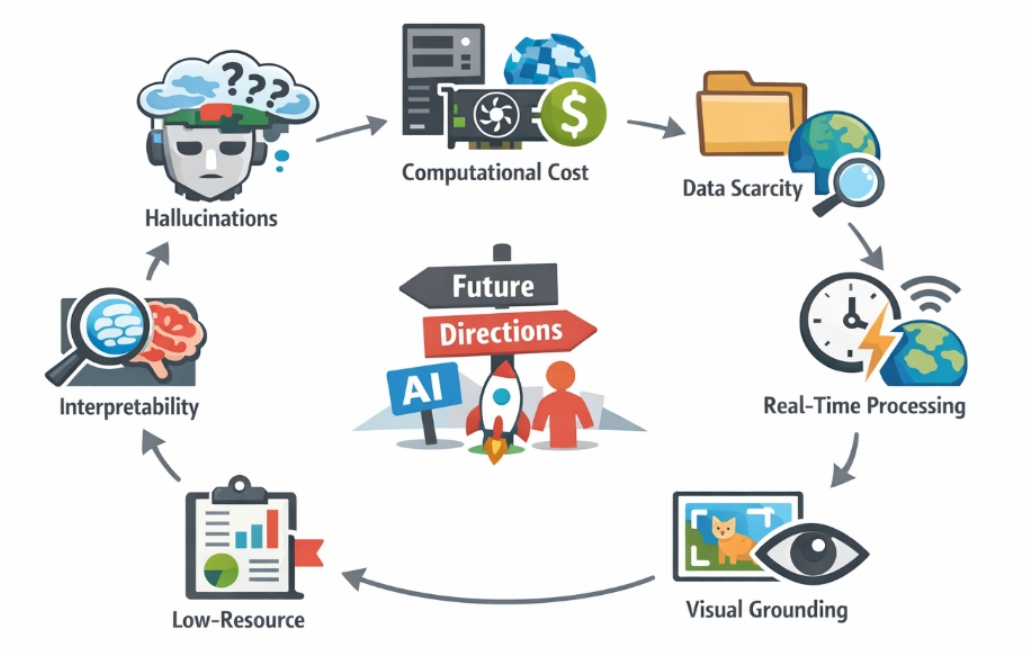

Fig. 4. Overview of key challenges and future directions for LLM-enhanced multimodal visual question answering, highlighting hallucination, efficiency, data scarcity, real-time processing, visual grounding, and interpretability in a unified framework.

- **Context-Aware Interventions**: Navigating the state of cognitive flow in AI-augmented reasoning necessitates context-aware AI interventions that adapt based on type, timing, and scale to maintain or restore flow [131].

E. Visual Grounding and Interpretability

For MLLMs to be trustworthy, they must not only provide correct answers but also explain their reasoning and accurately ground their understanding in the visual input.

- **Improved Grounding Mechanisms**: While visual grounding has significantly advanced, challenges remain in generalizing beyond detection [72]. Approaches like VLM-Grounder [70] and DEGround [71] are making strides in zero-shot and embodied 3D visual grounding.
- **Interpretable-by-Design Models**: DCLUB [132] introduces the Dynamic Clue Bottleneck Model, designed to factor model decisions into intermediate human-legible visual clues before generating an answer, providing inherent interpretability.
- **Addressing Misconceived Reasons**: Earlier work showed that VQA models often work for the "wrong reasons," exploiting dataset biases rather than true visual grounding [12]. Future work needs to ensure that MLLM performance gains stem from genuine multimodal understanding. The use of generative imagination can also elevate machine translation performance [133].

F. Future Directions

Based on the current challenges, several promising future directions emerge:

- **More Robust and Efficient Multimodal Architectures**: Developing architectures that can handle diverse modalities (e.g., speech, non-verbal cues beyond current vision/text) and integrate them more seamlessly. Approaches like MTAG [134] use graph-based models for unaligned human multimodal language sequences, offering interpretable frameworks. Research on multimodal LLM logical reasoning [59] and video models with domain-specific fine-grained heuristics [96] will further refine understanding. The goal of "less is more" in vision representation compression for efficient video generation is also relevant [135].
- **Advanced In-Context Learning for MLLMs**: Leveraging visual in-context learning can significantly enhance the adaptability, reasoning capabilities, and sample efficiency of Large Vision-Language Models, allowing them to learn new tasks and concepts directly from examples within the prompt [136].
- **Continual Learning and Adaptation for MLLMs**: Further research into effectively integrating new knowledge and adapting to new domains without catastrophic forgetting, as explored in [82], [86], [87]. This includes approaches for routing networks [84] and task-agnostic learning [88].
- **Embodied AI and Human-AI Interaction**: Exploring how MLLMs can be integrated into embodied agents for more intuitive and effective interaction with the physical world [73]. This also extends to interactive learning for LLM reasoning, where multi-agent systems enhance independent problem-solving abilities [137]. LLM-assisted visual analytics is another emerging field that will trans-

form capabilities through natural language interactions [138].

- **World Model Integration**: Moving towards more comprehensive world models that can simulate and predict dynamic environments, possibly using LLM-grounded video diffusion models [139]. This necessitates advanced capabilities in event correlation, script reasoning, and context-to-event understanding, often supported by pre-trained models and external knowledge graphs [140]–[142]. This includes developing medical LLMs with abnormal-aware feedback [143] and modular multi-agent frameworks for medical diagnosis [144].
- **Ethical AI and Bias Mitigation**: Continuously evaluating and mitigating biases in multimodal representations [112] and ensuring fairness [113] and trustworthiness in MLLM applications, especially regarding the potential for deception [61], [145].
- **Rethinking Visual Dependencies in Long-Context Reasoning**: Addressing how to efficiently and effectively process long visual contexts for reasoning tasks in MLLMs [146]. This includes developing robust reasoning frameworks capable of unraveling chaotic or complex long contexts [147] and fine-grained distillation for long document retrieval [148].
- **Specialized MLLMs**: Developing highly specialized MLLMs for niche scientific domains (e.g., astronomy [118], biomedical research [128], [149]) and engineering applications (e.g., insect recognition [150], defect recognition [151]).
- **Generative Multimodal Editing**: Expanding multimodal capabilities beyond understanding to include generation and editing of visual content guided by LLMs [152]–[155].
- **Bridging Speech and Text**: Enhancing ASR with Pinyin-to-Character pre-training in LLMs can open new avenues for spoken language interaction in visual QA [156].
- **Generalization Across Capabilities**: Further understanding "weak to strong generalization" for LLMs with multi-capabilities, as discussed in [157], will be key for models to flexibly handle diverse tasks.

These directions collectively point towards the development of more intelligent, adaptable, and trustworthy LLM-enhanced visual QA systems that can seamlessly integrate into various real-world applications.

VI. Conclusion

This survey has provided a comprehensive review of the transformative impact of Large Language Models (LLMs) on the fields of Image Question Answering (IQA) and Video Question Answering (VQA). We began by outlining the historical development of visual QA, from early feature-based and attention-driven models to the recognition of their limitations in complex reasoning and open-ended generation. We then delved into the foundational aspects of LLMs, highlighting their Transformer architecture, diverse pre-training objectives, and emergent capabilities in language understanding, generation, and intricate reasoning.

The core of our review categorized the architectural paradigms and methodologies that bridge LLMs with visual modalities. We discussed the rise of Multimodal Large Language Models (MLLMs) through Vision-Language Pre-training (VLP), encompassing unified-modal architectures like UNIMO [44] and BriVL [45], adapter-based approaches such as MAGMA [47] that preserve LLM integrity, and vision-guided generative models that inject visual information into GPLMs [49]–[51]. A significant emphasis was placed on Knowledge-Augmented and Retrieval-Augmented Generation (RAG), detailing multimodal RAG architectures [53], [54], the integration of explicit knowledge graphs [55], and dynamic/hierarchical retrieval strategies like QA-Dragon [56]. Furthermore, we explored advanced reasoning paradigms, including compositional reasoning [6], self-reflection and verification [5], [38], and causal/analogical reasoning [62], [63]. The growing importance of end-to-end multimodal learning and embodied AI, particularly in 3D visual grounding [70] and autonomous driving [67], was also highlighted, along with the critical aspect of continual learning for adaptability in dynamic environments [86].

A dedicated section scrutinized the evolving landscape of datasets, benchmarks, and evaluation metrics, emphasizing the shift towards tasks requiring complex reasoning (e.g., OK-VQA [55]), temporal understanding (e.g., Perception Test [92]), and conversational context (e.g., SIMMC 2.0 [100]). We discussed the emergence of LLM-as-a-Judge frameworks [102] and multi-dimensional evaluators [110] for more robust and human-aligned assessment.

Finally, we critically analyzed the prevalent challenges facing LLM-enhanced visual QA, including the persistent problem of hallucinations [106], [120], computational expense [35], data scarcity in low-resource settings [126], the demand for real-time processing [129], and the crucial need for improved visual grounding and interpretability [132]. We concluded by outlining promising future research directions, from developing more robust multimodal architectures and advancing continual learning capabilities to exploring embodied AI interactions and addressing ethical implications.

In summary, the integration of LLMs has profoundly reshaped the landscape of image and video question answering, offering unprecedented opportunities for building intelligent systems that can truly perceive, reason about, and interact with the visual world through natural language. While significant progress has been made, the journey towards fully capable, trustworthy, and efficient LLM-enhanced visual QA systems is ongoing, promising continued innovation and interdisciplinary research at the forefront of artificial intelligence.

References

[1] J. Singh, V. Ying, and A. Nutkiewicz, "Attention on attention: Architectures for visual question answering (VQA)," *CoRR*, vol. abs/1803.07724, 2018. [Online]. Available: http://arxiv.org/abs/1803.07724

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6077–6086. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[3] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," *CoRR*, vol. abs/1708.02711, 2017. [Online]. Available: http://arxiv.org/abs/1708.02711

[4] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1049–1065. [Online]. Available: https://aclanthology.org/2023.findings-acl.67/

[5] J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han, "Large language models can self-improve," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 1051–1068. [Online]. Available: https://aclanthology.org/2023.emnlp-main.67/

[6] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, and M. Lewis, "Measuring and narrowing the compositionality gap in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 5687–5711. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.378/

[7] F. Ke, J. Hsu, Z. Cai, Z. Ma, X. Zheng, X. Wu, S. Huang, W. Wang, P. D. Haghighi, G. Haffari, R. Krishna, J. Wu, and H. Rezatofighi, "Explain before you answer: A survey on compositional visual reasoning," *CoRR*, vol. abs/2508.17298, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2508.17298

[8] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Can knowledge graphs reduce hallucinations in LLMs? : A survey," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 3947–3960. [Online]. Available: https://aclanthology.org/2024.naacl-long.219/

[9] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *CoRR*, vol. abs/2306.08302, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.08302

[10] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren, "A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2763–2775. [Online]. Available: https://aclanthology.org/2022.acl-long.197/

[11] R. Cadène, C. Dancette, H. Ben-Younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases in visual question answering," *CoRR*, vol. abs/1906.10169, 2019. [Online]. Available: http://arxiv.org/abs/1906.10169

[12] R. Shrestha, K. Kafle, and C. Kanan, "Visual grounding methods for vqa are working for the wrong reasons!" *arXiv preprint arXiv:2004.05704v4*, 2020. [Online]. Available: http://arxiv.org/abs/2004.05704v4

[13] D. Patel, R. Parikh, and Y. Shastri, "Recent advances in video question answering: A review of datasets and methods," *CoRR*, vol. abs/2101.05954, 2021. [Online]. Available: https://arxiv.org/abs/2101.05954

[14] C. Wang, P. Liu, and Y. Zhang, "Can generative pre-trained language models serve as knowledge bases for closed-book QA?" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 3241–3251. [Online]. Available: https://aclanthology.org/2021.acl-long.251/

[15] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi, "Open-domain question answering goes conversational via question rewriting," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 520–534. [Online]. Available: https://aclanthology.org/2021.naacl-main.44/

[16] C. You, N. Chen, and Y. Zou, "Self-supervised contrastive cross-modality representation learning for spoken question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 28–39. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.3/

[17] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 547–564. [Online]. Available: https://aclanthology.org/2021.naacl-main.46/

[18] I. Yamada, A. Asai, and H. Hajishirzi, "Efficient passage retrieval with hashing for open-domain question answering," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, 2021, pp. 979–986. [Online]. Available: https://aclanthology.org/2021.acl-short.123/

[19] C. Sciavolino, Z. Zhong, J. Lee, and D. Chen, "Simple entity-centric questions challenge dense retrievers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 6138–6148. [Online]. Available: https://aclanthology.org/2021.emnlp-main.496/

[20] O. Ram, Y. Kirstain, J. Berant, A. Globerson, and O. Levy, "Few-shot question answering by pretraining span selection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 3066–3079. [Online]. Available: https://aclanthology.org/2021.acl-long.239/

[21] T. Zhao, X. Lu, and K. Lee, "SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 565–575. [Online]. Available: https://aclanthology.org/2021.naacl-main.47/

[22] Y. Xu, C. Zhu, R. Xu, Y. Liu, M. Zeng, and X. Huang, "Fusing context into knowledge graph for commonsense question answering," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 1201–1207. [Online]. Available: https://aclanthology.org/2021.findings-acl.102/

[23] Y. Sun, Q. Shi, L. Qi, and Y. Zhang, "JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 5049–5060. [Online]. Available: https://aclanthology.org/2022.naacl-main.372/

[24] P. Kapanipathi, I. Abdelaziz, S. Ravishankar, S. Roukos, A. Gray, R. Fernandez Astudillo, M. Chang, C. Cornelio, S. Dana, A. Fokoue, D. Garg, A. Gliozzo, S. Gurajada, H. Karanam, N. Khan, D. Khandelwal, Y.-S. Lee, Y. Li, F. Luus, N. Makondo, N. Mihindukulasooriya, T. Naseem, S. Neelam, L. Popa, R. Gangi Reddy, R. Riegel, G. Rossiello, U. Sharma, G. P. S. Bhargav, and M. Yu, "Leveraging Abstract Meaning Representation for knowledge base question answering," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 3884–3894. [Online]. Available: https://aclanthology.org/2021.findings-acl.339/

[25] Z. Li, S. Fan, Y. Gu, X. Li, Z. Duan, B. Dong, N. Liu, and J. Wang, "Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024,*

Vancouver, Canada, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 18 608–18 616. [Online]. Available: https://doi.org/10.1609/aaai.v38i17.29823

[26] B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, and S. Yih, "UniK-QA: Unified representations of structured and unstructured knowledge for open-domain question answering," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1535–1546. [Online]. Available: https://aclanthology.org/2022.findings-naacl.115/

[27] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: pre-training of generic visual-linguistic representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SygXPaEYvH

[28] J. Hu, M. Johnson, O. Firat, A. Siddhant, and G. Neubig, "Explicit alignment objectives for multilingual bidirectional encoders," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 3633–3643. [Online]. Available: https://aclanthology.org/2021.naacl-main.284/

[29] Z. He, G. Blackwood, R. Panda, J. J. McAuley, and R. Feris, "Synthetic pre-training tasks for neural machine translation," in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 8080–8098. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-acl.512

[30] R. Zhong, K. Lee, Z. Zhang, and D. Klein, "Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2856–2878. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.244/

[31] Y. Gu, X. Han, Z. Liu, and M. Huang, "PPT: Pre-trained prompt tuning for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8410–8423. [Online]. Available: https://aclanthology.org/2022.acl-long.576/

[32] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 1–9. [Online]. Available: https://aclanthology.org/2022.acl-short.1/

[33] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, and M. Sun, "Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2225–2240. [Online]. Available: https://aclanthology.org/2022.acl-long.158/

[34] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. Lee, "LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 5254–5276. [Online]. Available: https://aclanthology.org/2023.emnlp-main.319/

[35] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "LLMLingua: Compressing prompts for accelerated inference of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 13 358–13 376. [Online]. Available: https://aclanthology.org/2023.emnlp-main.825/

[36] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao, "Large language models are better reasoners with self-verification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 2550–2575. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.167/

[37] L. Pan, A. Albalak, X. Wang, and W. Wang, "Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 3806–3824. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.248/

[38] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 1827–1843. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.123/

[39] M. F. Anjum, "When reasoning beats scale: A 1.5b reasoning model outranks 13b llms as discriminator," *arXiv preprint arXiv:2505.03786v1*, 2025. [Online]. Available: http://arxiv.org/abs/2505.03786v1

[40] H. Zhang, X. Liu, and J. Zhang, "Extractive summarization via ChatGPT for faithful summary generation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 3270–3278. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.214/

[41] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, and S. Sitaram, "MEGA: Multilingual evaluation of generative AI," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 4232–4267. [Online]. Available: https://aclanthology.org/2023.emnlp-main.258/

[42] S. Xu, Y. Cao, Z. Wang, and Y. Tian, "Fraud detection in online transactions: Toward hybrid supervised–unsupervised learning pipelines," in *Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI 2025), Chengdu, China*, 2025, pp. 20–22.

[43] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu, "VLP: A survey on vision-language pre-training," *Int. J. Autom. Comput.*, vol. 20, no. 1, pp. 38–56, 2023. [Online]. Available: https://doi.org/10.1007/s11633-022-1369-5

[44] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 2592–2607. [Online]. Available: https://aclanthology.org/2021.acl-long.202/

[45] Y. Huo, M. Zhang, G. Liu, H. Lu, Y. Gao, G. Yang, J. Wen, H. Zhang, B. Xu, W. Zheng, Z. Xi, Y. Yang, A. Hu, J. Zhao, R. Li, Y. Zhao, L. Zhang, Y. Song, X. Hong, W. Cui, D. Hou, Y. Li, J. Li, P. Liu, Z. Gong, C. Jin, Y. Sun, S. Chen, Z. Lu, Z. Dou, Q. Jin, Y. Lan, W. X. Zhao, R. Song, and J.-R. Wen, "Wenlan: Bridging vision and language by large-scale multi-modal pre-training," *arXiv preprint arXiv:2103.06561v6*, 2021. [Online]. Available: http://arxiv.org/abs/2103.06561v6

[46] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si, "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 7241–7259. [Online]. Available: https://aclanthology.org/2022.emnlp-main.488/

[47] C. Eichenberg, S. Black, S. Weinbach, L. Parcalabescu, and A. Frank, "MAGMA – multimodal augmentation of generative models through adapter-based finetuning," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 2416–2428. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.179/

[48] H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang, "E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021,

pp. 503–513. [Online]. Available: https://aclanthology.org/2021.acl-long.42/

[49] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 3995–4007. [Online]. Available: https://aclanthology.org/2021.emnlp-main.326/

[50] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Singapore: Association for Computational Linguistics, 2023, pp. 543–553. [Online]. Available: https://aclanthology.org/2023.emnlp-demo.49/

[51] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 12 585–12 602. [Online]. Available: https://aclanthology.org/2024.acl-long.679/

[52] Y. Li, Z. Lan, and J. Zhou, "Text or pixels? it takes half: On the token efficiency of visual text inputs in multimodal llms," *CoRR*, vol. abs/2510.18279, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2510.18279

[53] W. Chen, H. Hu, X. Chen, P. Verga, and W. Cohen, "MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 5558–5570. [Online]. Available: https://aclanthology.org/2022.emnlp-main.375/

[54] M. M. Abootorabi, A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. S. Baghshah, and E. Asgari, "Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation," *arXiv preprint arXiv:2502.08826v3*, 2025. [Online]. Available: http://arxiv.org/abs/2502.08826v3

[55] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "KAT: A knowledge augmented transformer for vision-and-language," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 956–968. [Online]. Available: https://aclanthology.org/2022.naacl-main.70/

[56] Z. Jiang, P. Wu, X. Yuan, W. Fan, and Q. Li, "Qa-dragon: Query-aware dynamic rag system for knowledge-intensive visual question answering," *arXiv preprint arXiv:2508.05197v1*, 2025. [Online]. Available: http://arxiv.org/abs/2508.05197v1

[57] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," *arXiv preprint arXiv:2404.15406v2*, 2024. [Online]. Available: http://arxiv.org/abs/2404.15406v2

[58] H. Yuan, Y. Sun, Y. Li, T. Zhang, X. Deng, H. Ding, L. Qi, A. Wang, X. Li, and M.-H. Yang, "Visual reasoning tracer: Object-level grounded reasoning benchmark," *arXiv preprint arXiv:2512.05091v1*, 2025. [Online]. Available: http://arxiv.org/abs/2512.05091v1

[59] Y. Xiao, E. Sun, T. Liu, and W. Wang, "Logicvista: Multimodal LLM logical reasoning benchmark in visual contexts," *CoRR*, vol. abs/2407.04973, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.04973

[60] A. Azaria and T. Mitchell, "The internal state of an LLM knows when it's lying," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 967–976. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.68/

[61] H. Huan, M. Prabhudesai, M. Wu, S. Jaiswal, and D. Pathak, "Can llms lie? investigation beyond hallucination," *CoRR*, vol. abs/2509.03518, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.03518

[62] Y. Liu, W. Chen, G. Li, and L. Lin, "Causalvlr: A toolbox and benchmark for visual-linguistic causal reasoning," *CoRR*, vol. abs/2306.17462, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.17462

[63] S. Musker, A. Duchnowski, R. Millière, and E. Pavlick, "Llms as models for analogical reasoning," *arXiv preprint arXiv:2406.13803v3*, 2024. [Online]. Available: http://arxiv.org/abs/2406.13803v3

[64] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744v1*, 2016. [Online]. Available: http://arxiv.org/abs/1612.01744v1

[65] C. Hori and T. Hori, "End-to-end conversation modeling track in DSTC6," *CoRR*, vol. abs/1706.07440, 2017. [Online]. Available: http://arxiv.org/abs/1706.07440

[66] C. T. Lengerich and A. Y. Hannun, "An end-to-end architecture for keyword spotting and voice activity detection," *CoRR*, vol. abs/1611.09405, 2016. [Online]. Available: http://arxiv.org/abs/1611.09405

[67] Z. Guo, Y. Luo, L. Sha, D. Wang, P. Wang, C. Xu, and Y. Yang, "2nd place solution for CVPR2024 E2E challenge: End-to-end autonomous driving using vision language model," *CoRR*, vol. abs/2509.02659, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.02659

[68] Z. Li, W. Yao, Z. Wang, X. Sun, J. Chen, N. Chang, M. Shen, Z. Wu, S. Lan, and J. M. Álvarez, "Generalized trajectory scoring for end-to-end multimodal planning," *CoRR*, vol. abs/2506.06664, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2506.06664

[69] Z. Li, K. Li, S. Wang, S. Lan, Z. Yu, Y. Ji, Z. Li, Z. Zhu, J. Kautz, Z. Wu, Y.-G. Jiang, and J. M. Alvarez, "Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation," *arXiv preprint arXiv:2406.06978v4*, 2024. [Online]. Available: http://arxiv.org/abs/2406.06978v4

[70] R. Xu, Z. Huang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Vlm-grounder: A VLM agent for zero-shot 3d visual grounding," in *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 2024, pp. 3961–3985. [Online]. Available: https://proceedings.mlr.press/v270/xu25c.html

[71] Y. Zhang, D. Wu, H. Shi, Y. Liu, T. Wang, H. Fan, and X. Dong, "Grounding beyond detection: Enhancing contextual understanding in embodied 3d grounding," *arXiv preprint arXiv:2506.05199v2*, 2025. [Online]. Available: http://arxiv.org/abs/2506.05199v2

[72] L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu, "Towards visual grounding: A survey," *CoRR*, vol. abs/2412.20206, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.20206

[73] S. Incao, C. Mazzola, G. Belgiovine, and A. Sciutti, "A roadmap for embodied and social grounding in llms," *arXiv preprint arXiv:2409.16900v1*, 2024. [Online]. Available: http://arxiv.org/abs/2409.16900v1

[74] A. Lotfi, E. Fini, S. Bengio, M. Nabi, and E. Abbe, "Chain-of-sketch: Enabling global visual reasoning," *arXiv preprint arXiv:2410.08165v2*, 2024. [Online]. Available: http://arxiv.org/abs/2410.08165v2

[75] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 5666–5675. [Online]. Available: https://aclanthology.org/2021.acl-long.440/

[76] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 5301–5311. [Online]. Available: https://aclanthology.org/2021.acl-long.412/

[77] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 7617–7630. [Online]. Available: https://aclanthology.org/2023.acl-long.421/

[78] J. Cheng, I. Fostiropoulos, B. Boehm, and M. Soleymani, "Multimodal phased transformer for sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for

Computational Linguistics, 2021, pp. 2447–2458. [Online]. Available: https://aclanthology.org/2021.emnlp-main.189/

[79] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 7837–7851. [Online]. Available: https://aclanthology.org/2022.emnlp-main.534/

[80] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 756–767. [Online]. Available: https://aclanthology.org/2023.emnlp-main.49/

[81] L. Mathur and M. J. Mataric, "Introducing representations of facial affect in automated multimodal deception detection," in *ICMI '20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, K. P. Truong, D. Heylen, M. Czerwinski, N. Berthouze, M. Chetouani, and M. Nakano, Eds. ACM, 2020, pp. 305–314. [Online]. Available: https://doi.org/10.1145/3382507.3418864

[82] L. Peng, J. Elenter, J. Agterberg, A. Ribeiro, and R. Vidal, "Loranpac: Low-rank random features and pre-trained models for bridging theory and practice in continual learning," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: https://openreview.net/forum?id=bqv7M0wc4x

[83] V. Singh, A. Choromanska, S. Li, and Y. Du, "Wake-sleep energy based models for continual learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*. IEEE, 2024, pp. 4118–4127. [Online]. Available: https://doi.org/10.1109/CVPRW63382.2024.00415

[84] M. Collier, E. Kokiopoulou, A. Gesmundo, and J. Berent, "Routing networks with co-training for continual learning," *CoRR*, vol. abs/2009.04381, 2020. [Online]. Available: https://arxiv.org/abs/2009.04381

[85] J. Kim, J. Kim, and N. Kwak, "Stacknet: Stacking parameters for continual learning," *arXiv preprint arXiv:1809.02441v3*, 2018. [Online]. Available: http://arxiv.org/abs/1809.02441v3

[86] S. Srivastava, M. Y. Harun, R. Shrestha, and C. Kanan, "Improving multimodal large language models using continual learning," *arXiv preprint arXiv:2410.19925v2*, 2024. [Online]. Available: http://arxiv.org/abs/2410.19925v2

[87] Z. Shi, P. Liu, T. Su, Y. Wu, K. Liu, Y. Song, and M. Wang, "Densely distilling cumulative knowledge for continual learning," *arXiv preprint arXiv:2405.09820v1*, 2024. [Online]. Available: http://arxiv.org/abs/2405.09820v1

[88] H. Zhu, M. Majzoubi, A. Jain, and A. Choromanska, "TAME: task agnostic continual learning using multiple experts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*. IEEE, 2024, pp. 4139–4148. [Online]. Available: https://doi.org/10.1109/CVPRW63382.2024.00417

[89] M. Luo, Y. Zeng, P. Banerjee, and C. Baral, "Weakly-supervised visual-retriever-reader for knowledge-based question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 6417–6431. [Online]. Available: https://aclanthology.org/2021.emnlp-main.517/

[90] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2263–2279. [Online]. Available: https://aclanthology.org/2022.findings-acl.177/

[91] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual news: Benchmark and challenges in news image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 6761–6771. [Online]. Available: https://aclanthology.org/2021.emnlp-main.542/

[92] V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, J. Heyward, M. Malinowski, Y. Yang, C. Doersch, T. Matejovicova, Y. Sulsky, A. Miech, A. Fréchette, H. Klimczak, R. Koster, J. Zhang, S. Winkler, Y. Aytar, S. Osindero, D. Damen, A. Zisserman, and J. Carreira, "Perception test: A diagnostic benchmark for multimodal video models," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract-Datasets_and_Benchmarks.html

[93] H. Ye, H. Zhang, E. Daxberger, L. Chen, Z. Lin, Y. Li, B. Zhang, H. You, D. Xu, Z. Gan, J. Lu, and Y. Yang, "Mm-ego: Towards building egocentric multimodal llms for video qa," *arXiv preprint arXiv:2410.07177v2*, 2024. [Online]. Available: http://arxiv.org/abs/2410.07177v2

[94] K. Ishihara, K. Sasaki, T. Takahashi, D. Shiono, and Y. Yamaguchi, "STRIDE-QA: visual question answering dataset for spatiotemporal reasoning in urban driving scenes," *CoRR*, vol. abs/2508.10427, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2508.10427

[95] J. Y. Li, A. Jansen, Q. Huang, J. Lee, R. Ganti, and D. Kuzmin, "MAQA: A multimodal QA benchmark for negation," *CoRR*, vol. abs/2301.03238, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.03238

[96] T. Yu, K. Fu, S. Wang, Q. Huang, and J. Yu, "Prompting video-language foundation models with domain-specific fine-grained heuristics for video question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 2, pp. 1615–1630, 2025. [Online]. Available: https://doi.org/10.1109/TCSVT.2024.3475510

[97] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, and M. Gardner, "A dataset of information-seeking questions and answers anchored in research papers," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 4599–4610. [Online]. Available: https://aclanthology.org/2021.naacl-main.365/

[98] S. Mishra, A. Mitra, N. Varshney, B. Sachdeva, P. Clark, C. Baral, and A. Kalyan, "NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 3505–3523. [Online]. Available: https://aclanthology.org/2022.acl-long.246/

[99] M. Bartolo, T. Thrush, R. Jia, S. Riedel, P. Stenetorp, and D. Kiela, "Improving question answering model robustness with synthetic adversarial data generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 8830–8848. [Online]. Available: https://aclanthology.org/2021.emnlp-main.696/

[100] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4903–4912. [Online]. Available: https://aclanthology.org/2021.emnlp-main.401/

[101] O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, and O. Abend, "q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 7856–7870. [Online]. Available: https://aclanthology.org/2021.emnlp-main.619/

[102] X. Ho, J. Huang, F. Boudin, and A. Aizawa, "Llm-as-a-judge: Reassessing the performance of llms in extractive QA," *CoRR*, vol. abs/2504.11972, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2504.11972

[103] S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, and J. Gao, "GO FIGURE: A meta evaluation of factuality in summarization," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 478–487. [Online]. Available: https://aclanthology.org/2021.findings-acl.42/

[104] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in*

Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 7514–7528. [Online]. Available: https://aclanthology.org/2021.emnlp-main.595/

[105] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 12 076–12 100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/

[106] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 292–305. [Online]. Available: https://aclanthology.org/2023.emnlp-main.20/

[107] S. BN, H.-C. Shing, L. Xu, M. Strong, J. Burnsky, J. Ofor, J. R. Mason, S. Chen, S. Srinivasan, C. Shivade, J. Moriarty, and J. P. Cohen, "Fact-controlled diagnosis of hallucinations in medical text summarization," *arXiv preprint arXiv:2506.00448v1*, 2025. [Online]. Available: http://arxiv.org/abs/2506.00448v1

[108] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 15 607–15 631. [Online]. Available: https://aclanthology.org/2023.acl-long.870/

[109] X. Wang, X. Tang, X. Zhao, J. Wang, and J.-R. Wen, "Rethinking the evaluation for conversational recommendation in the era of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 10 052–10 065. [Online]. Available: https://aclanthology.org/2023.emnlp-main.621/

[110] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a unified multi-dimensional evaluator for text generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 2023–2038. [Online]. Available: https://aclanthology.org/2022.emnlp-main.131/

[111] X. Wang, H. Wang, and D. Yang, "Measure and improve robustness in NLP models: A survey," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 4569–4586. [Online]. Available: https://aclanthology.org/2022.naacl-main.339/

[112] C. Ross, B. Katz, and A. Barbu, "Measuring social biases in grounded vision and language embeddings," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 998–1008. [Online]. Available: https://aclanthology.org/2021.naacl-main.78/

[113] P. Delobelle, E. Tokpo, T. Calders, and B. Berendt, "Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1693–1706. [Online]. Available: https://aclanthology.org/2022.naacl-main.122/

[114] A. Shah, S. Singh, and S.-Y. Tao, "Feature extraction and evaluation for biomedical question answering," *arXiv preprint arXiv:2105.14013v1*, 2021. [Online]. Available: http://arxiv.org/abs/2105.14013v1

[115] A. Jararweh, M. Adams, A. Sahu, A. Mueen, and A. Anwar, "Llavul: A multimodal LLM for interpretable vulnerability reasoning about source code," *CoRR*, vol. abs/2509.17337, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.17337

[116] M. Takrouri, N. M. Cuadrado, and M. Takác, "Knowledge distillation from large language models for household energy modeling," *CoRR*, vol. abs/2502.03034, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.03034

[117] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, "AGIEval: A human-centric benchmark for evaluating foundation models," in *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 2299–2314. [Online]. Available: https://aclanthology.org/2024.findings-naacl.149/

[118] T. M. U. Collaboration, J. Audenaert, M. Bowles, B. M. Boyd, D. Chemaly, B. Cherinka, I. Ciucă, M. Cranmer, A. Do, M. Grayling, E. E. Hayes, T. Hehir, S. Ho, M. Huertas-Company, K. G. Iyer, M. Jablonska, F. Lanusse, H. W. Leung, K. Mandel, J. R. Martínez-Galarza, P. Melchior, L. Meyer, L. H. Parker, H. Qu, J. Shen, M. J. Smith, C. Stone, M. Walmsley, and J. F. Wu, "The multimodal universe: Enabling large-scale machine learning with 100tb of astronomical scientific data," *arXiv preprint arXiv:2412.02527v1*, 2024. [Online]. Available: http://arxiv.org/abs/2412.02527v1

[119] D. Till, J. Smeaton, P. Haubrick, G. Saheb, F. Graef, and D. Berman, "Teaming llms to detect and mitigate hallucinations," *CoRR*, vol. abs/2510.19507, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2510.19507

[120] Z. Dai, X. Li, S. Zhang, Y. Wu, and J. Li, "See different, think better: Visual variations mitigating hallucinations in lvlms," *CoRR*, vol. abs/2507.22003, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2507.22003

[121] H. Nguyen, Z. He, S. A. Gandre, U. Pasupulety, S. K. Shivakumar, and K. Lerman, "Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation," *arXiv preprint arXiv:2502.11306v1*, 2025. [Online]. Available: http://arxiv.org/abs/2502.11306v1

[122] H. Zheng, T. Xu, H. Sun, S. Pu, R. Chen, and L. Sun, "Thinking before looking: Improving multimodal LLM reasoning via mitigating visual hallucination," *CoRR*, vol. abs/2411.12591, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.12591

[123] S. Ratnakar, A. Talasila, R. Chamadiya, N. Agarwal, and V. K. Doifode, "Beyond QA pairs: Assessing parameter-efficient fine-tuning for fact embedding in llms," *CoRR*, vol. abs/2503.01131, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.01131

[124] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, and P. Nakov, "Generating answer candidates for quizzes and answer-aware question generators," *CoRR*, vol. abs/2108.12898, 2021. [Online]. Available: https://arxiv.org/abs/2108.12898

[125] S. Changpinyo, D. Kukliansy, I. Szpektor, X. Chen, N. Ding, and R. Soricut, "All you may need for VQA are image captions," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1947–1963. [Online]. Available: https://aclanthology.org/2022.naacl-main.142/

[126] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 2545–2568. [Online]. Available: https://aclanthology.org/2021.naacl-main.201/

[127] H. S. Saadi, M. D. Bui, M. Sanz-Guerrero, and K. von der Wense, "JGU mainz's submission to the WMT25 shared task on llms with limited resources for slavic languages: MT and QA," *CoRR*, vol. abs/2509.22490, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.22490

[128] M. Xiao, X. Cai, C. Wang, and Y. Zhou, "m-kailin: Knowledge-driven agentic scientific corpus distillation framework for biomedical large language models training," *CoRR*, vol. abs/2504.19565, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2504.19565

[129] J. Kasai, K. Sakaguchi, Y. Takahashi, R. L. Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui, "Realtime QA: what's the answer right now?" in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb Abstract-Datasets_and_Benchmarks.html

[130] S. Q. Liu, Z. Ji, Y. Wang, and Z. Zhang, "Real time state monitoring and fault diagnosis system for motor based on labview," *CoRR*, vol. abs/1806.09998, 2018. [Online]. Available: http://arxiv.org/abs/1806.09998

[131] D. Dissanayake and S. Nanayakkara, "Navigating the state of cognitive flow: Context-aware AI interventions for effective reasoning support," *CoRR*, vol. abs/2504.16021, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2504.16021

[132] X. Fu, B. Zhou, S. Chen, M. Yatskar, and D. Roth, "Dynamic clue bottlenecks: Towards interpretable-by-design visual question answering," *arXiv preprint arXiv:2305.14882v2*, 2023. [Online]. Available: http://arxiv.org/abs/2305.14882v2

[133] Q. Long, M. Wang, and L. Li, "Generative imagination elevates machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 5738–5748. [Online]. Available: https://aclanthology.org/2021.naacl-main.457/

[134] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 1009–1021. [Online]. Available: https://aclanthology.org/2021.naacl-main.79/

[135] Y. Zhou, J. Zhang, G. Chen, J. Shen, and Y. Cheng, "Less is more: Vision representation compression for efficient video generation with large language models," 2024.

[136] Y. Zhou, X. Li, Q. Wang, and J. Shen, "Visual in-context learning for large vision-language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 15 890–15 902.

[137] H. Lin, S. Cao, S. Wang, H. Wu, M. Li, L. Yang, J. Zheng, and C. Qin, "Interactive learning for LLM reasoning," *CoRR*, vol. abs/2509.26306, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.26306

[138] M. Hutchinson, R. Jianu, A. Slingsby, and P. Madhyastha, "Llm-assisted visual analytics: Opportunities and challenges," *CoRR*, vol. abs/2409.02691, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2409.02691

[139] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "Llm-grounded video diffusion models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=exKHibougU

[140] Y. Zhou, T. Shen, X. Geng, G. Long, and D. Jiang, "Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2559–2575.

[141] Y. Zhou, X. Geng, T. Shen, G. Long, and D. Jiang, "Eventbert: A pre-trained model for event correlation reasoning," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 850–859.

[142] Y. Zhou, X. Geng, T. Shen, J. Pei, W. Zhang, and D. Jiang, "Modeling event-pair relations in external knowledge graphs for script reasoning," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

[143] Y. Zhou, L. Song, and J. Shen, "Improving medical large vision-language models with abnormal-aware feedback," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 12 994–13 011. [Online]. Available: https://aclanthology.org/2025.acl-long.636/

[144] ——, "MAM: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration," in *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 25 319–25 333. [Online]. Available: https://aclanthology.org/2025.findings-acl.1298/

[145] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 4454–4470. [Online]. Available: https://aclanthology.org/2023.acl-long.244/

[146] Y. Zhou, Z. Rao, J. Wan, and J. Shen, "Rethinking visual dependency in long-context reasoning for large vision-language models," *arXiv preprint arXiv:2410.19732*, 2024.

[147] Y. Zhou, X. Geng, T. Shen, C. Tao, G. Long, J.-G. Lou, and J. Shen, "Thread of thought unraveling chaotic contexts," *arXiv preprint arXiv:2311.08734*, 2023.

[148] Y. Zhou, T. Shen, X. Geng, C. Tao, J. Shen, G. Long, C. Xu, and D. Jiang, "Fine-grained distillation for long document retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 732–19 740.

[149] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A collection of open-source pretrained large language models for medical domains," in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 5848–5864. [Online]. Available: https://aclanthology.org/2024.findings-acl.348/

[150] Q. Wang, C. Wang, Z. Lai, and Y. Zhou, "Insectmamba: State space model with adaptive composite features for insect recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[151] Q. Wang, H. Hu, and Y. Zhou, "Memorymamba: Memory-augmented state space model for defect recognition," *arXiv preprint arXiv:2405.03673*, 2024.

[152] Y. He, Z. Liu, J. Chen, Z. Tian, H. Liu, X. Chi, R. Liu, R. Yuan, Y. Xing, W. Wang, J. Dai, Y. Zhang, W. Xue, Q. Liu, Y. Guo, and Q. Chen, "Llms meet multimodal generation and editing: A survey," *CoRR*, vol. abs/2405.19334, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.19334

[153] Y. Zhou and G. Long, "Style-aware contrastive learning for multi-style image captioning," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 2257–2267.

[154] Y. Zhou, W. Tao, and W. Zhang, "Triple sequence generative adversarial nets for unsupervised image captioning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7598–7602.

[155] C. Wang, Y. Zhou, Q. Wang, Z. Wang, and K. Zhang, "Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13 391–13 397.

[156] Y. Yang, Y. Peng, E. S. Chng, and X. Zhong, "Bridging speech and text: Enhancing ASR with pinyin-to-character pre-training in llms," in *14th IEEE International Symposium on Chinese Spoken Language Processing, ISCSLP 2024, Beijing, China, November 7-10, 2024*, Y. Qian, Q. Jin, Z. Ou, Z. Ling, Z. Wu, Y. Li, L. Xie, and J. Tao, Eds. IEEE, 2024, pp. 646–650. [Online]. Available: https://doi.org/10.1109/ISCSLP63861.2024.10800477

[157] Y. Zhou, J. Shen, and Y. Cheng, "Weak to strong generalization for large language models with multi-capabilities," in *The Thirteenth International Conference on Learning Representations*, 2025.