Toward Multimodal Agent Intelligence: Perception, Reasoning, Generation and Interaction

Zixuan Zhou, Maycon Leone de Melo, Tatiane Araújo Rios
Federal University of Bahia

Abstract: The pursuit of Artificial General Intelligence (AGI) necessitates the development of agents that can understand and interact with the world in a manner akin to humans. A cornerstone of this endeavor is multimodal agent intelligence, which equips agents with the ability to process, comprehend, and act upon information from a multitude of sensory channels, such as vision, language, and audio. This survey provides a comprehensive overview of the field, charting a course through the core components required to build such sophisticated systems. We begin by establishing the foundations of multimodal intelligence, defining key concepts and tracing its evolution. We then delve into the three pillars of agent capability: Multimodal Perception, which covers how agents see, hear, and read the world; Multimodal Reasoning and Learning, which explores how they think, infer, and acquire new knowledge from diverse data streams; and Multimodal Generation and Interaction, which examines how they communicate and create content across different modalities. By systematically reviewing state-of-the-art techniques, benchmark datasets, and architectural paradigms in each of these areas, we map the current landscape of research. Finally, we synthesize the major open challenges, including robustness, interpretability, data scarcity, and ethical considerations, and propose promising future directions. This survey aims to serve as a valuable resource for researchers and practitioners, illuminating the path toward creating more capable, collaborative, and human-like intelligent agents.

Keywords: Multimodal Agent Intelligence, Multimodal Perception, Multimodal Reasoning, Multimodal Generation, Human–AI Interaction, Interpretability, Survey

1. Introduction

The evolution of Artificial Intelligence (AI) has been marked by a progressive shift from systems capable of solving narrow, specific problems to those that exhibit more general, human-like cognitive abilities [1]. A critical element of this evolution is the move beyond unimodal data processing. Humans perceive and understand the world through a rich tapestry of sensory inputs, vision, sound, touch, and language, which are seamlessly integrated to form a coherent understanding of reality. To create truly intelligent agents, AI systems must mirror this ability, processing and synthesizing information from multiple modalities to achieve a deeper, more contextualized comprehension [2]. This is the central premise of multimodal agent intelligence.

The growing importance of this field is driven by the limitations of unimodal systems. An agent relying solely on text may miss the nuance conveyed by a speaker's tone of voice or facial expression [3]. Similarly, a vision-only system may fail to understand the context of a scene without accompanying linguistic descriptions [4]. Multimodal agents, by contrast, can leverage the complementary and redundant information across modalities to build more robust, accurate, and comprehensive models of the world [5]. This capability is essential for a wide range of applications, from advanced robotics and autonomous systems that must navigate complex, dynamic environments [6] to sophisticated conversational agents that can engage in natural, empathetic human-computer interaction [7].

The development of such agents is not merely an engineering challenge; it pushes the boundaries of our understanding of intelligence itself. It requires us to tackle fundamental questions about knowledge representation, reasoning, learning, and communication [8]. For instance, how can an agent adapt its existing knowledge to novel problems or changing contexts, a process known as Creative Problem Solving (CPS) [8]? How can we design systems that learn continually from a stream of multimodal data without catastrophically forgetting past knowledge [9, 10]? Furthermore, as these agents become more powerful and autonomous, we must address the profound ethical implications of their deployment, ensuring they are aligned with human values and operate safely and fairly [11].

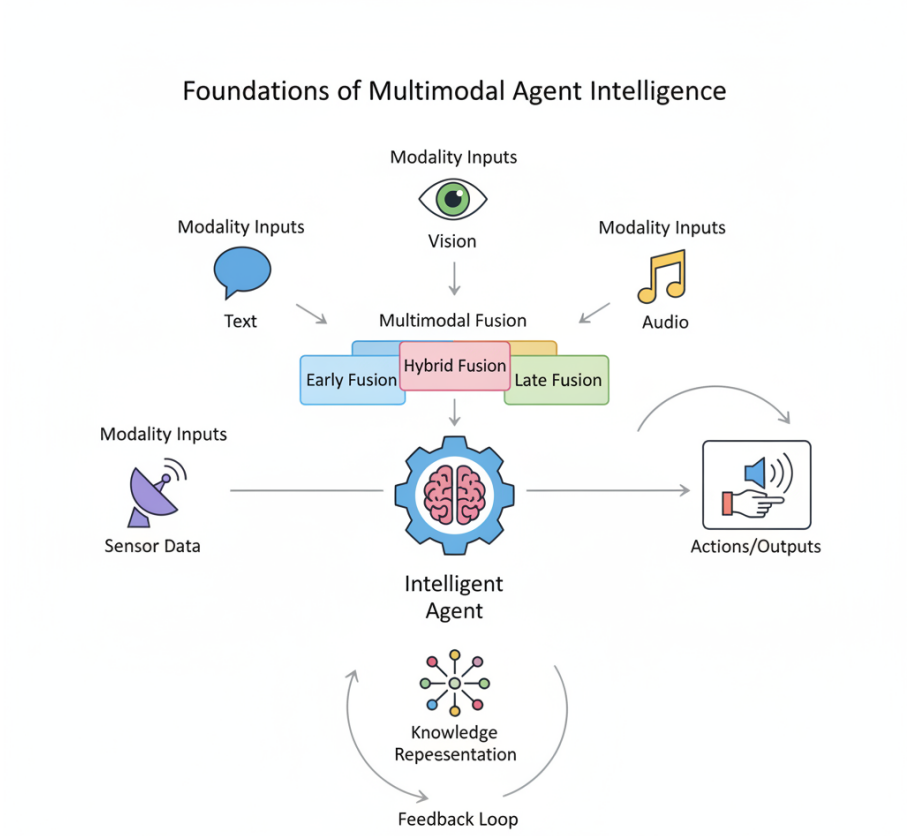

Fig. 1. Foundations of Multimodal Agent Intelligence: An intelligent agent processes diverse modality inputs through various fusion mechanisms to build unified knowledge representations, enabling effective actions and a continuous feedback loop with its environment.

This survey aims to provide a structured and comprehensive review of the core components of multimodal agent intelligence. We organize our discussion around four key pillars: the foundational concepts that underpin the field, the perceptual capabilities that allow agents to sense the world, the reasoning and learning mechanisms that enable them to understand it, and the generative and interactive abilities that allow them to communicate and create within it. We will explore the progression from foundational theories to cutting-edge models, including the recent rise of large-scale foundation models that are reshaping the landscape of AI [12, 13]. By synthesizing a vast body of literature, this paper will outline the current state of the art, identify critical challenges, and chart a course for future research toward the ultimate goal of building truly intelligent multimodal agents.

2. Foundations of Multimodal Agent Intelligence

Before delving into the specific architectures and capabilities of modern multimodal agents, it is strictly necessary to establish a robust theoretical baseline. This section aims to disambiguate the terminology that permeates the field, clarifying what constitutes an ”agent” within the artificial intelligence landscape, rigorously defining ”modality” and ”multimodality,” and deconstructing the critical mechanisms of ”multimodal fusion.” Furthermore, we trace the conceptual evolution from unimodal processing to complex multimodal paradigms, highlighting the architectural principles and mathematical representations that underpin these intelligent systems.

2.1. Defining the Intelligent Agent

At its core, an intelligent agent is conceptualized as an autonomous entity situated within an environment. It operates through a continuous feedback loop: perceiving the state of the world through sensors and influencing that state through actuators to achieve specific objectives [8]. This definition underscores the agent’s role not merely as a passive data processor, but as an active participant in its surroundings. The nature of ”intelligence” governing these agents is itself a subject of extensive debate. While classical perspectives often rely on the Turing

test, modern research seeks more formalized, quantifiable measures of general intelligence [14, 15]. Recently, a significant paradigm shift has been proposed, moving away from defining intelligence solely as competence in predefined tasks. Instead, researchers advocate for viewing intelligence as a formative, "open-ended" process of self-organization. In this view, termed "open-ended intelligence," agents are not just optimizing for a static metric but are individuated and evolved through continuous, formative interactions with their environment [1]. In this context, a "modality" represents a distinct channel of communication or perception, such as text, vision (images/video), audio, or varying sensor data (e.g., LiDAR, depth, thermal). Consequently, "multimodality" is the discipline of engineering systems capable of ingesting, processing, and synthesizing information from two or more such disparate channels. The transition from unimodal to multimodal agents represents a leap towards mimicking human-like perception, where holistic understanding is derived from the convergence of senses.

2.2. Mechanisms of Multimodal Fusion

The cornerstone of effective multimodal processing is *multimodal fusion*, the methodological approach to integrating information from distinct modalities to enhance prediction, classification, or generation tasks. Fusion strategies are generally taxonomized by the stage in the processing pipeline at which the integration occurs [2]. **Early Fusion (Feature-Level):** This strategy involves the integration of raw data or low-level feature representations immediately at the input stage. By concatenating features before they pass through the main processing layers, the model is compelled to learn cross-modal correlations from the very beginning [16]. While powerful for capturing low-level dependencies (e.g., synchronizing lip movement with speech audio), early fusion often suffers from the curse of dimensionality and requires data to be highly synchronized. **Late Fusion (Decision-Level):** Conversely, late fusion trains independent models for each modality. These unimodal experts generate separate predictions, which are subsequently aggregated at the final stage using techniques such as averaging, weighted voting, or more sophisticated ensemble methods like agglomerative clustering [17]. This approach offers modularity and flexibility, allowing the use of the best specific architecture for each data type, but potentially overlooks the complex, non-linear interactions between modalities that occur at the feature level. **Hybrid and Intermediate Fusion:** To bridge the gap between early and late strategies, hybrid fusion (often referred to as intermediate or deep fusion) has emerged as a dominant paradigm. Here, information is fused at various depths within a deep neural network, typically using mechanisms like cross-attention or shared representation layers [16]. This allows the system to learn interactions at multiple levels of abstraction. Beyond these standard categories, the field has seen diverse methodological innovations tailored to specific constraints. For computational efficiency, low-rank fusion techniques have been proposed, which decompose high-dimensional weight tensors to model latent multimodal interactions without the prohibitive cost of over-parameterization [18]. In signal processing contexts, such as multi-sensor filter design, researchers have explored arithmetic average density fusion to robustly handle noisy inputs [19]. The search for optimal fusion has even extended to biologically inspired models, such as algorithms mimicking the human immune system's information processing capabilities [20]. Interestingly, the principles of fusion are so fundamental that parallel concepts are investigated in disparate fields, such as the magneto-inertial fusion energy domains, illustrating the broad scientific relevance of combining distinct energy or information sources [21].

2.3. Multimodal Knowledge Representation

For an agent to reason effectively, it requires a robust *knowledge representation*—a way to encode diverse data streams into a unified, semantically meaningful format. A primary goal is to construct shared embedding spaces where semantically related concepts from different modalities (e.g., an image of a cat and the word "cat") are mapped to proximal points [22]. Graph-based representations have gained prominence for their ability to model complex, relational data. By structuring information as nodes and edges, agents can capture dependencies between entities and events that linear vectors might miss [23, 24]. Specifically, Knowledge Graphs (KGs) provide a structured, symbolic layer of "world knowledge" that complements the statistical, subsymbolic patterns learned by neural networks, leading to more robust neuro-symbolic systems [25]. On a more theoretical level, researchers have explored rigorous algebraic structures to formalize these representations. Polynomial representations [26] and generalized path algebras [27] offer mathematical frameworks for encoding sequences and transitions. Furthermore, specific algebraic constructs, such as hom-Lie algebras [28] and hyper loop algebras [29], have been studied to capture complex symmetries and conservation laws within data representations. Concepts from Clifford theory, such as glider representations, provide additional tools for analyzing the structural properties of these latent spaces [30]. In the linguistic domain, learning effective cross-lingual sentence representations remains a foundational challenge, enabling agents to transfer knowledge across language barriers and unifying textual modalities [31]. Finally, as these representations become increasingly complex, the interpretability and evaluation of learned features have become critical active research areas. It is vital to decipher exactly what information—spurious correlations or genuine causal links—is encoded within the agent's latent states [32]. Ultimately, these foundational pillars—the

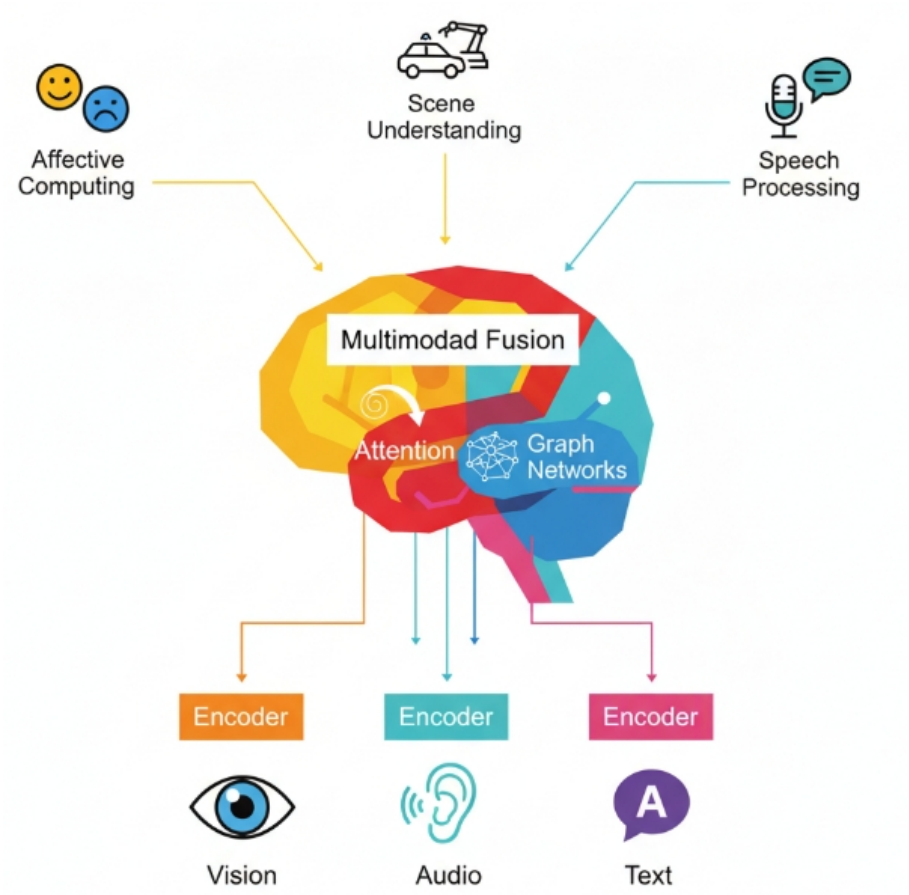

Fig. 2. Multimodal Perception: Integrating diverse sensory inputs (Vision, Audio, Text) through advanced fusion strategies for robust understanding in applications like Affective Computing, Scene Understanding, and Speech Processing.

definition of agency, the mechanics of fusion, and the rigor of representation—collectively enable the creation of systems capable of perceiving, reasoning, and interacting within a complex, multimodal world [33].

3. Multimodal Perception

Perception serves as the fundamental gateway through which intelligent agents interpret and interact with their surroundings. While unimodal systems are limited by the physical constraints of a single sensor, multimodal perception mimics the biological capability of synthesizing heterogeneous sensory inputs—such as vision, language, and audio—to construct a comprehensive and coherent understanding of the world. This section provides an in-depth exploration of multimodal perception mechanisms, transitioning from the foundational architectures of feature extraction and fusion to advanced applications. We examine how these systems overcome the noise and ambiguity inherent in individual modalities to achieve breakthroughs in domains ranging from affective computing and semantic scene understanding to robust speech processing.

3.1. Foundations: Feature Extraction and Fusion Architectures

The efficacy of a multimodal perception pipeline is heavily dependent on two critical stages: the extraction of high-fidelity representations from individual modalities and the subsequent fusion strategy used to integrate them.

3.1.1. Unimodal Representation Learning

Modern perception systems rely on specialized encoders to handle the distinct statistical properties of different data types. In the visual domain, the evolution from hand-crafted features to deep learning has been transformative. State-of-the-art approaches now predominantly utilize deep convolutional neural networks (CNNs) or Vision Transformers (ViTs) to extract hierarchical spatial-temporal features from images and video streams [34]. Parallel advancements in Natural Language Processing (NLP) have provided powerful tools for textual perception; pre-

trained language models (PLMs) such as BERT and its variants generate contextualized embeddings that capture subtle semantic nuances [35]. Similarly, in the auditory domain, features are no longer limited to basic spectral analysis; raw waveforms and spectrograms are now processed through specialized neural encoders to capture prosody and acoustic environments [36].

3.1.2. Advanced Fusion Strategies

Once unimodal representations are obtained, the core challenge lies in fusing them effectively. The optimal fusion strategy often depends on the alignment and granularity of the data. Recent research has moved beyond simple concatenation towards sophisticated architectural designs. For instance, in the complex realm of autonomous driving, the IS-Fusion framework demonstrates the importance of hierarchical integration by explicitly incorporating both instance-level and scene-level multimodal information for 3D object detection [37].

To address specific sensor characteristics, researchers have developed specialized modules. A notable example is the Angle-based perception module, explicitly designed to optimize the fusion of visible and infrared imagery by accounting for their geometric disparities [38]. Furthermore, general-purpose architectures continue to evolve, with attention-based frameworks leading the way in dynamically weighting modality importance [2]. However, fusion is not a "one-size-fits-all" solution; the challenge of handling imperfectly registered data—common in medical imaging—suggests that the optimal fusion point (early, intermediate, or late) remains highly model- and task-specific [16].

3.2. Affective Computing and Human Analysis

One of the most profound applications of multimodal perception is the analysis of human affect. Recognizing emotion is an inherently multimodal process, as human expression is distributed across facial micro-expressions, vocal intonations, body language, and linguistic content.

3.2.1. Context and Structure in Emotion Recognition

Research consistently shows that leveraging multiple modalities yields superior performance over unimodal baselines. For example, in high-stakes scenarios like deception detection, combining facial affect representations with other cues has proven highly discriminative [3]. Beyond simple feature combination, structural modeling has become crucial. Graph Convolutional Networks (GCNs) have been successfully deployed to model the complex dependencies across different modalities and speakers in conversational settings [39].

A significant trend in recent literature is the shift from context-agnostic recognition to context-aware understanding. Systems like EMERSK and SAFER argue that emotion cannot be isolated from its environment; they explicitly incorporate situational knowledge and background scene context to refine facial emotion recognition predictions [40, 41].

3.2.2. Fusion Mechanisms for Sentiment and Reasoning

The technical approaches to fusing affective signals are diverse. Traditional yet effective methods include local-global ranking fusion [42] and hybrid systems combining deep neural networks with Bayesian classifiers to handle uncertainty [43]. More recently, novel architectures have emerged to capture fine-grained interactions, such as Double Multi-Head Attention systems [44] and efficient Mamba-based fusion networks [45].

The advent of Large Language Models (LLMs) has further revolutionized this field. Models like Emotion-LLaMA represent a leap towards "reasoning" agents, integrating audio, visual, and textual inputs to not only recognize emotions but also explain them [46]. This progress builds upon a rich foundation of multimodal sentiment analysis research, which focuses on detecting valence and affectual states from text [47] and diverse multimodal sources. Numerous works have explored various facets of this challenge, from interaction-aware representations to unified pre-training frameworks [48–55].

3.3. Scene Understanding and Domain-Specific Perception

Beyond human-centric tasks, multimodal perception is critical for robustly interpreting complex physical scenes and abstract social events.

3.3.1. Physical World Perception

In robotics and autonomous systems, safety relies on redundancy. Fusing data from complementary sensors—such as cameras (rich texture) and LiDAR (accurate depth)—is essential for tasks like 3D object detection, semantic segmentation, and navigation [2]. Domain-specific constraints often dictate the choice of algorithms. For instance, in precision agriculture, where computational resources may be limited, techniques combining Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVM) have been effectively applied for early disease

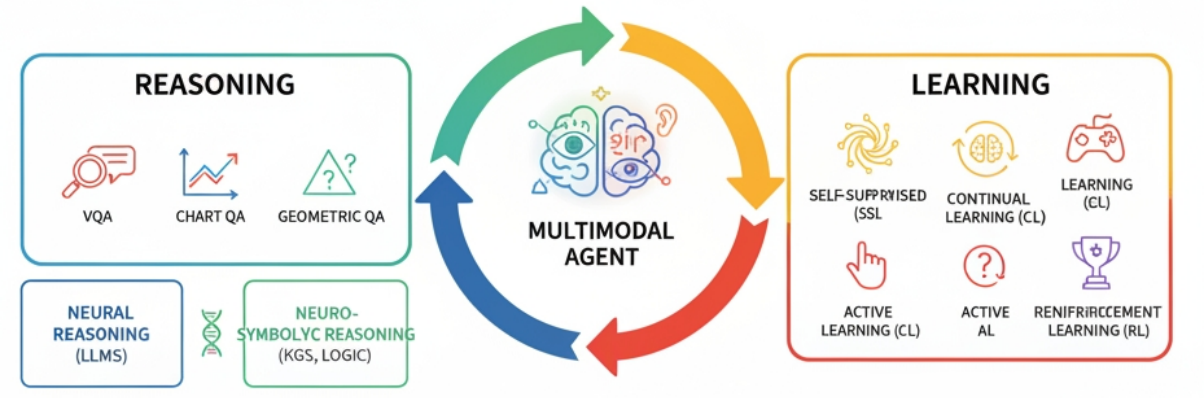

Fig. 3. Multimodal Agent Intelligence: Interplay of Advanced Reasoning and Adaptive Learning Strategies

detection [56]. Similarly, in medical imaging, the fusion of anatomical information from CT and MRI scans enables organ segmentation with a precision unattainable by single-modality methods [57]. Emerging wearable technologies also leverage multimodal fusion, utilizing inertial and mechanomyography sensors to reconstruct facial activity without the privacy concerns or occlusion issues of cameras [58].

3.3.2. Social Media and Abstract Content

The scope of perception extends to the digital realm, where agents must interpret abstract content. The detection of fake news [59], the identification of sarcasm [60,61], and the recognition of propaganda in internet memes [62] all benefit significantly from fusing visual imagery with textual cues. In these tasks, the visual modality often serves as a grounding mechanism to verify or contradict the textual claims.

3.4. Audio-Visual Speech Processing

The interaction between sight and sound is a distinct domain where multimodal perception has driven transformative improvements, particularly in challenging acoustic environments.

Audio-visual speech enhancement has shown remarkable success by using visual cues (lip movements) to separate speech from noise, effectively solving the ”cocktail party” problem [63,64]. This principle has been extended to increasingly complex setups, such as incorporating ultrasound tongue images to resolve phonetic ambiguities [65]. The integration of visual information is now considered crucial for multi-channel speech separation, dereverberation, and robust recognition in noisy, real-world scenarios [66].

Advancements in this field also cover signal capture and synthesis. Generative models have been employed to enhance the quality of body-conducted speech capture, compensating for the loss of high-frequency content [67]. Novel neural architectures like dCoNNear aim to provide artifact-free processing for closed-loop audio systems, ensuring high perceptual quality [68]. Furthermore, multimodal approaches enable semantic learning from untranscribed data; visually grounded models can learn to associate spoken captions with images, thereby enabling capabilities like semantic speech retrieval without relying on textual transcriptions [69].

4. Multimodal Reasoning and Learning

Once an agent has perceived the world through its various sensors, it must process this information to infer knowledge, make decisions, and learn from its experiences. This section explores the cognitive layer of multimodal agent intelligence, focusing on how agents reason over and learn from heterogeneous data. We will cover different reasoning paradigms, from symbolic and neural to hybrid neuro-symbolic approaches, and their application in challenging tasks like question answering. Additionally, we will discuss advanced learning strategies, including self-supervised, continual, and active learning, which are crucial for building agents that can adapt and generalize in dynamic environments.

4.1. Complex Multimodal Reasoning Tasks

Multimodal reasoning requires the agent to go beyond simple pattern recognition and perform complex inferential steps that connect information across modalities. This capability is rigorously tested through various benchmarks, with Question Answering (QA) serving as a primary evaluation testbed. In these tasks, an agent must integrate disparate data streams to derive a correct answer. Visual Question Answering (VQA) [34], for instance, requires

the synthesis of visual semantics from an image with the linguistic intent of a natural language question. However, the field has moved towards even more intricate domains. Reasoning over charts involves understanding data trends and structural relationships [70], while geometric problem solving requires rigorous logical deduction based on visual figures [71]. Furthermore, in document intelligence, agents must navigate and reason over long-form, visually rich documents to extract relevant information [72].

A key technical challenge in these domains is effective cross-modal grounding. To address this, recent models have focused on synergizing motion and appearance features to reason about temporal events in videos [73]. Others have enhanced reasoning depth by retrieving and integrating relevant knowledge from external sources, bridging the gap between implicit visual cues and explicit world knowledge [74–76]. The complexity increases further in Conversational QA settings, where the agent must resolve co-references and maintain context consistency across multiple turns of dialogue [77]. Beyond standard QA, researchers are also exploring specialized reasoning capabilities, such as extracting spatial relationships from text descriptions [78] and performing temporal reasoning to infer implicit events and timelines [79].

4.2. Neuro-Symbolic and Neural Reasoning Paradigms

To tackle these complex reasoning tasks, various computational paradigms have been explored. Purely neural approaches, particularly Large Language Models (LLMs), have demonstrated impressive zero-shot reasoning abilities. By utilizing advanced prompting techniques like Chain-of-Thought (CoT) and the more robust Plan-and-Solve (PS) prompting strategies, LLMs can decompose complex problems into intermediate steps [80]. Despite these successes, pure neural models often struggle with structured data manipulation, long-term logical consistency, and hallucination.

This limitation has catalyzed the rise of neuro-symbolic approaches, which aim to combine the robust learning and generalization capabilities of neural networks with the interpretability and explicit reasoning power of symbolic systems. These hybrid frameworks offer a ”best of both worlds” solution. For instance, integrating Knowledge Graphs (KGs) with language models has been shown to significantly enhance QA performance by providing structured background knowledge [81]. Similarly, symbolic logic can be employed to guide generative models, ensuring that the outputs adhere to specific physical laws or safety constraints [82].

The versatility of neuro-symbolic systems is evident across various domains. They have been successfully developed for complex spatio-temporal reasoning tasks [83], Commonsense QA where logical inference is paramount [84], and improving compositional generalization in language understanding [85]. In specialized high-stakes fields like medical imaging, these systems provide a promising avenue for domain generalization by integrating expert clinical knowledge into the learning process [86]. Furthermore, to address scalability and noise issues inherent in large-scale ontologies, the development of robust reasoners for formal languages, such as RDF and Description Logic, remains an active and critical area of research [87].

4.3. Advanced Learning Strategies: Adaptation and Efficiency

Beyond static reasoning capabilities, the ability to learn efficiently and adaptively is paramount for intelligent agents operating in the wild. Self-Supervised Learning (SSL) has emerged as a powerful paradigm for learning rich, generalized representations from vast amounts of unlabeled data. In SSL, a model learns by solving a ”pretext” task—such as predicting a missing part of the input or contrasting augmented views of the same data. This approach has proven highly effective across modalities. For example, it has been applied to learn robust representations from raw audio by jointly leveraging synchronized audio and visual signals [36]. In computer vision, SSL techniques facilitate unsupervised image clustering [88] and optimize dense contrastive learning to identify better positive pairs, thereby improving object detection performance [89].

The benefits of SSL extend to various downstream applications. It enhances model robustness under transfer learning scenarios [90] and serves as a critical auxiliary task to boost performance in few-shot learning settings where labeled data is scarce [91]. Furthermore, SSL has been utilized to pre-train controllers for efficient Neural Architecture Search [92]. However, it is important to note that the utility of SSL representations often depends on the alignment between the pretext task and the specific downstream environment [93]. Recent advances have also shown that combining contrastive self-supervision with consistency regularization yields state-of-the-art results in semi-supervised learning [94].

As agents are deployed in real-world settings, they encounter non-stationary data distributions and evolving tasks. Continual Learning (CL), or lifelong learning, addresses the challenge of learning sequentially without suffering from ”catastrophic forgetting” of previously acquired knowledge. Various strategies have been proposed to mitigate this phenomenon. Regularization-based methods, such as those employing energy-based models, help preserve important parameters [10], while meta-learning approaches aim to learn representations that are inherently robust to forgetting [9]. Replay-based methods leverage generative models to reproduce past data, maintaining a memory of previous distributions [95]. On the architectural front, solutions include parameter stacking

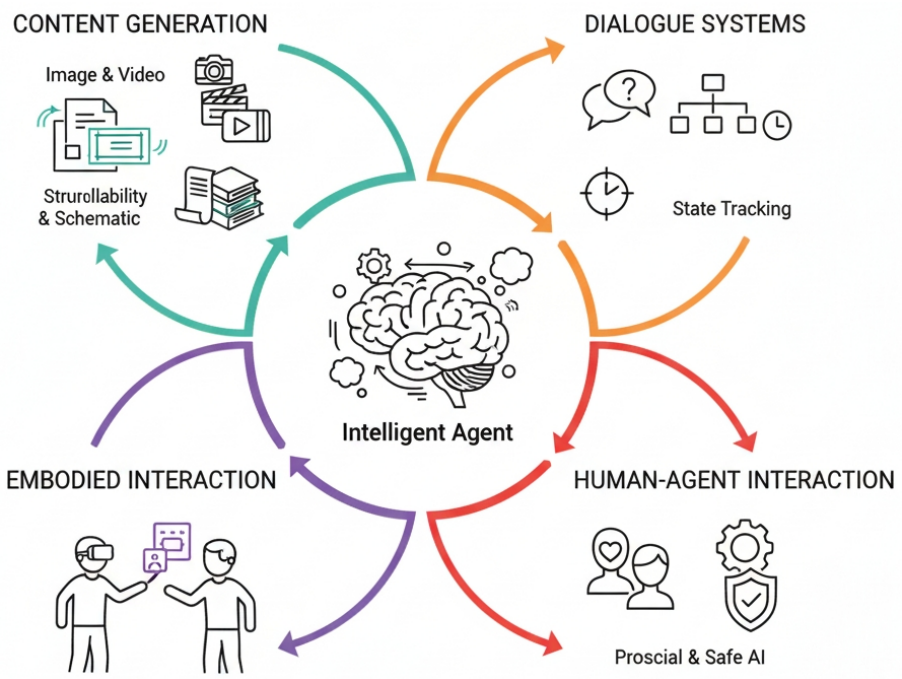

Fig. 4. An overview of Multimodal Generation & Interaction, illustrating the interconnected pillars of advanced content creation, intelligent dialogue systems, and immersive, prosocial human-agent collaboration.

to segregate task-specific knowledge [96] or utilizing sparse routing networks with multiple experts to specialize in different tasks dynamically [97, 98]. Applying CL methods is particularly valuable when integrating new modalities into existing foundation models, such as adding vision capabilities to an LLM while minimizing the degradation of its linguistic performance [99]. To advance this field, the development of standardized reference architectures [100] and comprehensive benchmarks [101] remains crucial.

Finally, other learning paradigms play essential roles in specific contexts. Active Learning optimizes the training process by selecting the most informative data points for human labeling, thereby reducing annotation costs [102]. Meanwhile, Reinforcement Learning (RL) enables agents to learn through interaction, proving essential for mastering complex navigation behaviors [6] and solving distinct challenges like low-resource relation extraction through policy optimization [103].

5. Multimodal Generation and Interaction

The ultimate test of an intelligent agent lies not merely in its capacity to perceive and reason about the external world, but fundamentally in its ability to actively participate within it through coherent action and communication. This section delves into the output-oriented frontiers of multimodal agents, focusing on two pivotal capabilities: the generation of high-fidelity, contextually grounded content across diverse modalities, and the orchestration of natural, collaborative interactions with human users. We explore the rapid evolution of generative paradigms—from static image synthesis to dynamic video production—and examine the sophisticated mechanisms underpinning modern dialogue systems. Furthermore, we highlight the principles of human-agent interaction (HAI), emphasizing how multimodality serves as the cornerstone for creating systems that are not only intelligent but also intuitive, trustworthy, and socially aligned.

5.1. Advanced Multimodal Content Generation

Multimodal generation involves the synthesis of content in a target modality conditioned on input from source modalities, a process that requires capturing complex cross-modal correlations.

Image and Video Synthesis One of the most transformative advancements has occurred in text-to-image synthesis, where models have evolved from generating simple object-centric scenes to producing highly realistic and artistically complex imagery. Recent innovations have begun to transcend traditional noise-based diffusion boundaries. For instance, novel flow matching techniques have emerged that learn direct mappings between modality

distributions, effectively obviating the need for iterative noise-based generation and complex conditioning mechanisms, thereby streamlining the generative process [104]. This generative paradigm is progressively expanding into the temporal dimension, addressing the challenges of video grounding and generation. Unlike static images, video synthesis requires maintaining temporal consistency and causal logic. Researchers have made significant strides in this area, developing models capable of aligning textual queries with video segments and generating coherent video content from natural language descriptions [105–108].

Structured and Schematic Generation Beyond visual pixels, multimodal generation also encompasses the creation of structured narratives and layouts. In the domain of creative writing, systems can now perform story generation guided by visual inputs, weaving plotlines that are semantically aligned with a sequence of images [109]. Similarly, for visually rich document understanding, models have demonstrated the ability to generate structured document layouts, effectively arranging text and images to optimize information presentation [110]. To tackle the generation of long, coherent text passages, strategies that mimic human planning have proven effective; specifically, the progressive generation of content by first establishing a high-level plan or keyword sequence allows for better discourse management [111].

Controllability and Agentic Frameworks The utility of generative models is maximized when outputs can be precisely tailored to user intent. Foundational generative capabilities are being adapted to create customized information that directly addresses specific user needs [112]. Furthermore, to enforce strict adherence to logical or lexical constraints during generation, neuro-logic decoding algorithms have been introduced to guide the decoding process without requiring model retraining [113]. More recently, the field has moved towards agentic frameworks that handle complex, multi-step instruction-based image generation. These agents can reason about tools and intermediate steps, representing a significant leap towards controllable and iterative creative tools [114].

5.2. Goal-Oriented and Knowledgeable Dialogue Systems

A crucial application of multimodal interaction is the domain of conversational agents. Modern dialogue systems, predominantly built upon Large Language Models (LLMs), aim to transcend simple question-answering to engage in open-domain, knowledgeable, and goal-oriented discourse.

Knowledge Grounding and Retrieval A persistent challenge in neural dialogue generation is the tendency to ”hallucinate” facts. To mitigate this, Retrieval-Augmented Generation (RAG) has become a standard paradigm. By fetching relevant documents from external corpora to inform the generated response, agents can provide answers that are factually accurate and information-rich [115].

Control and State Tracking Beyond factual accuracy, controlling the stylistic and semantic attributes of the generated dialogue is essential for specific applications. Techniques using semantic exemplars—prototypical responses that guide the model—have been shown to effectively steer the model's output towards desired goals or personas [116]. In the context of task-oriented dialogue (e.g., booking flights or technical support), maintaining an accurate belief state is critical. This task, known as dialogue state tracking, has been successfully reframed as a language modeling problem, where schema-driven prompting allows the model to track user intent and slot values dynamically [117]. Additionally, incorporating formal semantic representations, such as Abstract Meaning Representation (AMR), enhances the model's ability to grasp explicit core semantic knowledge, leading to more robust language understanding [118].

5.3. Immersive and Prosocial Human-Agent Interaction

Human-agent interaction is fundamentally multimodal; effective communication relies on the seamless integration of verbal and non-verbal cues within a shared environment.

Situated and Embodied Interaction To move interaction beyond text boxes, agents must understand the physical world. Datasets and benchmarks like SIMMC 2.0 are being actively developed to train agents for task-oriented dialogues grounded in immersive, multimodal environments (e.g., VR or AR settings), enabling them to co-observe and discuss objects with users [119].

Cognitive and Social Alignment The ability to generate responses with appropriate emotional tone, backed by corresponding facial expressions or gestures, represents the next frontier in interaction. Moreover, to foster deeper reasoning, frameworks that encourage divergent thinking—such as simulating multi-agent debates—have

been proposed to lead to more robust and creative problem-solving outcomes in dialogue [120]. Crucially, as agents become more capable, safety becomes paramount. Building prosocial backbones for conversational agents is critical to ensure they interact in a helpful, harmless, and honest manner, actively avoiding toxic or biased outputs [121].

Evaluation Challenges Finally, reliable human assessment of these open-domain dialogue systems remains a significant bottleneck. The subjectivity of conversation quality necessitates the development of robust evaluation methodologies that can measure true engagement, coherence, and utility, moving beyond superficial automated metrics [122]. The ultimate goal is to create agents that are not just intelligent tools, but collaborative and socially aware partners in human endeavors.

6. Challenges, Future Directions, and Conclusion

Despite the remarkable progress in multimodal agent intelligence, fueled by the advent of large foundation models and advanced fusion techniques, the journey toward creating systems with truly human-like cognitive capabilities is fraught with significant challenges. Current systems, while impressive in specific benchmarks, often lack the flexibility, robustness, and ethical grounding required for widespread real-world deployment. This final section provides a comprehensive analysis of the current landscape, detailing key open problems such as data scarcity, interpretability, robustness, and ethical alignment. We then propose a roadmap of promising future research directions, ranging from neuro-symbolic architectures to the pursuit of causal reasoning. We conclude by reiterating the transformative potential of this field.

6.1. Current Challenges

Data Scarcity, Quality, and Modality Imbalance. One of the most persistent bottlenecks in training robust multimodal agents is **data scarcity and quality**. While the volume of data on the internet is vast, high-quality, aligned multimodal data remains a scarce resource. Large-scale datasets scraped from the web are often plagued by noise, irrelevant information, and weak alignment between modalities [123], which can severely hinder model convergence and performance. This issue is exacerbated when moving beyond general domains into specialized fields. For instance, in the medical domain, privacy regulations and the high cost of expert annotation lead to a critical shortage of labeled data for tasks like medical visual question answering or report generation [124, 125]. Similarly, in industrial settings, anomaly detection systems often struggle with incomplete modalities due to sensor failures or transmission errors [126].

Furthermore, the ”long-tail” problem is significant in multilingual multimodal learning. Most current datasets are heavily skewed towards English, resulting in suboptimal performance for low-resource languages and hindering the global applicability of these agents [127–129]. This linguistic disparity extends to speech and document understanding tasks, where diverse language coverage is essential [130, 131]. Although weakly supervised [75] and semi-supervised [130] learning methods offer partial mitigation by leveraging unlabelled data, the need for curated, high-quality multimodal benchmarks remains paramount. Recent initiatives, such as the creation of the MULTIMODAL UNIVERSE—a massive scientific dataset comprising 100TB of astronomical data—represent a significant step toward addressing data scarcity in specialized scientific domains [132].

Robustness, Generalization, and Hallucination. Another major hurdle is ensuring **robustness and generalization** in dynamic environments. Multimodal models often demonstrate high performance on in-distribution test sets but suffer catastrophic performance drops when faced with real-world distribution shifts or adversarial perturbations. This fragility necessitates rigorous research into domain adaptation and generalization techniques, which aim to make models robust across varying environmental conditions, sensor configurations, and contextual shifts [133]. A related challenge is modality robustness; models can inadvertently learn to over-rely on a dominant modality (e.g., text) while ignoring others (e.g., audio or vision), making them brittle when the dominant modality is noisy or missing [134].

Moreover, the generative nature of modern large vision-language models (LVLMs) introduces the phenomenon of ”object hallucination,” where agents confidently describe objects or events that are not present in the input media. This hallucination problem underscores a fundamental gap between statistical language modeling and grounded multimodal understanding [135]. Additionally, models frequently achieve high accuracy for the ”wrong reasons,” exploiting spurious correlations in the training data rather than learning genuine causal relationships. For example, a model might classify an image as a ”bedroom” simply because it detects a bed, without understanding the spatial layout or other contextual cues [136].

Interpretability and Explainability. As multimodal agents are increasingly deployed in high-stakes applications, **interpretability and explainability** become critical requirements. In fields like clinical decision-making, a "black box" prediction is insufficient; practitioners need to understand the rationale behind a diagnosis to trust and verify the system's output [137]. However, as models grow in complexity and parameter count, their decision-making processes become more opaque and difficult to trace. Recent frameworks like GLIMPSE attempt to bridge this gap by providing holistic, cross-modal explanations that detail how different input signals contributed to the final output [138]. Other approaches focus on making the reasoning process explicit by integrating external knowledge graphs, thereby allowing the model to provide evidence-based justifications [74]. Nevertheless, communicating the internal state of a super-intelligent system remains an ultimate challenge; as agents surpass human capabilities, their compressed internal representations might appear as unintelligible noise to human observers, complicating the alignment of their objectives with human intent [139].

Ethical Alignment and Social Responsibility. Finally, **ethical considerations** are paramount as agents become integral to societal infrastructure. We must ensure these systems are fair, private, and aligned with human values. Large language models, which often serve as the cognitive core of multimodal agents, are prone to inheriting and amplifying harmful biases present in their training data [140]. Mitigating these biases requires comprehensive methods to measure and correct unfairness across all modalities. Additionally, the use of synthetic data is growing, necessitating frameworks for fast and fair synthetic data generation to avoid feedback loops of bias [141]. Ethical data collection is also a concern, particularly when involving vulnerable populations; strict guidelines must be established to protect privacy and consent [142]. Beyond data, we must consider physical safety; for embodied agents like assistive robots, developing ethical hazard analysis methods is crucial to prevent physical harm [143]. Implementing ethically aligned design principles throughout the industrial lifecycle of these agents is a practical and necessary step toward building trustworthy AI [144].

6.2. Future Research Directions

Looking ahead, several research avenues hold the promise of overcoming current limitations and unlocking new capabilities for multimodal agents.

Advanced Foundation Models and Lifelong Learning. The development of **multimodal foundation models** continues to accelerate. Future work will likely focus on more efficient methods for adaptation, parameter-efficient fine-tuning, and structured generation to handle complex tasks with lower computational overhead [133, 145]. Furthermore, as agents operate in the wild, they must possess **lifelong and continual learning** capabilities. Unlike static models, future agents must adapt to new environments and learn new concepts over long time spans without suffering from catastrophic forgetting of previously acquired knowledge [9]. Integrating these advanced models with **federated learning** paradigms offers a promising solution to privacy concerns and data silos, enabling agents to learn collaboratively from decentralized data sources while preserving user privacy [13].

Neuro-Symbolic Integration and Reasoning. To achieve robust generalization, the synergy between connectionist learning and symbolic reasoning—known as **neuro-symbolic AI**—is a key direction. While deep learning excels at pattern recognition, symbolic systems provide the logical structure needed for complex reasoning and planning. Hybrid architectures that combine these strengths can lead to agents that are both learnable and logically consistent [84, 86]. This is closely tied to the goal of improving **cross-modal alignment**. Future frameworks must move beyond simple contrastive learning (which aligns representations globally) to sophisticated associative learning methods that can perform fine-grained alignment and reasoning across modalities [146, 147].

Common Sense, Causality, and Social Intelligence. Finally, the pursuit of true **common sense and social intelligence** remains the grand challenge of the field. Current models struggle with "physical common sense" (understanding how objects interact) and "social common sense" (understanding human intent). Future research must focus on models that can reason about causality rather than just correlation [148]. This includes understanding complex event sequences, scripts, and temporal dependencies [149], as well as modeling the correlation between multimodal events over time [150, 151]. Agents equipped with these capabilities will be able to navigate complex human interactions more naturally and effectively.

6.3. Conclusion

In conclusion, multimodal agent intelligence represents a vibrant and rapidly advancing frontier in artificial intelligence. The convergence of perception, reasoning, and interaction across text, vision, and audio is enabling the creation of systems that are more capable, adaptable, and aligned with human needs than ever before. While

significant challenges regarding data quality, robustness, interpretability, and ethics remain, the ongoing research outlined in this survey is steadily paving the way for a future where intelligent agents can act as seamless, trustworthy partners in our daily lives, augmenting human intellect and creativity to solve the world's most complex problems.

References

1. David Weinbaum and Viktoras Veitas. Open ended intelligence: The individuation of intelligent agents. *arXiv preprint arXiv:1505.06366v2*, 2015.
2. Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, Rongtao Xu, and Shibiao Xu. Multimodal fusion and vision-language models: A survey for robot vision. *arXiv preprint arXiv:2504.02477v3*, 2025.
3. Leena Mathur and Maja J Matarić. Introducing representations of facial affect in automated multimodal deception detection. *arXiv preprint arXiv:2008.13369v1*, 2020.
4. Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online, 2021. Association for Computational Linguistics.
5. Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *arXiv preprint arXiv:2210.17444v3*, 2022.
6. Marvin Chancán and Michael Milford. Robot perception enables complex navigation behavior via self-supervised learning. *arXiv preprint arXiv:2006.08967v1*, 2020.
7. Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
8. Evana Gizzi, Lakshmi Nair, Sonia Chernova, and Jivko Sinapov. Creative problem solving in artificially intelligent agents: A survey and framework. *arXiv preprint arXiv:2204.10358v1*, 2022.
9. Khurram Javed and Martha White. Meta-learning representations for continual learning. *arXiv preprint arXiv:1905.12588v2*, 2019.
10. Shuang Li, Yilun Du, Gido M. van de Ven, and Igor Mordatch. Energy-based models for continual learning. *arXiv preprint arXiv:2011.12216v3*, 2020.
11. Alexis Roger, Esma Aïmeur, and Irina Rish. Towards ethical multimodal systems. *arXiv preprint arXiv:2304.13765v3*, 2023.
12. Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562v5*, 2023.
13. Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Bo Zhao, Liping Yi, Alysa Ziying Tan, Yulan Gao, Anran Li, Xiaoxiao Li, Zengxiang Li, and Qiang Yang. Advances and open challenges in federated foundation models. *arXiv preprint arXiv:2404.15381v4*, 2024.
14. Shane Legg and Marcus Hutter. Tests of machine intelligence. *arXiv preprint arXiv:0712.3825v1*, 2007.
15. Samuel Allen Alexander. Intelligence via ultrafilters: structural properties of some intelligence comparators of deterministic legg-hutter agents. *arXiv preprint arXiv:1910.09721v2*, 2019.
16. Lucas W. Remedios, Han Liu, Samuel W. Remedios, Lianrui Zuo, Adam M. Saunders, Shunxing Bao, Yuankai Huo, Alvin C. Powers, John Virostko, and Bennett A. Landman. Influence of early through late fusion on pancreas segmentation from imperfectly registered multimodal mri. *arXiv preprint arXiv:2409.04563v1*, 2024.
17. Evgenii Razinkov, Iuliia Saveleva, and Jiři Matas. Alfa: Agglomerative late fusion algorithm for object detection. *arXiv preprint arXiv:1907.06067v1*, 2019.
18. Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. Low rank fusion based transformers for multimodal sequences. *arXiv preprint arXiv:2007.02038v1*, 2020.
19. Tiancheng Li, Yan Song, Enbin Song, and Hongqi Fan. Arithmetic average density fusion – part i: Some statistic and information-theoretic results. *arXiv preprint arXiv:2110.01440v4*, 2021.
20. Jamie Twycross and Uwe Aickelin. Information fusion in the immune system. *arXiv preprint arXiv:1003.1598v1*, 2010.
21. C. L. Nehl, R. J. Umstattd, W. R. Regan, S. C. Hsu, and P. B. McGrath. Retrospective of the arpa-e alpha fusion program. *arXiv preprint arXiv:1907.09921v2*, 2019.
22. Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, Alessandro Calefati, and Faisal Shafait. Do cross modal systems leverage semantic relationships? *arXiv preprint arXiv:1909.01976v1*, 2019.
23. Vítor Lourenço and Aline Paes. Learning attention-based representations from multiple patterns for relation prediction in knowledge graphs. *arXiv preprint arXiv:2206.04801v1*, 2022.
24. Bongseok Lee and Yong Suk Choi. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

25. Emanuele Bastianelli, Domenico Bloisi, Roberto Capobianco, Guglielmo Gemignani, Luca Iocchi, and Daniele Nardi. Knowledge representation for robots through human-robot interaction. *arXiv preprint arXiv:1307.7351v2*, 2013.
26. Jiuzu Hong and Oded Yacobi. Polynomial representations and categorifications of fock space. *arXiv preprint arXiv:1101.2456v3*, 2011.
27. Shouchuan Zhang and Yao-Zhong Zhang. Structures and representations of generalized path algebras. *arXiv preprint arXiv:math/0402188v5*, 2004.
28. Yunhe Sheng. Representations of hom-lie algebras. *arXiv preprint arXiv:1005.0140v4*, 2010.
29. Dijana Jakelic and Adriano Moura. Finite-dimensional representations of hyper loop algebras over non-algebraically closed fields. *arXiv preprint arXiv:0711.0795v4*, 2007.
30. Frederik Caenepeel and Fred Van Oystaeyen. Clifford theory for glider representations. *arXiv preprint arXiv:1603.02493v4*, 2016.
31. Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836v4*, 2018.
32. Antonia Gogoglou, C. Bayan Bruss, and Keegan E. Hines. On the interpretability and evaluation of graph representation learning. *arXiv preprint arXiv:1910.03081v1*, 2019.
33. Ana Durica, John Booth, and Ivana Drobnjak. Towards multimodal representation learning in paediatric kidney disease. *arXiv preprint arXiv:2511.13637v1*, 2025.
34. Peiyuan Chen, Zecheng Zhang, Yiping Dong, Li Zhou, and Han Wang. Enhancing visual question answering through ranking-based hybrid training and multimodal fusion. *arXiv preprint arXiv:2408.07303v2*, 2024.
35. Yan Ling, Jianfei Yu, and Rui Xia. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159, Dublin, Ireland, 2022. Association for Computational Linguistics.
36. Abhinav Shukla, Stavros Petridis, and Maja Pantic. Learning speech representations from raw audio by joint audiovisual self-supervision. *arXiv preprint arXiv:2007.04134v1*, 2020.
37. Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. *arXiv preprint arXiv:2403.15241v1*, 2024.
38. Xiaopeng Liu, Yupei Lin, Sen Zhang, Xiao Wang, Yukai Shi, and Liang Lin. Angularfuse: A closer look at angle-based perception for spatial-sensitive multi-modality image fusion. *arXiv preprint arXiv:2510.12260v1*, 2025.
39. Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin 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)*, pages 5666–5675, Online, 2021. Association for Computational Linguistics.
40. Mijanur Palash and Bharat Bhargava. Emersk – explainable multimodal emotion recognition with situational knowledge. *arXiv preprint arXiv:2306.08657v1*, 2023.
41. Mijanur Palash and Bharat Bhargava. Safer: Situation aware facial emotion recognition. *arXiv preprint arXiv:2306.09372v1*, 2023.
42. Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. *arXiv preprint arXiv:1809.04931v1*, 2018.
43. Luca Surace, Massimiliano Patacchiola, Elena Battini Sönmez, William Spataro, and Angelo Cangelosi. Emotion recognition in the wild using deep neural networks and bayesian classifiers. *arXiv preprint arXiv:1709.03820v1*, 2017.
44. Federico Costa, Miquel India, and Javier Hernando. Double multi-head attention multimodal system for odyssey 2024 speech emotion recognition challenge. *arXiv preprint arXiv:2406.10598v1*, 2024.
45. Zanxu Wang and Homayoon Beigi. Quality-controlled multimodal emotion recognition in conversations with identity-based transfer learning and mamba fusion. *arXiv preprint arXiv:2511.14969v1*, 2025.
46. Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161v2*, 2024.
47. Saif M. Mohammad. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. *arXiv preprint arXiv:2005.11882v2*, 2020.
48. Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
49. Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu 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*, pages 756–767, Singapore, 2023. Association for Computational Linguistics.
50. Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu 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)*, pages 7617–7630, Toronto, Canada, 2023. Association for Computational Linguistics.
51. Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages

2282–2294, Seattle, United States, 2022. Association for Computational Linguistics.

52. Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, Online, 2021. Association for Computational Linguistics.
53. Junyan Cheng, Iordanis Fostiropoulos, Barry Boehm, and Mohammad Soleymani. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2447–2458, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
54. Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng 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)*, pages 5301–5311, Online, 2021. Association for Computational Linguistics.
55. Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal sentiment detection based on multi-channel graph neural networks. 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)*, pages 328–339, Online, 2021. Association for Computational Linguistics.
56. Yousef Alhwaiti, Muhammad Ishaq, Muhammad Hameed Siddiqi, Muhammad Waqas, Madallah Alruwaili, Saad Alanazi, Asfandyar Khan, and Faheem Khan. Early detection of late blight tomato disease using histogram oriented gradient based support vector machine. *arXiv preprint arXiv:2306.08326v3*, 2023.
57. Supriti Mulay, Deepika G, Jeevakala S, Keerthi Ram, and Mohanasankar Sivaprakasam. Liver segmentation from multimodal images using hed-mask r-cnn. *arXiv preprint arXiv:1910.10504v1*, 2019.
58. Hymalai Bello, Luis Alfredo Sanchez Marin, Sungho Suh, Bo Zhou, and Paul Lukowicz. Inmyface: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition. *arXiv preprint arXiv:2302.04024v1*, 2023.
59. Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, Online, 2021. Association for Computational Linguistics.
60. Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland, 2022. Association for Computational Linguistics.
61. Saroj Basnet, Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanoji, and Marcos Zampieri. Evaluating open-source vision-language models for multimodal sarcasm detection. *arXiv preprint arXiv:2510.11852v1*, 2025.
62. Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. 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)*, pages 6603–6617, Online, 2021. Association for Computational Linguistics.
63. Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-PIneda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. *arXiv preprint arXiv:1908.02590v3*, 2019.
64. Julius Richter, Simone Frintrop, and Timo Gerkmann. Audio-visual speech enhancement with score-based generative models. *arXiv preprint arXiv:2306.01432v1*, 2023.
65. Rui-Chen Zheng, Yang Ai, and Zhen-Hua Ling. Incorporating ultrasound tongue images for audio-visual speech enhancement. *arXiv preprint arXiv:2309.10455v2*, 2023.
66. Guinan Li, Jiajun Deng, Mengzhe Geng, Zengrui Jin, Tianzi Wang, Shujie Hu, Mingyu Cui, Helen Meng, and Xunying Liu. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition. *arXiv preprint arXiv:2307.02909v1*, 2023.
67. Julien Hauret, Thomas Joubaud, Véronique Zimpfer, and Éric Bavu. Configurable eben: Extreme bandwidth extension network to enhance body-conducted speech capture. *arXiv preprint arXiv:2303.10008v2*, 2023.
68. Chuan Wen, Guy Torfs, and Sarah Verhulst. dconnear: An artifact-free neural network architecture for closed-loop audio signal processing. *arXiv preprint arXiv:2501.04116v3*, 2025.
69. Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *arXiv preprint arXiv:1710.01949v2*, 2017.
70. Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics.
71. Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online, 2021. Association for Computational Linguistics.
72. Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online, 2021. Association for Computational Linguistics.
73. Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video 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 1: Long Papers), pages 6167–6177, Online, 2021. Association for Computational Linguistics.

74. Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng 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*, pages 956–968, Seattle, United States, 2022. Association for Computational Linguistics.
75. Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
76. Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William 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*, pages 5558–5570, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
77. Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas 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*, pages 520–534, Online, 2021. Association for Computational Linguistics.
78. Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, 2021. Association for Computational Linguistics.
79. Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online, 2021. Association for Computational Linguistics.
80. Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada, 2023. Association for Computational Linguistics.
81. Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online, 2021. Association for Computational Linguistics.
82. Jacob K. Christopher, Michael Cardei, Jinhao Liang, and Ferdinando Fioretto. Neuro-symbolic generative diffusion models for physically grounded, robust, and safe generation. *arXiv preprint arXiv:2506.01121v1*, 2025.
83. Jae Hee Lee, Michael Sioutis, Kyra Ahrens, Marjan Alirezaie, Matthias Kerzel, and Stefan Wermter. Neuro-symbolic spatio-temporal reasoning. *arXiv preprint arXiv:2211.15566v2*, 2022.
84. Alessandro Oltramari, Jonathan Francis, Filip Ilievski, Kaixin Ma, and Roshanak Mirzaee. Generalizable neuro-symbolic systems for commonsense question answering. *arXiv preprint arXiv:2201.06230v1*, 2022.
85. Danial Kamali, Elham J. Barezi, and Parisa Kordjamshidi. Nesycoco: A neuro-symbolic concept composer for compositional generalization. *arXiv preprint arXiv:2412.15588v1*, 2024.
86. Midhat Urooj, Ayan Banerjee, Farhat Shaikh, Kuntal Thakur, and Sandeep Gupta. Single domain generalization in diabetic retinopathy: A neuro-symbolic learning approach. *arXiv preprint arXiv:2509.02918v1*, 2025.
87. Gunjan Singh, Sumit Bhatia, and Raghava Mutharaju. Neuro-symbolic rdf and description logic reasoners: The state-of-the-art and challenges. *arXiv preprint arXiv:2308.04814v1*, 2023.
88. Evgenii Zheltonozhskii, Chaim Baskin, Alex M. Bronstein, and Avi Mendelson. Self-supervised learning for large-scale unsupervised image clustering. *arXiv preprint arXiv:2008.10312v2*, 2020.
89. Yunsung Lee, Teakgyu Hong, Han-Cheol Cho, Junbum Cha, and Seungryong Kim. Houghcl: Finding better positive pairs in dense self-supervised learning. *arXiv preprint arXiv:2111.10794v1*, 2021.
90. Andrius Ovsianas, Jason Ramapuram, Dan Busbridge, Eeshan Gunesh Dhekane, and Russ Webb. Elastic weight consolidation improves the robustness of self-supervised learning methods under transfer. *arXiv preprint arXiv:2210.16365v1*, 2022.
91. Nathaniel Simard and Guillaume Lagrange. Improving few-shot learning with auxiliary self-supervised pretext tasks. *arXiv preprint arXiv:2101.09825v1*, 2021.
92. Kwanghee Choi, Minyoung Choe, and Hyelee Lee. Pretraining neural architecture search controllers with locality-based self-supervised learning. *arXiv preprint arXiv:2103.08157v1*, 2021.
93. Evan Racah and Christopher Pal. Supervise thyself: Examining self-supervised representations in interactive environments. *arXiv preprint arXiv:1906.11951v1*, 2019.
94. Byoungjip Kim, Jinho Choo, Yeong-Dae Kwon, Seongho Joe, Seungjai Min, and Youngjune Gwon. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480v1*, 2021.
95. Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Continual state representation learning for reinforce-

ment learning using generative replay. *arXiv preprint arXiv:1810.03880v3*, 2018.
96. Jangho Kim, Jeesoo Kim, and Nojun Kwak. Stacknet: Stacking parameters for continual learning. *arXiv preprint arXiv:1809.02441v3*, 2018.
97. Mark Collier, Efi Kokiopoulou, Andrea Gesmundo, and Jesse Berent. Routing networks with co-training for continual learning. *arXiv preprint arXiv:2009.04381v1*, 2020.
98. Haoran Zhu, Maryam Majzoubi, Arihant Jain, and Anna Choromanska. Tame: Task agnostic continual learning using multiple experts. *arXiv preprint arXiv:2210.03869v2*, 2022.
99. Shikhar Srivastava, Md Yousuf Harun, Robik Shrestha, and Christopher Kanan. Improving multimodal large language models using continual learning. *arXiv preprint arXiv:2410.19925v2*, 2024.
100. Tom Diethe, Tom Borchert, Eno Thereska, Borja Balle, and Neil Lawrence. Continual learning in practice. *arXiv preprint arXiv:1903.05202v2*, 2019.
101. Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
102. Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
103. Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. Gradient imitation reinforcement learning for low resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2746, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
104. Qihao Liu, Xi Yin, Alan L. Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 2755–2765. Computer Vision Foundation / IEEE, 2025.
105. Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
106. Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
107. Yucheng Zhou, Jihai Zhang, Guanjie Chen, Jianbing Shen, and Yu Cheng. Less is more: Vision representation compression for efficient video generation with large language models, 2024.
108. Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics.
109. Yucheng Zhou and Guodong Long. Multimodal event transformer for image-guided story ending generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3434–3444, 2023.
110. Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. 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)*, pages 2579–2591, Online, 2021. Association for Computational Linguistics.
111. Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online, 2021. Association for Computational Linguistics.
112. Qingyao Ai, Jingtao Zhan, and Yiqun Liu. Foundations of genir. *arXiv preprint arXiv:2501.02842v1*, 2025.
113. Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States, 2022. Association for Computational Linguistics.
114. Yucheng Zhou, Jiahao Yuan, and Qianning Wang. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*, 2025.
115. Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
116. Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, and Amy Pavel. Controlling dialogue generation with semantic ex-

emplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3018–3029, Online, 2021. Association for Computational Linguistics.

117. Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
118. Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. Semantic representation for dialogue modeling. 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)*, pages 4430–4445, Online, 2021. Association for Computational Linguistics.
119. Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak 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*, pages 4903–4912, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
120. Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA, 2024. Association for Computational Linguistics.
121. Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
122. Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland, 2022. Association for Computational Linguistics.
123. Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. 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)*, pages 182–189, Online, 2021. Association for Computational Linguistics.
124. Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
125. Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland, 2022. Association for Computational Linguistics.
126. Wenbo Sui, Daniel Lichau, Josselin Lefèvre, and Harold Phelippeau. Incomplete multimodal industrial anomaly detection via cross-modal distillation. *arXiv preprint arXiv:2405.13571v4*, 2024.
127. Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online, 2021. Association for Computational Linguistics.
128. Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. 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)*, pages 827–838, Online, 2021. Association for Computational Linguistics.
129. Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, 2021.
130. Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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)*, pages 993–1003, Online, 2021. Association for Computational Linguistics.
131. Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland, 2022. Association for Computational Linguistics.
132. The Multimodal Universe Collaboration, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael

Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, and John F. Wu. The multimodal universe: Enabling large-scale machine learning with 100tb of astronomical scientific data. *arXiv preprint arXiv:2412.02527v1*, 2024.

133. Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592v4*, 2025.
134. Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States, 2022. Association for Computational Linguistics.
135. Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics.
136. Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. 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)*, pages 6153–6166, Online, 2021. Association for Computational Linguistics.
137. Casey C. Bennett and Kris Hauser. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *arXiv preprint arXiv:1301.2158v1*, 2013.
138. Guanxi Shen. Glimpse: Holistic cross-modal explainability for large vision-language models. *arXiv preprint arXiv:2506.18985v3*, 2025.
139. Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. *arXiv preprint arXiv:2110.01835v1*, 2021.
140. Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, 2021. Association for Computational Linguistics.
141. Weijie Xu, Jinjin Zhao, Francis Iannacci, and Bo Wang. Ffpdg: Fast, fair and private data generation. *arXiv preprint arXiv:2307.00161v1*, 2023.
142. Angus Addlesee and Pierre Albert. Ethically collecting multi-modal spontaneous conversations with people that have cognitive impairments. *arXiv preprint arXiv:2009.14361v1*, 2020.
143. Catherine Menon, Austen Rainer, Patrick Holthaus, Gabriella Lakatos, and Silvio Carta. Ehazop: A proof of concept ethical hazard analysis of an assistive robot. *arXiv preprint arXiv:2406.09239v1*, 2024.
144. Erika Halme, Mamia Agbese, Hanna-Kaisa Alanen, Jani Antikainen, Marianna Jantunen, Arif Ali Khan, Kai-Kristian Kemell, Ville Vakkuri, and Pekka Abrahamsson. Implementation of ethically aligned design with ethical user stories in smart terminal digitalization project: Use case passenger flow. *arXiv preprint arXiv:2111.06116v1*, 2021.
145. Franz Louis Cesista. Multimodal structured generation: Cvpr's 2nd mmfm challenge technical report. *arXiv preprint arXiv:2406.11403v2*, 2024.
146. Zhiyuan Ma, Jianjun Li, Guohui Li, and Kaiyan Huang. Cmal: A novel cross-modal associative learning framework for vision-language pre-training. *arXiv preprint arXiv:2410.12595v1*, 2024.
147. Yucheng Zhou and Guodong Long. Improving cross-modal alignment for text-guided image inpainting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3445–3456, 2023.
148. Minh Tran Phu and Thien Huu Nguyen. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online, 2021. Association for Computational Linguistics.
149. Yucheng Zhou, Xiubo Geng, Tao Shen, Jian Pei, Wenqiang Zhang, and Daxin Jiang. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.
150. Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. Eventbert: A pre-trained model for event correlation reasoning. In *Proceedings of the ACM Web Conference 2022*, pages 850–859, 2022.
151. Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin 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)*, pages 2559–2575, 2022.