Multimedia Cognition and Evaluation in Open Environments

Wei Feng*
Department of Computer Science and Technology, Tsinghua University
fw22@mails.tsinghua.edu.cn

Haoyang Li*
Department of Computer Science and Technology, Tsinghua University
lihy218@gmail.com

Xin Wang†
Department of Computer Science and Technology, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Xuguang Duan
Zi Qian
Department of Computer Science and Technology, Tsinghua University
dxg18@mails.tsinghua.edu.cn
qian-z20@mails.tsinghua.edu.cn

Wu Liu
JD Explore Academy
liuwu1@jd.com

Wenwu Zhu†
Department of Computer Science and Technology, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

ABSTRACT

Within the past decade, a plethora of emerging multimedia applications and services has catalyzed the production of an enormous quantity of multimedia data. This data-driven epoch has significantly propelled the trajectory of advanced research in various facets of multimedia, including image/video content analysis, multimedia search and recommendation systems, multimedia streaming, and multimedia content delivery among others. In parallel to this, the discipline of cognition, has embarked on a renewed trajectory of progression, largely attributing its remarkable success to the revolutionizing advent of machine learning methodologies. This concurrent evolution of the two domains invariably presents an intriguing question: What happens when multimedia meets cognition? To decipher this complex interplay, we delve into the concept of Multimedia Cognition, which encapsulates the mutual influence between multimedia and cognition. This exploration is primarily directed toward three crucial aspects. Firstly, the way multimedia and cognition influence each other, prompting theoretical developments towards multiple intelligence and cross-media intelligence. More important, cognition reciprocates this interaction by infusing novel perspectives and methodologies into multimedia research, which can promote the interpretability, generalization ability, and logical thinking of intelligent systems in open environments. Last but not least, these two aspects form a loop in which multimedia and cognition interactively enhance each other, bringing a new research problem, so that the proper evaluation for multimedia cognition in open environments is important. In this paper, we discuss what and how efforts have been done in the literature and share our insights on research directions that deserve further study to produce potentially profound impacts on multimedia cognition and evaluation in open environments.

*Both authors contributed equally to this research.
†Corresponding authors.

McGE '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0278-5/23/10.
https://doi.org/10.1145/3607541.3616823

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; *Question answering*; • **Computing methodologies** → *Computer vision.*

KEYWORDS

multimedia cognition, multimedia evaluation, open environments

ACM Reference Format:
Wei Feng, Haoyang Li, Xin Wang, Xuguang Duan, Zi Qian, Wu Liu, and Wenwu Zhu. 2023. Multimedia Cognition and Evaluation in Open Environments. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3607541.3616823

1 INTRODUCTION

Multimedia was first conceptualized as an innovative union of multiple media formats in 1960s, which has been developed for decades in both academia and industry [7, 28]. As it stands today, Multimedia has come to be comprehensively delineated as an interactive amalgamation of various electronic media, encompassing video, image, audio, and text components [48]. This definition, as currently encapsulated by the user-editable resource, Wikipedia, captures its dynamic and multifaceted nature. Moreover, the field of Artificial Intelligence (AI) emerged on the academic research horizon in the 1950s [40], and since then, it has witnessed significant advancements in a multitude of methodologies. Initially, these two critical research domains followed distinct paths, operating largely independently of each other. However, the proliferation of diverse multimedia data types has propelled the development of machine learning techniques, leading to the discovery of practical models capable of processing a wide range of real-world multimedia information. Consequently, these advancements have paved the way for the application of AI in various real-world scenarios. Therefore, multimedia intelligence through exploring the mutual influences between multimedia and AI has been proposed and widely studied [60].

Despite the notable success of multimedia intelligence, the existing literature ignores more realistic scenarios that the developed theories and approaches deployed in the wild, i.e., in the

open environments, leading to suboptimal performances and poor interpretability, generalization ability, logical thinking ability for intelligent systems in open environments [30, 31, 34, 57, 59].

More recently, cognition intelligence [37] as a research area for human-like artificial intelligence (AI), has also significantly attracted attention recently. It has seen the advent of multiple methodologies, inclusive but not limited to symbolic reasoning [9, 35], probabilistic models in Bayesian networks [12, 42], biological inspiration in the form of evolutionary algorithms [6], and more recently, the deep learning approach, a human and neural inspired paradigm neural-symbolic reasoning [14, 54] that has revolutionized the landscape of AI.

Multimedia cognition, which combines the strength of multimedia and cognition, has become a promising research direction and attracted an increasing number of interests from the community, spanning over a variety of machine learning methodologies and applications recently. On the one hand, the widespread availability of multimedia data has led to the emergence of numerous multimodal applications, including audio-visual speech recognition [1, 50], image/video captioning [45, 53, 55], and visual question answering [4, 17], etc. On the other hand, cognition research focusing on studying perception and reasoning can enhance the human-like reasoning characteristics in multimedia, resulting in more inferrable multimedia [60]. Facing opportunities as well as challenges, we believe it is the right time to review and promote the studies of multimedia cognition approaches and their evaluation, especially in real-world open environments. As a result, the convergence of multimedia and cognition gives rise to multimedia intelligence, creating a loop in which multimedia and cognition interact with one another, thus generating mutual influence and enhancement, which can now be applied to a wide range of real-world open-environment scenarios.

In this paper, we provide comprehensive and systematic descriptions of multimedia cognition in open environments. Firstly, we present some key theories of multimedia cognition in open environments, including multiple intelligence theory, and cross-media intelligence. Then, we summarize the existing methodologies into three categories based on the effectiveness in multimedia cognition tasks, i.e., open-environment disentanglement approaches for interpretability, open-environment invariant learning approaches for generalization ability, and open-environment reasoning approaches for logical thinking ability, and elaborate representative approaches in each category. Last but not least, we propose comprehensive multimedia cognition evaluations specifically for open environments as well as some experimental results, which could shed light on further research of multimedia cognition.

2 MULTIMEDIA COGNITION THEORIES IN OPEN ENVIRONMENTS

In this section, we provide a systematic summary of the multimedia cognition theories, which mainly consist of two parts, i.e., multiple intelligence theory and cross-media intelligence theory.

2.1 Multiple Intelligence Theory

In the field of Cognitive science and pedagogy, scientists have proposed a variety of intelligence theories to describe and evaluate the

Figure 1: The framework of multiple intelligence

level of human intelligence, such as unified intelligence theory, dual intelligence theory [25], ternary intelligence theory, and theory of multiple intelligences. Different intelligence theories have their own advantages and disadvantages as well as different scopes of use. Among them, Theory of multiple intelligences, proposed by Howard Earl Gardner in 1987, emphasizes that agents have multiple intelligences, which can more comprehensively evaluate the development potential and intelligence level of agents [16]. In the Theory of multiple intelligences shown in Figure 1, the intelligence of agents includes eight different types [49]:

- Verbal–Linguistic Intelligence: Refers to people's ability to utilize, understand, and express language. This intelligence is manifested in oral expression, written expression, and language learning, along with understanding, rhetoric, and speech.
- Musical Intelligence: Refers to people's ability to understand, create, and perform music. This intelligence is manifested in music perception, memory, expression and creation.
- Logical/Mathematical Intelligence: Refers to people's ability to analyze, reason, and solve problems. This intelligence is manifested in the learning and application of mathematics and science, logical thinking, problem-solving, and reasoning abilities.
- Visual/Spatial Intelligence: Refers to people's ability to process spatial information, image thinking, and imagination. This intelligence is manifested in the abilities of art, architecture, design, image processing, and geography.
- Bodily/Kinesthetic Intelligence: Refers to people's ability to exercise, coordinate, and control their bodies. This intelligence is manifested in sports, dance, handicrafts, etc.
- Intrapersonal Intelligence: Refers to people's ability of understanding themselves, self-control, and self-reflection. This intelligence is manifested in aspects such as self-awareness, emotional management, goal setting, and self-evaluation.
- Interpersonal Intelligence: Refers to people's ability in interpersonal communication and understanding of others. This intelligence is manifested in social skills, leadership, empathy and cooperation ability.
- Naturalistic Intelligence: Refers to the ability to understand and apply nature, including identifying and classifying living and non-living organisms, understanding natural laws, and mastering natural skills.

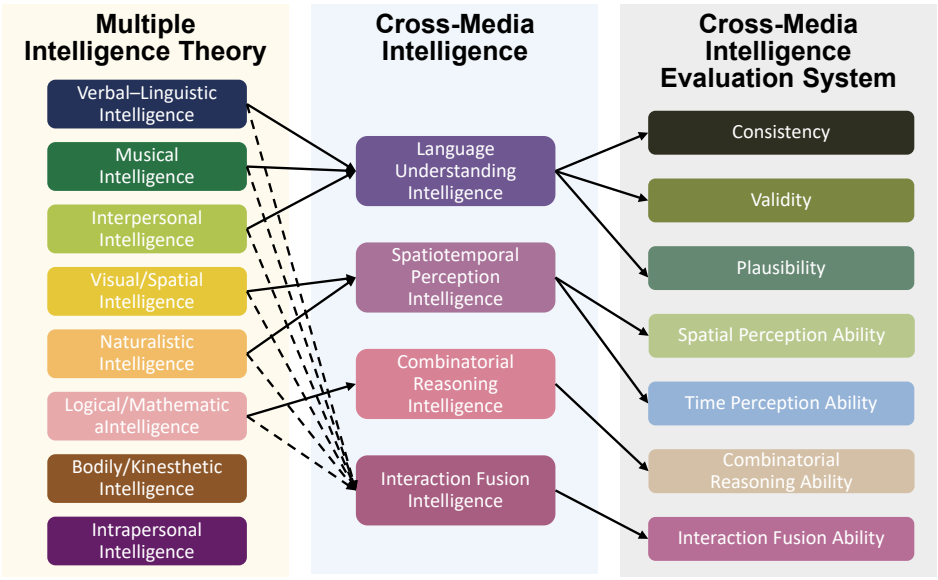

Figure 2: The framework of cross-media intelligence: basis, concept, and evaluation. Each key concept in cross-media intelligence involves some multiple intelligence theories, which are denoted by the arrows. And these key concept in cross-media intelligence is evaluated by some specific aspects, which are also represented by the arrows. Note that the arrows with the dashed line mean Interaction Fusion Intelligence might involve partial but not total connected theories from multiple intelligence.

2.2 Cross-Media Intelligence

In the Theory of Multiple Intelligence, the different types of intelligence above are not independent of each other, but interrelated and intertwined, forming the cross-media intelligence shown in Figure 2. Each agent also has strengths and weaknesses in different types of intelligence. Corresponding to the field of cross-media intelligence, we have summarized **four types** of intelligence that cross-media intelligence should have, as follows:

- Language Understanding Intelligence: Refers to the ability of an agent to process abstract language symbols (including words, musical symbols, and other symbols). This intelligence is manifested in Natural language processing [41], context question dialogue, cross-modal visual question answering [56] and other capabilities. The measurement of language understanding intelligence includes the traditional objective and subjective evaluation standards such as Perplexity and Crowd Sourcing. In this evaluation system, we further emphasize the "consistency", "legitimacy" and "credibility" of language understanding and Smart lock understanding results. These indicators not only measure the accuracy of language understanding results, but also examine the rationality, legitimacy and credibility.
- Spatial Perception Intelligence: Spatial perception mainly examines the ability of intelligent agents to correctly perceive semantics in images (video frames). This ability is closely related to the natural observation intelligence and visual-spatial intelligence of human intelligent agents, manifested in the ability of cross-media intelligent models in computer vision recognition, scene analysis, and other aspects. The specific measurement indicator is the intersection and union ratio between the attention region and the real region of the model under non-strong supervision.
- Time Perception Intelligence: Time perception ability, is the ability of an agent to understand context and causal order above the spatial dimension. When asked about specific concepts and problems, we hope that the agent can accurately locate the specific contextual video frames involved in the concepts and problems [27]. The specific measurement indicator is the intersection and union ratio between the time attention frame segments and the real frame segments of the model under non-strong supervision.
  Due to the fact that research on time perception and spatial perception often takes place simultaneously, we generally refer to both time perception intelligence and spatial perception intelligence as **spatiotemporal perception intelligence**.
- Combinatorial Reasoning Intelligence: Combinatorial reasoning intelligence is closely related to the logical and mathematical intelligence of human intelligence. We hope that agents can combine different modal information and execute reasoning according to the correct steps, thereby generating the correct syntax or reasoning tree. The specific measurement indicator is the consistency between the inference sequence syntax tree generated by the model and the real inference tree.
- Interaction Fusion Intelligence: Interaction fusion ability is an important difference between cross-media intelligence and other single-media intelligence. Cross-media intelligence not only needs to understand single-mode information, but also needs to align, interact, and fuse information from different modes. We use the accuracy of the model in the typical task of visual reasoning to measure the interaction fusion ability of the model.

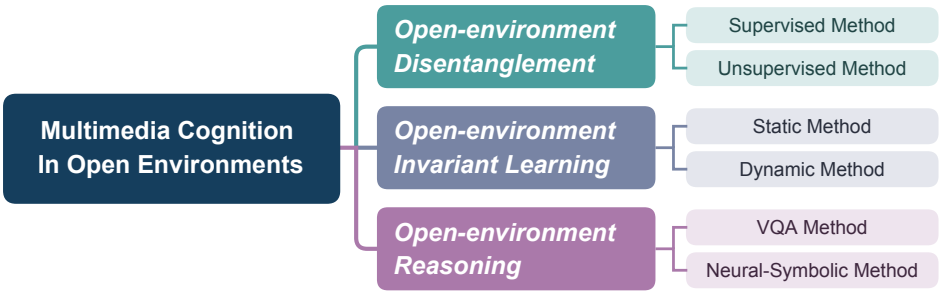

Figure 3: Overview of methodology categorization of multimedia cognition in open environments

3 MULTIMEDIA COGNITION APPROACHES IN OPEN ENVIRONMENTS

In this section, we discuss three main branches of techniques for multimedia cognition in open environments, which includes open environment disentanglement, invariant learning, and reasoning. In general, disentangled multimedia representation learning can separate the explanatory factors of variations behind the data, enhancing interpretability. And invariant learning aims to capture invariant relations between the entities and the labels, so that the generalization ability in open environments can be largely improved. Finally, reasoning techniques are also important for multimedia cognition, since they can assess things rationally by applying logic based on new or existing information when making a decision in open environments. Now we will describe these techniques in detail. The framework of multimedia cognition methods in open environments is shown in Figure 3.

3.1 Open-environment Disentanglement

The domain of disentangled representation learning has attracted significant interest for representation learning in open environments, particularly within the scope of multimedia representation learning [20, 47]. The primary objective of this discipline is to construct representations capable of isolating the causal factors behind data variations. It has been empirically demonstrated that these types of representations exhibit greater resilience to intricate variations, leading to enhanced generalization capabilities and improved robustness against adversarial attacks in open environments. In addition, disentangled representations inherently possess superior interpretability in open environments. Nevertheless, the task of learning representations that can disentangle latent factors is a relatively uncharted territory in the literature in terms of multimedia cognition in open environments. This prominent gap in knowledge underscores the need for further investigation into this crucial aspect of representation learning. Here we talk about two mainstream methods including disentanglement with and without labels (i.e., supervised and unsupervised disentanglement).

3.1.1 Disentanglement with Labels. The genesis of a real-world multimedia entity is typically a product of intricate interactions involving numerous latent factors. Traditional deep learning algorithms applied to multimedia data often overlook the interconnected nature of these latent elements, leading to the derived representations being both non-robust and challenging to interpret. Despite this, the task of generating representations that disentangle these latent factors remains largely uncharted territory within the realm of deep learning literature, which can serve as the basis of multimedia cognition. DisenGCN [38] introduces a novel approach termed as the disentangled graph convolutional network, which is designed to learn disentangled representations for structured entities. The primary innovation is an ingenious neighborhood routing mechanism. This system possesses the ability to dynamically pinpoint the latent factor potentially entities responsible for the creation of an edge between an entity and its adjacent instances. In response, it allocates the neighboring node to a specific channel, engineered to extract and convolute features that are peculiar to the identified factor. The convergence properties of this routing mechanism are not left to chance.

3.1.2 Disentanglement without Labels. Besides the supervised method above, the disentanglement without labels also receives much attention. The recent surge in interest and impressive results of self-supervised learning applications in open environments cannot go unnoticed in the realm of multimedia representation learning. Nevertheless, the creation of authentic-world multimedia entities is often an outcome of intricate interactions among an abundance of latent factors. Existing self-supervised learning approaches also tend to take a holistic perspective, inadvertently disregarding the intermingled nature of these latent elements. Consequently, this oversight results in subpar learned representations for downstream tasks, and their interpretability is considerably challenged in open environments. In the work of [32], the authors unveil a novel methodology dubbed Independence Promoted Disentangled Graph Contrastive Learning (IDGCL). This pioneering approach capitalizes on the self-supervision paradigm to learn disentangled representations for structured entities. Specifically, the method first discerns the latent factors intrinsic to the input graph and subsequently generates its factorized representations. It presents a unique factor-wise discrimination objective, implemented in the style of contrastive learning. This design is instrumental in compelling the factorized representations to independently encapsulate the expressive information emanating from distinct latent factors. To augment the independence amongst these representations, it adopts the Hilbert-Schmidt Independence Criterion [46] to eradicate dependencies that may exist among different representations. This criterion is seamlessly integrated within the self-supervised framework as a regularizer, thereby enhancing its efficiency. Finally, the disentanglement for the structured entities can largely improve the effectiveness of the representation learning in open environments.

3.2 Open-environment Invariant Learning

In addition to the disentanglement techniques, invariant learning is also an effective technique for multimedia cognition in open environments. We mainly talk about invariant learning for static and dynamic entities.

3.2.1 Invariant Learning for Static Entities. The capacity of multimedia representation learning in delivering effective results is clearly demonstrated when training and testing graph data are drawn from an identical distribution. Nevertheless, under conditions of distribution shifts, a significant proportion of existing methodologies exhibit limitations in their ability to generalize. It's noteworthy that the concept of invariant learning [5], which derives its foundational principles from causality, offers theoretical assurances of generalization under distribution shifts that widely exist in open environments, and has proven successful in practical scenarios [2, 10, 18, 26]. However, its implementation in the context of graph learning (GIL) [33], which is specifically designed to learn generalized representations under conditions of distribution shifts. The proposed methodology ingeniously captures invariant relationships between predictive structural entity information and their corresponding labels across different latent environments. This is achieved by jointly optimizing the specially designed modules. The robustness and generalization in open environments of the proposed method are guaranteed by theoretical validations.

3.2.2 Invariant Learning for Dynamic Entities. Besides, multimedia entities can be changing in open environments, leading to the research for dynamic entities. Taking the structural entities as examples, the prowess of dynamic graph neural networks in prediction tasks has been showcased through their ability to leverage both structural and temporal dynamics. However, an evident shortfall of the current dynamic graph neural networks is their lack of resilience against distribution shifts, which are the naturally occurring phenomenon in dynamic graphs in open environments. This limitation arises chiefly because the patterns exploited by DyGNNs tend to be variants with respect to labels under distribution shifts. Therefore, in the research work of [58], the authors pioneer an approach aimed at managing spatiotemporal distribution shifts in dynamic graphs. The approach revolves around the identification and utilization of invariant patterns - structures and features with predictive capabilities that remain stable despite distribution shifts. It faces two primary challenges: firstly, the discovery of complex variant and invariant spatiotemporal patterns in dynamic graphs, which incorporate fluctuating graph structures and node features. Secondly, the management of spatiotemporal distribution shifts using the unearthed variant and invariant patterns.

To tackle these challenges, the authors introduce a novel model Disentangled Intervention-based Dynamic graph Attention networks (DIDA). The proposed approach is adept at managing spatiotemporal distribution shifts in dynamic graphs by unearthing and fully exploiting invariant spatiotemporal patterns. This is achieved in three stages. Firstly, a disentangled spatiotemporal attention network is proposed to capture the variant and invariant patterns. Secondly, a spatiotemporal intervention mechanism is designed to generate multiple interventional distributions by sampling and reassembling variant patterns across neighborhoods and time stamps, thereby negating the spurious influences of variant patterns. Lastly, the authors introduce an invariance regularization term with the aim to minimize prediction variances in the intervened distributions, ensuring our model can generate predictions based on invariant patterns with stable predictive abilities, and hence effectively manage distribution shifts. The robustness and generalization ability in open environments of the proposed method is validated through experiments on three real-world datasets and a synthetic dataset, outperforming state-of-the-art baselines under distribution shifts. This research is the first study of spatiotemporal distribution shifts in dynamic graphs in open environments.

3.3 Open-environment Reasoning

Based on the tasks faced by Theory of multiple intelligences and Cross-Media Intelligence in an open environment, different multimedia reasoning models were proposed for solving specific problems, and corresponding theoretical systems were established.

3.3.1 Visual Question Answering. Visual Question Answering (VQA) aims to answer natural language questions of given images, and its task is normally free-form and open-ended [4, 52]. In the open and uncertain environment, artificial intelligence research faces challenges such as weak correspondence between separate texts and non-textual objects, or difficulty in understanding and reasoning based on the huge dimensions of corpora and image databases. In specific scenarios, VQA tasks can be divided into related subtasks such as TextVQA and VideoQA. For example, text visual question answering (TextVQA) aims to answer questions related to textual content present in the images [8]. To address TextVQA tasks, Liang et al. [36] proposed a novel multi-modal contextual graph neural network (MCG) model, which is able to capture the connections between visual characteristics of scene texts and non-textual objects in images. After encoding scene texts into richer features containing textual, visual and positional features, this model represents the visual relations between scene texts and non-textual objects using a contextual graph neural network, which outperforms the baseline approaches. For general scenarios and other specific scenarios, Perceptual Visual Reasoning (PVR) with Knowledge Propagation model and Dynamic Spatio-Temporal modular Network (DSTN) model were also proposed for VQA tasks [29, 43].

3.3.2 Neural-symbolic learning. Neural-symbolic learning aims to integrate the perceptual capabilities of neural perception and the reasoning abilities of symbolic logic [15]. Logic symbols are Constructed language symbols used to express logic forms and logic operations in logic. As a type of image symbol, they are widely used in the field of logical reasoning. Although traditional symbolic logical reasoning approaches have well-established theories and diverse applications for managing discrete logic, they are not specifically devised to handle semantic data, including raw images and text. Early neural symbol learning focuses on combining these two modules and optimizing them in isolation, which would be difficult to obtain the global optimal results. Thus, Duan et al. proposed DeepLogic [11], a joint learning model of neural perception and logical reasoning, which contains a deep-logic module (DLM) [39] and a deep&logic optimization (DLO) algorithm. Through this model, the perception component offers guidance for acquiring logic rules,

while the logic formulas obtained from the logical reasoning component serve as supervision for neural perception learning. Without using pre-existing tools, this proposed DeepLogic framework demonstrates superior performance over DNN-based baselines by a considerable margin and surpasses other strong baselines.

4 MATHEMATICAL DEFINITION OF EVALUATION SYSTEM

Specifically, the intelligent evaluation system includes five dimensions to measure the model's cross-media intelligence ability, namely language understanding ability, time perception ability, spatial perception ability, combination reasoning ability, and interaction fusion ability.

4.1 Language Understanding Intelligence

Language understanding intelligence includes three indicators, namely consistency, validity, and plausibility. These three indicators were first proposed by Husdon et al. in [24] to measure the language understanding ability of intelligent agents, and many subsequent works have continued to use these three indicators to measure the language related ability of their models [13, 43, 44].

4.1.1 Consistency. Consistency is used to measure the consistency between the answers generated by a model when facing different problems. We calculate this indicator using the following method.

Firstly, for a set of question-answer pairs (q, a), we define a derived question-answer set $E_q = \{(q_1, a_1), (q_2, a_2), \ldots, (q_n, a_n)\}$, where any set of derived question-answer pairs $(q_i, a_1), i = 1, 2, \ldots, n$ can be inferred from the original question-answer pair (q, a). For example, as shown in Figure 4, given the original question-answer pairs: "Is there a red apple on the right side of the white plate?" - "Yes", we can obtain the derived question-answer pairs: "Is the plate on the left side of the apple?" - "Yes", "Is there a plate on the left side of the red fruit?" - "Yes", etc. For each model that correctly answers the question $q \in Q$, we extend it to a derived question-answer set E_q, we measure the model's impact on the accuracy Acc_q of the questions contained in E_q , we calculate the consistency C using the following formula:

$$C = \frac{\sum_{q \in Q} Acc_q}{|Q|}, \tag{1}$$

where Q represents the set of correctly answered questions, and $|Q|$ represents the size of the set.

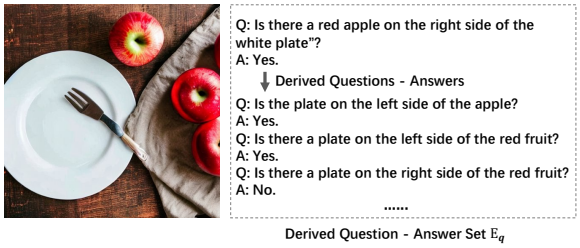

Figure 4: Example of consistency. All derived question-answer pairs are generated from the first original question-answer pair. If a model with high consistency answers one original question correctly, then it should not give wrong answers to the derivative questions.

4.1.2 Validity. Validity is used to check whether a given answer is within the scope of the question. For example, as shown in Figure 5, answering a question about a certain color. For a question, if the answer given by the model is within the scope of the question, we believe that the answer to the question is legal, otherwise the answer is illegal. We use the average legality rate as the numerical value of this indicator, as shown in the following formula:

$$C = \frac{\sum_{q \in Q} Valid_q}{|Q|}, \tag{2}$$

where $Valid_q$ represents whether the answer to question q is legal, Q represents all sets of questions, and $|Q|$ represents the size of the set.

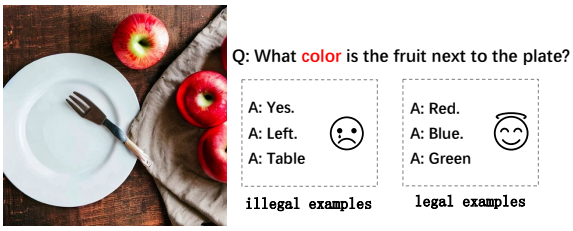

Figure 5: Example of validity. Regardless of right or wrong, answers that meet the requirements of the question are considered legal, otherwise illegal.

4.1.3 Plausibility. Plausibility is used to measure the overall level of model mastery of general knowledge. For example, as shown in Figure 6, when asked about the color of an apple, if the model provides an answer in red, green, or yellow, we believe the answer is trustworthy, otherwise we believe the answer is untrustworthy. We record the trusted answer set for question q as A_q, we calculate whether the answer appears at least once in the entire dataset and is related to the subject of the question. For question q, if the answer given by the model is in the trusted answer set A_q, we believe the answer is credible, otherwise the answer is not credible. We use the average credibility rate as the numerical value of this indicator, as shown in the following formula:

$$C = \frac{\sum_{q \in Q} Plause_q}{|Q|}, \tag{3}$$

where $Plause_q$ represents whether the answer to question q is trustworthy, Q represents all sets of questions, and $|Q|$ represents the size of the set.

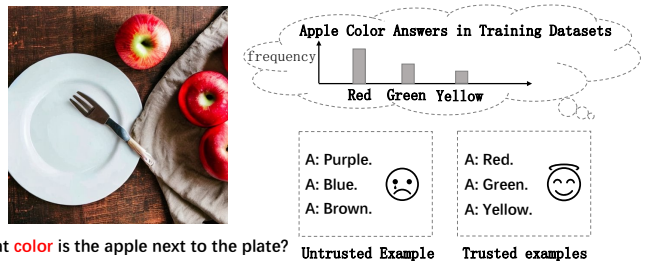

Figure 6: Example of plausibility. A question may have multiple answers, and all answers that have been statistically correlated with the question subject in the dataset are considered trusted, while others are considered untrusted.

4.2 Spatial Perception Intelligence

We hope that a model has the ability to correctly perceive semantic-related local regions in images (or video frames). Anderson et al. proposed in [3] that models need to have the ability to correctly focus attention on semantically related local regions. In addition, many works have incorporated attention mechanisms in model design to make models more focused on fine-grained semantically related regions [29, 43]. Specifically, as shown in Figure 7, when asked the question "What color of clothing does a girl wear?", we hope that the model can focus its attention on local areas related to the "clothes the girl wears" [29]. For each problem, we record the true attention value of the image area related to the problem as V^G , and the attention value generated by the model as V. We use the Intersection over Union (IOU) of the two to measure the spatial perception ability of the model [24]. We calculate the intersection ratio I using the following formula:

$$I = \frac{\sum_{i=0}^{N} V_i^G \times V_i}{N}, \tag{4}$$

where $N = w \times h$ represents the number of pixels in the image, w represents the length of the image, and h represents the width of the image.

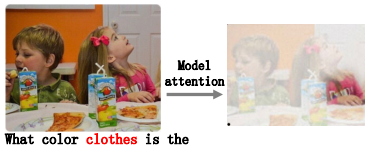

Figure 7: Example of spatial perception ability. We hope the model focuses on where the question subject refers.

4.3 Time Perception Intelligence

We hope that a model has the ability to correctly perceive semantic-related fragments in videos. Huang et al. proposed in [22] that the model should focus its attention on semantically related frames, and fine-grained attention should facilitate the model's subsequent logical reasoning. Specifically, as shown in Figure 8, when asked "What color of clothing does the woman cooking in the video wear?", we hope that the model can correctly locate the frame where the "woman cooking" exists. For certain semantics, we measure the time perception ability of the model by detecting the intersection over Union (IOU) between the frame where the semantics exist and the frame where the semantics actually exist. We calculate the intersection ratio I using the following formula:

$$I = \frac{\sum_{i=0}^{N} V_i^G \times V_i}{N}, \tag{5}$$

where N represents the number of frames in the video, $V \in \mathrm{R}^{N\times 1}$ represents the probability of the model detecting the existence of semantic frames, and $V^G \in \mathrm{R}^{N\times 1}$represents the probability of the frame where the semantic truly exists.

Figure 8: Example of time perception ability. We hope the model focuses on when (or which frames) the question subject refers to.

4.4 Combinatorial Reasoning Intelligence

For a combinatorial problem that requires multi-step reasoning, we hope that the model can infer according to the correct reasoning steps. As shown in Figure 9 which provides a combinatorial inference problem and its corresponding inference graph examples, we hope that the model has the ability to generate correct syntax trees or inference graphs [29, 43]. For a syntax tree, we can use the Reverse Polish notation to express it as a sequence structure. Therefore, we record the sequence syntax tree structure generated by the model as $T = [t_0, t_1, \ldots, t_n]$, and the real sequence syntax tree structure as $T^G = [t_0^G, t_1^G, \ldots, t_n^G]$. We use the average sequence accuracy index to measure the model's combinatorial reasoning ability. Among them, the sequence accuracy $SeqAcc_q$ is the ability to generate accurate syntax trees by calculating the model in the test set, when $T = T^G$, we have $t_i = t_i^G, \forall i \in [0, 1, \ldots, n]$. We believe that the model produces an accurate syntax tree. At this point, the indicator S can be calculated using the following formula:

$$S = \frac{\sum_{q\in Q} SeqAcc_q}{|Q|}, \tag{6}$$

where Q represents all problem sets, and $|Q|$ represents the size of the set.

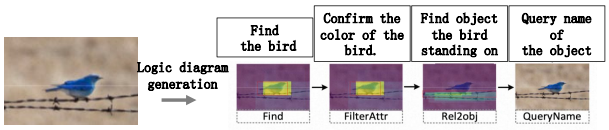

Figure 9: Example of combinatorial reasoning ability. We hope the model can generate sequential logical inference structures step by step for combinatorial reasoning.

4.5 Interaction Fusion Intelligence

The interactive fusion ability reflects the model's ability to process cross-modal information. In visual Q&A tasks, this ability mainly reflects the model's ability to generate correct answers, and we use accuracy indicators to measure this ability [3, 21, 23, 29]. Calculate using the following formula:

$$S = \frac{\sum_{q\in Q} Acc_q}{|Q|}, \tag{7}$$

where Acc_q represents whether the answer to question q is correct, Q represents all sets of questions, and $|Q|$ represents the size of the set.

5 EXPERIMENTS AND RESULTS

Based on the measurement dimensions mentioned above, We present the results of existing models such as Perceptual Visual Reasoning (PVR) with Knowledge Propagation model and Dynamic Spatio-Temporal modular Network (DSTN) model, and their baseline models in Table 1.

5.1 Datasets

The detailed evaluations were conducted on the GQA dataset [24] and the AGQA dataset [19], respectively, to measure the performance of recently proposed models from the five measurement dimensions mentioned above. The GQA dataset was proposed by Hudson et al. in 2019 and is a real image Q&A dataset containing massive inference questions [24]. This dataset was built with the development of their strong question engine that creates diverse

Measurement Dimensions	Performance	Representative Baselines	Multimedia Cognition Approaches
Language Understanding Ability	Consistency	83.64 (N2NMNs [21])	85.85 (PVR [29])
	Validity	96.29 (N2NMNs [21])	96.47 (PVR [29])
	Plausibility	84.57 (Bottom-Up [3])	84.96 (PVR [29])
Time Perception Ability	IOU	36.30(Random [43])	59.70 (DSTN [43])
Spatial Perception Ability	IOU	88.29 (MAC [23])	97.44 (PVR [29])
Combinatorial Reasoning Ability	Consistency	-	99.84 (DSTN [43])
Interaction Fusion Ability	Accuracy	55.44 (N2NMNs [21])	57.33 (PVR [29])

Table 1: Experimental results. We report the ability results of existing cross-media cognitive methods in the evaluation system on the Multimedia Cognition Approaches set, while their baselines' results are on the Representative Baselines set. Measurement indicators are labeled on the Measurement Dimensions and Performance set. 'Random' means the method which randomly generates predictions for lack of available baselines.

reasoning questions through graph structures of the Visual Genome scene. Correctly answering the questions in this dataset requires the model to have good language understanding ability, spatial perception ability, combined reasoning ability, and interactive fusion ability. The AGQA dataset [19] was proposed by M Grunde McLaughlin et al. in 2021 and is a real video Q&A dataset that involves a large number of spatiotemporal reasoning problems. It contains 192M unbalanced question-answer pairs for 9.6K videos and a balanced subset of 3.9M question-answer pairs.

5.2 Perceptual Visual Reasoning

Novel module-based methods face challenges such as inadequate explainability and logical inference capabilities. Undoubtedly, the gap between these early investigations and actual human reasoning is still substantial. To compensate for the lack of sufficient explainability and logical inference in traditional VQA research, Li et al. proposed a module-based approach called the Perceptual Visual Reasoning (PVR) model for real-world visual reasoning [29]. The PVR model addresses the real-world visual reasoning problem by breaking down a given question into multiple interconnected sub-tasks and progressively addressing these sub-tasks. Firstly, a collection of neural modules was designed for specific functionalities such as localizing relevant visual regions, performing logical inference and generating answers. Each module is capable of incorporating external guidance information to specialize its functionality. Secondly, a tree-based modular layout of hierarchical neural modules for reasoning was developed that integrates low-level visual perception and high-level logic inference within a unified framework. Finally, modules in the layout would be dynamically formed into a modular neural network. After each module in the network received outputs from its child modules and generated output to its parent module, the final answer of the VQA task would be obtained from the topmost module. These design choices promote an understandable and compositional reasoning process, helping the PVR model produces transparent, explainable intermediate results.

5.3 Dynamic Spatio-Temporal modular Network

As an extension of VQA, Video Question Answering (VideoQA) aims to correctly answer questions given the related videos instead of static images. Compared with VQA tasks, VideoQA requires more reasoning operations due to the massive dataset on both temporal and spatial scales [51]. In 2022, the novel dynamic spatio-temporal modular network (DSTN) model was proposed [43], as the initial approach using modular neural networks in VideoQA for interpretable video reasoning in real-life situations. Specifically, the proposed DSTN model first utilizes a hierarchical logic structure to decompose the given question systematically into multiple subtasks. These sub-tasks encompass essential concepts such as object, subject, relation, location, action, temporal order, and duration. Secondly, to address diverse sub-tasks, multiple modules would be introduced with distinct functionalities that encompass temporal and spatial localization, logical reasoning, relation exploration, and more. These modules are flexibly combined into a modular network with a hierarchical logical structure, enabling enhanced logical reasoning capabilities. In the end, the integrated modular neural network concurrently processes textual features and visual features in a progressive approach to produce the final answer of the VideoQA task. Additionally, it carries out comprehensive experiments to showcase the benefits of DSTN employing diverse metrics and configurations, and analyzes the performance of distinct modules to validate their rationale and the overall model's interpretability.

Based on the results of the baseline model in VQA-related tasks, we can summarize that in the multimedia cognition field, the PVR model and DSTN model demonstrate leading levels of language understanding ability, time perception ability, spatial perception ability, combination reasoning ability, and interaction fusion ability.

6 CONCLUSION

In this paper, we reveal the convergence of multimedia and cognition in open environments. We present the novel concept of *Multimedia Cognition* which explores the co-influence between multimedia and cognition via presenting the related theories, methodologies, and practical evaluations. With the development of large language models in the past few years, the interpretability, generalization ability, and logical thinking of intelligent systems in open environments can be largely improved by the novel multimedia cognition approaches, which would promote the further development of AI.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2018), 8717–8727.

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems* 34 (2021), 3438–3450.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[6] Thomas Bäck and Hans-Paul Schwefel. 1993. An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation* 1, 1 (1993), 1–23.

[7] Atta Badii, David Fuschi, Ali Khan, and Adedayo Adetoye. 2009. Accessibility-by-design: a framework for delivery-context-aware personalised media content re-purposing. In *HCI and Usability for e-Inclusion: 5th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2009, Linz, Austria, November 9-10, 2009 Proceedings 5*. Springer, 209–226.

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4291–4301.

[9] Rodney A Brooks. 1981. Symbolic reasoning among 3-D models and 2-D images. *Artificial intelligence* 17, 1-3 (1981), 285–348.

[10] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning*. PMLR, 1448–1458.

[11] Xuguang Duan, Xin Wang, Peilin Zhao, Guangyao Shen, and Wenwu Zhu. 2022. DeepLogic: Joint Learning of Neural Perception and Logical Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4321–4334.

[12] Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine learning* 29 (1997), 131–163.

[13] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 6616–6628.

[14] Artur d'Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088* (2019).

[15] Artur S d'Avila Garcez, Krysia Broda, and Dov M Gabbay. 2002. *Neural-symbolic learning systems: foundations and applications*. Springer Science & Business Media.

[16] Howard Gardner. 1987. The theory of multiple intelligences. *Annals of dyslexia* (1987), 19–35.

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[18] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[19] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11287–11297.

[20] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230* (2018).

[21] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 804–813.

[22] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11021–11028.

[23] Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067* (2018).

[24] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[25] Scott Barry Kaufman. 2009. *Beyond general intelligence: The dual-process theory of human intelligence*. Yale University.

[26] Masanori Koyama and Shoichiro Yamaguchi. 2020. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883* (2020).

[27] Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. 2023. Curriculum multi-negative augmentation for debiased video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[28] Baochun Li, Zhi Wang, Jiangchuan Liu, and Wenwu Zhu. 2013. Two decades of internet video streaming: A retrospective view. *ACM transactions on multimedia computing, communications, and applications (TOMM)* 9, 1s (2013), 1–20.

[29] Guohao Li, Xin Wang, and Wenwu Zhu. 2019. Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th acm international conference on multimedia*. 530–538.

[30] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022. OOD-GNN: Out-of-Distribution Generalized Graph Neural Network. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[31] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987* (2022).

[32] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2022. Disentangled Graph Contrastive Learning With Independence Promotion. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[33] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2022. Learning Invariant Graph Representations for Out-Of-Distribution Generalization. In *Advances in Neural Information Processing Systems*.

[34] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2023. Invariant Node Representation Learning under Distribution Shifts with Multiple Latent Environments. *ACM Transactions on Information Systems* (2023).

[35] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. 2018. Symbolic graph reasoning meets convolutions. *Advances in neural information processing systems* 31 (2018).

[36] Yaoyuan Liang, Xin Wang, Xuguang Duan, and Wenwu Zhu. 2021. Multi-modal contextual graph neural network for text visual question answering. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 3491–3498.

[37] Huimin Lu, Yujie Li, Min Chen, Hyoungseop Kim, and Seiichi Serikawa. 2018. Brain intelligence: go beyond artificial intelligence. *Mobile Networks and Applications* 23 (2018), 368–375.

[38] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International conference on machine learning*. PMLR, 4212–4221.

[39] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori. 2019. Integrating learning and reasoning with deep logic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 517–532.

[40] Marvin Minsky. 1961. Steps toward artificial intelligence. *Proceedings of the IRE* 49, 1 (1961), 8–30.

[41] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18, 5 (2011), 544–551.

[42] Judea Pearl. 2011. Bayesian networks. (2011).

[43] Zi Qian, Xin Wang, Xuguang Duan, Hong Chen, and Wenwu Zhu. 2022. Dynamic spatio-temporal modular network for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4466–4477.

[44] Ruixue Tang and Chao Ma. 2020. Interpretable Neural Computation for Real-World Compositional Visual Question Answering. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part III 3*. Springer, 89–101.

[45] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[46] Tinghua Wang, Xiaolu Dai, and Yuze Liu. 2021. Learning with Hilbert–Schmidt independence criterion: A review and new perspectives. *Knowledge-based systems* 234 (2021), 107567.

[47] Xin Wang, Hong Chen, and Wenwu Zhu. 2021. Multimodal disentangled representation for recommendation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[48] Wikibooks. 2023. Introduction to Computer Information Systems/Multimedia — Wikibooks, The Free Textbook Project. https://en.wikibooks.org/w/index.php?title=Introduction_to_Computer_Information_Systems/Multimedia&oldid=4233927 [Online; accessed 20-July-2023].

[49] Wikipedia. 2023. Theory of multiple intelligences. https://en.wikipedia.org/wiki/Theory_of_multiple_intelligences [Online; accessed 20-July-2023].

[50] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, 2 (2013), 153–163.

[51] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. 2003. VideoQA: question answering on news video. In *Proceedings of the eleventh*

ACM international conference on Multimedia. 632–641.

[52] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3480–3491.

[53] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515.

[54] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems* 31 (2018).

[55] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4584–4593.

[56] Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition* 108 (2020), 107563.

[57] Ziwei Zhang, Xin Wang, Zeyang Zhang, Peng Cui, and Wenwu Zhu. 2021. Revisiting Transformation Invariant Geometric Deep Learning: Are Initial Representations All You Need? *arXiv preprint arXiv:2112.12345* (2021).

[58] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. 2022. Dynamic graph neural networks under spatio-temporal distribution shift. *Advances in Neural Information Processing Systems* 35 (2022), 6074–6089.

[59] Zeyang Zhang, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2022. Learning to solve travelling salesman problem with hardness-adaptive curriculum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9136–9144.

[60] Wenwu Zhu, Xin Wang, and Wen Gao. 2020. Multimedia intelligence: When multimedia meets artificial intelligence. *IEEE Transactions on Multimedia* 22, 7 (2020), 1823–1835.