Computer Science Review 48 (2023) 100548

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Visual language integration: A survey and open challenges

Sang-Min Park [a,1], Young-Gab Kim [b,*]

[a] Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
[b] Department of Computer and Information Security, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 28 July 2021
Received in revised form 10 October 2022
Accepted 21 February 2023
Available online 2 March 2023

Keywords:
Multimodal learning
Multi-task learning
End-to-end learning
Embodiment
Visual language interaction

ABSTRACT

With the recent development of deep learning technology comes the wide use of artificial intelligence (AI) models in various domains. AI shows good performance for definite-purpose tasks, such as image recognition and text classification. The recognition performance for every single task has become more accurate than feature engineering, enabling more work that could not be done before. In addition, with the development of generation technology (e.g., GPT-3), AI models are showing stable performances in each recognition and generation task. However, not many studies have focused on how to integrate these models efficiently to achieve comprehensive human interaction. Each model grows in size with improved performance, thereby consequently requiring more computing power and more complicated designs to train than before. This requirement increases the complexity of each model and requires more paired data, making model integration difficult. This study provides a survey on visual language integration with a hierarchical approach for reviewing the recent trends that have already been performed on AI models among research communities as the interaction component. We also compare herein the strengths of existing AI models and integration approaches and the limitations they face. Furthermore, we discuss the current related issues and which research is needed for visual language integration. More specifically, we identify four aspects of visual language integration models: multimodal learning, multi-task learning, end-to-end learning, and embodiment for embodied visual language interaction. Finally, we discuss some current open issues and challenges and conclude our survey by giving possible future directions.

Contents

* Correspondence to: Department of Computer and Information Security, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea.
E-mail addresses: wiyard@korea.ac.kr (S.-M. Park), alwaysgabi@sejong.ac.kr (Y.-G. Kim).
[1] *Sang-Min Park* is with the Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea.

https://doi.org/10.1016/j.cosrev.2023.100548
1574-0137/

1. Introduction

An artificial intelligence (AI) model precepts various types of objects, inferences recognized perceptions and generates expressions based on representation in multiple forms (e.g., rap generation [1], image-to-image transfer [2], story writing [3], game playing [4], and dancing [5]). AI agents understand better human intentions with several AI models based on the current information and the previous history. They also make novel and creative responses based on the agent's own persona to enhance interactions. Fast multimodal recognition [6], self-supervised learning [7], and persona generation [8] enhance the visual language interaction between humans and agents. However, the purpose of each AI model is segmented and specified for each task rather than focused on a comprehensive interaction with humans. Therefore, we must study integrated agents that perform a seamless and consistent interaction by integrating fragmented AI models. A model that grows in size with an improved performance requires more computing power and more complicated designs to train than before. This increases the complexity of each model and requires more paired data, making model integration difficult. Therefore, increasing interest is being spent on multimodal learning that integrates various modals, multi-task learning that performs various tasks as one model, and the end-to-end (E2E) model that processes multiple consecutive tasks at once.

The integration is also actively discussed in the visual language domain. From the early days of deep learning, a vision model has been developed with stable performance and various applications in image classification and image synthesis. In the language domain, with the recent advent of bidirectional encoder representations from transformers (BERT) [9] and generative pre-trained transformer 3 (GPT-3) [10], pre-trained models perform good results for various downstream tasks (e.g., question and answer (QA), natural language inference). Due to the success in the vision and language domain, interest in visual language integration that connects and simultaneously utilizes the two domain inputs is growing. Since each domain model has been developed based on an interoperable solution (e.g., attention, Transformer), integration is performed more easily than before, but the visual language integration has problems and limitations (e.g., concurrency, paired training data). Therefore, domain knowledge for each domain and an extensive understanding of scalable solutions are required for visual language integration.

In the future, the AI model needs more studies that understand complex sentiments and interactively express them rather than just performing modal integration and conversion. The response needs a more expressive and artistic activity like music and dances, not only for the purpose of conveying meaning. In addition, a persona is used to keep the context transferred in a consistent style, and novelty, which is a critical evaluation element for generation, makes an expression creative. The consistency and diversity of novel expressions immerse humans more and induce better interaction.

Based on these concepts summarized so far, we classify the visual language integration into functional classification with representation learning, decision making, and generative interaction, as shown in Fig. 1.

Representation learning extracts features and represents them in vector space for visual language tasks. It is an advantage over using the raw input directly because it simplifies the input and takes advantage of the similarity with the distance vector. Decision making deals with how to infer and make a decision necessary to complete visual language tasks. Lastly, generative interaction was extended to an embodiment way to express the result or response for the visual language tasks. Rule-based interactions provide a more stable answer, but it requires a lot of manual work and has limitations in expression, so we focus on generative interactions.

We divide each step (i.e., representation learning, decision making, and generative interaction) into four aspects with information integration methodologies: multimodal learning, multi-task learning, end-to-end learning, and embodiment [11]. It is similar to the basic steps in the natural language processing domain (i.e., understanding, analyzing, and generating text). Multimodal learning simultaneously learns multiple modals (e.g., vision, language) in the same space [12–24]. Multimodal learning is important because people get information compounded from multiple modals (e.g., context, pose, tone of voice) rather than unimodal (e.g., instruction). The multimodal representation mainly deals with vision and language rather than other modals (e.g., sound). This survey deals with inference between the multimodal input and diverse multimodal expressions. When a plurality of tasks is required to perform a process, performing the plurality of tasks with one model is more advantageous than using each model. In particular, it affects coherence and complexity in comprehensively performing modal tasks of different characteristics (e.g., visual language). Multi-task learning obtains prior knowledge from common characteristics of tasks by processing multiple tasks simultaneously [25–27]. It has the advantage of reasoning with relatively little data in decision making. In addition, pre-trained models are used for various downstream tasks with fine-tuning and few-shot learning as multi-task learning. Since multi-task learning focuses on processing multiple tasks in one model in parallel, integrating sequential processes is also required in the visual language domain. E2E simplifies sequential tasks into one model or pipeline instead of cascade models [28–34]. The advantage of this approach is that it reduces the complexity of the tasks and information losses incurred in cascade models. As the interest in robots and RL agents grows, research that includes actions (e.g., embodied question and answer (EQA) [35–37] and visual language navigation (VLN) [38–41]) is increasing. Because one dimension of tasks (i.e., action)

Table 1
Organization for visual language integration.

	Multimodal learning	Multi-task learning	End-to-end learning	Embodiment
3. Representation Learning	Modal alignment Cross-modal attention Scene graph	Pre-trained language model Joint representation model Multimodal pre-trained model	Acoustic signal understanding Domain knowledge understanding Spoken language understanding	Co-reference resolution Anaphora resolution Exophora resolution
4. Decision Making	Multihop reasoning Relational reasoning Graph reasoning	Multi-agent RL Imagination-augmented RL Language-grounded RL	Meta RL Graph RL Planning RL	Episodic memory Intrinsic motivation Theory of mind
5. Generative Interaction	Multimodal interaction Persona generation Multimodal story	Multi-task interaction Knowledge distillation Visual QA	End-to-End interaction Sequential integration Spoken language translation	Embodied interaction Embodied QA Visual language navigation
6. Discussion Challenges	Fast modal conversion	Few-shot learning	Life-long learning	Immersive interaction

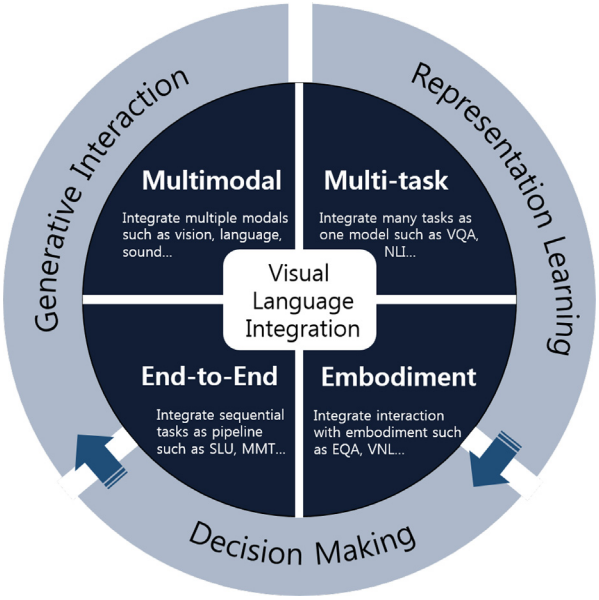

Fig. 1. Diagram of visual language integration with multimodal, multi-task, end-to-end learning, and embodiment.

increases, the process is more complex than other recognition and tasks. However, the more complex the processes, the greater the integration effect. Embodiment induces user immersion by considering the agent as an entity and interacting with the environment through a physical body within that environment, not just a virtual program [42–45]. It needs a lot of integration from sensor data to body motion.

This survey describes the current issues and applicable methodologies in detail based on the latest paper. Each of the integration mentioned above is covered in some other surveys [7,46]; however, these surveys deal with visual language integration only to complete tasks (e.g., visual question and answer (VQA) [47–55], EQA, VLN). Although they issued the problem of composing tasks that require integration at present, they have not broadly dealt with the necessary technologies as perspectives of future expansion. To the best of our knowledge, no existing work has dealt with comprehensive integration and broader perspectives from classification to neuroscience concepts for integration in the aspect of an embodied visual language interaction. Therefore, in this research, we present a detailed taxonomy and a comprehensive survey on the integration for embodied visual language interaction. We classify the study of visual language integration as a matrix that consists of three sequential steps (i.e., representation learning, decision making, and generative interaction) and four components (i.e., multimodal learning, multi-task learning, end-to-end learning, and embodiment), as depicted in Table 1. For each element, the significant features (e.g., multimodal pre-trained model) are selected, and the feature has some items (e.g., VideoBERT [56], ViLBERT [57]) as depicted in Fig. 2. Fig. 2 illustrates the schematic tree diagram of visual language integration. Table 1 shows the organization for visual language integration, and Table 2 presents a list of the main acronyms.

The remainder of this article is organized as follows. Section 2 summarizes the approaches of visual language integration as four components. Section 3 provides the representative learning of the perception layer, which is the first of the three main stages. Section 4 summarizes various methods of inference and decision-making based on the representation received through the perception layer. Section 5 deals with various generative interactions to effectively represent inference information. Section 6 presents and comprehensively discusses technologies that lack limitations and study in an aspect of the implementation of visual language integration. Finally, Section 7 concludes the paper.

2. Approaches to visual language integration

This section summarizes visual language integration from four perspectives: multimodal learning, multi-task learning, end-to-end learning, and embodiment. Four perspectives are used to investigate each step (i.e., from representation learning to generative interaction) in Sections 3–5. We show a hierarchical view of each approach and summarize the latest research for further understanding.

Multimodal learning. Humans do not interpret only the meaning of utterances when they communicate with others. The ambiguity of the language is compensated for to determine the speaker's underlying intention based on the direct and indirect representations of the surrounding environment. For example, emotional recognition, which is the beginning of emotional interaction, uses multimodal fusion to compensate for the shortage of context in textual information [6]. For vision, instead of merely recognizing objects, detailed information is obtained to interpret the accurate meaning of the image by identifying the text in the image. From video captioning to video generation, multimodal learning generates it beyond the limitations of media. Multimodal learning has been studied in various directions with unimodal integration and expansion in Fig. 3.

Multimodal effectiveness has constraints due to the performance deterioration issue caused by the inconsistency between modals. Unimodal models have a similar performance compared to multimodal training in some studies [12]. Elliott et al. [13] divided the method of analyzing images in multimodal into three types: double-attention mechanisms, global image vector, and element-wise multiplication. Through experiments on each influence, they insist that there was a performance improvement when using multimodal, but there was no significant difference

Fig. 2. The schematic tree diagram of visual language integration.

Table 2
List of main acronyms.

Acronym	Full Form	Acronym	Full Form
aGCN	Attention based graph convolution network	MMT	Multimodal machine translation
AGI	Artificial general intelligence	Monet	Multi-object network
APM	Actions per minute	NMT	Neural machine translation
AR	Augmented reality	PBT	Population based training
ASR	Automatic speech recognition	PDDL	Planning domain definition language
BERT	Bidirectional encoder representations from transformer	PLM	Pre-trained language model
Clevr	Bompositional language and elementary visual reasoning	PPDDL	Probabilistic planning domain definition language
CNN	Convolutional neural networks	R-CNN	Regions with convolutional neural networks
DCN	Deep circuit network	RDDL	Relational dynamic influence diagram language
DSTC	Dialogue system technology challenge	RL	Reinforcement learning
E2E	End-to-end	RNN	Recurrent neural networks
EVLI	Embodied visual-language interaction	SAC	Soft actor–critic
FVTA	Focal visual-text attention network	SLAM	Simultaneous localization and mapping
GAN	Generative adversarial network	SLU	Spoken language understanding
GNN	Graph neural networks	SNS	Social network service
GPT	Generative pre-training	UCB	Upper confidence limit
HPO	Hyperparameter optimization	VGDS	In video-grounded dialogue systems
HRL	Hierarchical reinforcement learning	VIN	Value repetition network
KB	Knowledge base	VLN	Vision-language navigation
KG	Knowledge graph	VPLM	Video pre-trained model
LSTM	Long short-term memory models	VR	Virtual reality
MARL	Multi-agent reinforcement learning	CDWE	Convolution–deconvolution word embedding

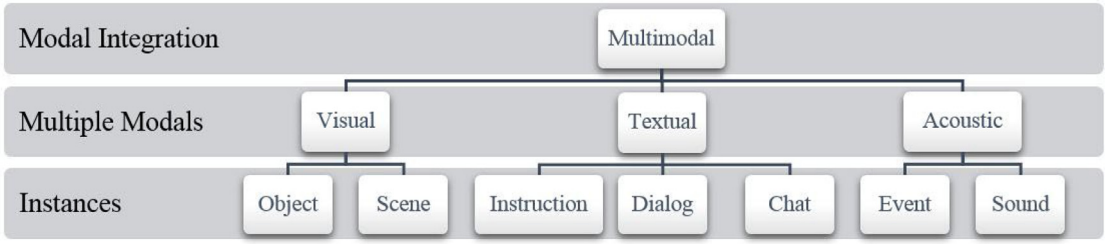

Fig. 3. Hierarchical view of multimodal learning with instances, multiple modals, and modal integration.

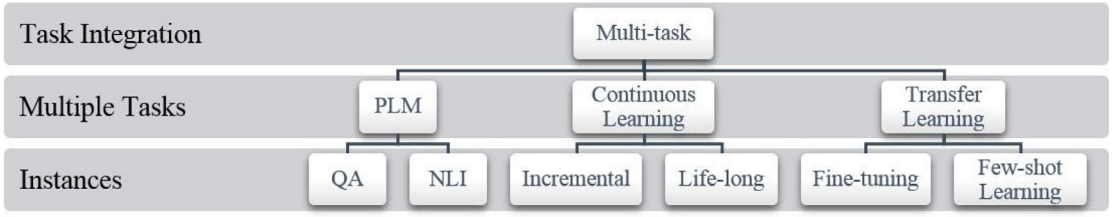

Fig. 4. Hierarchical view of multi-task learning with instances, multiple tasks, and task integration.

in the result using only textual information. In the multimodal translation, the dataset Multi30k [58] for measuring the performance of the translation due to the visual context is too simple and consists of short sentences, so it sufficiently renders the context even with a unimodal context.

Conversely, some studies [14,15] show that multimodal compensates for the insufficient context of unimodal for generating a better response or obtaining a more robust result against noise. The multimodal translation is one of the examples for the expansion of unimodal multilingual translation, which is the most representative task in the language domain. Delbrouck et al. [16] showed improved performance in multimodal translation when Flickr30K Entities was used to translate English sentences and images into German with attention models. Caglayan et al. [12] also showed that the translation's robustness is improved by utilizing various visual inputs classified as color deprivation, entity masking, progressive masking, and visual sensitivity. Elliot et al. [14] proposed an adversarial evaluation that is resistant to noise by comparing appropriate and inappropriate visual data when using sentences and visual context.

In terms of multimodal methodology, a convolutional neural network (CNN), a vision-based technology, is used to process languages more simply and faster than a recurrent neural network (RNN). Shuang et al. [59] proposed that Convolution–Deconvolution Word Embedding (CDWE) extends the deconvolution, widely used in vision, to generate word embeddings. Conversely, language models and transformers are commonly used in image recognition and generation. As a similar solution is used as a cross-modal, heterogeneity for integration between modals has gradually reduced, such as spoken language translation, which connects a text decoder to the automatic speech recognition (ASR) encoder.

From a data scalability perspective, many paired data are required to deal with various modals spontaneously. However, the multimodal dataset has a small-scale problem and lack of diversity due to difficulty in data collection and the limitation of annotation. To solve this problem, Zadeh et al. [17] proposed Dynamic Fusion Graph (DFG), a multimodal fusion for sentiment analysis with dataset CMU-MOSEI composed of 1000 speakers and 250 topics. As mentioned earlier, as the number of modal dimensions increases, the number of paired data, the batch size dimension, and the modal synchronization are the main limitations of multimodal training.

The multimodal integration is divided into early fusion, late fusion, and pre-trained models depending on the modal integration stage. Early fusion has the advantage of the computational amount being reduced by fusing the modal early; however, the detailed information of each modal is lost, and the paired training data are required. On the contrary, late fusion is based on the high-performance results determined in each modal; hence, the computational burden is small, and the paired data are not as necessary as those in early fusion. Zhang et al. [60] used late fusion to explicitly investigate the influence of each function, considering three types of visual, spatial, and semantic. The visual object is classified with CNN, and the spatial feature has the coordinates of two objects that encode the spatial layout, and finally, the semantic function utilizes the class labels of two objects rather than predicates. However, late fusion has a problem that cannot be comprehensively judged when the results are received in each modal conflict. For more accurate analysis, Liang et al. [18] proposed relative sentiment analysis, named multimodal local–global ranking fusion (MLRF) for a complex combination of visual and acoustic. They did not merely classify emotions as scalar values but ranked after measuring an increase or decrease of emotional intensity for the partial video segment. In another way, pre-trained models (e.g., VideoBERT) use large-scale training data, so modal information is used in a comprehensive way.

Multi-task learning. For tasks with insufficient training data, the joint training of similar tasks was recently conducted to obtain a better performance based on common language characteristics. PLM After its creation, a large pre-trained language model (PLM) (e.g., BERT, GPT-2, and GPT-3) is used in downstream tasks by applying fine-tuning or few-shot learning [61]. This PLM is trained with mass data; hence, it shows a stable and high performance in downstream tasks. The training requires massive data and computing resources as the PLM size increases. However, the model of various downstream tasks obtains a better performance with a relatively small size by using the PLM. For example, T5, which integrates a textual task in one model, is proposed to process translation, QA, and so on [62]. In addition, the conditional transformer language generates various forms of sentences (e.g., wiki, horror, and humor) from prefixes without retraining [63].

Multi-task learning in Fig. 4 compensates for tasks with insufficient data by spontaneously training various tasks. The types of multi-task learning are divided into PLM, continuous learning, and transfer learning, depending on the learning time of the

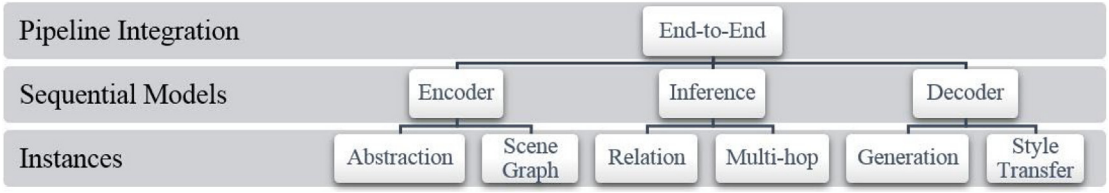

Fig. 5. Hierarchical view of end-to-end learning with instances, sequential model, and pipeline integration.

model. While PLM uses a model learned for multi-tasks defined in advance, continuous learning increasingly updates models by overlapping to an existing model [9,10]. In addition, transfer learning reflects and utilizes domain knowledge in the existing PLM model to process tasks that are not defined in advance [46,64].

For multilingual translation, a model that spontaneously translates various languages with multi-task learning achieves relatively high performance, even with a few resource pairs. Domhan et al. [25] proposed joint training for mass unpaired language and a small number of language pairs to improve neural machine translation (NMT) performance. This multi-task learning also has a model regularization effect, making it robust against noise and surpassing ensemble systems. For this reason, multi-task learning is used in various fields that lack actual human data to compensate. For example, it is difficult to obtain real-world dialog among humans, but the multi-task dialog is more difficult. To compensate for this problem, Budzianowski et al. [65] proposed a dataset MULTIWOZ classified for various domains and topics in terms of belief tracking, dialogue act, and response generation for crowd-sourcing-based structure, analysis, and collection procedures.

Reinforcement learning (RL) is closest to artificial general intelligence (AGI) than other methods, and studies on multi-tasks are actively conducted. Since RL is deficient in sample efficiency and is difficult to generalize, there are many studies on generalization to train spontaneously and inference generally in multi-tasks (e.g., Atari 57). Hessel et al. [26] insist that multi-tasks are more complicated than single tasks because multi-tasks models are balanced between various tasks under limited representation capacity. They proposed a scale-invariant actor–critic algorithm that updates experience in terms of reward scale, reward sparsity, and agent competency. Besides, their agent surpasses average-performance humans by a single training policy for Atari 57, which is regarded as various sequential decision tasks, using additional tasks, such as instant reward prediction and automatic encoding pixel control. On the contrary, unlike a human, when multi-task learning is used to perform various tasks, it encounters catastrophic forgetting that erases the parameter values trained in the previous task. It is relatively easy to use in similar tasks but easily over-fitting when the target domain data are insufficient.

End-to-End learning. Multimodal learning represents various modal inputs as one representation, while multi-task learning trains various tasks in one model. Representative examples of multimodal learning include the multimodal sentiment analysis that analyzes human emotions in terms of face, sound, and utterance in dialog tasks and VQA that spontaneously processes visual information and textual queries. Meanwhile, a representative example of multi-task learning is a pre-trained model used for downstream tasks (e.g., QA and natural language inference). Besides, the E2E integration of various modules is actively studied according to the sequential time and process flow, as depicted in Fig. 5. For example, E2E learning is developing in the form of Tacotron [66], which converts modal from text to voice. Furthermore, Translatotron [67] translates from voice input to voice output with a sequential process. The cascaded model is limited to features classified through feature engineering for each model; thus, the dimension of representation is reduced. Compared to the cascaded model, the E2E model has the advantage of utilizing most of the input without losing data in the intermediate process. For example, Translatotron performs an interpretation, including the foreign language's original pronunciation and sentiment meaning. It also has the advantage of responding in a voice form that reflects the prosody of the actual speaker.

Various advantages have been mentioned, but the performance of the E2E model is inferior to a model optimized for each module. Therefore, many studies [41,68] have been proposed in which the performance is improved by augmenting the training data to compensate for insufficient data or by providing an appropriate intermediate task to use for supplementary loss. Besides, a domain adaptation is needed to create dialog responses, even with small data in an environment where domain data are insufficient, because a large dataset is required to respond with consistently and grammatically correct sentences. So, Li et al. [69] proposed E2E multiple domain task and fast domain adaptation with Convlab [70] in dialogue system technology challenge-8 (DSTC-8). As such, the E2E model has the advantage of performing various sequence modules at once as one model while facing the problems of bottleneck and inconsistency between individual constituent elements.

Embodiment. The multimodal, multi-task, and temporal E2E models describe the development direction of the current AI model. In terms of interaction, many studies have performed tasks, such as image classification and language inference, based on the recognized vision and language. However, research on integrated interactive solutions is still insufficient. For example, the interactive dialog remains a challenging domain because it needs to understand and reflect its emotions. The embodiment has an essential meaning in interaction because a difference exists between a virtual agent and a visible person. While the purpose of a VQA is to answer a text question about a given image, the EQA performs a task analyzing the sensor information obtained by the embodied agent through proactive navigation. For example, to answer questions about a car's color far away, agents recognize and respond by moving proactively based on prior knowledge of where the car is and what route to go [42]. Moreover, it is necessary to have the ability to recognize complex situations based on various input sensors, as shown in Fig. 6 and comprehensively interpret the current situation using prior knowledge based on short- and long-term memory. These EQA tasks have recently expanded in the form of a dialog, in which the agent supplements by querying the oracle for insufficient information to perform the task.

Another advantage of the embodiment is that it is expressed more effectively in the form of gestures and dances using the body. Embodiment enhances the interaction between humans and agents further; however, it is not easily spread due to the lack of detailed simulation tools to reduce the gap between the real world and simulation.

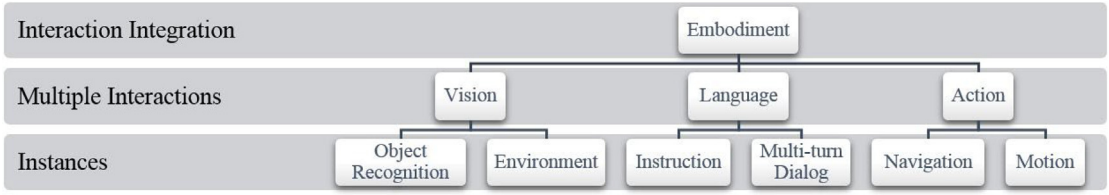

Fig. 6. Hierarchical view of embodiment with instances, multiple interactions, and interaction integration.

3. Representation learning

The beginning of visual language interaction is accurate recognition. Representation is important in effectively expressing the recognized result in a form suitable for task execution. We must consider the alignment between modals and the long dependency of a certain period to integrate these representations comprehensively. Modal alignment normalizes the different sampling ratios and resolutions of various modals. Furthermore, when the number of dimensions increases proportional to the number of input modals, cross-modal attention reduces the computation amount in multimodal learning. In addition, the agent adaptively creates a representation according to the downstream task because there are tasks that require fast processing as abstractions (e.g., multi-agent game) and require fine-grained inference (e.g., scene graphs). Scene graphs are a great way to utilize relationship information structurally between objects and body parts [71–76]. In particular, it attracts attention as an effective method for reducing recognition (e.g., abnormal body movements, body overlap). Like an implicit representation of language as a high-level goal for RL, the complexity of multi-tasks is reduced when several components are integrated into one system. In multi-task learning, a pre-trained model (e.g., BERT encoder [9]) simplifies the input by representing the input in common representation space. Joint representation encodes multiple modals in the same representation space. It is advantageous for multi-tasks that require simultaneous inputs from different modal. Furthermore, a multimodal pre-trained model jointly trains multiple modals from the beginning of the process (i.e., representation learning). It provides a low-complexity approach to multi-task learning of vision and natural language.

In addition, E2E training reduces information loss, cascaded errors, and bottlenecks caused by model inconsistency. In the past, it was necessary to design each model (e.g., feature extraction, acoustic model) to change an acoustic signal into a representation. E2E processing of these submodules reduces complexity and model size. Also, domain knowledge-based methods are used to increase the accuracy since the meaning of the acoustic representation varies depending on the task and context. Furthermore, with the emergence of research that simultaneously processes voice signal conversion and the meaning of the voice signal with the E2E model, more accurate recognition is possible due to the voice representation reflecting the natural language meaning.

From the embodiment point of view, humans communicate through gestures and expressions, which are non-verbal elements as well as language. Co-reference resolution serves to link different words representing arbitrary entities in natural language understanding. Anaphora resolution enables a more accurate understanding of the meaning by finding the noun corresponding to the pronoun in the representation of the previous sentence. Furthermore, Exophora Resolution resolves the combined representation of vision and language for objects pointed to by humans. It converts the pointed object into a word and performs the instruction by combining it with natural language instruction. Exophora resolution is scalable that handles more complex commands (e.g., instruction with pointing objects) in VLN and EQA tasks.

Thus, representation learning is closely related to visual language integration. This section focuses on studies of scene graph, language-grounded RL, spoken language understanding (SLU), and Exophora resolution to effectively construct a representation in Fig. 7.

3.1. Representation learning: Multimodal

How to combine information from other modals without losing detailed information is an important problem in multimodal; therefore, the alignment for integrating modal representation is frequently studied. However, the long-term dependency, different frequencies, and training complexity are the major constraints of alignment. To solve the alignment constraints, modal alignment, cross-modal attention, and scene graph are used. As shown in Fig. 8, multimodal learning simplified the representation with modal alignment and cross-modal attention by converting the elements of the image into a scene graph.

Modal alignment. Social network services (SNS) studies perform multimodal learning to complement the meaning of using different modals (e.g., YouTube and Instagram). Text and images in the SNS post do not have the same meaning but contain more complex meanings through meaning multiplication [19,20,77]. Multimodal learning is especially effective when the meaning of image and text are divergent. As described previously, the multimodal used in various domains, including SNS, has several problems due to the diversity of modality. Each modal is simultaneously collected over time in the perception layer, but sometimes the modal information becomes noise due to conflicting results in other modals.

When multimodal integration, a temporal mismatch occurs because the resolution and frequency of each modal are different. It depends on the model input and implementation, so modal alignment is important to solve the inconsistency problem of multimodal learning. To solve the limitations of alignment, several alternative studies [6,15,21] use low-level fusion and a long short-term memory model (LSTM).

Cross-modal attention. LSTM is an RNN-based method that reflects past information using an implicit representation but has a limitation that cannot reflect long-term global modality. To solve the long-term dependency problem of LSTM, an attention-based multimodal transformer is proposed in multimodal sentiment analysis and video-grounded dialogue systems [21]. The multimodal transformer model adjusts potential cross-reaction to fuse multimodal training. A semi-supervised non-linear mapping based on cosine similarity is also used to compare contextualized sentence embedding of text language model and feature map of CNN layer in the same space [22].

Scene graph. Modal alignment and cross-modal attention play an important role in multimodal learning. The method of independently predicting and processing each element takes processing time in proportion to the number of objects, and the higher the resolution, the more burdensome. In particular, in an environment (e.g., game, home) where a large number of various objects

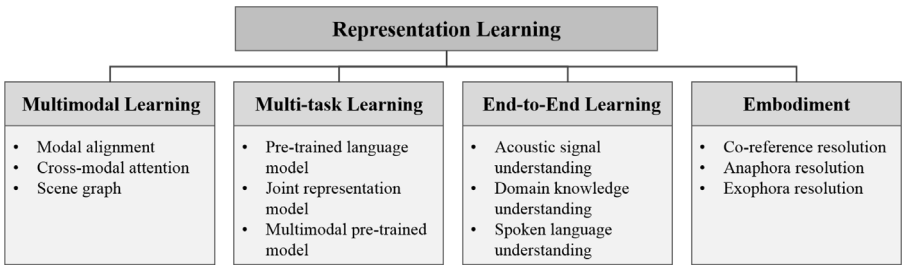

Fig. 7. Component view of the representation learning.

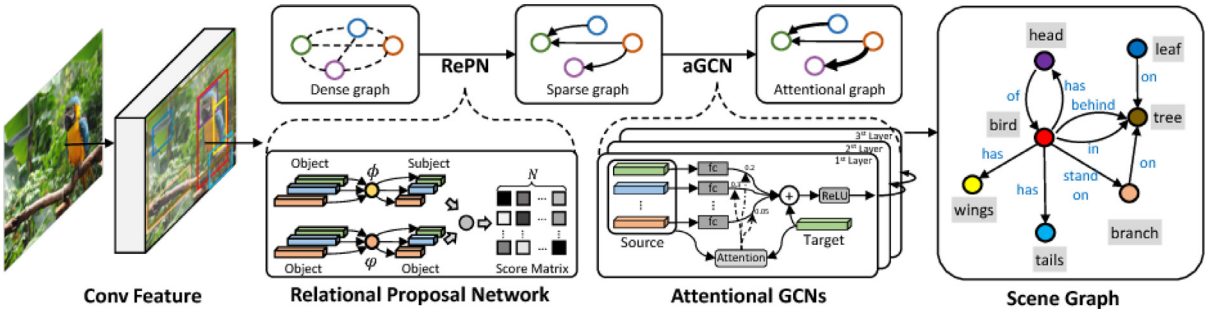

Fig. 8. The pipeline of relational proposal network with Graph R-CNN and scene graph [76]; The Region Proposal Network finds the object region. Relational Proposal Network (RePN) uses the found object region as a node and selects edges connecting the node. Attentional Graph Convolutional Network (aGCN) creates a scene graph by combining contextual information from neighboring nodes and giving weights to edges with importance.

need to be processed quickly, the relationship between objects is used to narrow the scope of recognition objects and increase the accuracy. Recently, for fine-grained image analysis, scene graph studies have been conducted that analyze each component of the image as a node and use the relationship as a link. Using the scene graph, the information of vision is described in more structural detail [71]. For example, when a scene graph is generated to analyze the object interaction of an image element in a surveillance video, high anomaly detection performance is obtained [72].

The scene graph interprets the behavioral meaning more accurately by reflecting the body's structural characteristics when the human body overlaps. CNN and global context encoding are used to capture asymmetric dependencies and context patterns between objects in real-time multi-party 3D motion capture and pose estimation [73,74]. Although it is able to capture the real-time 3D motion of difficult scenes with a single-color camera and separate the body structure of humans (e.g., shaking hands), there are still limitations to capturing very close interactions (e.g., hugs).

To compensate for the limitation of these scene graphs, GAN generates robust training data, and attention is used to extract important features from the mass data. GAN compensates with the generated image using label-free data and creates various graphs of semantic consistency in a large-scale text [75]. In addition, the scene graph has a problem in that the scale increases unrealistically depending on the number of objects. An attention-based graph convolution network is used to find the potential relation of the second-order between objects in the image by calculating the relativity score between pairs of objects [76]. Attention has the advantage of capturing context information between objects and relations with object node extraction, relation edge pruning, and graph context integration.

3.2. Representation learning: Multi-task

The domain where multi-tasking is most widely used is the language domain. The PLM, which is made of tremendous learning data and computing power, is breaking the state-of-the-art record of many natural language tasks (e.g., BERT, GPT-3). It is expanding its scope of use from relatively simple QA tasks to more complex tasks (e.g., visual dialogs). Natural language as a representation is implicit; hence, it is used in a multi-task (e.g., language-grounded RL) employed as a reward for reinforcement learning [78,79]. To solve the alignment complexity of multimodal, research on pre-trained models (e.g., videoBERT) has recently emerged, which reduced the complexity of learning multimodal and multi-task.

Joint representation model. Joint representation processes input data in the same space in a task, including multiple modals. An audio–visual training model uses the joint representation to separate objects and sounds. The multimodal clustering network is used to synchronously cluster the multimodal vectors of the convolution map in the shared space and train E2E to match the image and sound with maximum margin loss [23]. DSTC8's Audio Visual Scene-Aware Dialog (AVSD) is a task that generates responses for video chatting with the ESC-50 dataset consisting of 50 categories and 50 five-second audio clips [80]. Performance is improved when simultaneously performed in text and video tasks, but performance deteriorates when summary text is not used [24].

Multimodal pre-trained model. Joint representation training for each task requires more resources and is more complex than single-task representation training. Therefore, PLM based on joint representation training provides a positive effect in reducing training time and broadening the scope of representation even in multimodal tasks. Integrated PLMs are used for various tasks, and combinations of PLMs are used for each task's characteristics and different update cycles. With the success of PLM in the language domain, research in multimodal PLM has also increased. Especially in the case of video, the video pre-trained model (VPM) that jointly trains multimodal data of vision and text is effectively used in downstream tasks with low complexity (e.g., VideoBERT [56], ViLBERT [57]). VPM performs visual question response, visual commonsense reasoning, reference representation, and caption-based image searches. BERT-based VPM

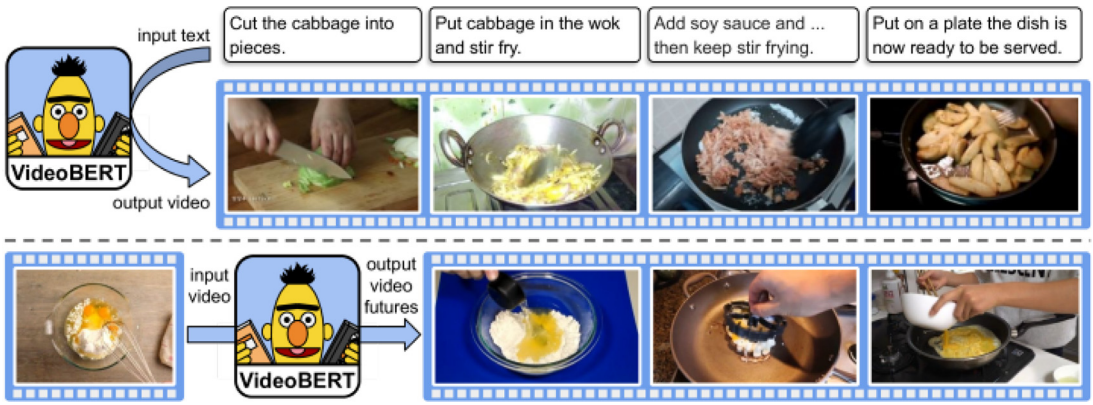

Fig. 9. Sequence flow of VideoBERT model [56]; (A) Recipe Illustration is the task of generating a video when given a textual cooking recipe. (B) Future frame prediction is the task of creating the next step into a video.

performs vector/quantization of video data and trains bidirectional joint distribution for visual and language token sequences in Fig. 9. It distinguishes individual objects and attributes in a spatially detailed visual representation in real-world applications (e.g., cooking). In a multimodal analysis, the vision domain has a weaker contextual relation than the text-domain. Vision models are composed of deeper networks, so a more detailed design is required for text training. These video-based BERT models are not yet widely used like language, but it is a necessary technology to reduce the load of downstream tasks and increase accuracy with pre-training representation.

3.3. Representation learning: End-to-end

From a multimodal perspective, research and interest in sound are less than language or vision. The tagging process is more complex than other tasks and is not used directly like language or vision; therefore, the sound is mainly used to enrich the environmental context or handle specific events for indirect purposes. It is used in VLN to limit the talker's location or distinguish where the agent's current location. In the game, the sound of an event is triggered as an intent for a specific action; however, so far, it is naive. Previously, the voice signal was converted into text through ASR and used as an intent through natural language processing. However, an SLU model has been proposed to analyze the intention from voice signals without conversion.

E2E learning and hierarchical learning are used to reduce the limitations of structural approaches (e.g., scene graph). Multiple-scale situational context is estimated in detail using a hierarchical structure with a semi-supervised method. Hierarchical learning serves to simplify the E2E model by hierarchically constructing shared representation (e.g., sub-graph). A sub-graph scene generation clusters the object pairs into sub-graphs by clustering and shares representations [81]. E2E learning is also used to capture the interaction between objects in the input image and generate the entire graph with related embedding [82].

Spoken language understanding. With the recent increase in interest in SLUs, attempts have been made to understand and translate meanings without interpreting voices as text. However, two major problems in the SLU are still being encountered: robustness to ASR errors and data scarcity of newly expanded domains. To recover ASR errors, language models and domain ontology, including domain knowledge and linguistic information, are also used [83]. SLU uses the information of persona or pitch that is lost while converting from voice signal to text, so it is possible to grasp a more accurate meaning that the previous conversion method does not have.

3.4. Representation learning: Embodiment

Coreference resolution is the task of finding expressions that refer to the same entity within a sentence. Anaphora resolution is a task that analyzes the words of the previous sentence pointed to by pronouns [84,85]. In a conversation consisting of short sentences, anaphora resolution is necessary to understand the conversation. Furthermore, exophora resolution performs an embodied instruction (e.g., come here) using additional information (e.g., hand pointing).

Anaphora resolution. Anaphora resolution and co-reference resolution are used for inferring cross-references in questions and dialogs [88,89]. Recently, this anaphora has gone beyond simple sentence-level analysis and are widely used in multimodal content (e.g., video) and SNS service (e.g., Twitter) [86,87]. Visual cross-reference in movies localizes the characters and learns the character's ground by associating descriptions and visual shapes. Cross-referencing relations on Twitter mostly arise from the answer and pronoun anaphora. It requires external knowledge, so it is difficult for the conversation agent to solve it alone.

Exophora resolution. Furthermore, Exophora applying the concept of embodiment, specifies the object by voice, text, and action pointing. A person points an object in a non-verbal form by a hand pointing it instead of language. If anaphora simply connects meaning between texts, in the case of Exophora resolution, specific instructions are performed from a multimodal interaction perspective by conferencing motion and speech.

3.5. Discussion: Representation learning

This section summarizes methods for making a multimodal representation (see Table 3). Scene graphs are an excellent approach to compensate for the explainable properties that have emerged as limitations of the neural net model. Various attempts have been made to classify bodies in overlapping situations and predict human poses behind walls using scene graphs.

In addition, the generated training data are used to increase the recognition accuracy, and a point network recognizes the entire object with a part of the view. These recent methods of solving multi-tasks with E2E and multimodal data have lowered the technical barrier of integration. However, in terms of integration, temporal order consideration, state-specific rewards, and multi-level commands are important tasks to be solved. Language is a unique abstraction used to define rewards in a language-grounded RL clearly. A study of Exophora resolution, which is a non-language instruction, also requires more studies but is challenging to develop because it requires visual, auditory,

Table 3
Descriptions and comparisons of representation learning for visual language integration.

	Topics	Papers	Approaches	Techniques	Models	Descriptions
3.1	Modal alignment	Libovický et al. (2018) [20]	Video summary	Social analysis	Content F1	Content F1 metric evaluation; summarize task to measure suitability
		Kruk et al. (2019) [19]	Multiplication	Classification	MDID	Multimodal Document Intent Dataset (MDID); divided into situational classification and semiotic classification
	Cross-modal attention	Le et al. (2019) [34]	Attention dialog response	VGDS	MTN	Multimodal Transformer Networks (MTN) in Video-Grounded Dialogue Systems (VGDS)
		Tsai et al. (2019) [21]	Low-level fusion	Pairwise attention	MulT	Cross-modal attention-based Multimodal Transformer (MulT); adjusts cross-reaction to fuse multimodal
		Akbari et al. (2019) [22]	Multimodal attention	Image-sentence	DCN function map	Deep Convolutional Neural Network (DCN), Semi-supervised nonlinear mapping
	Scene graph	Newell et al. (2017) [82]	Associative embeddings	Visual genome	Pixels-to-Graphs	E2E learning the related embedding; generate the entire graph by capturing the interaction
		Zellers et al. (2018) [73]	Body structure	Visual genome	SMN	Stacked Motif Network (SMN); encoding a global context and capturing asymmetric dependencies
		Yang et al. (2018) [76]	Quick scaling	Relational analysis	REPN	Relational Proposal Network (REPN); calculates the relativity score between pairs of objects
		Li et al. (2018) [81]	Subgraph	Visual genome	SMP	Spatial-weighted Message Passing (SMP); Subgraph-based scene graph generation; reduced by clustering the relation
		Qi et al. (2019) [75]	Semi-supervised	Graph generation	KE-GAN	Knowledge Embedded Generative Adversarial Networks (KE-GAN); generates graphs of semantic consistency in text
		Mehta et al. (2020) [74]	Selective skip connection	3D motion capture	SelecSLS Net	Multi-party 3D motion capture and pose estimation in real-time for human joints
3.2	Joint representation model	Hu et al. (2019) [23]	Clustering	Audio–visual	DMC	Deep Multimodal Clustering (DMC); Unsupervised audio–visual training model; separate objects and sounds
		Li et al. (2021) [24]	Language generation	Scene-aware dialog	Universal multimodal transformer	Learns joint representations
	Multimodal Pre-trained model	Sun et al. (2019) [56]	Pretrained model	Video LM	VideoBERT	Joint visual language model; vector quantization
		Lu et al. (2019) [57]	Pretrained model	Visual language	ViLBERT	BERT-based multimodal stream model
3.3	Spoken language understanding	Li et al. (2019) [83]	Acoustic and domain knowledge	SLU	Robust SLU	Dictionary function in domain ontology; recover from ASR errors
3.4	Anaphora resolution	Rohrbach et al. (2017) [86]	Video ground	Cross-referencing	Joint attention	Attention and two-level clustering; to learn the character's ground for visual cross-referencing
		Akta et al. (2018) [87]	Pronoun anaphora	Cross-referencing	–	Cross-referencing relationships; sharing the experience with the pronoun anaphora in Twitter
		Niu et al. (2019) [88]	Visual attention	Cross-referencing	RvA	Recursive Visual Attention (RvA); for inferring cross-references between questions and dialog records
		Sukthanker et al. (2020) [89]	Anaphora	Co-ref. resolution	–	Data and metrics of anaphora resolution and co-reference resolution

and 3D spatial mapping. Therefore, studies on a general-purpose simulation are needed in advance for use in various interactive environments.

4. Decision making

Deep RL is successful in making decisions about sophisticated games (e.g., Atari and Go), but actual decision making requires inference from complex visual observations. The AI agent performs tasks (e.g., classification, reasoning, and translation) using visible target objects and implicit prior knowledge, short/long-term memory, and history in Fig. 10. The types of reasoning vary from relational reasoning, causal inference, and physical reasoning. The inference in symbolic AI requires a formal representation form, such as first-order logic, strict integrity, and computing power [90]. However, with the recent development of studies [91–94] that apply graphs to neural nets, the barrier to reasoning has been lowered. In addition, causal inference has been conducted in a neural net in conjunction with the boom of explainable AI. Among various inferences, this paper deals with multi-hop reasoning, which infers multiple hops required to answer in-depth questions as a natural language task, and relational reasoning, which uses relationships between entities, which is important in object recognition in the recent vision series. Recently, for more structural relational reasoning, graph reasoning has been used.

In performing complex tasks, the agent must plan the tasks to be executed according to the context. In addition, considering that the interest in multi-agents has recently increased, studies [95,96] have been conducted to perform tasks by defining the relationship between agents into collaboration and hostility. The multi-agent RL enhances through knowledge sharing, memory, abstraction, and language-grounding. In multi-tasks, it is difficult to generate training data for each task, so interest

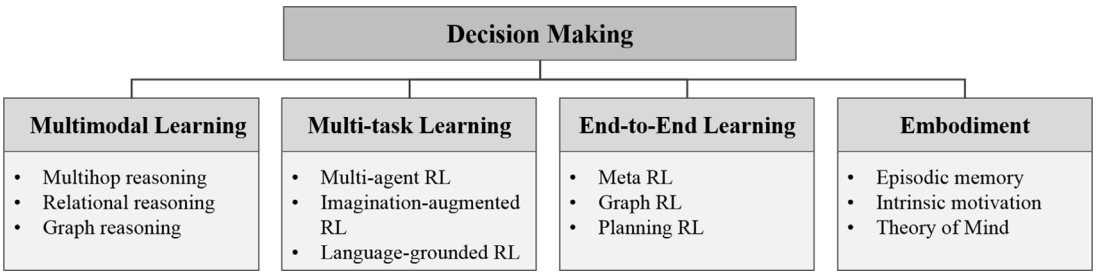

Fig. 10. Component view of decision making.

in effective data utilization (e.g., imagination-augmented RL) is growing. In addition, language-grounded RL performs complex tasks (e.g., VLN) by simultaneously dealing with natural language tasks and movement tasks [97–99].

In decision making, E2E learning is a way to reduce the system's complexity. Meta RL is recently used mainly for generalization by increasing the sample efficiency, and it is widely used to find optimal parameters and structures with E2E learning [100–103]. With recent advances in graph-based methods, Graph RL captures interactions between agents in a cooperative multi-agent environment as a relational representation [104–107]. Solving complex tasks that consist of many subtasks requires a sequential plan to cover each task.

Experimentally proven concepts have recently been used in RL by reinterpreting psychology or neuroscience. Conversely, some studies have proven concepts and experiments of the neural net through neuroscience. An embodiment means an individual that occupies space with physical reasoning. Therefore, understanding and reasoning of physics is the ability required for the agent, and a benchmark for physical reasoning is proposed, including classical mechanics puzzles in a 2D physical environment [108].

An embodiment considers a physical action (e.g., the movement of a robot). This is similar to complex decision making, and various psychological approaches (e.g., episodic memory, intrinsic motivation) are conceptually used. Episodic memory [109–113] restores previous important events, and intrinsic motivation [114,115] explains a series of behaviors seen in the behaviors in-depth. In particular, ToM [116,117] becomes an essential basis for making more appropriate decisions by understanding other agents' situations in a multi-agent environment.

4.1. Decision making: Multimodal

Reasoning derives new information based on the given information. In a multimodal, each modal may have the same information but complements the information lacking in other modals. In terms of methodology, multi-hop reasoning, relational reasoning, and graph reasoning have been raised recently as issues. As shown in Fig. 11, relational reasoning performs tasks by extracting relations among inputs and has an advantage over visual QA [118].

Multihop reasoning. Recently, multi-hop reasoning of graph neural networks (GNN) has been used to create new knowledge in visual and language [119–123]. A multi-layer multiplex graph neural network is used to capture object representations and multiple relations in multiple panel diagram inference tasks [119]. It summarizes the graph extracted from the task diagram and selects the most probable response among candidates. Graph-based iterative search method with wiki graphs is used to find responses to web- scale multi-hop open-domain questions, which is difficult to find relation [120]. A policy-based agent with a continuous state shows better performance for multi-hop relation path training than a path ranking-based algorithm and knowledge graph [121]. In the visual dialog that considers semantic dependencies between dialog entities, the dialog entity is considered the observed node, and the response to the question is displayed as a missing value node. A differentiated graph neural network with an expectation maximization (EM) algorithm is used for both the basic dialog structure and the missing node values based on the current question and dialog history [122].

Relational reasoning. Relational embedding jointly represents the connection between related objects. Link prediction is performed by training the low-level representations of entities and relations. To improve the interpretation for KG reasoning, modeling the relation embedding parameterized with discrete values reduces the solution space [124]. Relation generation improves by explicitly modeling interdependencies between object instances in the scene graph [125]. In addition, the performance of relational reasoning is improved by encoding a global context and a geometric layout.

Multimodal relational reasoning is used for VQA, which considers combining the interactions between questions and images. The relation is changed to atomic inference primitives with rich vectorial representations to improve visual and question interactions [126]. Constructive reasoning about quantity, comparison, and relation is used for semantic diversity, composition, and visual reasoning of language caption [127], and a multi-level attention mechanism is used for video inference [128].

Graph reasoning. Since the graph serves as a knowledge repository in conjunction with KB, it is important to effectively utilize encoding, sampling, and utilization in visual language interaction. GCN is a representative model for training the representation of attribute graphs. Graph reasoning trains a fixed representation of an entity in multiple relational graphs, which is generalized to infer unseen entity relations during inference. To improve graph reasoning, various approaches are proposed [91]. Various layer sampling is used to alleviate neighboring explosion problems during mini-batch training. A graph sampling-based induction, which has the advantage of separating sampling from back-propagation, improves efficiency and accuracy [92]. An attention-based method is used to combine graph neighbors of entities adaptively and to learn the query-dependent representation of entities [129].

Graphs are also used to build dependency trees to infer relations in the unstructured text [93,94]. The localized first-order approximation of spectral graph convolution is used to select a convolution architecture to semi-supervised training of graph-structured data [93]. Attention-guided graph convolutional networks directly obtain the entire dependency tree as input instead of the rule-based hard pruning strategy [94]. It automatically trains how to participate in related substructures useful for relation extraction tasks selectively. Loose-structured open text knowledge descriptions are used for automatic KG construction instead of template methods for common sense knowledge graphs [130].

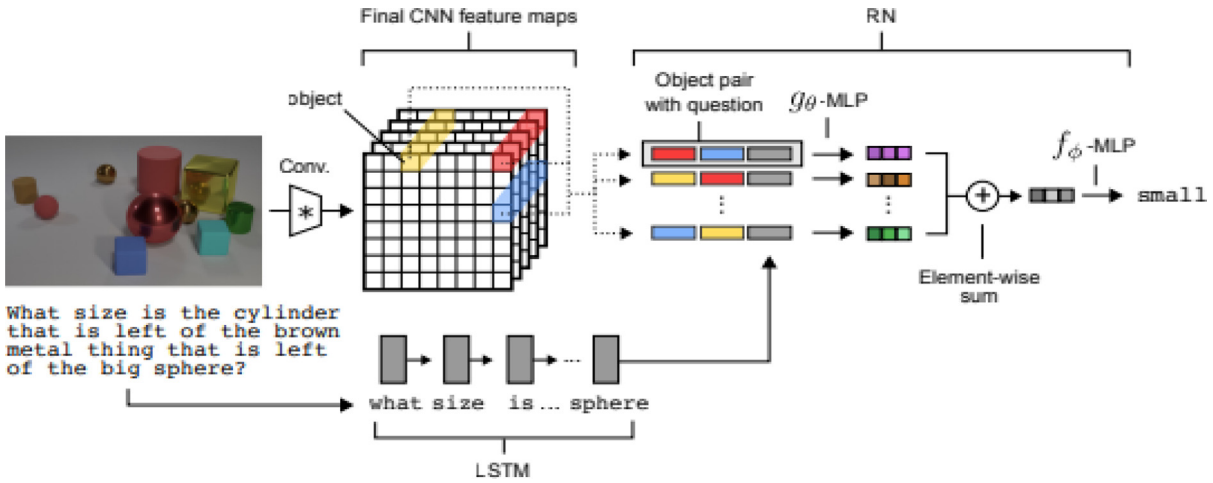

Fig. 11. Flow diagram of a relational network with CNN feature maps and LSTM [118]; Questions with relational information are encoded through LSTM, and images are encoded through CNN. Pairs are created by cascading the encoded question and the representation of the object. After that, the final answer is obtained through an element-wise sum.

4.2. Decision making: Multi-task

RL is a method in which an agent achieves a goal by determining the sequential action. The agent is the subject who performs the action and receives the reward as a result of the action. It gradually optimizes action to get closer to the goal through trial and error. RL is divided into model-based learning and model-free learning. Because RL is an inefficient sample solution, model-free RL has the disadvantage of taking a long time to train. In order to solve these shortcomings, a pre-defined model is used to supplement the cold start problem and induce rapid convergence. However, in the case of a model-based method, it is difficult to converge if the model is different from the current task.

RL is also divided into on-policy and off-policy. The on-policy trains an algorithm using the deterministic output of the target policy, whereas the off-policy indirectly utilizes a stored distribution to train an RL agent. Since the on-policy only utilizes the current exploration data, it is difficult to converge due to the lack of training data and local maxima. To compensate for this problem, the off-policy is to use previous data as training data. The off-policy has the advantage of faster convergence because it has more training data than the on-policy, but it is difficult to converge properly when there is a domain gap with the target. Accordingly, methods such as episodic memory, world model, and language-grounding RL have been proposed to solve the problem of the RL sample inefficiency and sparse reward.

Learning complex tasks from scratch in MARL is impractical due to the sample complexity, so it is common to reuse knowledge gained from previous experience or other agents. Da Silva et al. [46] defined classification for the knowledge reuse problem and proposed the knowledge reuse methods between agents. For stable training, the Proximal Policy Optimization (PPO) for learning boundaries within a range or Actor–Critic for updating policy and value function parameters is used as a baseline. Engstrom et al. [68] examined the results of code-level optimization and algorithm augmentation described by supplementary details through case studies on two popular algorithms, PPO and Trust Region Policy Optimization (TRPO).

To solve the problem of difficult policy training due to environmental complexity between many agents, Long et al. [131] proposed an evolutionary population curriculum based on curriculum training that gradually increases the population of training agents in stages. In a reversible and reconfigurable environment, the agent performs a series of tasks to propose tasks, and the next agent undoes or repeats each task [132]. The agent automatically generates a search curriculum and reduces the number of training episodes. Diversity all you need (DIAYN) model learns a useful skill without a reward function, just as humans explore their environment without supervision [133]. DIAYN acquires skills by maximizing the information-theoretical target using the maximum entropy policy.

Multi-agent RL. The relation between multi-agents is divided into collaboration, competition, and oracle relations. In order to effectively use such a relation in multi-agent, it is necessary to introduce concepts such as Theory of Mind (ToM), intrinsic motivation, and heterogeneous competition. Personal innovation is shared with other humans in the form of collective knowledge through communication. However, in order to perform such communication, the agent needs domain expertise and solves the complex problem due to the long-term dependency of training.

Some agents learn when and what to advise with peer-to-peer training in a collaborative multi-agent RL and then use the advice received to improve local training [134]. Because these roles are not fixed, it is an advantage that agents learn how to request and provide advice at the appropriate moment, taking the role of the student and teacher, respectively. Through experiential comparison, not only does the training speed increase, but the agent also learns how to adjust the task that the other methods failed. Cooperative multi-agent RL requires a distributed policy, but there is a limitation of the agent's behavior coordination in a complex environment. The agent reconstructs parts of each other's observations to create common knowledge in distributed, collaborative multi-agent tasks.

In the hierarchical structure, the high level coordinates the agent group on the basis of common knowledge, and the low-level deals with small but potential knowledge. Common knowledge is used in probability matrix games and complex variance adjustment tasks in StarCraft II unit management [135]. A probabilistic actor–critic algorithm is used for learning hierarchical policy trees [96]. In collaboration, it is also essential to study how to believe from the perspective of the other agent or human [136]. Humans share potential minds like faith, and these social methods are important for recursive reasoning about the possible consequences of different human actions.

The agent cooperates and communicates with other agents to solve complex tasks in a partially cooperative environment. The card game Hanabi is an interesting task because it requires the Theory of Mind to reason from the other agent's point of view when observing the behavior of other actors. A probabilistic

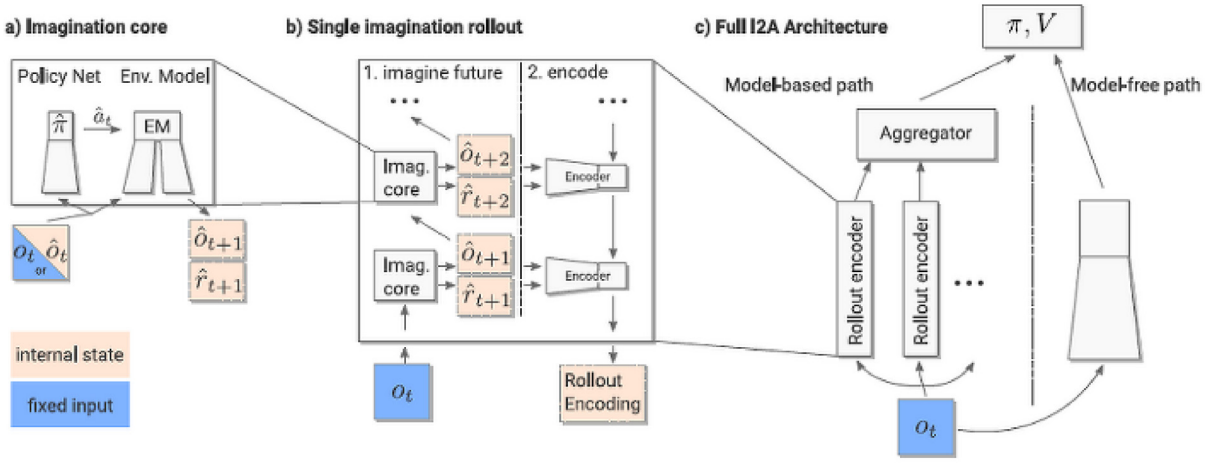

Fig. 12. The architecture and component of I2A with model-free RL and model-based RL [139]; By approximating the next state information to the current state and behavior, imagine future of virtual state information is generated using RNN. I2A consists of model-based and model-free approaches.

regression inference is used for an agent to explain how the other party will respond to future actions [137]. Each agent finds the best response and then uses the Varial Bayes to approximate the conditional policy of the other party to improve its policy. These methods show good performance in matrix games and differential games with the minor equilibrium that gradient-based methods cannot converge (e.g., Go, Poker and Dota).

Imagination augmented RL. Training data is augmented by predicting results for behaviors that have not been visited, such as simulation through a dream. The extracted functions of the world model are provided in a concise policy trained by evolution and operate in various environments [138]. The agent is fully trained in the environment generated by the internal world model, and the policy is transmitted back to the actual environment. As shown in Fig. 12, Racanière et al. [139] proposed Imagination-Augmented Agents (I2As) for deep RL that combine model-free RL and model-based RL. Unlike the other methods that define how the model reaches the policy, I2A trains a method of constructing an implicit plan with an arbitrary method by interpreting the prediction of the trained environment model.

Language-grounded RL. RL has limitations, such as inefficient sampling, difficult reward design, and time-consuming training. To compensate for these limitations, language is used as an effective method to abstract tasks or define goals that humans understand in the RL domain [95–99]. For example, 'reading' is a substitute to 'put books at eye-level, hands at the corners of books, eyes looking at books'. In addition to abbreviating the goals with implicitness, people easily understand language by breaking down difficult questions into several easy sub-questions. Some agents also break down instructions to a single skill level by imitating human behavior [140]. When an agent encounters an ambiguous situation, the agent clarifies the intention of instruction with a multi-turn dialog with oracle [141]. Answer in Questioner's Mind (AQM) agent asks a more consistent question to maximize information acquisition in a task-oriented dialog system [142].

Luketina et al. [143] separated RL into 'language-conditional RL' that interacts with language and 'language-supported RL' that facilitates training for methods that apply language to RL. In general, there are more studies on language-conditional RL, a way to get rewards from instructions, than language-supported RL. To use language in RL, the agent needs to solve the symbol ground problem, which requires explicit knowledge of the language, object, and action relations. In order to perform the language description based on the actual action target, Language Enhanced Exploration (LE2) model collected the description from the social partner and jointly trained the language conditional reward function and the goal-conditioned policy [144]. Language instructions are used to form an intermediate reward for Montezuma's Revenge [145]. In a similar game, the high-level behavior of the agent is similar, so the language helps the agent understand environments and training policy [146]. Knowledge graph A2C (KG-A2C) is an expandable exploration method using language behavior and dynamic knowledge graphs to infer game states in a template-based task space [147].

4.3. Decision making: End-to-end

For effective multi-tasking, one must plan, whether it is direct or indirect, considering sequence and association. Multi-agent planning is a complex and highly considered task because it must consider multi-agents' collaboration and hostility (e.g., games). To solve this problem, researches [148,149] on methods (e.g., reuse of knowledge) are being conducted. Symbolic planning defines the command concept using prior knowledge and symbolic description logic (e.g., PDDL [150]). The ability to plan long sequence tasks is influenced by the prediction model accuracy. Classic symbolic planners utilize pre-defined symbol rules and symbolic states with limited practical application. Planning is also used for E2E training of images and the generation of long sentences [151,152].

Meta RL. Since RL with low sampling efficiency requires a lot of data, meta-learning is performed to fine tune a pre-trained model and reinforce a pre-trained policy. Unlike the general deep neural net, meta-learning learns a new method and adapts to a new environment with a few training data [100]. Meta-learning is divided into few-shot learning, architecture search, and hyper-parameter optimization (HPO). Meta training is used for an unseen task after training various tasks to extract general knowledge. In this way, a task that lacks training data is improved by sharing a parameter or optimization program. Meta-learner with sub-task graph inference utilizes sub-task sets and dependencies that are unknown to the agent instead of directly training meta-policies for few-shot learning [101].

The model-agnostic meta-runner acquires meta-learning parameters from similar tasks and adapts to new tasks in the same distribution with little gradient updates. A gated propagation network improves training with attention and gating in

a graph that explains the relation by propagating messages between prototypes of other classes and updating them in the memory of other classes [102]. Multimodal MAML (MMAML) modulates meta-trained pre-parameters to enable rapid adaptation and improved training for multiple mode distribution [103].

Graph RL. Graphs are used to understand and collaborate between agent interactions in multi-agent RL. It is difficult to learn an abstract representation of interaction in a dynamic environment where the agent moves continuously and the neighborhood changes rapidly. A graph convolution RL captures interactions between agents as a relation representation and improves cooperative relations with the temporal regularization and consistency in a multi-agent environment [104]. Offline RL optimization minimizes the execution cost of the calculation graph in the optimization compiler and generalizes on the unseen graph without additional training [105]. Weisfeiler–Lehman graphs are used to analyze the expressive power of GNNs [106]. The node's representation vector is calculated by recursively aggregating the representation vector of neighboring nodes. A graph is also used to represent the intermediate state of the generated output in the code generation task [107].

Planning RL. The symbolic model works independently for each problem instance, so it is not possible to transfer the experience to a new instance. However, the neural net model has the advantage of transferring training. Some planning RLs combine two advantages of neural net and symbolic. A regression planning network (RPN) has the advantage of being able to directly generate long-term symbolic plans with high-dimensional observations [148]. Domain-independent probability planners make quick plans by using MDP descriptions in argument languages (e.g., PDDL [150], RDDL [149]). A neural network model plans from the task target in reverse order and generates successive intermediate targets reaching the current observations. A simulated 3D kitchen environment is suitable for evaluating planning RL tasks because it is complex visual scenes and contains various objects. Xu et al. [149] proposed an independent transmission algorithm for the MDP planning represented in RDDL. A shared embed space for state and state–action pairs is used to utilize the symbolic state configuration of the domain through RDDL. Zero-shot transmission is possible without a domain simulator and a new instance training by training the RL agent in the shared embed space.

4.4. Decision making: Embodiment

A neural net is a model created by taking a motif from neuroscience and has a conceptual match with neuroscience. Interaction is also gaining complementary effects from psychology. Attempts have been made to solve the problems of the neural net based on the concepts and experimental results of psychology and neuroscience. In particular, Theory of Mind, inductive bias, and intrinsic motivation have been effective methods in the embodied visual language interaction [153–157]. Model-based RL expands a recognition model and compares it with human mental simulation. Neural-symbolic visual question answering (NS-VQA) system first creates a structured scene representation from the image, recovers the trace from the question, and finally gets the answer from the scene representation [47].

Because of the strong inductive bias that the voice delivers to the model, adding text tasks to the voice significantly improves image search performance compared to training alone [64, 158]. Using knowledge-based classifiers that explicitly track belief states makes it difficult to use them in the real world due to non-vocabulary word problems. An external symbolic knowledge base is used to explicitly use the structure of the belief state to provide better induction bias [159].

Dabney et al. [160] provided an integration framework to understand the representation of reward and value in the brain by explaining rich empirical phenomena using dopamine's reward prediction error theory. They assume that the brain represents possible rewards as a probability distribution rather than as a single scalar and that various future results are represented spontaneously in parallel.

Episodic memory. Humans store important episodes in episodic memory and use them when in the attentional situation. Episode memory tracks functional and structural interactions between brain regions, especially the hippocampus [109]. Episode memory is declarative memory that contains information related to the time and place of acquisition, unlike memories of meaning. Gradient episodic memory reduces forgetting by transferring previous knowledge to evaluate model training on continuous data [110].

Memory is the process of reproducing an episode and is related to autonomous consciousness and subjective sense of time. Gershman et al. [111] proposed that RL is related to procedural and semantic memory from a recognition point of view and is selected from a behavioral value or world model that is gradually extracted from many experiences. Differentiable neural computer (DNC) is a representative example using episodic memory and is suitable for dealing with episodes and continuous data. Conversational memory network uses contextual information in the dialog history to recognize the sentiment of the utterance in the video [112]. An explicit memory tracker investigates whether decisions are made with the conditions of the rule in a dialog machine-reading [113].

Intrinsic motivation. In an environment with sparse feedback, the method of training target-oriented behavior cannot train the value function due to insufficient exploration. A motivated agent solves a problem with a method of exploring new behaviors on its own rather than solving the problem directly. In general, the experience of novelty and surprise improves training and memory rather than state curiosity. Oudeyer et al. [114] explained how psychology and neuroscience conceptualize curiosity and intrinsic motivation so that it is an intrinsic reward with the brain's novelty, complexity, and information scale. For the motivation inherent in the synergy task, Chitnis et al. [115] provided an incentive to the agent to perform the joint task because the agent cannot achieve it if the agent acts on its own. This approach is more efficient than the surprise-based intrinsic motivational approach in sparse reward tasks.

Theory of Mind. It is important to consider whether the agent is friendly or not in order to consider the reaction to the agent's actions in the future. Theory of Mind (ToM) broadly refers to the ability of humans to represent the mental state of other humans, including desire, belief, and intention [116]. Rabinowitz et al. [161] designed a Theory of Mind neural network, a ToMnet, that builds a model of an agent by observing behavior using meta-learning. With a strong pre-trained model of agent behavior and a small number of behavioral observations, it is possible to bootstrap rich predictions about agent characteristics and mental states. Melhart et al. [117] examined how the emotional mind theory of gameplay influences behavioral recognition, performance, and frustrating behavior in face sentiment recognition tasks.

4.5. Discussion: Decision making

Section 5 focuses on the inference and decision for visual language integration (see Table 4). A graph decomposes images into scene graphs rather than abstract representations and analyzes them with clear and detailed features. In conjunction with the recent trend of explainable AI, the relations between objects are

Table 4
Descriptions and comparisons of decision making for visual language integration.

	Topics	Papers	Approaches	Techniques	Models	Descriptions
4.1	Multihop reasoning	Zhu et al. (2019) [123]	Multi-hop relational	Natural language	GP-GNN	Generate the parameters of graph neural network (GP-GNN)
		Wang et al. (2020) [119]	Multi-layer multiplex	Diagrammatic reasoning	MXGNet	Multiple layer graph neural network; multiple panel diagram inference tasks
		Asai et al. (2020) [120]	Sequential retrieval	Visual Dialog	Learning-to-retrieve	Takes into account semantic dependencies between dialog entities
	Relation reasoning	Woo et al. (2018) [125]	Interdependence modeling	Visual genome	LinkNet	Explicitly modeling interdependencies between instances
		Xu et al. (2019) [124]	Knowledge graph	Link prediction	DihEdral	Embedding scoring function for link prediction
	Graph reasoning	Gordon et al. (2011) [91]	Common-sense	Personal Stories	Causal reasoning	Automated common sense causal reasoning
		Zeng et al. (2020) [92]	Graph sampling	Graph training	GraphSAINT	Graph sampling-based induction training; mini-batch
4.2	Reinforcement learning	Da Silva et al. (2019) [46]	Transfer learning	Knowledge reuse	MAS	Multiagent Systems (MAS); a solution classification; the knowledge reuse problem
		Engstrom et al. (2019) [68]	Case study,	Code-level optimizations	PPO/TRPO	Code-level optimization and algorithm augmentation
		Eysenbach et al. (2019) [133]	Maximum entropy policy	Reward function	DIAYN	Diversity is all you need (DIAYN); learn a useful skill without a reward function
	Multi-agent	Omidshafiei et al. (2019) [134]	Learning to teach	Multi-agent coordination	LeCTR	Learning to Coordinate and Teach Reinforcement (LeCTR); teach in a multi-agent environment
		Vinyals et al. (2019) [135]	General purpose	Multi-agent	AlphaStar	General-purpose training to solve StarCraft
		Schroeder de Witt et al. (2019) [96]	Actor critic	Cooperative agent	MACKRL	Multi-agent common knowledge RL (MACKRL); probabilistic actor–critic algorithm
		Wen et al. (2019) [137]	Variational Bayes	Nash equilibrium	PR2	Probabilistic recursive reasoning (PR2); probabilistic regression inference framework
	Imagination-augmented RL	Racanière et al. (2017) [139]	Combination	Interpret predictions	I2As	Imagination-Augmented Agents (I2As); combine model-free RL and model-based RL
		Ha et al. (2018) [138]	Uncertainty	Model-based RL	World model	Unsupervised RNN; generate the environment
	Language-grounding RL	Das et al. (2017) [95]	Cooperative	Image guessing	Visual dialog agent	Cooperative image guessing game between two RL agents; dialog
		Mendez et al. (2019) [97]	Domain-agnostic	Task-oriented dialog	Action embeddings	Behavior embedding; capturing a general-purpose structure
		Luketina et al. (2019) [143]	Survey	Decision making	Language informed RL	Divide language-conditional RL and language-assisted RL
		Lair et al. (2019) [144]	Social interactions	Language grounding	LE2	Language Enhanced Exploration (LE2); motivation-based target search
		Goyal et al. (2019) [145]	Intermediate language rewards	Rewards shaping	LEARN	LanguagE-Action Reward Network (LEARN); intermediate reward using language instructions; Montezuma's Revenge
		Wu et al. (2020) [98]	Sequentially regulation	Video under-standing	TSP-PRL	Tree-Structured Policy-based Progressive RL (TSP-PRL); sequentially adjusts temporal boundaries
4.3	Meta RL	Pérez-Rúa et al. (2019) [100]	Generic search space	Multimodal classification	Sequential exploration	Efficient sequential model-based search
		Sohn et al. (2020) [101]	UCB	Meta-learner	MSGI	Meta training with sub-task graph inference (MSGI); Upper confidence bound (UCB)
	Graph RL	Jiang et al. (2020) [104]	Cooperation	Multi-agent	Graph convolutional	Capture interactions between agents
	Planning RL	Xu et al. (2019) [149]	Learning-to-plan	Planning	RPN	Regression planning network (RPN)
		Asai et al. (2020) [148]	Neuro-symbolic	State transition	Heuristics	Combines neural net and symbolic; domain-independent heuristics
4.4	Phycology & neuroscience	Yi et al. (2018) [47]	Neural-symbolic	VQA	NS-VQA	Combines in-depth representation training; solves complex inference tasks; CLEVR
		Hamrick et al. (2019) [154]	Comparison	Mental simulation	Model-based RL	Expand a recognition model
	Episodic memory	Moscovitch et al. (2016) [109],	Survey	Episode memory	Hippocampus	Track functional interactions between brain regions
		Lopez-Paz et al. (2017) [110]	Continual learning	Forgetting	GEM	Gradient Episodic Memory (GEM); continuous training model that reduces forgetting
	Intrinsic motivation	Chitnis et al. (2020) [115]	Intrinsic motivation	Sparse reward	Incentivize agents	Studies the role of intrinsic motivation as a search bias for RL
	Theory of Mind	Rabinowitz et al. (2018) [161]	Theory of Mind	Sally–Anne test	ToMnet	Mind neuron theory network (ToMnet)
		Melhart et al. (2020) [117]	Frustration	Gameplaying	Emotional	Emotional mind theory of gameplay influence

decomposed into triples and used to infer the relation between image features. Although many studies have been conducted to reduce the inference complexity, obtaining training data in various domains required for the inference remains difficult.

The study of multi-agent training is still limited to the domain. It is developing around collaboration or simple tasks; hence, it is necessary to challenge a general multi-agent, which covers various tasks [162]. Reinforcement training technologies (e.g., DIYAN [133]) are particularly good studies that solve the limitations of the intrinsic motivation-based RL by discovering new skills through autonomous training.

Introducing psychological concepts (e.g., ToM) is a positive trend in computer science. For example, emotions are defined as scalar values and quickly classified by simple calculations, but emotional analysis in psychology considers more diverse and complex factors. This fusion serves as a basis for reducing the gap between human and agent interaction.

A recent study on dopamine showed that the distributed RL was similarly performed in the human brain, and the expectation of reward differed depending on the agent's personality. However, as in the case of back-propagation, neural nets do not always match human processing processes; therefore, selective application and interpretation are necessary.

5. Generative interaction

Generative interaction plays a role in converting or expressing representations according to the target task. Just as human interaction is not unimodal, expressions must be multimodal with appropriate timing and harmonious intensity of expression. Persona gives personality and enables a richer interaction on a dialog or image caption. For example, the decoder performs the downstream task of the PLM and expresses complicated content in a simple manner, such that a five-year-old child understands. The entertaining element is used as an intermediate step to make the task and expand the rich interaction effectively. Since the dialogue is not a simple combination of single turns of dialog, a hierarchical and continuous approach (e.g., multimodal story) that considers the plot is required to maintain a natural and long dialogue. PLM has the advantage of providing fluent answers, but in order to perform QA tasks based on accurate facts, multi-task interaction (i.e., retrieving and processing KB information) is required. Also, knowledge distillation is used to reduce inference time and model size for the purpose of performing multiple tasks within an appropriate response time. In VQA tasks, which respond to visual information in natural language, multi-task learning is required to generate an answer by processing visual and natural language recognition together. E2E also plays an important role in generative interactions. Since the generative model is more complex and requires more resources than the rule-based method, it is important to introduce E2E that lowers the complexity. E2E is also used to recognize speech signals, understand their meaning, and express them in natural language by integrating sequential operations. Furthermore, E2E in spoken language translation is a good example of reducing cascade errors and enabling natural generative responses with nuances [30,32].

Whereas multimodal interactions involve facial expressions, embodied interactions emphasize hand and body expressions. These interactions require an understanding of the body and a complex generative model. Similarly, an EQA agent performs more active tasks (e.g., movement) than VQA to find insufficient information [35–37]. Furthermore, in VLN tasks, some agents request and receive information through interaction with users and oracles [38–41]. In such a complex task, generative interactions are more effective because as the number of elements increases and the environment becomes more diverse, it is difficult to make pre-defined interactions. It evolved from the VQA to VLN; thus, the expansion of the embodiment adds novelty to the interaction beyond the fusion of vision, sound, and text. Fig. 13 shows the component view of generative interaction.

5.1. Generative interaction: Multimodal

Studies on content generation, such as generating images using GAN and sentences using the GPT-3, have recently been actively conducted [163]. Image generation has evolved from creating a real image or a 2D image into creating a 3D image using hallucinated color and depth. Efficiently rendering with motion parallax, the content and structure synthesized around the depth boundary show a less artificial appearance [164].

In the language domain, BERT and GPT-3 are used to generate the response converted to the style of the dialog according to the form of the previous dialog. In addition, studies are being conducted to generate sentences that meet conditions using pre-defined keywords. Conditional mask language modeling fine-tunes the bidirectional characteristics of BERT for consistent sentence generation [165]. A Plug and Play Language Model (PPLM) performs controllable language generation for various topics and sentiment styles [166]. The generation model is effective in multimodal expression. A multimodal story is provided in place of a textual story, and the interaction is delivered immersive based on rich expression through a multimodal interaction. Not only the multi-directionality of expression but also a comprehensive embodied expression and the persona dialog that gives personality to the agent have been proposed recently.

Multimodal interaction. Multimodal interaction is an expression in the form of language, picture, song, dance, movement, and so on. Conditional text-to-image generation focused primarily on generating a single image from the condition. A system that repeatedly generates images according to continuous language input or feedback requires interaction between concepts in feedback history. It makes it much more difficult than a first-stage generation task. A multimodal agent generates a background, adds a new object, and applies a simple transformation to an object [167]. Briot et al. [168] proposed a methodology based on five dimensions for deep learning analysis to generate music content, including acoustic features (e.g., melody, polyphony, or accompaniment) and control models (e.g., feed-forward, repetitive feed-forward, sampling).

Persona generation. As mentioned previously, in constructing the dialog, agents have responded based on training data without personality. Because this monotony makes it difficult to maintain a long dialog, studies have been proposed to maintain a consistent dialog pattern by introducing the concept as a persona that views the agent as an entity. Tacotron, voice synthesis, learns the potential embedding space of prosody derived from a reference acoustic representation containing the desired prosody [28,169]. Shuster et al. [170] defined personal captioning as the integration of controllable styles and personality traits because standard image caption tasks (e.g., COCO [171] and Flickr30k [172]) are based on factual information and do not reflect their apparent personality.

In order for the dialog system to continue for a longer and more human dialog, personas must be used for an empathetic dialog system [173,174]. However, in order to create a large persona representation via language, there is a lack of dialog data. Personal Dialog is a large multi-turn dialog dataset based on a sequential conditional GAN that includes various characteristics (e.g., age, gender, location, interest tags) from multiple speakers [8]. A story generation that focuses on text-type control based on persona is also proposed [175]. Dinculescu et al. [176]

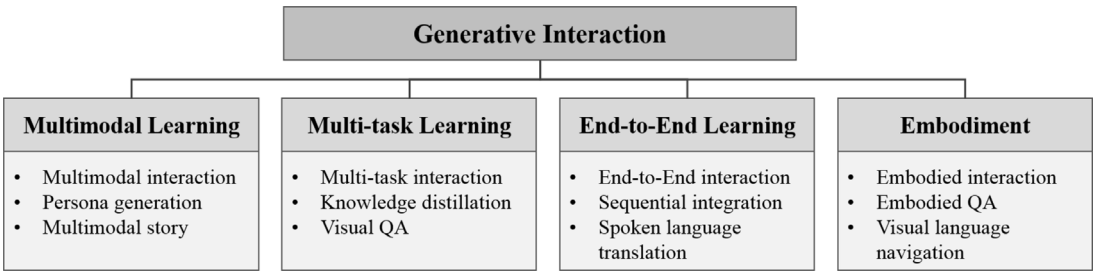

Fig. 13. Component view of generative interaction.

proposed an approach to quickly train a small personalized model to control a larger pre-trained latent variable model.

Multimodal story. There are some studies that visualize stories beyond images or sentences or create counterfactual stories [3,177–180]. StoryGAN is a story-to-image sequence generation model that visualizes stories sequentially with a context encoder and two discriminators by generating image sequences one by one for each sentence [3]. Unlike video generation, story visualization does not focus on the continuity of the generated image but focuses more on dynamics and character consistency. The counterfactual story is rewriting the story with minimal modifications so that the task is compatible with the given anti-reflective events [177]. Counterfactual reasoning predicts with a causal narrative chain, semi-constant constancy, and a conditional language generation model how the alternative event caused other results, unlike the actual event. Cross-assignment uses the arrangement of latent representation by assuming the distribution of potential contents shows good performance in three tasks of modifying sentiment, word substitution decryption, and word order recovery [178].

5.2. Generative interaction: Multi-task

VQA is a task for inferring the relations between individual objects by fusion and attention from images and questions. It recognizes objects or text on the screen and uses them as information to respond to questions [181]. Knowledge distillation and modeling are proposed to improve the VQA performance.

Multi-task interaction. Descriptive social intelligence modeling is proposed to investigate open-ended questions and reasons about the social situation through numerical supervision [182]. ConVQA is a dataset to evaluate the consistency of VQA quantitatively by generating logically consistent QA pairs for the observable facts of the image and collecting common sense annotated by humans [48]. YouMakeup VQA Challenge 2020 is a benchmark for understanding segmented behavior and the visual effects of various actions in domain-specific videos (e.g., makeup lecture videos) [49].

Knowledge distillation. To solve a multimedia question, the agent needs to view the entire collection contained in a series of photos or videos and identify the snippet supporting the response. Focal Visual-Text Attention network (FVTA) provides group inference of visual question responses with visual and text sequence information [183]. Fact-based visual question response (FVQA) subsequently reduces the large set of facts until one entity is predicted as the answer to the question-image pair using the entity graph and GCN [184]. Chandrasekaran et al. [50] proposed a human-in-a-loop approach that treats the model as a black box to analyze whether the description actually makes the VQA model more predictable by humans.

Visual QA. The fundamental way to improve VQA performance is to make vision recognition accurate. CNN and self-attention for language input improve the visual processing of VQA [51]. Compensation images that have similar image pairs with two other responses to the question are collected to balance the VQA dataset [49]. Look, Read, Reason & Answer (LoRRA) predicts a response that is deduced from the sentence found in the image of the TextVQA dataset [53]. Class activation mapping (e.g., Grad-CAM) is used as a discriminator and a supervisor to describe the network's performance in visual description [54]. A unified vision language pre-training model using a shared multilayer transformer and a self-interest mask for VQA and vision language generation [55].

5.3. Generative interaction: End-to-end

E2E training is spreading in various fields, including visual, language, and audio recognition. E2E is often used to facilitate applications to reduce the model and task complexity. Using Translatotron is an excellent example of effectively reducing the number and time of models by integrating tasks performed separately into E2E.

End-to-End interaction. In the visual domain, a deep circuit network (DCN) with attention finds image areas that are considered important for object recognition in an E2E manner [185]. Ferreira et al. [186] provided a comparison between the neural pipeline and E2E method with Gated-Recurrent Units (GRU) and Transformer for RDF triple-based text generation. Since the E2E dialog system with a monolithic neural structure is trained only with input–output remarks, it is difficult to explain the reason. Ham et al. [29] proposed an E2E neural structure for a dialog system that solves the above problems in the human evaluation of DSTC8.

Sequential integration. ASR performance deteriorates when training data is insufficient in a real user environment or when training data and test data do not match. E2E modeling method, SLU shows effective performance in low ASR accuracy situations for a cloud-based modular dialog system. Neural Module Network (NMN) parses questions into language sub-structures and solves each subtask [30]. An end-to-end module network (N2NMN) directly predicts and infers the network layout for each instance by generating network structure literary demonstration and using downstream task loss [31]. A memory fusion network (MFN) is used for multi-view sequential training that explicitly explains the interactions between structures and continuous models over time in speaker sentiment analysis [187].

Spoken language translation. Audio signals are converted into text through transcription or n-best hypothesis-based recognition results. After that, natural language processors classify text into domains, intentions, and slots for downstream tasks. A comprehensive E2E training system for colloquial understanding infers meaning directly from audio functions without intermediate text representation. E2E solution, Translatotron, as depicted in Fig. 14, is used to reduce the complexity of current processes and maximize information without loss [32].

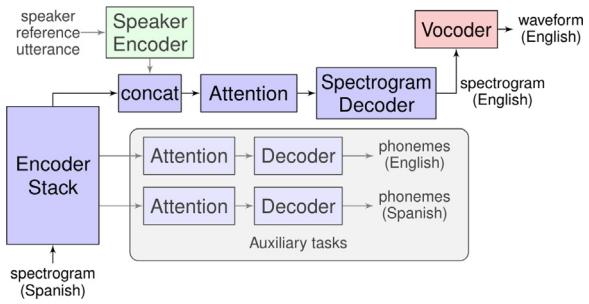

Fig. 14. Workflow and component diagram of Translatotron with encoders and decoders [67]; Translatotron changes from Spanish spectrogram to English waveform in E2E. This procedure composes of several decoders, encoders, and auxiliary tasks.

5.4. Generative interaction: Embodiment

The representative examples of embodied interactions are EQA and VLN. EQA is a newly defined task in which an agent searches for a real environment from a self-centered recognition and answers a user's question. VLN is the task of directly following instructions for navigation [43,188]. These two approaches have the advantage of being able to process modules separated into several tasks within a single framework.

Thus, the information transfer between modules is handled without loss, and coping with new environments becomes possible based on richer information.

Embodied expression. Non-language behaviors, such as gestures, facial expressions, body postures, and language cues, compensate for or clarify language messages. Modeling these behaviors, especially three-dimensional interactions, improves telepresence in the form of an avatar. In creating such a personalized avatar, the dynamics between the voice and body posture of the avatar and the interpersonal relationship are modeled with the interlocutor in the dialog. A Dyadic Residual-Attention Model (DRAM) integrates intrapersonal (monadic) and interpersonal (dyadic) dynamics using selective attention to generate body pose sequences according to interpersonal audio and body poses [33]. Chan et al. [4] proposed a model that transmits the performance to an amateur target in a few minutes, given the source video of a dancing human. To transmit motion, the pose is extracted from the source subject as an intermediate representation, and the trained pose-to-shape mapping is applied to generate the target subject as video-to-video.

Embodied QA. EQA requires complex detailed conditions for navigation, as shown in Fig. 15, but lacks detailed information about the environment, resulting in poor response and navigation accuracy performance. Segmented visual attention extracts local semantic functions, and bottom-up visual attention is used for VQA subtasks according to the guidance of semantic functions. The high-speed video segmentation framework and a feature fusion strategy are used to guide navigator training without additional computational costs [35]. QA-Multi-Target EQA (MT-EQA) transforms a given question into an executable sub-program sequentially and guides the agent to various locations by learning how to select observations along the path [36]. Wijmans et al. [37] instantiated the query response for a large-scale navigation task in a real environment and bridged the gap between simulation and the real world with behavioral mimics and inflection weighting. For a better understanding of the visual object, a point cloud is used to restore a part of the obscured object.

Visual language navigation. VLN is the task of finding directions or following language instructions, as depicted in Fig. 16. There are Habitat [44], HANNA [141], and Unity ML-agent as representative platforms. Habitat-Sim is a 3D simulator capable of processing sensors, and Habitat-API is a modular library for E2E VLN development [44]. SplitNet separates visual recognition and policy training based on the Habitat platform [45]. SplitNet explicitly recognizes the training target for visual exploration by integrating and decomposing the optional tasks and the optional training. Visual-Teach-and-Repeat (VT & R) and SnapNav navigate with only a snapshot of the environment with direction guidance [189]. The maps used for VLN are various, such as photo-based MINOS, Gibson, small maze, outdoor city, and forest.

Discriminative Particle Filter RL (DPFRL) encodes the differentiated particle filter, which is differentially updated with E2E training for decision making and modeling in complex observations [190]. The language command should generally be inferred according to the perceptual context because it only identifies a few high-level decisions and landmarks rather than full low-level motor behavior. A high-level controller provides direction commands, and a low-level controller provides real-time control and obstacle avoidance. Imitation Learning with Indirect Intervention (I3L) requests by specifying only the high-level final target because the requester does not know how to move to the target object [38]. Hierarchical decomposition and module training avoid the high sample complexity of E2E learning and robustly state estimation errors in local policy. Active Neural simultaneous localization and mapping (SLAM) is a modular and hierarchical approach to learning policy for exploring 3D environments [191]. Chaplot et al. [192] designed a topology representation of a space that utilizes semantic information effectively and provides approximate geometric inference for long-distance VLN tasks in an unseen environment.

There are ambiguous situations that are difficult to interpret, relying solely on visual information and language instructions. When the agent recognizes that it cannot solve itself, the task is re-executed by querying the oracle of the teacher–student concept. Chi et al. [39] proposed a model that requests the user's help based on the pre-defined confidence threshold of the next motion prediction. The lack of data is a critical problem in language-based exploration because humans take a lot of time and money to demonstrate the language interactions. Wang et al. [27] proposed a multi-task search model that generalizes language-based search tasks (e.g., Dialog from Navigation (NDH) [40]). Data scarcity is solved through multi-task training, and training parameters are shared and joint training across various tasks. In addition, the interleaved multi-task data sampling strategy is adopted to prevent the shared model from being dominated by one task. Fried et al. [41] provided a model to synthesize new instructions for data augmentation, implement practical reasoning, and evaluate how well a candidate sequence of operations describes an instruction. Since the VLN metric mostly focuses on reaching the shortest distance, there remain two important problems: how to apply the trained policy in an unseen scenario and implement the system in a real environment.

5.5. Discussion: Generative interaction

This section summarizes the generative interactions for visual language integration (see Table 5). Image generation using GAN and Transformer has already reached a level difficult for humans to distinguish, and video is becoming quite sophisticated. Aside from the side effects of generative models, such as fake videos, debates about how to distinguish fakes and ethics are also becoming important issues. GPT-based language generation also naturally generates long paragraphs for each domain. However, studies on embodied interactions, in which each element of technology is fused into a consistent story or naturally expressed, are still limited.

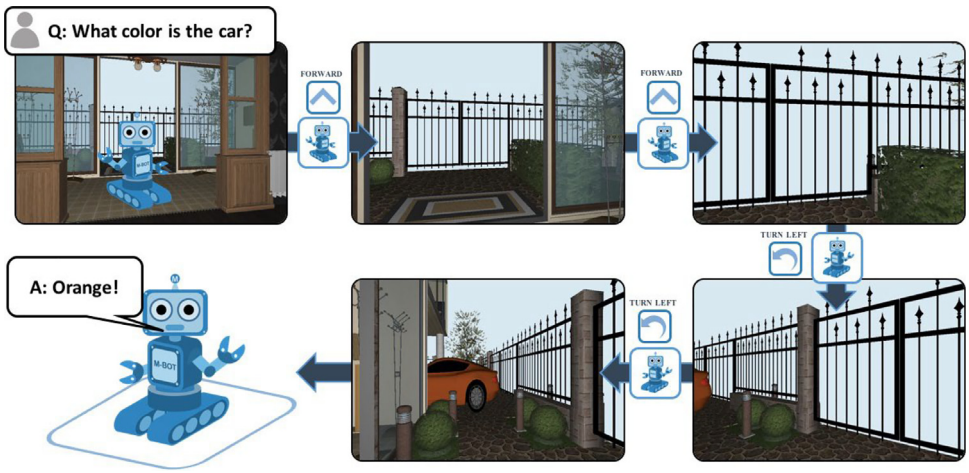

Fig. 15. Example scenario of EQA which answers the question of car color [37]; To answer the question about the color of the car, the agent has to move to a garage for insufficient information. After that, the agent answers the question based on the visual clue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

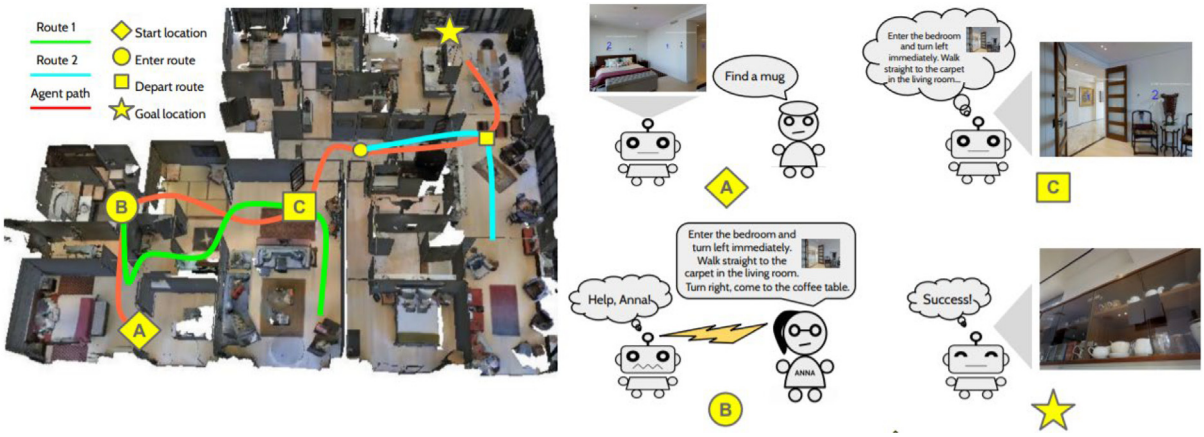

Fig. 16. Example of visual language navigation, HANNA [141]; (A) User gives an instruction 'find a mug'. (B) Oracle provides information of location and trajectory to an agent. (C) The agent follows the instruction to reach the goal.

Persona has been used for longer and more natural dialogs than before, but long-term conversations, conflict responses, and question-oriented dialog configurations are seen clearly with limitations as sustainable general solutions. Humans basically have multi-persona and express them differently depending on time and place; thus, a study on more complex modeling techniques that consider the situation is necessary.

The more an approach compresses space and time in various tasks (e.g., Translatotron), the simpler it is to make interaction possible. Compared to other models, Translatotron has the advantage of solving foreign language pronunciation or error overlap. However, many tasks must be solved in the future for commercial use because performance is low for every single task.

Lastly, in performing comprehensive interactions (e.g., dialog manager and RL task), it is important to have a simulator (e.g., a game) that easily visualizes multiple tasks. As another approach, a machine learning tool (e.g., UnityML) that configures simulation and machine learning together is a powerful simulator in terms of expandability. Virtual reality (VR) and augmented reality (AR) expand the game-like interaction model, bringing it a step forward with an embodied visual language interaction.

6. Discussion and open challenges

As depicted in Table 6, we divide the visual language integration into three steps: (1) representation learning, which collects information through the sensor; (2) decision making, which comprehensively decides; and (3) the final generative interaction. We analyzed each step by separating it into visual, textual, acoustic, multimodal, and embodiment. We organized tasks and modals in terms of the concepts and implemented models in Table 2. For example, Monet performs object classification without generating a bounding box for every object, and BERT is a representative pre-trained auto-encoding model in the language domain [193]. In the sound domain, Translatotron, which translates to voice-to-voice, is a representative example, and Habitat is a representative model of an embodied model.

The unimodal interaction aims to communicate the intent with dialogs clearly. However, in the multimodal interaction, each modal is integrated with other modals; thus, the modal representation is abstracted or specified in a structural form (e.g., a graph). Furthermore, the embodied visual language interaction, which adds the embodied concept to multimodal learning, extends the dimension in the concept of a physical interaction space. For example, the embodied visual language interaction is hierarchically divided into a global agent that deals with the overall workflow and a sub-agent that performs each given task. In detail, the language sub-agent updates the dialog to the instance level depending on the constantly changed dialog's context and state. The acoustic sub-agent composes music to indirectly express context and generates musical instrument sounds by converting music

Table 5
Descriptions and comparisons of generative interactions of visual language integration.

	Topics	Papers	Approaches	Techniques	Models	Descriptions
5.1	Multimodal interaction	El-Nouby et al. (2019) [167]	Conditioned generation	Text-to-image	Recurrent image generation	Generate a background, add a new object, and apply a simple transformation
		Briot et al. (2019) [168]	Survey	Music generation	Five dimensions	A methodology based on five dimensions
		Pidhorskyi (2020) [163]	Companion encoder	Face image generation	ALAE	Autoencoder based Adversarial Latent Autoencoder (ADAE)
		Shih et al. (2020) [164]	Layered Depth Image	3D photo synthesis	Inpainting model	Converting a single RGB-D input image into a new 3D photo
	Persona generation	Skerry-Ryan et al. (2018) [28]	Voice prosody	Voice synthesis	Tacotron	Expands the Tacotron voice synthesis
		Shuster et al. (2019) [170]	Personal subtitles	Image caption	Personality captions	Controllable style and personality traits
		Zheng et al. (2019) [8]	Pos. emotion elicitation	Dialog	MC-HRED	Personal Dialog; persona-aware attention and persona-aware bias
		Lubis et al. (2020) [174]	Emotion	Dialog	Empathetic model	The dialog system framework using a hybrid n-gram and neural network
	Multimodal story	Shen et al. (2017) [178]	Unpaired	Style transfer	Cross-alignment	Performing style transfer using the arrangement of latent representation
		Li et al. (2019) [3]	Visualization	Story	StoryGAN	Story-to-image-sequence generation
		Qin et al. (2019) [177]	Reverse reasoning	Story	Counterfactual story rewriting	Rewriting the story with minimal modifications
5.2	Multi-task interaction	Zadehet et al. (2019) [182]	Social intelligence	QA	Social-IQ	Modeling social intelligence through numerical supervision
		Ray et al. (2019) [48]	Consistency	VQA	ConVQA	Dataset ConVQA and metric
		Chen et al. (2020) [49]	Benchmark	VQA	Cross-modal alignments	Question-response task; YouMakeup VQA Challenge 2020; Cross-modal semantic alignments
	Knowledge distillation	Liang et al. (2018) [183]	Snippet identification	VQA	FVTA	Focal Visual-Text Attention network (FVTA)
		Chandrasekaran et al. (2018) [50]	Human in a loop	VQA	Human-interpretable	Small personalized model to control a larger pre-trained latent variable model
	Visual QA	Goyal et al. (2017) [52]	Benchmark	VQA	VQA v2.0	Balancing the VQA dataset by collecting compensation images
		Delbrouck et al. (2019) [51]	E2E features extraction	VQA	CNN augmented	CNN augmented with self-attention
		Zhou et al. (2019) [55]	Pretraining model	Vision-language	Unified VLP	Unified vision language pre-training (VLP)
5.3	End-to-End interaction	Qian et al. (2017) [30]	E2E modeling	SLU	ASR-free model	E2E modeling without ASR for SLU
		Linsley et al. (2019) [185]	Large-scale experiment	Image categorization	ClickMe	Deep convolutional network (DCN); ClickMe
		Ferreira et al. (2019) [186]	E2E generation	Data-to-text	GRU, Transformer	A systematic comparison between the neural pipeline and E2E data-text method
	Sequential integration	Hu et al. (2017) [31]	Direct prediction	QA	N2NMN	End-to-end module network (N2NMN); learn to reason by directly predicting instance
	Spoken language translation	Serdyuk et al. (2018) [32]	E2E ASR learning	SLU	Audio recognition	Comprehensive training system for colloquial understanding
5.4	Embodied expression	Ahuja et al. (2019) [33]	Intrapersonal, Interpersonal	Non-verbal behaviors	DRAM	Dyadic Residual-Attention Model (DRAM); integrates intrapersonal and interpersonal dynamics
		Chan et al. (2019) [4]	Pose representation	Video-to-video	Pose-to-appearance	Video-to-video translation using pose as an intermediate representation
	Embodied QA	Luo et al. (2019) [35]	Visual attention	EQA	Segmentation based model	Segmented visual attention mechanism
		Yu et al. (2019) [36]	Multiple targets	EQA	MT-EQA	VQA module-based architecture; QA-Multi-Target EQA (MT-EQA)
	VL navigation	Fried et al. (2018) [41]	Context Inference	Low-level behaviors	Speaker–Follower	Synthesizing new instructions for data reinforcement
		Wang et al. (2019) [27]	Multi-task	Data sparsity	Generalized model	Multiple task search model
		Nguyen et al. (2019) [38]	Imitation Learning	VLN	I3L	Indirect Intervention (I3L)
		Savva et al. (2019) [44]	Platform	VLN	Habitat	Habitat-Sim; photorealistic 3D simulation
		Gordon et al. (2019) [45]	Selective learning	VLN	Splitnet	Decoupling visual perception and policy learning; auxiliary tasks
		Chaplot et al. (2020) [191]	Modular and hierarchical	VLN	Active Neural SLAM	Flexibility with respect to input modalities
		Xie et al. (2020) [189]	Snapshot	VLN	SnapNav	Few snapshots of the environment

Table 6
Task and models of visual language integration in the aspect of representation learning, decision making, and generative interaction.

	Modals	Representation learning	Decision making	Generative interaction
Tasks	Visual	Semantic segmentation, Image classification, Object detection, Scene parsing, Depth estimation, Facial recognition, Object tracking	Recognition, Object classification, Visual reasoning,	Image generation, Constrained image generation, Super-resolution, Domain adaptation, Style transfer, Image inpainting
	Textual	Word embedding, Sentence embedding, Language modeling, Text classification, Named entity recognition, Dependency parsing,	Natural language inference, Entailment, Relation classification, Relational reasoning, Reading comprehension, Information retrieval, Coreference resolution, Entity linking	Mud game generation, story generation, Chat, Question generation, Text summarization
	Acoustic	Endpoint detection, noise reduction, Speech recognition, Speech separation, Keyword spotting, Audio classification, Sound event detection	Speaker recognition, Speaker verification, Acoustic scene classification, Environmental sound classification	Sound generation, Music generation, Speech synthesis, Music generation, Audio generation
	Multimodal	Multimodal fusion, Video object segmentation, Visual object tracking	Multimodal inference Video prediction, Scene classification, Video object segmentation	Multimodal response generation, Image captioning
	Embodied	Pose estimation, Action classification, Eye-tracking	Description, Question, Activity recognition, Gaze estimation, Gesture recognition	Action, Movement
Models	Visual	Monet, Decatron, EfficientNet, Resnet, VGG, Mask R-CNN	Scene graph	Geneva, Webtoon, BiGGAN. StyleGAN, COCO-GAN, RCAN
	Textual	BERT, XLNET	ALBERT, SemBERT, ERNIE	Counterfactual, DialoGPT, PPLM, CTRL, GPT2, GPT3
	Acoustic	Seepch2Vec, ContextNet	CLEAR	TTS, Tacotron, WaveNet
	Multimodal	VideoBERT, Vilbert	MUREL, DANs	MEGAN, UniViLM
	Embodied	Worldmodel VIV-PoseNet	Midlevel, splitnet, Diayn, SAC, Clevr, I2A	Multi-agent, Minirts

scores according to the characteristics of each instrument. The visual sub-agent uses a generative adversarial network (GAN) to represent the agent's face according to the current sentiment. The embodied agent shows a human-like movement through RL-based imitation learning.

From a conceptual point of view, the global agent creates an overall plot of a long story like a movie and defines each persona of the agents. The agent performs actions according to the previously generated policy, reacts according to the situation that occurs while performing, and updates the policy based on the reward of the response.

Conventional approaches (e.g., one-to-one mapping, ensemble dialog manager, and state machine) do the same with a limited representation. In addition, the user's exceptional action is handled in a pre-defined form according to manually generated rules. However, these methods lack the depth of interaction; therefore, the user is easily accustomed to iterating several times and does not feel an immersive and lasting experience. In addition, it cannot provide an integrated service of the entire AI models because it was provided in a restricted form due to the limitation of the amount of process computation. Hence, it provides a service with low quality of detail or limited freedom.

The following subsections discuss in-depth the abstraction techniques for dealing with various multimodal data and few-shot learning that effectively performs multi-task learning. In addition, we look at the points necessary for the E2E training through life-long learning and discuss the differences between embodied interaction and conventional interaction.

6.1. Multimodal: Fast modal conversion

Each unimodal model size for the visual language integration model varies from 20 million parameters to 170 billion parameters and becomes larger to optimize the performance of each task. In this situation, the model integration is not a simple task when considering the computation resources and the paired data required for training. The data required for multimodal training basically increases in dimension in proportion to the number of unimodal. Furthermore, when expanded to multi-agents, more data processing is required in proportion to the number of agents. In particular, fast object recognition, decision making, and rapid reaction are necessary because the actions per minute in StarCraft corresponding to the number of actions per minute reaches 100 to 200.

Recognizing many objects using individual object detection increases the amount of computation in proportion to the number of objects, so there is an attempt to reduce the computational burden using the abstraction concept. Especially, some studies (e.g., World Model [138] and Monet [193]) abstract the multiple objects to representation for fast object recognition and efficient training. The abstracted sentiment is used through these abstraction methods to make expression effective at the decoder by directly using representation instead of scalar value or label.

Abstraction is also used in various forms for representation in reinforcement learning. Kulkarni et al. [194] proposed a Transporter to discover object key points or landmarks and train object representation useful for control and RL. The consistent long-term tracking enables high sample efficiency with key point coordinates in the control area of the RL and enables deep search without external reward because it greatly reduces the search space. To capture a potential representation of state or environment without supervised reward, Anand et al. [195] proposed Spatiotemporal Deep Infomax (SP-DIM), a self-directed state representation training that uses the spatial characteristics of visual observation in an RL environment. It captures small objects and shows robustness in noise on the benchmark, consisting of 22 Atari 2600 games based on the cascade training environment.

Language is ambiguous because it provides abstraction in a form that minimizes the energy consumption of the speaker about the situation but can express the intention of the speaker implicitly. Jiang et al. [196] proposed using a language that is flexibly applied to various goals, fast training, and generalization of combination as an abstraction to solve the problem of difficulty in abstraction generalized abstraction in hierarchical RL. They use

human-interpretable high-level language instructions as high-level policy and combine low-level policy with various high-level targets without retraining.

For a rich interaction with various modals, multimodal learning aligns multiple modal data using fusion and abstraction. Multimodal pre-trained models are recently used for complex downstream tasks. The complexity of context recognition is reduced by constructing scene graphs for the scene entities. We investigated herein the study of multi-hop reasoning for relations between entities based on the constructed graph. From an expressive point of view, we summarized the study of generating responses that reflected facial expressions and persona using multimodal generative models. Furthermore, more immersive expressions are provided to users using multi-dimensional interaction, such as multimodal stories and embodied expressions.

6.2. Multi-task: Few-shot learning

The multi-task model is simplified by integrating methods that process each task individually, but a detailed context is sometimes ignored. Moreover, the analysis of continuous interaction is difficult because the values of the sensors are normalized for a certain period and are defined as discrete labels. Accordingly, studies have been conducted on reducing the modal inconsistency problem and the model training burden of downstream tasks by creating a pre-trained model that jointly trains a large amount of vision and language training data (e.g., videoBERT) [34]. However, fine-tuning tasks to be applied to downstream tasks, especially mobile tasks, are burdensome depending on the task because they still require much data and computing power. Therefore, studies on few-shot learning, which does not need retraining, are necessary.

The dialog engine was previously divided into a goal-driven dialog model and a chit-chat model, but the current dialog is expanding into an area that simultaneously supports multi-task learning. The performance of the E2E model is mostly insufficient to support the multi-task domain; thus, previous solutions combine various dialog engines with an ensemble and select the engine suitable for the situation. However, the ensemble has limitations on model composition, and creating a consistent response is difficult. Miller et al. [197] provided an integrated framework, ParlAI, for training and testing dialog models with multi-task training, data collection, human evaluation, and online RL. ParlAI performs various tasks in the same interface with dialog datasets (e.g., SQuAD, bAbI task, MCTest, WikiQA, QACNN, QADailyMail, CBT, bAbI dialog box, Ubuntu, OpenSubtitles, and VQA) [197]. As mentioned earlier, to reduce the problems of multi-task learning, tasks and interfaces are needed to simplify with pre-trained models and integrated frameworks.

Previous multi-task training used ensemble methods with various models to solve similar problems in the same domain. Due to the recent advances in the PLM, we observed an increase in the number of studies for fine-tuning and few-shot learning PLM in downstream work. In addition, research on how to directly utilize other modals (e.g., language-based RL) is being conducted by using language instructions as a direct reward [198].

6.3. End-to-end: Life-long learning

E2E training is the process of integrating multiple models into one model in sequential order. The range for time is a continuous task or episode unit that contains the concept of life-long learning from the perspective of an individual's history. Life-long learning is a series of learning how to continuously adapt and interact with the environment in a sequence without forgetting what has been previously learned [199]. Online learning requires knowledge fusion to learn from sequentially presented data streams and is effective in autonomous agents or continuous tasks. Life-long learning has the advantage of being a sustainable model, but it also has the catastrophic forgetting problem of overwriting previously trained parameters. To solve this problem, the agent maintains storage, such as the gradient episode memory, to store a previous event or utilizes some of the previous data for the current task to not forget the last task.

Planning is an important task in performing multi-tasks by connecting various tasks. More studies on RL-based planning have been made than on conventional symbolic planning. Meta-learning, knowledge reuse, and research on multi-agent systems are needed to compensate for the sample inefficiencies and limitations of RL-based planning. The VQA is a representative example of multimodal multi-task training that spontaneously processes the image recognition function of vision and query response of language. E2E learning is used for complex tasks (e.g., planning) to reduce the complexity of the VQA task according to various inputs.

Life-long learning is widely used for vision, dialog, RL, and transfer training. An example of vision-based life-long learning, Yoo et al. [200] proposed a model to improve the accuracy of life-long learning for CNN based on knowledge subscription. Because these models do not use continuous training data or episode memory and each delta model is independent, it essentially avoids the catastrophic forgetting problem. As an example of life-long learning in a language domain, Sun et al. [201] proposed a language model that spontaneously trains sample generation for new tasks. Since a pseudo sample is generated in advance during training, additional generator and gradient episodic memory are not required. Continuous and Interactive Learning of Knowledge (CILK) is a triple-based model that continuously learns and infers new knowledge in a dialog [195]. In the case of persona chat, the agent does not generate a new response every time from scratch but instead responds based on pre-defined hobbies. However, since these dialog models are limited to short dialogs, it is necessary to apply life-long learning for persona agents that maintain dialogs consistently based on long-term conversation history.

Since life-long learning has to deal with a lot of data, methods such as meta-learning and batch parallelization are needed to train efficiently. A Meta Markov decision process (MDP) optimizes the search for exploration strategies for new but related tasks in RL-based transfer learning tasks and enables continuous translation by utilizing the previous experience [202]. BatchEnsemble that parallelized inside and outside the model like a normal ensemble solves catastrophic forgetting by training various tasks with shared weight in sequential order without accessing previous task data [203].

Lastly, in life-long learning RL, an efficient sampling mechanism is required for continuous online adaptation according to changes in environment, limitations, tasks, and agents [204]. Probabilistic motion planning through online adaptation adapts to the new environment in a few seconds for the KUKA LWR arm using a training signal that mimics the mental regeneration strategy and motivational signal recognition dissonance to enhance the experience [205].

6.4. Embodiment: Immersive interaction

A common problem with many generative agents is that the generated responses are repeated or have an opposite response without continuity with the previous response. Because the training loss of the generative model is mostly based on similarity and fitness, it shows a naive and boring response by selecting

a repetition of a similar expression or a commonly used response. Therefore, the novelty of an agent response is a metric for measuring whether the agent is interactive.

Generative models for interaction are used not only in conversation but also in embodied objects and environments. Recently, photo-realistic visual language generation has made it possible to easily convert natural high-resolution images guided by language (e.g., CLIP [206–208] and diffusion model [209–211]). It is able to compensate for insufficient context by expressing explanations during conversation as images. Furthermore, a text-to-video generation is able to create a virtual environment in the metaverse and a video guided by scenario and plot [212–214]. These photo-realistic visualized objects and natural environments help embodied interaction in multimodal manners.

The embodied interaction makes a conversation more fluent and robust to the unseen environment. The power of embodiment is quite strong, making the conversation more immersive. Rather than the agent acting as a previously defined state machine to perform various actions, it is necessary to have a memory that implicitly stores the previous history and an online generation model that matches the given persona. Generative studies based on the accumulated user's actions have recently been conducted to express various methods, such as 'AI dungeon 2', followed by mud game-type stories and 'Dance now' expressed as dance with continuous actions [4,5]. The embodied agent operates a sequential integration with E2E rather than a pre-defined combination to give an immersive experience different from a simple statement machine or a behavior graph.

Finally, the embodiment is an important factor that will be the basis for AI agents and human interaction in the future [215]. By using the embodied interaction, the agent performs various expressions (e.g., gestures) and tasks, even for non-language instructions (e.g., Exophora). It also serves as a good platform for utilizing psychology and neuroscience (e.g., inductive bias and Theory of Mind). VLN is a task that experimentally develops this embodiment along with other multimodal and requires various trials and studies.

6.5. Discuss: Visual language integration

In this section, we discuss the concepts and approaches for visual language integration. The concept of fusion and abstraction used to represent multimodal integration is summarized in terms of vision, language, and RL. Abstraction has the advantage of processing multimodal but needs a hierarchical approach to solve the ambiguity of the abstracted representation.

From the multi-task perspective, we examine the fine-tuning model suitable for each task according to the PLM development that processes various downstream tasks in one model. Fine-tuning has also recently been burdened as data grows; hence, it is preferable to use few-shot learning. Although the PLM shows a good performance for similar language tasks, overcoming the conventional ensemble model in the multi-domain dialog is still difficult. Therefore, studies on multimodal PLM must be continued.

Catastrophic forgetting is an old problem of multi-task learning, which is a more critical problem in life-long learning. Consistent long-term training is important in the development of sustainable AI rather than a single response. The scope of the interaction has expanded beyond the expression in the dialog to time, gesture, and space. Agents can actively express using various multimodal generation to have a rich interaction with users.

One of the reasons why the embodied visual language interaction is difficult is the disparity integration with the heterogeneity of various scales of element technologies. To overcome these difficulties, it is necessary to have in-depth knowledge of the detailed characteristics of element technologies and broadly analyze the similarities and differences between the technologies. In recent years, integration has become more difficult, with the speed of the AI model development becoming faster; hence, increasing demand for surveys and comprehensive research has been observed.

Some research institutes have proposed several simulation tools, but these are often limited to specific tasks and are less versatile. Another problem is that a cross-evaluation between studies cannot be performed because each research community uses its own tools. In terms of methodology, this problem can be supplemented through competition and standardization. In terms of technology, a tool that can be used more universally and can handle various inputs and low-level perceptions (e.g., sound and lip-reading) must be developed. Interest in detailed data augmentation (e.g., texture and shading) in a virtual simulation is also necessary.

It is difficult to define rewards in reinforcement learning and evaluate the novelty in generative learning; thus, a comprehensive evaluation metric is important in visual language integration. VLN evaluates the success rate and the minimum distance to reach the target. While being relatively simple, clear metrics can be useful for each task. Challenges in evaluating immersive interactions with simple conversation time or subjective user ratings must be solved.

Immersive interaction requires interdisciplinary research in psychology and neuroscience, but it is currently being used only in conceptual areas. For example, sentiment analysis is an important factor in interaction, but in the practical domain, a small number of classification features with scalar values are dealt with instead of a classification based on various hypotheses dealt with within psychology. In addition, cognitive psychology and neuroscience are used to interpret AI techniques, but explainable AI that can explain the cause of consequences and causal inference that can infer causal relationships must be improved. These interdisciplinary studies are needed for deeper and more explainable interactions.

Vision model and language model are evolving based on solutions suitable for each modal characteristic. On the one hand, the methods used in vision (e.g., GAN, contrastive learning) are sometimes used in natural language, and the methods used first in natural language are used in vision (e.g., PLM, Transformer). Therefore, the interaction of solutions for these modals increases the consistency of integration. Studies of the integration methodology that are able to generalize and explain the evolutionary patterns of visual-language integration are needed (e.g., separation–integration, thesis-antithesis-synthesis) because the integration of these various methods is able to be combined in various ways.

7. Conclusions

In this research, we classified visual language integration into three steps: representation learning, decision making, and generative interaction. We also described the required technology and limitations for multimodal learning, multi-task learning, end-to-end learning, and an embodiment for each step. AI models require much training data and computing power to learn each model, and visual language integration models need additional efforts to integrate component models. Obtaining paired data for multimodal and multi-task learning is difficult and expensive; therefore, unsupervised or self-learning is being focused on and developed.

We summarized the visual language integration in terms of four integration methodologies. Representatively, multimodal learning starts with visual captioning that describes a simple

image as text and challenges the visual dialog through the VQA. It has recently been expanded to visual storytelling. Multi-task learning has increased interest and study with the PLM development, but problems, such as catastrophically forgetting, still need to be solved. End-to-end learning is a good opportunity for reducing system complexity and develops in the form of a spoken language translation. The embodiment shows more growth than before but is still in the simulation stage due to physical limitations. It is used as an embodied conversational agent by extension based on multi-turn and dialog history.

These component technologies and methodologies for visual language integration are evolving but are still limited and require large-scale improvements for immersive interactions. Providing immersive interactions is critical in extending the latest advances in AI technology to more tasks that users can actually benefit from. Therefore, embodied visual language interactions are still an open issue and require further investigation. In terms of user benefits, visual language integration is utilized to construct and evolve large virtual systems (e.g., Metaverse). In particular, AR and VR are used as visual language interfaces in the virtual world, allowing users to experience life like a movie.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2012635).

References

[1] Y. Pang, J. Lin, T. Qin, Z. Chen, Image-to-image translation: Methods and applications, IEEE Trans. Multimed. (2021) http://dx.doi.org/10.1109/TMM.2021.3109419.

[2] T. Jumneanbun, S. Sae-Lao, P. Paliyawan, R. Thawonmas, K. Sookhanaphibarn, W. Choensawat, Rap-style comment generation to entertain game live streaming, in: 2020 IEEE Conference on Games (CoG), 2020, pp. 706–707, http://dx.doi.org/10.1109/CoG47356.2020.9231636.

[3] Y. Li, Z. Gan, et al., Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6329–6338, http://dx.doi.org/10.1109/CVPR.2019.00649.

[4] C. Chan, S. Ginosar, T. Zhou, A.A. Efros, Everybody dance now, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5933–5942, http://dx.doi.org/10.1109/ICCV.2019.00603.

[5] AI Dungeon, https:play.aidungeon.io (consulted in 2022).

[6] Y. Jiang, W. Li, M.S. Hossain, M. Chen, A. Alelaiwi, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, Inf. Fusion 53 (2020) 209–221, http://dx.doi.org/10.1016/j.inffus.2019.06.019.

[7] A. Jaiswal, A.R. Babu, M.Z. Zadeh, D. Banerjee, F. Makedon, A survey on contrastive self-supervised learning, Technologies 9 (1) (2021) 2, http://dx.doi.org/10.3390/technologies9010002.

[8] Y. Zheng, G. Chen, M. Huang, S. Liu, X. Zhu, Persona-aware dialogue generation with enriched profile, in: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019.

[9] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[10] L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences, Minds Mach. 30 (4) (2020) 681–694.

[11] S.M. Park, Y.G. Kim, Visual language navigation: a survey and open challenges, Artif. Intell. Rev. (2022) 1–63, http://dx.doi.org/10.1007/s10462-022-10174-9.

[12] O. Caglayan, P. Madhyastha, L. Specia, L. BarraultOzan, Probing the need for visual context in multimodal machine translation, in: NAACL-HLT (1), 2019, pp. 4159–4170, http://dx.doi.org/10.18653/v1/N19-1422.

[13] D. Elliott, S. Frank, L. Barrault, F. Bougares, L. Specia, Findings of the second shared task on multimodal machine translation and multilingual image description, in: Proceedings of the Second Conference on Machine Translation, 2017, pp. 215–233, http://dx.doi.org/10.18653/v1/W17-4718.

[14] D. Elliott, Adversarial evaluation of multimodal machine translation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2974–2978.

[15] M. Ehatisham-Ul-Haq, A. Javed, M.A. Azam, H.M. Malik, A. Irtaza, I.H. Lee, M.T. Mahmood, Robust human activity recognition using multimodal feature-level fusion, IEEE Access 7 (2019) 60736–60751.

[16] J.B. Delbrouck, S. Dupont, An empirical study on the effectiveness of images in multimodal neural machine translation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 910–919.

[17] A. Zadeh, P. Pu, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2018, http://dx.doi.org/10.18653/v1/P18-1208.

[18] P.P. Liang, A. Zadeh, L.P. Morency, Multimodal local–global ranking fusion for emotion recognition, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 472–476, http://dx.doi.org/10.1145/3242969.3243019.

[19] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, A. Divakaran, Integrating text and image: Determining multimodal document intent in instagram posts, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 4622–4632.

[20] J. Libovický, S. Palaskar, S. Gella, F. Metze, Multimodal abstractive summarization of open-domain videos, in: Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL) NIPS 2018, Montreal, Canada, 2018.

[21] Y.H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, p. 6558, http://dx.doi.org/10.18653/v1/P19-1656.

[22] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, S.F. Chang, Multi-level multimodal common semantic space for image-phrase grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12476–12486.

[23] D. Hu, F. Nie, X. Li, Deep multimodal clustering for unsupervised audiovisual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9248–9257.

[24] Z. Li, Z. Li, J. Zhang, Y. Feng, J. Zhou, Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 2476–2483.

[25] T. Domhan, F. Hieber, Using target-side monolingual data for neural machine translation through multi-task learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1500–1505.

[26] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, H. van Hasselt, Multi-task deep reinforcement learning with popart, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3796–3803, (1).

[27] X. Wang, V. Jain, E. Ie, W.Y. Wang, Z. Kozareva, S. Ravi, Natural language grounded multitask navigation, natural language grounded multitask navigation, in: NeurIPS Visually Grounded Interaction and Language (ViGIL), 2019.

[28] R.J. Skerry-Ryan, E. Battenberg, et al., Towards end-to-end prosody transfer for expressive speech synthesis with tacotron, in: International Conference on Machine Learning, 2018, pp. 4693–4702.

[29] D. Ham, J.G. Lee, Y. Jang, K.E. Kim, End-to-end neural pipeline for goal-oriented dialogue system using GPT-2, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 583–592.

[30] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, E. Tsuprun, Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system, in: 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU, IEEE, 2017, pp. 569–576.

[31] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, Learning to reason: End-to-end module networks for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 804–813.

[32] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, Y. Bengio, Towards end-to-end spoken language understanding, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 5754–5758.
[33] C. Ahuja, S. Ma, L.P. Morency, Y. Sheikh, To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations, in: 2019 International Conference on Multimodal Interaction, 2019, pp. 74–84.
[34] H. Le, D. Sahoo, N. Chen, S. Hoi, Multimodal transformer networks for end-to-end video-grounded dialogue systems, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 5612–5623.
[35] H. Luo, G. Lin, Z. Liu, F. Liu, Z. Tang, Y. Yao, SegEQA: Video segmentation based visual attention for embodied question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9667–9676.
[36] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T.L. Berg, D. Batra, Multi-target embodied question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6309–6318.
[37] E. Wijmans, S. Datta, et al., Embodied question answering in photorealistic environments with point cloud perception, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6659–6668.
[38] K. Nguyen, D. Dey, C. Brockett, B. Dolan, Vision-based navigation with language-based assistance via imitation learning with indirect intervention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12527–12537.
[39] T.C. Chi, M. Shen, M. Eric, S. Kim, D. Hakkani-tur, Just ask: An interactive learning framework for vision and language navigation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 2459–2466, (3).
[40] J. Thomason, M. Murray, M. Cakmak, L. Zettlemoyer, Vision-and-dialog navigation, in: Conference on Robot Learning, PMLR, 2020, pp. 394–406.
[41] D. Fried, R. Hu, et al., Speaker-follower models for vision-and-language navigation, in: Advances in Neural Information Processing Systems, 2018, pp. 3318–3329.
[42] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, D. Batra, Embodied question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1–10.
[43] S.D. Morad, R. Mecca, R.P. Poudel, S. Liwicki, R. Cipolla, Embodied visual navigation with automatic curriculum learning in real environments, IEEE Robot. Autom. Lett. 6 (2) (2021) 683–690.
[44] M. Savva, A. Kadian, et al., Habitat: A platform for embodied ai research, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9339–9347.
[45] D. Gordon, A. Kadian, D. Parikh, J. Hoffman, D. Batra, Splitnet: Sim2sim and task2task transfer for embodied visual navigation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1022–1031.
[46] F.L. Da Silva, A.H.R. Costa, A survey on transfer learning for multiagent reinforcement learning systems, J. Artificial Intelligence Res. 64 (2019) 645–703.
[47] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J. Tenenbaum, Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 1039–1050.
[48] A. Ray, K. Sikka, A. Divakaran, S. Lee, G. Burachas, Sunny and dark outside?! improving answer consistency in VQA through entailed question generation, in: 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 2019, pp. 5860–5865.
[49] S. Chen, W. Wang, L. Ruan, L. Yao, Q. Jin, YouMakeup VQA challenge: Towards fine-grained action understanding in domain-specific videos, in: CVPR LVVU Workshop, 2020.
[50] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, D. Parikh, Do explanations make VQA models more predictable to a human? in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, Brussels, Belgium, 2018, pp. 1036–1042.
[51] J.B. Delbrouck, A. Maiorca, N. Hubens, S. Dupont, Modulated self-attention convolutional network for VQA, in: ViGIL@NeurIPS, 2019.
[52] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, 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, 2017, pp. 6904–6913.
[53] A. Singh, V. Natarajan, et al., Towards vqa models that can read, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8317–8326.
[54] B. Patro, V. Namboodiri, Explanation vs attention: A two-player game to obtain attention for VQA, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11848–11855, (07).
[55] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, J. Gao, Unified vision-language pre-training for image captioning and vqa, Assoc. Adv. Artif. Intell. 34 (07) (2019) 13041–13049.
[56] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert: A joint model for video and language representation learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7464–7473.
[57] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019) 13–23.
[58] Multi30k, https://github.com/multi30k/dataset (consulted in 2022).
[59] K. Shuang, Z. Zhang, J. Loo, S. Su, Convolution–deconvolution word embedding: An end-to-end multi-prototype fusion embedding method for natural language processing, Inf. Fusion 53 (2020) 112–122.
[60] J. Zhang, K. Shih, A. Tao, B. Catanzaro, A. Elgammal, An interpretable model for scene graph generation, in: Advances in Neural Information Processing Systems, 2018.
[61] K. Ethayarajh, How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 55–65.
[62] Text-to-text transfer transformer, https:github.com/google-research/text-to-text-transfer-transformer (consulted in 2022).
[63] Salesforce CTRL, https:github.com/salesforce/ctrl (consulted in 2022).
[64] L.I. Xuhong, Y. Grandvalet, F. Davoine, Explicit inductive bias for transfer learning with convolutional networks, in: International Conference on Machine Learning, 2018, pp. 2825–2834.
[65] P. Budzianowski, T.H. Wen, B.H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, M. Gašić, Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgiumm, 2018, pp. 5016–5026.
[66] Y.A. Chung, Y. Wang, W.N. Hsu, Y. Zhang, R.J. Skerry-Ryan, Semi-supervised training for improving data efficiency in end-to-end speech synthesis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 6940–6944.
[67] Y. Jia, R.J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, Y. Wu, Direct speech-to-speech translation with a sequence-to-sequence model, in: Interspeech, 2019.
[68] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry, Implementation matters in deep RL: A case study on PPO and TRPO, in: International Conference on Learning Representations, ICLR, 2019.
[69] J. Li, B. Peng, et al., Results of the multi-domain task-completion dialog challenge, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, Eighth Dialog System Technology Challenge Workshop, Vol. 7, 2020.
[70] S. Lee, Q. Zhu, et al., ConvLab: Multi-Domain End-to-End Dialog System Platform, ACL, Florence, Italy, 2019, pp. 64–69.
[71] R. Herzig, M. Raboh, G. Chechik, J. Berant, A. Globerson, Mapping images to scene graphs with permutation-invariant structured prediction, in: Advances in Neural Information Processing Systems, 2018, pp. 7211–7221.
[72] N.F. Chen, Z. Du, K.H. Ng, Scene graphs for interpretable video anomaly classification, in: Advances in Neural Information Processing Systems, 2018.
[73] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.
[74] D. Mehta, O. Sotnychenko, et al., XNect: Real-time multi-person 3D motion capture with a single RGB camera, ACM Trans. Graph. 1 (2020) 39 (4).
[75] M. Qi, Y. Wang, J. Qin, A. Li, KE-GAN: Knowledge embedded generative adversarial networks for semi-supervised scene parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5237–5246.
[76] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph r-cnn for scene graph generation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 670–685.
[77] D. Camacho, A. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, E. Cambria, The four dimensions of social network analysis: An overview of research methods, applications, and software tools, Inf. Fusion 63 (2020) 88–120.
[78] V. Chen, A. Gupta, K. Marino, Ask your humans: Using human instructions to improve generalization in reinforcement learning, in: ICLR 2020, 2020.
[79] Z. He, S. Sunkara, ActionBert: Leveraging user actions for semantic understanding of user interfaces, in: AAAI Conference on Artificial Intelligence (AAAI-21), 2021, pp. 5931–5938.

[80] H. Alamri, C. Hori, T.K. Marks, D. Batra, D. Parikh, Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7, in: DSTC7 at AAAI2019 Workshop. Vol. 2, 2018.
[81] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, X. Wang, Factorizable net: an efficient subgraph-based framework for scene graph generation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 335–351.
[82] A. Newell, J. Deng, Pixels to graphs by associative embedding, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 2168–2177.
[83] H. Li, C. Liu, S. Zhu, K. Yu, Robust spoken language understanding with acoustic and domain knowledge, in: 2019 International Conference on Multimodal Interaction, 2019, pp. 531–535.
[84] M. Poesio, R. Stuckardt, Y. Versley, Anaphora Resolution, Springer, 2016.
[85] S. Shekhar, U. Kumar, U. Sharma, To reduce the multidimensionality of feature set for anaphora resolution algorithm, in: Ambient Communications and Computer Systems, Springer, Singapore, 2018, pp. 437–446.
[86] A. Rohrbach, M. Rohrbach, S. Tang, S. Joon Oh, B. Schiele, Generating descriptions with grounded and co-referenced people, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4979–4989.
[87] B. Aktaş, T. Scheffler, M. Stede, Anaphora resolution for Twitter conversations: An exploratory study, in: Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference, 2018, pp. 1–10.
[88] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, J.R. Wen, Recursive visual attention in visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6679–6688.
[89] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Inf. Fusion 59 (2020) 139–162.
[90] M. Garnelo, M. Shanahan, Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, Curr. Opin. Behav. Sci. 29 (2019) 17–23.
[91] A.S. Gordon, C.A. Bejan, K. Sagae, Commonsense causal reasoning using millions of personal stories, in: Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.
[92] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, V. Prasanna, Graphsaint: Graph sampling based inductive learning method, in: ICRL 2020, 2020.
[93] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR 2017, 2017.
[94] Z. Guo, Y. Zhang, W. Lu, Attention guided graph convolutional networks for relation extraction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, Italy, 2019, pp. 241–251.
[95] A. Das, S. Kottur, J.M. Moura, S. Lee, D. Batra, Learning cooperative visual dialog agents with deep reinforcement learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2951–2960.
[96] C. Schroeder de Witt, J. Foerster, G. Farquhar, P. Torr, W. Boehmer, S. Whiteson, Multi-agent common knowledge reinforcement learning, in: Advances in Neural Information Processing Systems, 2019, pp. 9927–9939.
[97] J.A. Mendez, A. Geramifard, M. Ghavamzadeh, B. Liu, Reinforcement learning of multi-domain dialog policies via action embeddings, in: 3rd Workshop on Conversational AI Today's Practice Tomorrow's Potential, NeurIPS, 2019.
[98] J. Wu, G. Li, S. Liu, L. Lin, Tree-structured policy based progressive reinforcement learning for temporally language grounding in video, Assoc. Adv. Artif. Intell. 34 (7) (2020) 12386–12393.
[99] J. Wang, Y. Zhang, T.K. Kim, Y. Gu, Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system, in: ICLR 2021, 2021.
[100] J.M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, F. Jurie, Mfas: Multimodal fusion architecture search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6966–6975.
[101] S. Sohn, H. Woo, J. Choi, H. Lee, Meta reinforcement learning with autonomous inference of subtask dependencies, in: ICLR 2020, 2020.
[102] L. Liu, T. Zhou, G. Long, J. Jiang, C. Zhang, Learning to propagate for graph meta-learning, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.
[103] R. Vuorio, S.H. Sun, H. Hu, J.J. Lim, Multimodal model-agnostic meta-learning via task-aware modulation, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 1–12.
[104] J. Jiang, C. Dun, T. Huang, Z. Lu, Graph convolutional reinforcement learning, in: ICLR 2020, 2020.
[105] A. Paliwal, F. Gimeno, V. Nair, Y. Li, M. Lubin, P. Kohli, O. Vinyals, Reinforced genetic algorithm learning for optimizing computation graphs, in: ICLR 2020, 2020.
[106] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? in: ICLR 2018, 2018.
[107] M. Brockschmidt, M. Allamanis, A.L. Gaunt, O. Polozov, Generative code modeling with graphs, in: ICLR 2018, 2018.
[108] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, R. Girshick, Phyre: A new benchmark for physical reasoning, in: Advances in Neural Information Processing Systems 2019, 2019.
[109] M. Moscovitch, R. Cabeza, G. Winocur, L. Nadel, Episodic memory and beyond: the hippocampus and neocortex in transformation, Annu. Rev. Psychol. 67 (2016) 105–134.
[110] D. Lopez-Paz, M.A. Ranzato, Gradient episodic memory for continual learning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 6470–6479.
[111] S.J. Gershman, N.D. Daw, Nathaniel reinforcement learning and episodic memory in humans and animals: an integrative framework, Annu. Rev. Psychol. 68 (2017) 101–128.
[112] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2018, p. 2122.
[113] Y. Gao, C.S. Wu, et al., EMT: Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading, ACL, 2020.
[114] P.Y. Oudeyer, J. Gottlieb, M. Lopes, Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies, Prog. Brain Res. 229 (2016) 257–284, Elsevier.
[115] R. Chitnis, S. Tulsiani, S. Gupta, A. Gupta, Intrinsic motivation for encouraging synergistic behavior, in: ICLR 2020, 2020.
[116] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1 (4) (1978) 515–526.
[117] D. Melhart, G.N. Yannakakis, A. Liapis, I feel i feel you: A theory of mind experiment in games, KI-Künstl. Intell. 34 (1) (2020) 45–55.
[118] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 4967–4976.
[119] D. Wang, M. Jamnik, P. Lio, Abstract diagrammatic reasoning with multiplex graph networks, in: ICRL 2020, 2020.
[120] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, C. Xiong, Learning to retrieve reasoning paths over wikipedia graph for question answering, in: ICLR 2020, 2020.
[121] W. Xiong, T. Hoang, W.Y. Wang, DeepPath: A reinforcement learning method for knowledge graph reasoning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 564–573.
[122] Z. Zheng, W. Wang, S. Qi, S.C. Zhu, Reasoning visual dialogs with structural and partial observations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6669–6678.
[123] H. Zhu, Y. Lin, Z. Liu, J. Fu, T.S. Chua, M. Sun, Graph neural networks with generated parameters for relation extraction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019 Association for Computational Linguistics, 2019, pp. 1331–1339.
[124] C. Xu, R. Li, Relation embedding with dihedral group in knowledge graph, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, Italy, 2019, pp. 263–272.
[125] S. Woo, D. Kim, D. Cho, I.S. Kweon, Linknet: Relational embedding for scene graph, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 558–568.
[126] R. Cadene, H. Ben-Younes, M. Cord, N. Thome, Murel: Multimodal relational reasoning for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1989–1998.
[127] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, Y. Artzi, A corpus for reasoning about natural language grounded in photographs, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6418–6428.
[128] Y.W. Chu, K.Y. Lin, C.C. Hsu, L.W. Ku, Multi-Step Joint-Modality Attention Network for Scene-Aware Dialogue System, Association for the Advancement of Artificial Intelligence, 2020.
[129] T. Bansal, D.C. Juan, S. Ravi, A. McCallum, A2N: Attending to neighbors for knowledge graph inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4387–4392.
[130] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, COMET: Commonsense transformers for automatic knowledge graph construction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 4762–4779.
[131] Q. Long, Z. Zhou, A. Gupta, F. Fang, Y. Wu, X. Wang, Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learning (Diss.), Robotics Institute, 2020.
[132] S. Sukhbaatar, Z. Lin, I. Kostrikov, G. Synnaeve, A. Szlam, R. Fergus, Intrinsic motivation and automatic curricula via asymmetric self-play, in: ICLR 2018, 2018.

[133] B. Eysenbach, A. Gupta, et al., Diversity is all you need: Learning skills without a reward function, in: ICLR 2019, 2019.
[134] S. Omidshafiei, D.K. Kim, et al., Learning to teach in cooperative multiagent reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6128–6136.
[135] O. Vinyals, I. Babuschkin, et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354.
[136] H. Hu, J.N. Foerster, Simplified action decoder for deep multi-agent reinforcement learning, in: ICLR 2020, 2020.
[137] Y. Wen, Y. Yang, R. Luo, J. Wang, W. Pan, Probabilistic recursive reasoning for multi-agent reinforcement learning, in: ICLR 2019, 2019.
[138] D. Ha, J. Schmidhuber, Recurrent world models facilitate policy evolution, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 2455–2467.
[139] S. Racanière, T. Weber, Imagination-augmented agents for deep reinforcement learning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5694–5705.
[140] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.
[141] K. Nguyen, H. Daumé III, Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 684–695.
[142] S.W. Lee, T. Gao, S. Yang, J. Yoo, J.W. Ha, Large-scale answerer in questioner's mind for visual dialog question generation, in: ICLR 2019, 2019.
[143] J. Luketina, N. Nardelli, et al., A survey of reinforcement learning informed by natural language, in: International Joint Conferences on Artificial Intelligence, Vol. 57, 2019, http://dx.doi.org/10.24963/ijcai.2019/880.
[144] N. Lair, C. Colas, R. Portelas, J.M. Dussoux, P.F. Dominey, P.Y. Oudeyer, Language grounding through social interactions and curiosity-driven multi-goal learning, in: ViGIL@NeurIPS 2019, 2019.
[145] P. Goyal, S. Niekum, R.J. Mooney, Using natural language for reward shaping in reinforcement learning, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 2385–2391, http://dx.doi.org/10.24963/ijcai.2019/331.
[146] K. Narasimhan, R. Barzilay, T. Jaakkola, Grounding language for transfer in deep reinforcement learning, J. Artificial Intelligence Res. 63 (2018) 849–874.
[147] P. Ammanabrolu, M. Hausknecht, Graph constrained reinforcement learning for natural language action spaces, in: ICRL, 2020.
[148] M. Asai, C. Muise, Learning neural-symbolic descriptive planning models via cube-space priors: The voyage home (to STRIPS), in: IJCAI, 2020, pp. 2676–2682.
[149] D. Xu, R. Martín-Martín, D.A. Huang, Y. Zhu, S. Savarese, L.F. Fei-Fei, Regression planning networks, in: Advances in Neural Information Processing Systems, 2019, pp. 1319–1329.
[150] M. Fox, D. Long, PDDL2. 1: An extension to PDDL for expressing temporal planning domains, J. Artificial Intelligence Res. 20 (2003) 61–124.
[151] L. Reed, S. Oraby, M. Walker, Can neural generators for dialogue learn sentence planning and discourse structuring? in: Proceedings of the 11th International Natural Language Generation Conference, 2018 Association for Computational Linguistics, 2018, pp. 284–295.
[152] A.N. Bajpai, S. Garg, Transfer of deep reactive policies for mdp planning, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.
[153] J.B. Hamrick, K.R. Allen, V. Bapst, T. Zhu, K.R. McKee, J.B. Tenenbaum, P.W. Battaglia, Relational inductive bias for physical construction in humans and machines, in: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2018), 2018.
[154] J.B. Hamrick, Analogues of mental simulation and imagination in deep learning, Curr. Opin. Behav. Sci. 29 (2019) 8–16.
[155] C. Davis, L. Bulat, A. Vero, E. Shutova, Modelling visual properties and visual context in multimodal semantics, in: Advances in Neural Information Processing Systems, 2018.
[156] T.D. Kulkarni, K. Narasimhan, A. Saeedi, J. Tenenbaum, Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation, in: Advances in Neural Information Processing Systems, Vol. 29, 2016, 3862-3690.
[157] M. Shridhar, X. Yuan, M.A. Côté, Y. Bisk, A. Trischler, M. Hausknecht, Alfworld: Aligning text and embodied environments for interactive learning, in: ICLR 2021, 2021.
[158] G. Chrupała, Symbolic inductive bias for visually grounded learning of spoken language, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, Italy, 2019, pp. 6452–6462, http://dx.doi.org/10.18653/v1/P19-1647.
[159] L. Shu, P. Molino, M. Namazifar, B. Liu, H. Xu, H. Zheng, G. Tur, Incorporating the structure of the belief state in end-to-end task-oriented dialogue systems, in: 2nd Workshop on Conversational AI At Neural Information Processing Systems, Vol. 32, 2018.
[160] W. Dabney, Z. Kurth-Nelson, N. Uchida, C.K. Starkweather, D. Hassabis, R. Munos, M. Botvinick, A distributional code for value in dopamine-based reinforcement learning, Nature (2020) 1–5.
[161] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S.A. Eslami, M. Botvinick, Machine theory of mind, in: Proceedings of the 35 th International Conference on Machine Learning, Vol. 80, PMLR, Stockholm, Sweden, 2018, pp. 4218–4227.
[162] X. Puig, T. Shu, et al., Watch-and-help: A challenge for social perception and human-AI collaboration, in: ICLR 2021, 2021.
[163] S. Pidhorskyi, D.A. Adjeroh, G. Doretto, Adversarial latent autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14104–14113.
[164] M.L. Shih, S.Y. Su, J. Kopf, J.B. Huang, 3D photography using context-aware layered depth inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8028–8038.
[165] Y.C. Chen, Z. Gan, Y. Cheng, J. Liu, J. Liu, Distilling the Knowledge of BERT for Text Generation, Association for Computational Linguistics, 2019, pp. 5043–5053.
[166] S. Dathathri, A. Madotto, Plug and play language models: a simple approach to controlled text generation, in: ICLR 2020, 2020.
[167] A. El-Nouby, S. Sharma, et al., Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10303–10311, http://dx.doi.org/10.1109/ICCV.2019.01040.
[168] J.P. Briot, G. Hadjeres, F.D. Pachet, Deep Learning Techniques for Music Generation, Vol. 10, Springer, 2019.
[169] M. Fazel-Zarandi, S. Biswas, R. Summers, A. Elmalt, A. McCraw, M. McPhilips, J. Peach, Towards personalized dialog policies for conversational skill discovery, in: The 3rd Conversational AI Workshop, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019.
[170] K. Shuster, S. Humeau, H. Hu, A. Bordes, J. Weston, Engaging image captioning via personality, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12516–12526.
[171] COCO dataset, https://cocodataset.org/ (consulted in 2022).
[172] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015, pp. 2641–2649, http://dx.doi.org/10.1109/ICCV.2015.303.
[173] Y. Ma, K.L. Nguyen, F.Z. Xing, E. Cambria, A survey on empathetic dialogue systems, Inf. Fusion 64 (2020) 50–70.
[174] N. Lubis, S. Sakti, K. Yoshino, S. Nakamura, Dialogue model and response generation for emotion improvement elicitation, in: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019.
[175] K. Chandu, S. Prabhumoye, R. Salakhutdinov, A.W. Black, My way of telling a story: Persona based grounded story generation, in: Proceedings of the Second Workshop on Storytelling, 2019, pp. 11–21.
[176] M. Dinculescu, J. Engel, A. Roberts, MidiMe: Personalizing a MusicVAE model with user data, in: NeurIPS Workshop on Machine Learning for Creativity and Design 3.0, NIPS 2019, 2019.
[177] L. Qin, A. Bosselut, A. Holtzman, C. Bhagavatula, E. Clark, Y. Choi, Counterfactual story reasoning and generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5043–5053.
[178] T. Shen, T. Lei, R. Barzilay, T. Jaakkola, Style transfer from non-parallel text by cross-alignment, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 6833–6844.
[179] R. Benmalek, M. Khabsa, S. Desu, C. Cardie, M. Banko, Keeping notes: Conditional natural language generation with a scratchpad encoder, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4157–4167.
[180] S.M. Park, Y.G. Kim, Survey and challenges of story generation models - A multimodal perspective with five steps: Data embedding, topic modeling, storyline generation, draft story generation, and story evaluation, Inf. Fusion 67 (2021) 41–63, Elsevier.
[181] W. Zhang, J. Yu, H. Hu, H. Hu, Z. Qin, Multimodal feature fusion by relational reasoning and attention for visual question answering, Inf. Fusion 55 (2020) 116–126.
[182] A. Zadeh, M. Chan, P.P. Liang, E. Tong, L.P. Morency, Social-iq: A question answering benchmark for artificial social intelligence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8807–8817.
[183] J. Liang, L. Jiang, L. Cao, L.J. Li, A.G. Hauptmann, Focal visual-text attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6135–6143.
[184] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: Reasoning with graph convolution nets for factual visual question answering, in: Advances in Neural Information Processing Systems, 2018, pp. 2659–2670.

[185] D. Linsley, D. Shiebler, S. Eberhardt, T. Serre, Learning what and where to attend, in: Seventh International Conference on Learning Representations, Google, New Orleans, LA, 2019.
[186] T.C. Ferreira, C. van der Lee, E. Van Miltenburg, E. Krahmer, Neural data-to-text generation: A comparison between pipeline and end-to-end architectures, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 552–562.
[187] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.P. Morency, Memory fusion network for multi-view sequential learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5634–5641.
[188] M.K. Moghaddam, Q. Wu, E. Abbasnejad, J. Shi, Optimistic agent: Accurate graph-based value estimation for more successful visual navigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 3733–3742.
[189] L. Xie, A. Markham, N. Trigoni, Snapnav: Learning mapless visual navigation with sparse directional guidance and visual reference, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2020, pp. 1682–1688.
[190] X. Ma, P. Karkus, D. Hsu, W.S. Lee, N. Ye, Discriminative particle filter reinforcement learning for complex partial observations, in: ICLR 2020, 2020.
[191] D.S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, R. Salakhutdinov, Devendra Singh, et al., Learning to explore using active neural slam, in: ICLR 2020, 2020.
[192] D.S. Chaplot, R. Salakhutdinov, A. Gupta, S. Gupta, Neural topological SLAM for visual navigation, in: CVPR 2020, 2020, pp. 12875–12884.
[193] C.P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, A. Lerchner, MONet: Unsupervised scene decomposition and representation, 2019, CoRR abs/1901.11390.
[194] T.D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, V. Mnih, Unsupervised learning of object keypoints for perception and control, in: Advances in Neural Information Processing Systems, 2019, pp. 10724–10734.
[195] A. Anand, E. Racah, S. Ozair, Y. Bengio, M.A. Côté, R.D. Hjelm, Unsupervised state representation learning in atari, in: Advances in Neural Information Processing Systems, 2019, pp. 8769–8782.
[196] Y. Jiang, S.S. Gu, K.P. Murphy, C. Finn, Language as an abstraction for hierarchical deep reinforcement learning, in: Advances in Neural Information Processing Systems, 2019, pp. 9419–9431.
[197] A.H. Miller, W. Feng, et al., ParlAI: A dialog research software platform, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2017, pp. 79–84, http://dx.doi.org/10.18653/v1/D17-2014.
[198] F. Hill, O. Tieleman, T. von Glehn, N. Wong, H. Merzic, S. Clark, Grounded language learning fast and slow, in: ICLR 2021, 2021.
[199] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Inf. Fusion 58 (2020) 52–68.
[200] C. Yoo, B. Kang, M. Cho, SNOW: Subscribing to knowledge via channel pooling for transfer & lifelong learning of convolutional neural networks, in: International Conference on Learning Representations, 2019.
[201] F.K. Sun, C.H. Ho, H.Y. Lee, Lamol: Language modeling for lifelong language learning, in: ICLR 2020, 2020.
[202] S. Mazumder, B. Liu, S. Wang, N. Ma, Lifelong and interactive learning of factual knowledge in dialogues, in: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue 2019, 2019, pp. 21–31, http://dx.doi.org/10.18653/v1/W19-5903.
[203] F. Garcia, P.S. Thomas, A meta-mdp approach to exploration for lifelong reinforcement learning, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 5691–5700.
[204] Y. Wen, D. Tran, J. Ba, BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning, in: ICLR 2020, 2020.
[205] D. Tanneberg, J. Peters, E. Rueckert, Intrinsic motivation and mental replay enable efficient online adaptation in stochastic recurrent networks, Neural Netw. 109 (2019) 67–80.
[206] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[207] S. Shen, L.H. Li, H. Tan, et al., How much can CLIP benefit vision-and-language tasks? in: International Conference on Learning Representations, 2021.
[208] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, P. Fung, Enabling multimodal generation on CLIP via vision-language knowledge distillation, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 2383–2395.
[209] A. Nichol, P. Dhariwal, A. Ramesh, et al., Glide: Towards photorealistic image generation and editing with text-guided diffusion models, in: Proceedings of the 39th International Conference on Machine Learning, Vol. 162, PMLR, Baltimore, Maryland, USA, 2022.
[210] O. Avrahami, D. Lischinski, O. Fried, Blended diffusion for text-driven editing of natural images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18208–18218.
[211] S. Gu, D. Chen, J. Bao, et al., Vector quantized diffusion model for text-to-image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10696–10706.
[212] Text to Video: Early Access, Runway, https://www.youtube.com/watch?v=mYjfIc9xw90, (consulted in October 2022).
[213] Make-A-Video, Meta, https://makeavideo.studio/ (consulted in 2022).
[214] Phenaki, Google, https://phenaki.video/ (consulted in 2022).
[215] S.M. Park, Y.G. Kim, A metaverse: Taxonomy, components, applications, and open challenges, IEEE Access 10 (2022) 4209–4251, IEEE.