ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fcomp.2021.618069

Robot Concept Acquisition Based on Interaction Between Probabilistic and Deep Generative Models

Ryo Kuniyasu[1], *Tomoaki Nakamura*[1]*, *Tadahiro Taniguchi*[2] and *Takayuki Nagai*[3,4]

[1]Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, Tokyo, Japan, [2]College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan, [3]Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, Osaka, Japan, [4]Artificial Intelligence EXploration Research Center, The University of Electro-Communications, Tokyo, Japan

We propose a method for multimodal concept formation. In this method, unsupervised multimodal clustering and cross-modal inference, as well as unsupervised representation learning, can be performed by integrating the multimodal latent Dirichlet allocation (MLDA)-based concept formation and variational autoencoder (VAE)-based feature extraction. Multimodal clustering, representation learning, and cross-modal inference are critical for robots to form multimodal concepts from sensory data. Various models have been proposed for concept formation. However, in previous studies, features were extracted using manually designed or pre-trained feature extractors and representation learning was not performed simultaneously. Moreover, the generative probabilities of the features extracted from the sensory data could be predicted, but the sensory data could not be predicted in the cross-modal inference. Therefore, a method that can perform clustering, feature learning, and cross-modal inference among multimodal sensory data is required for concept formation. To realize such a method, we extend the VAE to the multinomial VAE (MNVAE), the latent variables of which follow a multinomial distribution, and construct a model that integrates the MNVAE and MLDA. In the experiments, the multimodal information of the images and words acquired by a robot was classified using the integrated model. The results demonstrated that the integrated model can classify the multimodal information as accurately as the previous model despite the feature extractor learning in an unsupervised manner, suitable image features for clustering can be learned, and cross-modal inference from the words to images is possible.

Keywords: concept formation, symbol emergence in robotics, probabilistic generative model, deep generative model, unsupervised learning, representation learning, cross-modal inference

OPEN ACCESS

Edited by:
Anna Esposito,
University of Campania 'Luigi
Vanvitelli, Italy

Reviewed by:
Attila Kovari,
University of Dunaújváros, Hungary
Juan Sebastian Olier,
Tilburg University, Netherlands

****Correspondence:***
Tomoaki Nakamura
tnakamura@uec.ac.jp

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *26 January 2021*
Accepted: *06 August 2021*
Published: *03 September 2021*

Citation:
Kuniyasu R, Nakamura T, Taniguchi T
and Nagai T (2021) Robot Concept
Acquisition Based on Interaction
Between Probabilistic and Deep
Generative Models.
Front. Comput. Sci. 3:618069.
doi: 10.3389/fcomp.2021.618069

1 INTRODUCTION

Not only multimodal clustering and cross-modal inference but also representation learning is critical for robots to form multimodal concepts from sensory data. We define multimodal categories that enable multimodal clustering and cross-modal inference as concepts. Cross-modal inference enables a robot to infer unobserved information from limited observed information, and this ability allows robots to respond to uncertain environments. Various models have been proposed for concept formation (Nakamura et al., 2014; Taniguchi et al., 2017). These models perform clustering and cross-modal inference based on the multimodal latent Dirichlet allocation (MLDA) (Nakamura

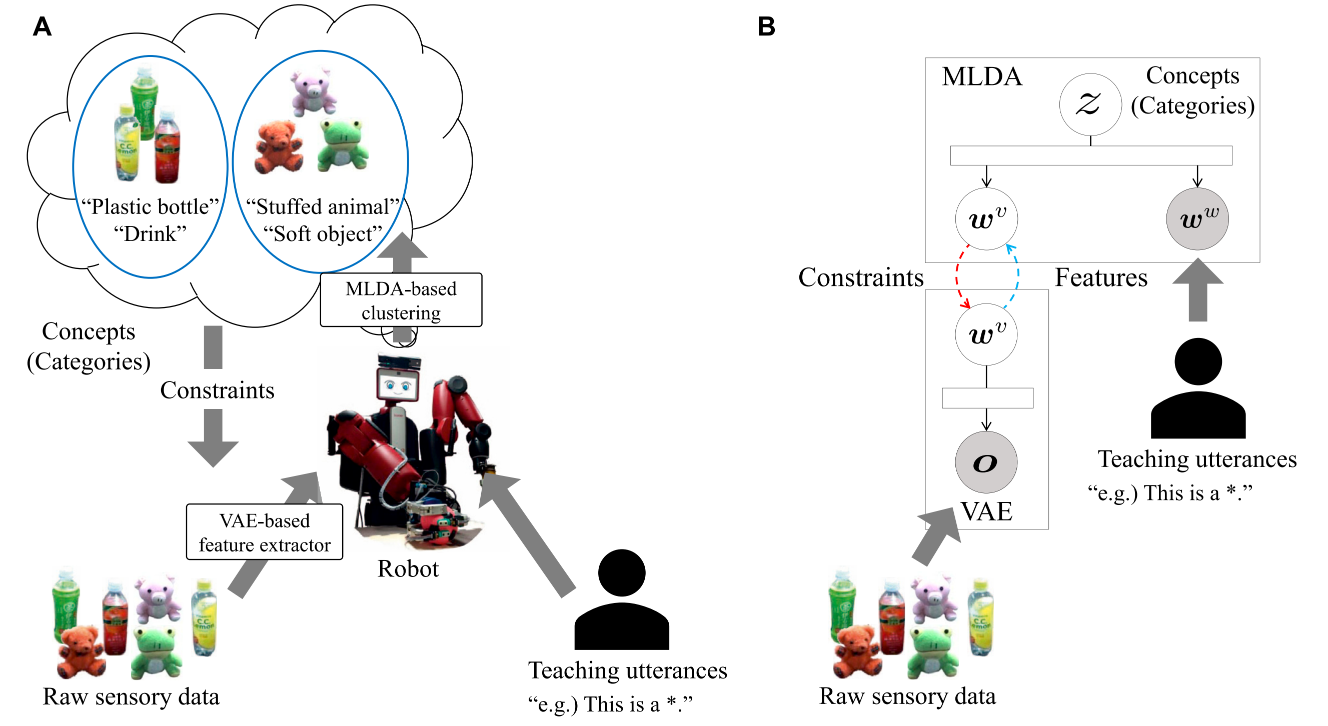

FIGURE 1 | Overview of study: **(A)** system and **(B)** corresponding integrated model proposed.

et al., 2009). However, pre-designed or pre-trained feature extractors are used, and representation learning is not performed simultaneously. Feature learning is an ability required for robots to obtain environment-dependent knowledge and adapt to the environment. Moreover, in these previous studies, the generative probabilities of the features extracted from sensory data could be predicted, but the sensory data could not be predicted in the cross-modal inference. Therefore, a method that can perform clustering, feature learning, and cross-modal inference of the sensory data is required for concept formation.

In this study, we focus on concept formation from image and word information, and propose a method to perform clustering and to learn suitable feature extractors simultaneously by integrating the MLDA-based concept formation and variational autoencoder (VAE)-based (Kingma and Welling, 2013) feature extractor. **Figure 1** presents an overview of this study. A robot forms concepts from the images captured by the camera and words given by the human user who teaches the object features. Furthermore, the formed concepts influence the learning of the feature extractor, making it possible to extract suitable features for concept formation in the environment.

Intelligent robots that learn from the information obtained from the environment and the humans (Ridge et al., 2010; Taniguchi et al., 2010; Mangin et al., 2015; Wächter and Asfour, 2015) have been developed in the field known as symbol emergence in robotics (Taniguchi et al., 2016, 2018). This field includes various research topics, such as concept formation (Nakamura et al., 2007; Ridge et al., 2010; Mangin et al., 2015), language acquisition (Mochihashi et al., 2009; Neubig et al., 2012), learning of interaction (Taniguchi et al., 2010), and segmentation and learning of actions (Wächter and Asfour, 2015; Nagano et al., 2019). The symbol emergence system was proposed by Taniguchi (Taniguchi et al., 2016, 2018), and it was inspired by the genetic epistemology proposed by Piaget (Piaget and Duckworth, 1970). In the symbol emergence system, robots self-organize into symbols from the multimodal sensory information obtained from the environment in a bottom-up manner. In this case, bottom-up means learning knowledge only from the sensory information obtained from the environment, without hand-crafting it or training it with artificial disambiguated information such as teacher signals and labels, which cannot be used in the human learning process. Furthermore, the symbols (e.g., language) are shared with others and the self-organization is influenced by the shared symbols through communication with others in a top-down manner. The symbols emerge from these bottom-up and top-down loops.

We believe that it is important for robots to learn such symbols autonomously and in an unsupervised manner in this loop to coexist with humans. Accordingly, we have proposed models that allow robots to acquire concepts and language by clustering the multimodal information that is obtained through sensors mounted on the robots in an unsupervised manner (Nakamura et al., 2014; Taniguchi et al., 2017).

However, these models are not truly bottom-up because suitable feature extractors for forming concepts are not obtained in a bottom-up manner, but provided manually.

Furthermore, although concept formation is an unsupervised process, the feature extractors include supervised learning, such as a pre-trained convolutional neural network (CNN). These models are based on the MLDA and the relationships among the modality features are learned. The model proposed in (Nakamura et al., 2014) enables robots to acquire object concepts and language simultaneously by mutual learning of the MLDA and language model from multimodal information. The multimodal information is composed of visual, auditory, and tactile information obtained by robots, as well as speech uttered by a human. They are obtained by observing, shaking, and grasping objects, and the human teaches object features through speech. The features are extracted from the visual, auditory, and tactile information by using the dense scale invariant feature transform (Vedaldi and Fulkerson, 2010), sigmoid function approximation, and the Mel-frequency cepstrum coefficient, respectively, which are used for the observation of the model. The model of (Taniguchi et al., 2017), combines the MLDA, simultaneous localization and mapping, the Gaussian mixture model (GMM), and language models to acquire the spatial concept and vocabulary. The human speech to describe the location, visual information of the location, and position of the robot are used to train the model. Similar to the previous method, the features are extracted from each set of information. For example, one of the observations of the model is the feature extracted by the pre-trained CNN known as Places205-AlexNet (Zhou et al., 2014).

In this study, we propose a method that enables unsupervised multimodal clustering, unsupervised feature learning, and cross-modal inference by integrating the MLDA-based concept formation and VAE-based feature extractor. The latent variables in the normal VAE are assumed to follow a Gaussian distribution and these cannot be used for MLDA observations because the observations are assumed to follow multinomial distributions. Jang et al. realized sampling from a categorical distribution using the Gumbel-Softmax distribution (Jang et al., 2017). However, the latent variables were sampled from a categorical distribution and not from a multinomial distribution. Therefore, we extend the VAE to the multinomial VAE (MNVAE), the latent variables of which follow a multinomial distribution, by modifying the method proposed in (Jang et al., 2017). Thereafter, we construct an integrated model of the MNVAE and MLDA and demonstrate that it can classify multimodal information that is composed of images and words. The integrated model performs clustering and learns features simultaneously by communicating the parameters computed in each model. The main contributions of this study are summarized as follows:

- We extend the VAE to the MNVAE, the latent variables of which follow a multinomial distribution, thereby enabling its combination with the MLDA.
- We demonstrate that the clustering performance can be improved and suitable features can be learned by interaction between the MNVAE and MLDA.
- We demonstrate that the cross-modal inference of images from words is possible while learning with a small dataset.

Moreover, we consider that the integration of the MLDA and representation learning is also significant in terms of leveraging past research. Although pre-designed or pre-trained feature extractors have been used, various studies (Attamimi et al., 2014; Nakamura and Nagai, 2017; Hagiwara et al., 2019; Miyazawa et al., 2019) in addition to the above have revealed the effectiveness of the MLDA for multimodal concept formation. That is, the integration of the MLDA and representation learning makes it possible to develop these studies further. Another advantage of the MLDA is that it can be interpreted easily and trained with a small-scale dataset, as indicated in our previous studies. Although it is possible to construct such a model using only deep neural networks (DNNs), it becomes more difficult to interpret the model as its size increases and a larger dataset is required for its training.

In the experiments, we demonstrated that suitable features for clustering images can be learned by the interaction between the MNVAE and MLDA in an unsupervised manner using our integrated model. Furthermore, the integrated model can generate images from the features that are estimated from the words using the MLDA. We used a multimodal dataset composed of images and teaching utterances, which were obtained by the robot observing the objects and the human teacher teaching the object features using speech (Aoki et al., 2016). Because the human teacher did not necessarily utter the words corresponding to the object labels, the teaching utterances included words that were not related to the labels, and speech recognition errors consequently occurred. Moreover, we assumed that the robot did not have a feature extractor for the images in advance. Therefore, the robot was required to learn important words and features in an unsupervised manner from this ambiguous dataset.

2 RELATED WORK

Stochastic models and non-negative matrix factorization have been used in several studies to learn the relationships among multimodal information (Ridge et al., 2010; Nakamura et al., 2014; Mangin et al., 2015; Taniguchi et al., 2017). However, in these works, the feature extractor was designed or learned in advance and it did not learn suitable features using only the training dataset.

Deep generative models such as the VAE and generative adversarial network (Goodfellow et al., 2014) have attracted increasing attention as methods to obtain features. These methods can deal with complicated high-dimensional data in an end-to-end unsupervised manner. In particular, the VAE encoder models the inference process of the latent variables from observations, whereas its decoder models the generative process of the observations from the latent variables. Therefore, the encoder can be used for feature extraction.

The VAE has also been extended to models that can learn the relationships among multimodal information (Suzuki et al., 2017; Wu and Goodman, 2018). In these studies, the features of multimodal information were extracted in an end-to-end manner using multiple encoders and decoders. However, a

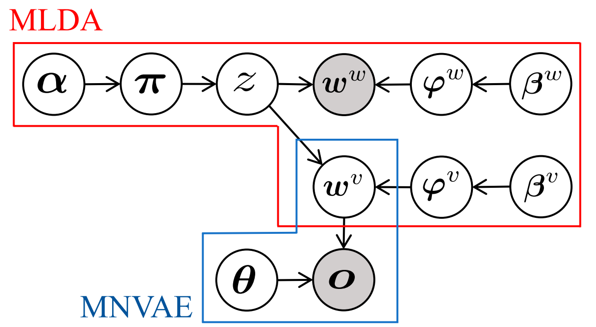

FIGURE 2 | Graphical representation of integrated model of MNVAE and MLDA.

TABLE 1 | Parameters of integrated model.

Parameter	Explanation
$\boldsymbol{O}$	Image observation
$\boldsymbol{W}^v$	Image feature
$\boldsymbol{\phi}$, $\boldsymbol{\theta}$	Parameters of MNVAE encoder and decoder
$\boldsymbol{W}^w$	Word observation
Z	Category
$\boldsymbol{\Pi}$	Parameter of multinomial distribution for category
$\boldsymbol{\varphi}^v$	Parameter of multinomial distribution for image feature
$\boldsymbol{\varphi}^w$	Parameter of multinomial distribution for word observation
$\boldsymbol{\alpha}$, $\boldsymbol{\beta}^v$, $\boldsymbol{\beta}^w$	Parameters of prior distributions

large amount of data was required to train the models, and these were evaluated using a large-scale dataset such as MNIST. It is very difficult to construct a large-scale multimodal dataset using sensors on the robot, and no such dataset exists at present. In our proposed method, the relationships among the multimodal data can be learned from a relatively small-scale multimodal dataset by integrating the MNVAE and MLDA, without designing a feature extractor.

Although the latent variables follow a Gaussian distribution in the normal VAE, models have been proposed in which other distributions were assumed (Jang et al., 2017; Srivastava and Sutton, 2017; Joo et al., 2019). These models made it possible to sample from the distribution, except for a Gaussian distribution, by formulating the sampling procedure into a differentiable form. Jang et al. proposed sampling from a Gumbel-Softmax distribution to obtain samples from a categorical distribution using the Gumbel-Max trick (Gumbel, 1954; Maddison et al., 2014), in which a Gumbel distribution was used instead of the indifferentiable argmax operation (Jang et al., 2017). In the Gumbel-Softmax distribution, the samples became one-hot vectors when the temperature parameter was lower; therefore, the generated samples could be treated as samples from a categorical distribution. Srivastava et al. used the Laplace approximation, whereby the parameters of a Dirichlet distribution were approximated by the parameters of a Gaussian distribution, making it possible to sample from a Dirichlet distribution by adding a softmax layer to the normal VAE (Srivastava and Sutton, 2017). Joo et al. applied parameter estimation of a multivariate Gamma distribution using an inverse Gamma cumulative distribution function approximation to enable sampling from a Dirichlet distribution (Joo et al., 2019). In this study, we used the Gumbel-Softmax distribution to obtain samples from a multinomial distribution.

Furthermore, clustering methods using features extracted by DNNs have been proposed (Huang et al., 2014; Xie et al., 2016; Yang et al., 2017; Abavisani and Patel, 2018; Hu et al., 2019). Huang et al. added new regularization terms to the feature extraction network to enable the learning of k-means-friendly feature extraction (Huang et al., 2014). Abavisani et al. used DNNs to compute a suitable similarity matrix as an input for spectral clustering (Ng et al., 2002) from multimodal data (Abavisani and Patel, 2018). Although these methods enabled the computation of suitable features for clustering, they did not consider the interaction between clustering and feature extraction. Xie et al. proposed a method to learn the feature extractor and perform clustering simultaneously by fine-tuning the pre-trained autoencoder using the Student's t-distribution (Maaten and Hinton, 2008) and Kullback–Leiblier (KL) divergence (Xie et al., 2016). Yang et al. and Hu et al. enabled the simultaneous learning of the feature extractor and clusters based on DNNs and k-means, respectively (Yang et al., 2017; Hu et al., 2019). Hu et al. dealt with the multimodal information of images and audio. Such deep clustering models enable the clustering of complicated data owing to the expressive power of deep neural networks. However, these are deterministic methods that focus only on the task of clustering and cannot perform probabilistic inference, generation, and prediction among multimodal information. Moreover, large datasets are required for training the models. It is difficult to construct such large datasets that are obtained by the robots.

3 INTEGRATED MODEL OF MULTINOMIAL VAE AND MULTIMODAL LATENT DIRICHLET ALLOCATION

Figure 2 presents a graphical illustration of the integrated model of the MNVAE and MLDA. In **Figure 2**, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}^v$, and $\boldsymbol{\beta}^w$ are the hyperparameters that determine the Dirichlet priors. Moreover, $\boldsymbol{\pi}$, $\boldsymbol{\varphi}^v$, and $\boldsymbol{\varphi}^w$ are the parameters of the multinomial distribution, whereas $\boldsymbol{\theta}$ is a parameter of the MNVAE. $\boldsymbol{o}$ and $\boldsymbol{w}^w$ are observations such as an image and a word included in teaching an utterance. $\boldsymbol{w}^v$ is a feature extracted from $\boldsymbol{o}$, and z is a category obtained by clustering $\boldsymbol{w}^v$ and $\boldsymbol{w}^w$. The parameters of the integrated model are listed in **Table 1**. We aim to extract the image feature $\boldsymbol{w}^v$ and to form the object category z from the object image $\boldsymbol{o}$ and word $\boldsymbol{w}^w$ in an unsupervised manner. In this section, we first explain the components of the integrated model, namely the MNVAE and MLDA, and subsequently describe the parameter estimation of the integrated model. Finally, we clarify the data generation by the MNVAE from a feature predicted by the MLDA.

3.1 Multinomial VAE

In the MNVAE, an observation $\boldsymbol{o}$ is compressed into an arbitrary-dimensional latent variable $\boldsymbol{w}^v$ through a neural network known as an encoder. Thereafter, $\boldsymbol{w}^v$ is reconstructed into the original dimensional value $\hat{\boldsymbol{o}}$ through a neural network known as a decoder. In this case, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ denote the parameters of the encoder and decoder, respectively. The parameters are learned such that $\hat{\boldsymbol{o}}$ becomes the same as $\boldsymbol{o}$. However, in the normal VAE, it is assumed that a prior follows a Gaussian distribution, and therefore it cannot share a latent variable with the MLDA, in which it is assumed that the observations follow multinomial distributions. To address this problem, the MNVAE computes λ_k, which are the parameters of the multinomial distribution $q_\phi(\boldsymbol{w}^v|\boldsymbol{o})$, where $k = 1, 2, \dots, K$, and K is the number of dimensions. Subsequently, sampling from the multinomial distribution is performed according to the following operation:

$$u_{n,k} \sim \mathrm{Uniform}(0, 1), \tag{1}$$

$$g_{n,k} = -\log\left(-\log\left(u_{n,k}\right)\right), \tag{2}$$

$$w^v_{n,k} = \frac{\exp\left(\left(\log\left(\lambda_k\right) + g_{n,k}\right)/\tau\right)}{\sum_{k=1}^{K}\exp\left(\left(\log\left(\lambda_k\right) + g_{n,k}\right)/\tau\right)}. \tag{3}$$

$$\boldsymbol{w}^v = \sum_{n=1}^{N}\boldsymbol{w}^v_n, \tag{4}$$

where N is the number of samplings and τ is the temperature. By setting τ to a small value, $\boldsymbol{w}^v_n$ becomes a one-hot vector.

3.2 Multimodal Latent Dirichlet Allocation

The MLDA is a model that extends the LDA (Blei et al., 2003) and can classify multimodal information. In the MLDA, it is assumed that the observations $\boldsymbol{w}^v = \{w^v_1, \dots, w^v_{N^v}\}$ and $\boldsymbol{w}^w = \{w^w_1, \dots, w^w_{N^w}\}$ are generated as follows, in which N^v and N^w are the total numbers of features for each observation.

- The category proportion $\boldsymbol{\pi}$ is determined by the Dirichlet prior $P\ (\boldsymbol{\pi}|\boldsymbol{\alpha})$:

$$\boldsymbol{\pi} \sim P(\boldsymbol{\pi}|\boldsymbol{\alpha}). \tag{5}$$

- The following procedure is repeated N^m times for $m \in \{v, w\}$ to generate the observations.

    - A category z is sampled from the multinomial distribution $P\ (z|\boldsymbol{\pi})$:

$$z \sim P(z|\boldsymbol{\pi}). \tag{6}$$

    - The observation $w^m_{n^m}$ corresponding to the category z is generated from the multinomial distribution $P(\boldsymbol{w}^m|\boldsymbol{\varphi}^m_z)$:

$$w^m_{n^m} \sim P\left(\boldsymbol{w}^m|\boldsymbol{\varphi}^m_z\right). \tag{7}$$

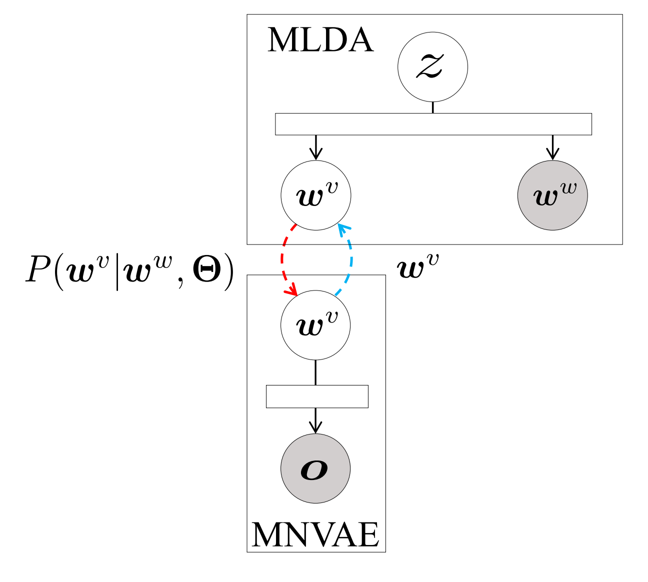

FIGURE 3 | Exchange of parameters between MNVAE and MLDA.

The observations $\boldsymbol{w}^v$ and $\boldsymbol{w}^w$ are classified by estimating the category z and parameters $\boldsymbol{\pi}$, $\boldsymbol{\varphi}^v$, and $\boldsymbol{\varphi}^w$ using Gibbs sampling in an unsupervised manner. Sampling is conducted such that the category z^{mij} is assigned to the ith feature of the modality $m \in \{v, w\}$ information $\boldsymbol{w}^m$ of the jth object from the conditional probability, where $\boldsymbol{\pi}$ and $\boldsymbol{\varphi}^m$ are marginalized out as follows:

$$P\left(z^{mij} = c|\boldsymbol{z}^{-mij}, \boldsymbol{w}^m, \boldsymbol{\alpha}, \boldsymbol{\beta}^m\right) \propto \left(N^{-mij}_{cj} + \alpha_c\right)\frac{N^{-mij}_{mw^m_i c} + \beta^m_i}{N^{-mij}_{mc} + D^m\beta^m_i}, \tag{8}$$

in which D^m is the dimension number of the observation of modality m. Moreover, a negative superscript indicates that the information is excluded. For example, z^{-mij} represents the remainder of the set of categories assigned to all objects, excluding z^{mij}. $N_{mw^m_i cj}$ is the number of times that category c is assigned to w^m_i, which is the ith feature of modality m of the jth object. Furthermore, $N_{mw^m_i c}, N_{cj}, N_{mc}$ are computed as follows:

$$N_{mw^m_i c} = \sum_j N_{mw^m_i cj}, \tag{9}$$

$$N_{cj} = \sum_{m, w^m_i} N_{mw^m_i cj}, \tag{10}$$

$$N_{mc} = \sum_{w^m_{i,j}} N_{mw^m_i cj}. \tag{11}$$

That is, $N_{mw^m_i c}$ is the number of times that category c is assigned to the ith feature w^m_i of modality m of all objects, N_{cj} is the number of times that category c is assigned to all modalities of the jth object, and N_{mc} is the number of times that category c is assigned to the observation $\boldsymbol{w}^m$ of the modality m of all objects. By means of repeated sampling according to **Eq. 8**, N^* converges to $\bar{N}_*$ and the parameters are computed as follows:

$$\hat{\varphi}_{mw^m_i c} = \frac{\bar{N}_{mw^m_i c} + \beta^m_i}{\bar{N}_{mc} + D^m\beta^m_i}, \tag{12}$$

$$\hat{\pi}_{cj} = \frac{\bar{N}_{cj} + \alpha_c}{N_j + C\alpha_c}, \tag{13}$$

where N_j is the total number of features of all modalities of the jth object and C is the number of categories. Finally, category z_j of the jth object can be determined as follows:

$$z_j = \text{argmax}_c \hat{\pi}_{cj}. \tag{14}$$

3.3 Parameter Estimation of Integrated Model

In this section, we discuss the parameter estimation of the integrated model. It is desirable for the model parameters illustrated in **Figure 2** to be optimized by directly maximizing the likelihood. However, it is difficult to estimate the latent variable $\boldsymbol{w}^v$ because it is shared between the MNVAE and MLDA. Therefore, in this study, the MNVAE and MLDA are learned independently, and all parameters are optimized approximately by exchanging the parameters computed in each model.

Figure 3 presents the exchange of parameters between the MNVAE and MLDA. The MNVAE compresses the observation $\mathbf{o}$ into $\boldsymbol{w}^v$ through the encoder and sends it to the MLDA. The MLDA considers $\mathbf{w}^v$ as an observation, and classifies $\boldsymbol{w}^v$ and observation $\boldsymbol{w}^w$ into category z. Thereafter, it sends the parameters of the distribution p $(\boldsymbol{w}^v|\boldsymbol{w}^w, \boldsymbol{\Theta})$ to the MNVAE, where $\boldsymbol{\Theta}$ is a set of parameters learned in the MLDA. Subsequently, the parameters of the MNVAE are optimized by the variational lower bound using the received parameters, as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}) = &-\gamma D_{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{w}^v|\boldsymbol{o}) \| P(\boldsymbol{w}^v|\boldsymbol{w}^w, \boldsymbol{\Theta})\right) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{w}^v|\mathbf{o})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{o}|\boldsymbol{w}^v)\right],\end{aligned} \tag{15}$$

where D_{KL} represents the KL divergence and γ is the weight thereon. In the integrated model, the MNVAE is initially optimized by using a uniform distribution because the order of the parameter update is MNVAE → MLDA → MNVAE → . . ., and p $(\boldsymbol{w}^v|\boldsymbol{w}^w, \boldsymbol{\Theta})$ has not yet been computed at this time. Thus, a latent space that is suitable for clustering is learned by regularizing $q_{\boldsymbol{\phi}}(\boldsymbol{w}^v|\boldsymbol{o})$, using p $(\boldsymbol{w}^v|\boldsymbol{w}^w, \boldsymbol{\Theta})$ as a prior.

We use the Symbol Emergence in Robotics tool KIT (SERKET) (Nakamura et al., 2018; Taniguchi et al., 2020) to implement the integrated model. SERKET is a framework for constructing a large-scale model composed of small models. SERKET supports independent learning in each module, the exchange of computed parameters between modules, and the optimization of the parameters of an entire model by means of the interaction among modules, as described above.

3.4 Cross-Modal Inference Using Integrated Model

It is possible to predict an observation from another observation using the learned parameters in the integrated model. When a new observation $\boldsymbol{w}^{w'}$ is obtained, the probability p $(w^v|\boldsymbol{w}^{w'}, \boldsymbol{\Theta})$ that the unobserved image feature w^v will be generated from the words $\boldsymbol{w}^{w'}$ can be computed. The image feature $\widehat{\boldsymbol{w}^v}$ is constructed by performing the following sampling from this distribution N times:

$$w^v \sim P(w^v|\boldsymbol{w}^{w\prime}, \boldsymbol{\Theta}), \tag{16}$$

$$\widehat{\boldsymbol{w}^v}[w^v] += 1. \tag{17}$$

The constructed $\widehat{\boldsymbol{w}^v}$ is input into the MNVAE decoder $p_{\boldsymbol{\theta}}(\boldsymbol{o}|\widehat{\boldsymbol{w}^v})$, and an image $\boldsymbol{o}'$ that is predicted from $\boldsymbol{w}^{w'}$ can subsequently be generated.

$$\boldsymbol{o}' \sim p_{\boldsymbol{\theta}}\left(\boldsymbol{o}|\widehat{\boldsymbol{w}^v}\right). \tag{18}$$

4 EXPERIMENTS

4.1 Dataset

In the experiments, we used the dataset that was used in (Aoki et al., 2016). This dataset was composed of images, tactile sensor values, and sound data obtained by the robot from objects and utterances given by a human who teaches features of the objects. The number of objects used was 499, which included a total of 81 categories, such as plastic bottles, candy boxes, stuffed animals, and rattles. We used only the image and utterance data from the dataset in the experiments. The images were obtained by extracting the object region using object detection from the scene acquired by the RGB camera. The images were resized to 144 × 120 because the image size differed according to the object size. We used these images as the observation $\boldsymbol{o}$. The teaching utterances were converted into strings using a phoneme recognizer and then divided into words in an unsupervised manner using the pseudo-online NPYLM (Araki et al., 2013). These words were converted into bag of words (BoWs) representations, and we used these BoWs as the observation $\boldsymbol{w}^w$.

4.2 Clustering and Representation Learning

We classified the multimodal dataset composed of images $\boldsymbol{o}$ and words $\boldsymbol{w}^w$ using the integrated model. We iterated the training of the integrated model by varying the initial value 10 times, and used the average result to evaluate the proposed method. **Figure 4** presents the network architecture of the MNVAE. We set $K = 32$, $N = 100$, and $\gamma = 1$ in the first learning and $\gamma = 10^4$ thereafter. Moreover, learning was performed with a batch size of 100 and 3,000 epochs. In this case, the initial value of τ was set to 1, and it was decayed at a rate of exp $(-10^{-3}t)$ for each epoch t. The MLDA was optimized by Gibbs sampling with 100 sampling times.

As a comparison method, the features extracted by the joint multimodal VAE (JMVAE) (Suzuki et al., 2017) were classified by the GMM. In the JMVAE, the encoder and decoder for the images have the same structure as those of the MNVAE; that is, the encoder has three convolutional layers and fully connected (FC) layers, whereas the decoder has an FC layer and seven deconvolutional layers. The number of dimensions of the latent variables was set to 32, as in the MNVAE. Refer to the **Appendix 1** for further details on the structure.

Figure 5 presents the latent variables compressed into two dimensions by t-SNE (Maaten and Hinton, 2008) for visualization. Each point represents one object and the color reflects its correct category. **Table 2** displays the classification

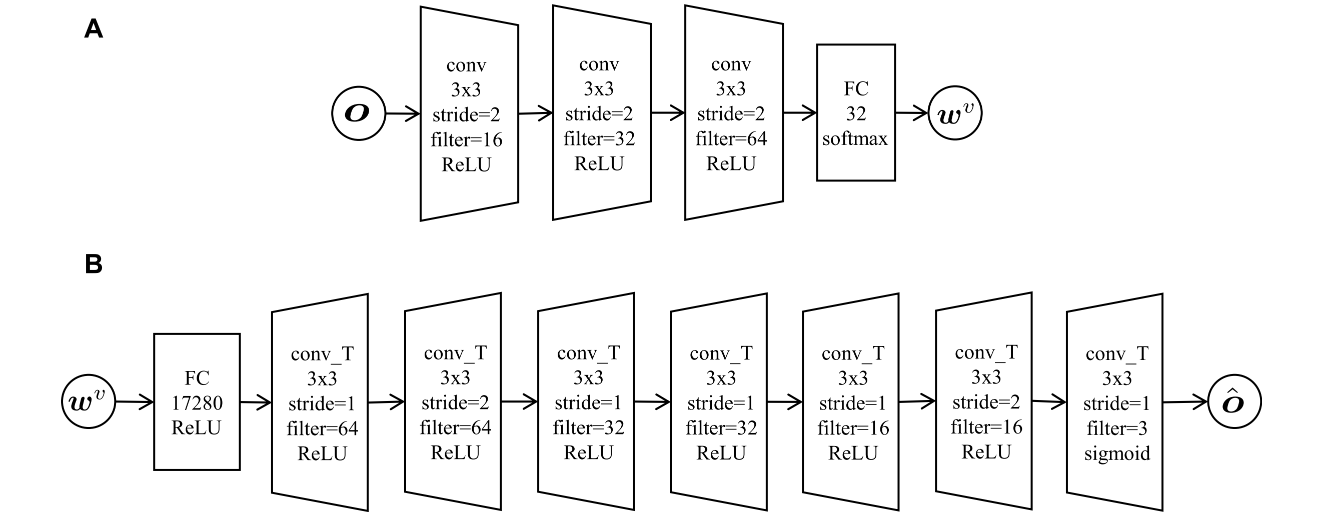
A
o
conv 3x3 stride=2 filter=16 ReLU
conv 3x3 stride=2 filter=32 ReLU
conv 3x3 stride=2 filter=64 ReLU
FC 32 softmax
w^v
B
w^v
FC 17280 ReLU
conv_T 3x3 stride=1 filter=64 ReLU
conv_T 3x3 stride=2 filter=64 ReLU
conv_T 3x3 stride=1 filter=32 ReLU
conv_T 3x3 stride=1 filter=32 ReLU
conv_T 3x3 stride=1 filter=16 ReLU
conv_T 3x3 stride=2 filter=16 ReLU
conv_T 3x3 stride=1 filter=3 sigmoid
$\hat{o}$

FIGURE 4 | Network architecture of MNVAE: **(A)** encoder consisting of three convolutional layers (conv) and a fully connected layer (FC) and **(B)** decoder consisting of an FC layer and seven deconvolutional layers (conv_T).

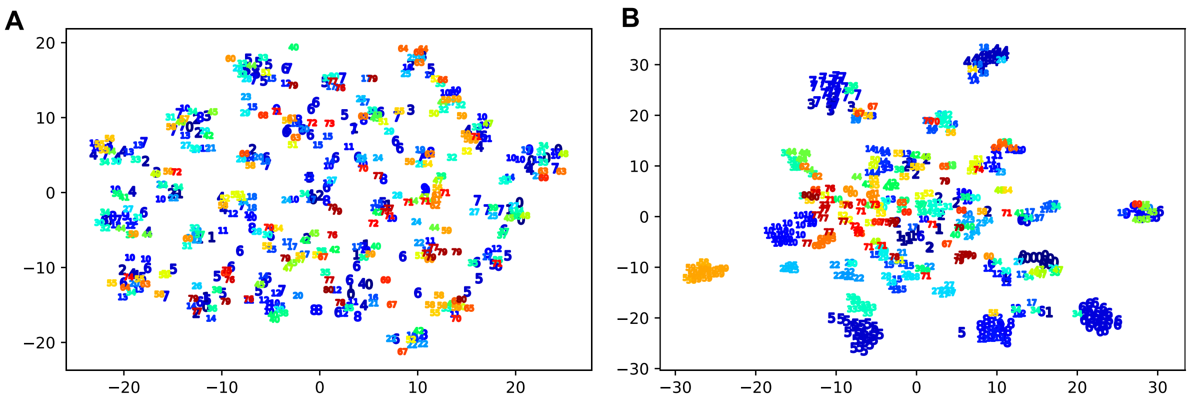

FIGURE 5 | Latent space after compression by t-SNE: **(A)** result of only MNVAE and **(B)** result of MNVAE + MLDA. The numbers at the data points represent the correct category index.

TABLE 2 | Classification accuracies and ARIs. "MLDA" used a pre-trained CNN for the feature extraction, and is therefore considered as the performance upper bound.

	Accuracy	ARI
MLDA	0.688 ± 0.014	0.613 ± 0.026
JMVAE + GMM	0.465 ± 0.013	0.271 ± 0.016
MNVAE–MLDA	0.491 ± 0.024	0.374 ± 0.031
MNVAE + MLDA	**0.660** ± **0.018**	**0.601** ± **0.032**

accuracies and adjusted rand indices (ARIs) (Hubert and Arabie, 1985). "MLDA" is the result of the MLDA classifying the words and image features extracted using the pre-trained CNN model[1] (Krizhevsky et al., 2012; Jia et al., 2014), as in the previous study (Aoki et al., 2016); "JMVAE + GMM" is the result of the GMM classifying the features extracted by the JMVAE; "MNVAE–MLDA" is the result of the classification by the integrated model without interaction; and "MNVAE + MLDA" is the result of the classification by the integrated model with 10 interactions. Moreover, the transition of the classification accuracy of the MNVAE + MLDA by exchanging the parameters is depicted in **Figure 6**. The horizontal axis represents the number of exchange iterations and the vertical axis indicates the accuracy. **Figure 5A** indicates that the data points of the same correct category were not close to one another when using only the MNVAE. However, they were located close together with the interaction between the MNVAE and MLDA, as illustrated in **Figure 5B**. In particular, the data points of categories 5, 6, and 59 were relatively well separated in the latent space. This was because the dataset exhibited a bias

[1] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

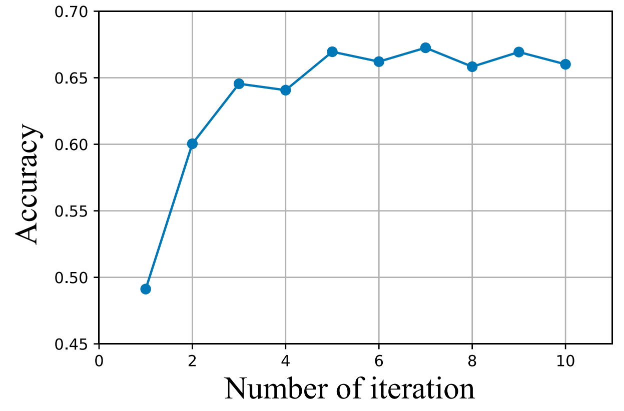

FIGURE 6 | Transition of classification accuracy of MNVAE + MLDA.

in the number of objects included in each category, and the numbers of objects in categories 5, 6, and 59 were relatively large compared to those of the other categories. Furthermore, **Figure 6** indicates that the classification accuracy was improved by iterating the exchange of parameters between the MNVAE and MLDA. Thus, the object category and word information affected the MNVAE through the interaction, and feature extraction that was suitable for clustering was obtained, thereby improving the classification accuracy. **Table 2** indicates that the classification accuracy of the unsupervised MNVAE + MLDA was similar to that of the MLDA with the supervised feature extraction and better than that of the JMVAE + GMM, which is likewise a method for unsupervised feature extraction and clustering. This suggests the effectiveness of the proposed learning method, in which the MNVAE and MLDA parameters are exchanged.

4.3 Cross-Modal Inference From Words to Images

This experiment generated images from words, and the generated images were evaluated to show that the proposed method can perform cross-modal inference. The moderate quality of the generated images may be attributed to the fact that the training dataset was not sufficiently large. However, it is considered that MNVAE + MLDA can

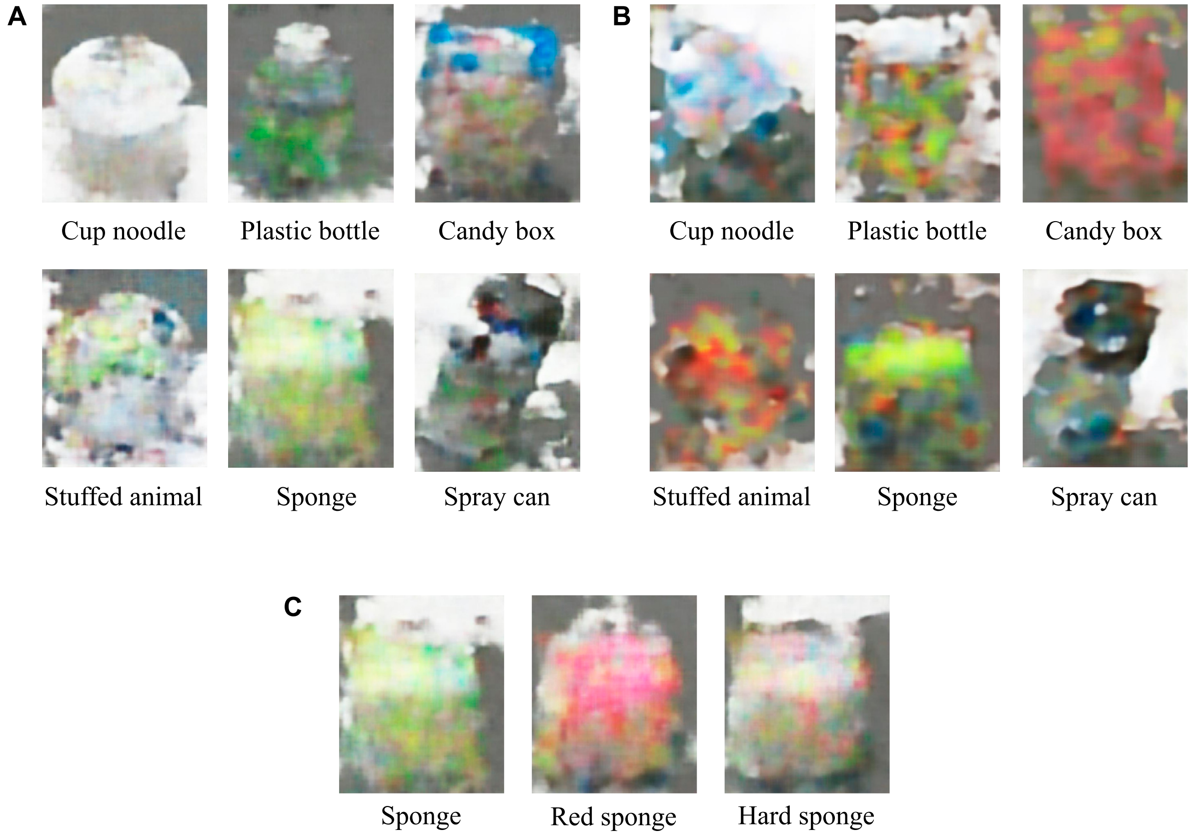

FIGURE 7 | Examples of images generated from words: **(A)** examples of MNVAE + MLDA, **(B)** examples of JMVAE, and **(C)** examples of images with changes owing to addition of adjectives.

TABLE 3 | Results of subjective evaluation of generated images. This table indicates the number of subjects who selected each method as generating more appropriate images representing the word.

Word	MNVAE + MLDA	JMVAE
Cup noodle	**18**	0
Plastic bottle	**17**	1
Candy box	**18**	0
Stuffed animal	**14**	4
Sponge	8	**10**
Spray can	**17**	1

generate more word-representative images than JMVAE because a latent space that is suitable for clustering is learned by regularizing $q_\phi(w^v|o)$ using $p(w^v|w^w, \Theta)$ in **Eq. 15**. Therefore, we conducted a subjective experiment to identify the method that can generate more word-representative images.

Images were generated from words using the MNVAE + MLDA learned in the clustering experiment. The words included "cup noodle", "plastic bottle", and "candy box". **Figure 7A,B** present examples of the images generated from the words using the trained MNVAE + MLDA and JMVAE. To evaluate the generated images, we showed 18 subjects a word and two images generated from the word by the MNVAE + MLDA and JMVAE, and asked them to choose which image was more appropriate to represent the word. This procedure was iterated for six words. The number of subjects who selected each method is displayed in **Table 3**. As illustrated in **Figure 7A**, the rough shapes of the objects were produced, although they were blurry. A white cylinder object was produced for "cup noodle" and a green bottle with a white cap was produced for "plastic bottle". The dataset images included numerous white cup noodles, green plastic bottles, and biscuits and other sweets contained in boxes, as well as yellow and blue sponges. Therefore, the images capturing these features were generated. In contrast, many of the images in **Figure 7B** were very collapsed and it was difficult to recognize them as the objects. This is because the dataset used in this experiment has only 499 data points, which is relatively small for training the JMVAE. Moreover, **Table 3** demonstrates that the MNVAE + MLDA-generated images were selected as the more word-representative images for most objects in the subjective experiment. Although the image generated from the sponge was comparable, this is because the sponge is a simple shaped object that the JMVAE could generate. According to these results, the MNVAE + MLDA was better able to perform cross-modal inference in the environment of this experiment.

Furthermore, it was confirmed that the generated images were changed by the addition of adjectives. Examples of this change are presented in **Figure 7C**. An image of a red sponge with a slight yellow tint was generated by adding "red" to "sponge" and "hard sponge" generated an image of a grayish sponge. It is believed that this was because "hard" was often taught for gray cans. Thus, the MNVAE + MLDA could learn suitable features from the small dataset, making it possible not only to compute the probability of generating the feature values, but also to generate the actual images.

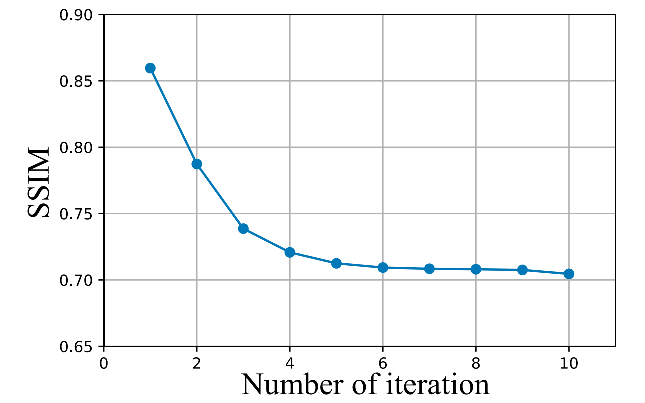

FIGURE 8 | Transition of SSIM.

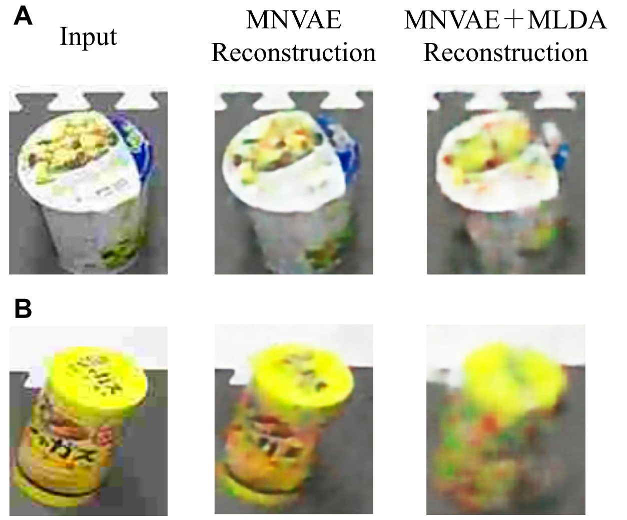

FIGURE 9 | Examples of reconstructed images: **(A)** cup noodle and **(B)** candy box. Left: input image, center: reconstructed image using only MNVAE, and right: reconstructed image using MNVAE + MLDA.

5 DISCUSSION

5.1 Image Reconstruction

We computed the structural similaritie (SSIM) (Wang et al., 2004) as a quantitative evaluation of the reconstructed images. The transition of the SSIM of the MNVAE + MLDA by exchanging the parameters is illustrated in **Figure 8**. The horizontal axis represents the number of exchange iterations and the vertical axis indicates the SSIMs. Moreover, **Figure 9** presents examples of images reconstructed by the MNVAE. It can be observed from **Figure 8** that the SSIM of the MNVAE + MLDA decreased with the interactions. This is because the latent variables represent not only the information for reconstructing the images, but also the features of the categories owing to the parameters received from the MLDA. Therefore, the SSIM

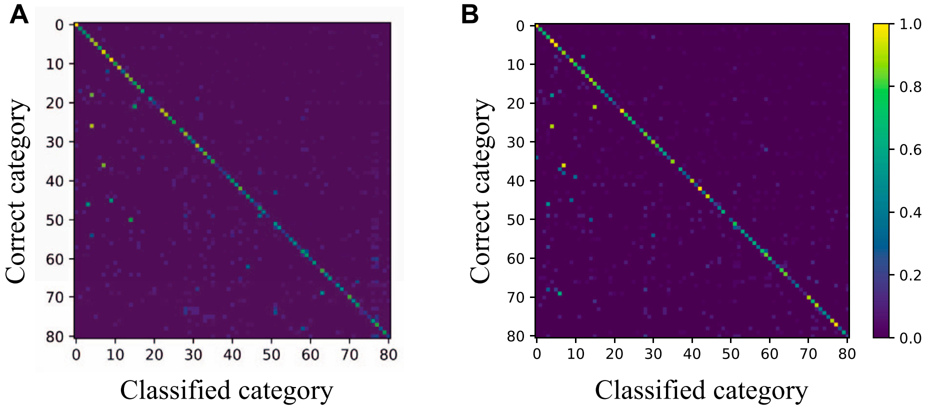

FIGURE 10 | Classification results (normalized confusion matrices): **(A)** MNVAE + MLDA and **(B)** MLDA.

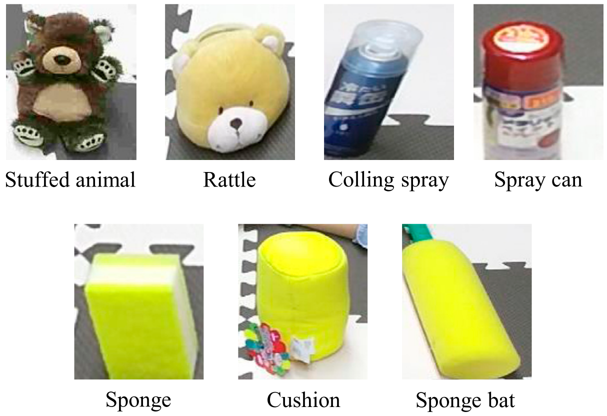

FIGURE 11 | Examples of objects in categories that were misclassified many times.

decreased and the reconstruction error increased. Although the SSIM was decreased, it was possible to distinguish objects from the reconstructed images, as indicated in **Figure 9**, suggesting that feature extraction that was suitable for clustering could be performed while capturing the image features for their reconstruction.

5.2 Classification Results

Figure 10 presents the classification results of the MNVAE + MLDA and MLDA. **Figures 10A,B** are normalized confusion matrices. The vertical axis is the index of the correct category and the horizontal axis is the index of the classified category. **Figure 10** indicates that stuffed animals (category 7) and rattles (category 36); cooling sprays (category 21) and spray cans (category 15); and sponges (category 4), cushions (Category 18), and sponge bats (category 26) were frequently misclassified in the MNVAE + MLDA and MLDA. Examples of the objects included in these categories are depicted in **Figure 11**. As illustrated in **Figure 11**, the features of these objects were quite similar. Furthermore, the features of the rattles were often taught by using utterances such as “this is a stuffed animal that makes sound”, whereas the features of the sponge, cushion, and sponge bat were taught by using utterances such as “this is soft”. Therefore, these categories had similar image features and similar words such as “stuffed animal”, “spray”, and “soft” were taught frequently, thereby increasing the misclassification. In (Aoki et al., 2016), tactile and sound information was used in addition to images and words, and the correct categories were determined on this basis. Hence, it is possible to decrease the misclassification by adding information that was not used in the experiment. For example, as the sounds of stuffed animals and rattles differ when shaking them, it is possible to classify them correctly using sound information. Comparing **Figures 10A,B**, it can be observed that the MNVAE + MLDA yielded slightly more misclassifications than the MLDA. The MLDA used the features extracted by the pre-trained CNN, which was effective. In contrast, in the MNVAE + MLDA, the feature extractor was learned at the same time by the interaction between the MNVAE and MLDA. As a result, we consider that several objects were misclassified because of biased words, and the features were also learned to represent those misclassifications.

5.3 Definition of Concepts

This paper uses a straightforward definition of concepts such as the multimodal categories that enable multimodal clustering and cross-modal inference, to present our outcome. Moreover, we deal with categories such as discretized and symbolic to compare them with the manually determined ground truth. This is because we focus on evaluating our proposed method from the perspective of engineering. By contrast, in cognitive science (CS), the characteristics of concepts have been studied (Olier et al., 2017), and we consider that our MLDA-based approach has the potential to replicate some of them.

One characteristic pointed out in CS is that a concept is not symbolic, but a distributed representation. In our approach, although we convert categories into the symbolic representation z by $\arg\max p\ (z|\mathbf{w}^v,\ \mathbf{w}^w)$, the categories can be represented by the distribution p $(z|\boldsymbol{w}^v, \boldsymbol{w}^w)$, which has a continuous parameter. Therefore, concepts are represented in a continuous space in MLDA, and thus it does not represent only symbolic and static categories.

It is also pointed out that embodiment is important and that concepts depend on action and context. An action can be easily introduced into MLDA by considering the action $\boldsymbol{a}$ to be a single modality as p $(z|\boldsymbol{w}^v, \boldsymbol{w}^w, \boldsymbol{a})$ (Fadlil et al., 2013). By connecting perception and action, it is possible to recall the unobserved perception from the action through a simulation p $(\boldsymbol{w}^v|\boldsymbol{a})$. Furthermore, we extend our MLDA to a time-series model, and action- and context-dependent concepts can be learned (Miyazawa et al., 2019) to a certain extent.

Although the MLDA-based approach cannot currently describe all cognitive phenomena, we consider MLDA to be a promising model. In the future, by integrating our previous studies (Fadlil et al., 2013; Miyazawa et al., 2019) and the proposals in this paper, we would like to develop a cognitive model that allows robots to learn by interacting with the environment.

6 CONCLUSIONS

We have proposed the MNVAE, which is an extension of the VAE, the latent variables of which follow a multinomial distribution. An integrated model of the MNVAE and MLDA was constructed, and the multimodal information was classified. The experiments demonstrated that the interaction between the MNVAE and MLDA could learn the features that are suitable for clustering. Moreover, the images representing the concepts acquired by the MLDA can be generated from the word information. Thus, we have revealed that combining the MNVAE and MLDA enables clustering and the learning of features from the information obtained from the environment in a truly bottom-up and unsupervised manner, without the need for a manually designed feature extractor, as used in previous studies.

We evaluated the integrated model using only image and word information. In the future, we will incorporate not only images and words, but also other modal information, such as tactile and auditory information. We will construct and evaluate a model that learns the feature extraction of the information of each modality using the MNVAE in an unsupervised manner. Furthermore, in this study, the BoW representation of strings recognized by a phoneme recognizer was used, and language knowledge (the language model) in the speech recognizer was not updated. We believe that acquiring such language knowledge from the information obtained by its own sensors in a bottom-up manner is important (Tangiuchi et al., 2019). Therefore, we will introduce mutual learning with the language model (Nakamura et al., 2014) into the integrated model, and we will construct a model that can learn the parameters of the speech recognizer simultaneously. Moreover, the experimental results showed that it was possible to learn from a small amount of data, but we found that certain categories could not be learned correctly. We consider that this is because the number of objects is small for learning these categories and the bias of the number of objects in each category is inevitable in the real environment. To achieve learning in such a biased environment, we believe that interactive and online learning is very important. The learner expresses its inner state by using its current knowledge and body, and the teacher changes his/her actions (e.g., the object presented next in the task used in this paper) by estimating the extent to which the learner could acquire knowledge from its feedback. Therefore, we plan to extend the proposed method to online learning to realize interactive learning.

DATA AVAILABILITY STATEMENT

The datasets used for this study can be found at our GitHub repository https://github.com/naka-lab/MLDA-MNVAE.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RK, TNak, TT, and TNag conceived, designed, and developed the research. RK and TNak performed the experiment and analyzed the data. RK wrote the manuscript with support from TNak, TT, and TNag. TNag supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by JST, Moonshot R and D Grant Number JPMJMS 2011.

REFERENCES

Abavisani, M., and Patel, V. M. (2018). Deep Multimodal Subspace Clustering Networks. *IEEE J. Sel. Top. Signal. Process.* 12, 1601–1614. doi:10.1109/jstsp.2018.2875385

Aoki, T., Nishihara, J., Nakamura, T., and Nagai, T. (2016). "Online Joint Learning of Object Concepts and Language Model Using Multimodal Hierarchical Dirichlet Process," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Daejeon, Korea, 2636–2642. doi:10.1109/iros.2016.7759410

Araki, T., Nakamura, T., and Nagai, T. (2013). "Long-Term Learning of Concept and Word by Robots: Interactive Learning Framework and Preliminary Results," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2280–2287. doi:10.1109/iros.2013.6696675

Attamimi, M., Fadlil, M., Abe, K., Nakamura, T., Funakoshi, K., and Nagai, T. (2014). "Integration of Various Concepts and Grounding of Word Meanings Using Multi-Layered Multimodal Lda for Sentence Generation," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 2194–2201. doi:10.1109/iros.2014.6942858

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Machine Learn. Res.* 3, 993–1022.

Fadlil, M., Ikeda, K.-i., Abe, K., Nakamura, T., and Nagai, T. (2013). "Integrated Concept of Objects and Human Motions Based on Multi-Layered Multimodal Lda," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2256–2263. doi:10.1109/iros.2013.6696672

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative Adversarial Nets," in Advances in neural information processing systems, Montreal, Canada, 2672–2680.

Gumbel, E. J. (1954). Statistical Theory of Extreme Values and Some Practical Applications. *NBS Appl. Mathematics Ser.* 33, 1–51.

Hagiwara, Y., Kobayashi, H., Taniguchi, A., and Taniguchi, T. (2019). Symbol Emergence as an Interpersonal Multimodal Categorization. *Front. Robot. AI.* 6, 134. doi:10.3389/frobt.2019.00134

Hu, D., Nie, F., and Li, X. (2019). "Deep Multimodal Clustering for Unsupervised Audiovisual Learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 9248–9257. doi:10.1109/cvpr.2019.00947

Huang, P., Huang, Y., Wang, W., and Wang, L. (2014). "Deep Embedding Network for Clustering," in 2014 22nd International conference on pattern recognition, Stockholm, Sweden, (IEEE), 1532–1537. doi:10.1109/icpr.2014.272

Hubert, L., and Arabie, P. (1985). Comparing Partitions. *J. Classification.* 2, 193–218. doi:10.1007/bf01908075

Jang, E., Gu, S., and Poole, B. (2017). "Categorical Reparameterization With Gumbel-Softmax," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017 (Conference Track Proceedings).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: Convolutional Architecture for Fast Feature Embedding," in ACM International Conference on Multimedia, Orlando, FL, 675–678.

Joo, W., Lee, W., Park, S., and Moon, I.-C. (2019). Dirichlet Variational Autoencoder. arXiv preprint arXiv:1901.02739 , 1–17.

Kingma, D. P., and Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114 , 1–14.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification With Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105. doi:10.1145/3065386

Maaten, L. v. d., and Hinton, G. (2008). Visualizing Data Using T-Sne. *J. Machine Learn. Res.* 9, 2579–2605.

Maddison, C. J., Tarlow, D., and Minka, T. (2014). A Sampling. *Adv. Neural Inf. Process. Syst.* 27, 3086–3094.

Mangin, O., Filliat, D., Ten Bosch, L., and Oudeyer, P.-Y. (2015). Mca-nmf: Multimodal Concept Acquisition With Non-Negative Matrix Factorization. *PloS one.* 10, e0140732. doi:10.1371/journal.pone.0140732

Miyazawa, K., Horii, T., Aoki, T., and Nagai, T. (2019). Integrated Cognitive Architecture for Robot Learning of Action and Language. *Front. Robot. AI.* 6, 131. doi:10.3389/frobt.2019.00131

Mochihashi, D., Yamada, T., and Ueda, N. (2009). "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," in Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, Suntec, Singapore, 100–108. doi:10.3115/1687878.1687894

Nagano, M., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I., and Takano, W. (2019). Hvgh: Unsupervised Segmentation for High-Dimensional Time Series Using Deep Neural Compression and Statistical Generative Model. *Front. Robot. AI.* 6, 115. doi:10.3389/frobt.2019.00115

Nakamura, T., Nagai, T., and Taniguchi, T. (2018). Serket: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model. *Front. Neurorobot.* 12, 25–16. doi:10.3389/fnbot.2018.00025

Nakamura, T., and Nagai, T. (2017). Ensemble-of-Concept Models for Unsupervised Formation of Multiple Categories. *IEEE Trans. Cogn. Developmental Syst.* 10, 1043–1057. doi:10.1109/TCDS.2017.2745502

Nakamura, T., Nagai, T., Funakoshi, K., Nagasaka, S., Taniguchi, T., and Iwahashi, N. (2014). "Mutual Learning of an Object Concept and Language Model Based on Mlda and Npylm," in IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 600–607. doi:10.1109/iros.2014.6942621

Nakamura, T., Nagai, T., and Iwahashi, N. (2007). "Multimodal Object Categorization by a Robot," in IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 2415–2420. doi:10.1109/iros.2007.4399634

Nakamura, T., Nagai, T., and Iwahashi, N. (2009). "Grounding of Word Meanings in Multimodal Concepts Using LDA," in IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 3943–3948. doi:10.1109/iros.2009.5354736

Neubig, G., Mimura, M., Mori, S., and Kawahara, T. (2012). Bayesian Learning of a Language Model From Continuous Speech. *IEICE Trans. Inf. Syst.* E95-D, 614–625. doi:10.1587/transinf.e95.d.614

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On Spectral Clustering: Analysis and an Algorithm. *Adv. Neural Inf. Process. Syst.* 14, 849–856.

Olier, J. S., Barakova, E., Regazzoni, C., and Rauterberg, M. (2017). Re-Framing the Characteristics of Concepts and Their Relation to Learning and Cognition in Artificial Agents. *Cogn. Syst. Res.* 44, 50–68. doi:10.1016/j.cogsys.2017.03.005

Piaget, J., and Duckworth, E. (1970). Genetic Epistemology. *Am. Behav. Scientist.* 13, 459–480. doi:10.1177/000276427001300320

Ridge, B., Skocaj, D., and Leonardis, A. (2010). "Self-Supervised Cross-Modal Online Learning of Basic Object Affordances for Developmental Robotic Systems," in IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 5047–5054. doi:10.1109/robot.2010.5509544

Srivastava, A., and Sutton, C. A. (2017). "Autoencoding Variational Inference for Topic Models," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2017). "Joint Multimodal Learning With Deep Generative Models," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings (OpenReview.net).

Taniguchi, A., Hagiwara, Y., Taniguchi, T., and Inamura, T. (2017). Online Spatial Concept and Lexical Acquisition With Simultaneous Localization and Mapping. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada. doi:10.1109/iros.2017.8202243

Tangiuchi, T., Mochihashi, D., Nagai, T., Uchida, S., Inoue, N., Kobayashi, I., et al. (2019). Survey on Frontiers of Language and Robotics. *Adv. Robotics.* 33, 700–730. doi:10.1080/01691864.2019.1632223

Taniguchi, T., Nakanishi, H., and Iwahashi, N. (2010). Simultaneous Estimation of Role and Response Strategy in Human-Robot Role-Reversal Imitation LearningThe 11th IFAC/IFIP/IFORS/IEA Symposium. *IFAC Proc. Volumes.* 43, 460–464. doi:10.3182/20100831-4-fr-2021.00081

Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., and Asoh, H. (2016). Symbol Emergence in Robotics: A Survey. *Adv. Robotics.* 30 (Issue 11), 706–728. doi:10.1080/01691864.2016.1164622

Taniguchi, T., Nakamura, T., Suzuki, M., Kuniyasu, R., Hayashi, K., Taniguchi, A., et al. (2020). Neuro-Serket: Development of Integrative Cognitive System Through the Composition of Deep Probabilistic Generative Models. *New Generation Comput.* 38, 1–26. doi:10.1007/s00354-019-00084-w

Taniguchi, T., Ugur, E., Hoffmann, M., Jamone, L., Nagai, T., Rosman, B., et al. (2018). Symbol Emergence in Cognitive Developmental Systems: a Survey. *IEEE Trans. Cogn. Developmental Syst.* 11, 494–516. doi:10.1109/TCDS.2018.2867772

Vedaldi, A., and Fulkerson, B. (2010). "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in ACM International Conference on Multimedia, New York, NY, 1469–1472.

Wächter, M., and Asfour, T. (2015). Hierarchical Segmentation of Manipulation Actions Based on Object Relations and Motion Characteristics. *Int. Conf. Adv. Robotics.*, 549–556. doi:10.1109/icar.2015.7251510

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 13, 600–612. doi:10.1109/tip.2003.819861

Wu, M., and Goodman, N. (2018). "Multimodal Generative Models for Scalable Weakly-Supervised Learning," in *Advances in Neural Information Processing Systems 31*. Editors S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, Canada: Curran Associates, Inc.), 5575–5585.

Xie, J., Girshick, R., and Farhadi, A. (2016). "Unsupervised Deep Embedding for Clustering Analysis," in International conference on machine learning, New York City, NY, 478–487.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. (2017). "Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering," in International conference on machine learning, Sydney, Australia (PMLR), 3861–3870.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning Deep Features for Scene Recognition Using Places Database. *Adv. Neural Inf. Process. Syst.* 27, 487–495.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX 1: NETWORK STRUCTURE OF JMVAE

In this section, we describe the network structure of the JMVAE used in the experiment as a comparison method. **Figure 12** presents the network structure of the JMVAE. The encoder and decoder for the images have the same structure as the MNVAE; that is, the encoder has three convolutional layers and FC layers, and the decoder has an FC layer and seven deconvolutional layers. The encoder and decoder for the words are composed of an FC layer (512 nodes). The structure of the joint encoder is shown in **Figure 12A** and multimodal values can be obtained through an FC layer (512 nodes), the input of which is the concatenated values of extracted features of images and words. The number of dimensions of the latent variables was set to 32, as in the MNVAE.

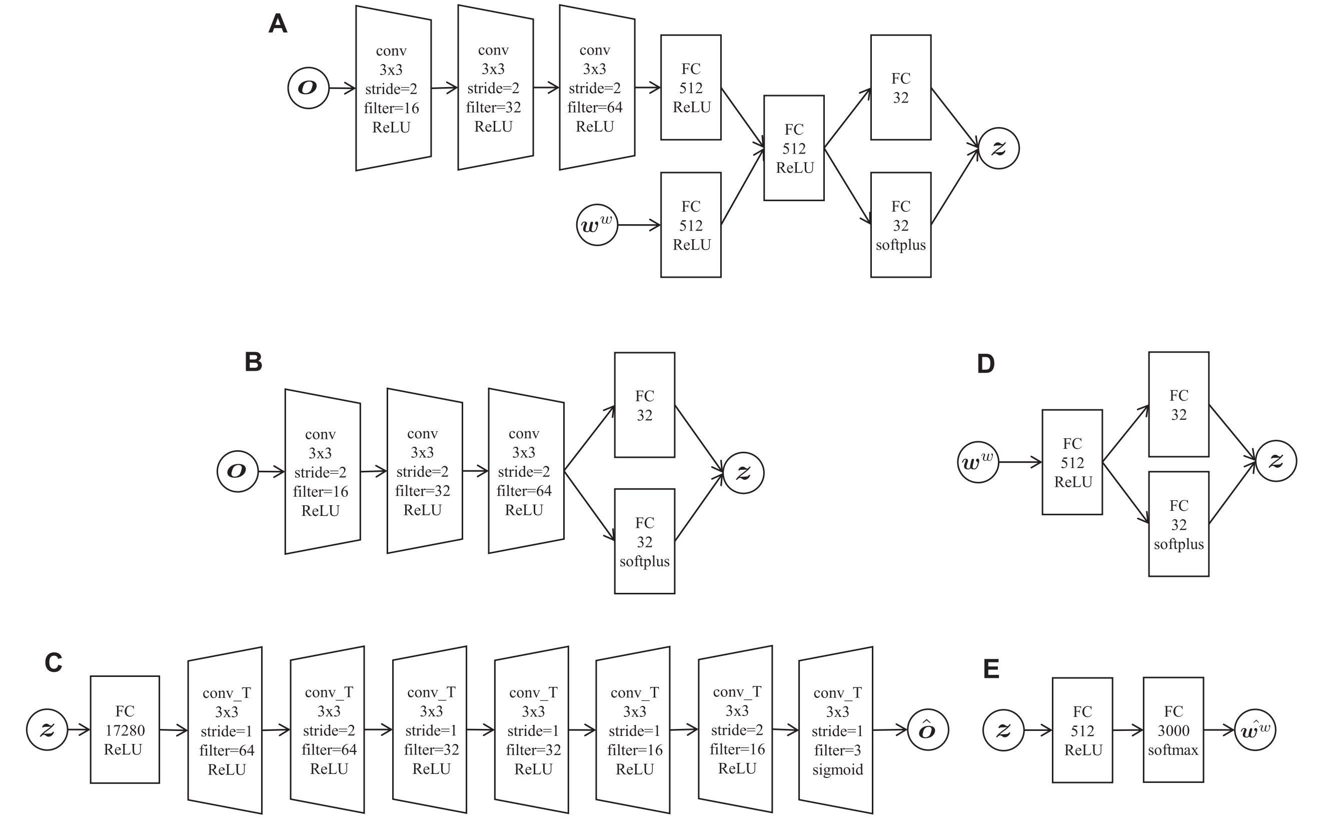
A
o
conv 3x3 stride=2 filter=16 ReLU
conv 3x3 stride=2 filter=32 ReLU
conv 3x3 stride=2 filter=64 ReLU
FC 512 ReLU
w^w
FC 512 ReLU
FC 512 ReLU
FC 32
FC 32 softplus
z
B
o
conv 3x3 stride=2 filter=16 ReLU
conv 3x3 stride=2 filter=32 ReLU
conv 3x3 stride=2 filter=64 ReLU
FC 32
FC 32 softplus
z
D
w^w
FC 512 ReLU
FC 32
FC 32 softplus
z
C
z
FC 17280 ReLU
conv_T 3x3 stride=1 filter=64 ReLU
conv_T 3x3 stride=2 filter=64 ReLU
conv_T 3x3 stride=1 filter=32 ReLU
conv_T 3x3 stride=1 filter=32 ReLU
conv_T 3x3 stride=1 filter=16 ReLU
conv_T 3x3 stride=2 filter=16 ReLU
conv_T 3x3 stride=1 filter=3 sigmoid
$\hat{o}$
E
z
FC 512 ReLU
FC 3000 softmax
$\hat{w^w}$

FIGURE 12 | Network structure of JMVAE used in experiment: **(A)** joint encoder, **(B)** image encoder, **(C)** image decoder, **(D)** word encoder, and **(E)** word decoder.