Cross-Modal Alignment Learning of Vision-Language Conceptual Systems

Taehyeong Kim¹, Hyeonseop Song¹, Byoung-Tak Zhang²
¹LG Electronics, ²Seoul National University
{taehyeong.kim,hyeonseop.song}@lge.com, btzhang@bi.snu.ac.kr

Abstract

Human infants learn the names of objects and develop their own conceptual systems without explicit supervision. In this study, we propose methods for learning aligned vision-language conceptual systems inspired by infants’ word learning mechanisms. The proposed model learns the associations of visual objects and words online and gradually constructs cross-modal relational graph networks. Additionally, we also propose an aligned cross-modal representation learning method that learns semantic representations of visual objects and words in a self-supervised manner based on the cross-modal relational graph networks. It allows entities of different modalities with conceptually the same meaning to have similar semantic representation vectors. We quantitatively and qualitatively evaluate our method, including object-to-word mapping and zero-shot learning tasks, showing that the proposed model significantly outperforms the baselines and that each conceptual system is topologically aligned.

1 Introduction

According to studies by psychologists and cognitive scientists, infants could categorize visual objects (hereafter objects) at the age of 3-4 months [28], segment words from fluent speech at the age of 7-8 months [12], and finally learn a language. How do infants learn a language and build their own conceptual systems [23]? Although it is not clearly known about these developmental processes, various studies have been conducted on the word learning mechanisms of infants [34], which could be a fundamental process for building vision-language conceptual systems.

Infants begin to learn the meaning of words by pairing heard words with visually present objects. It is not an easy process because the mapping of specific words to multiple objects in a visual scene is ambiguous [27]. For example, when their parents say the word “cup” when “spoon”, “cup”, “food”, or “plate” are placed on the table, infants may not know exactly which visual object it refers to. Nevertheless, typical infants know the meaning of some common nouns after the age of 6-9 months [1], which is a crucial stage for learning more words and acquiring language [34]. Although the mechanisms by which infants learn words and language remain unclear, the best-known mechanism is cross-situational learning (XSL) [8]. XSL is that even if it is ambiguous whether a word refers to a visual object in a scene, the mapping of words and visual objects becomes more and more accurate if sufficient information is obtained through continuous exposure to various situations.

Meanwhile, computer scientists have also studied methods for learning the mappings between objects and words in an unsupervised or weakly-supervised manner. The visual grounding (VG) seeks to find the target objects or regions in an image based on a natural language query [39]. Various methods [5, 21, 37] to solve the VG task have been studied, but they usually only map specific words or noun phrases and objects when an image and text are given as a pair, and it is not easy to generalize the relationship between objects and words. Aligned cross-modal representation learning also attempts to map between images and natural language [10, 29]. It builds embeddings that similarly represent
From the input sample, i) our model continuously learns object-object, word-word, and object-word co-occurrence statistics to construct cross-modal relational graph networks. ii) Based on the graph networks, it learns aligned semantic representations so that objects and words with conceptually the same meanings have similar representation vectors.

From these backgrounds, we propose a continuous learning method of vision-language conceptual systems, inspired by the word learning mechanisms. We assume that the learner’s object categorization and word segmentation abilities have already been developed; thus, we use image-level pseudo labels (or a pre-trained object detector) and text data. The proposed model builds cross-modal relational graph networks by continuously learning the co-occurrence statistics of objects and words from the image-text dataset based on the distributional hypothesis [7]. The distributional hypothesis states that linguistic items with similar distributions have similar meanings, which could be applied to objects in visual scenes as well as natural language. We use simple statistical rules to progressively construct object-relational and word-relational graph networks representing vision and language conceptual systems. Furthermore, we utilize the object-word co-occurrence statistics to align the two graph networks. These methods continuously reduce the ambiguity of object and word mapping, and gradually build vision-language conceptual systems (i.e. cross-modal relational graph networks). We additionally propose a novel method for learning aligned cross-modal representations based on the built relational graph networks. The proposed method trains the semantic representation vectors of objects and words in a self-supervised manner. We evaluate our model quantitatively and qualitatively, showing that the graph networks and semantic representations learned by the proposed methods are topologically aligned. An overview of the proposed methods is depicted in Figure 1.

Our main contributions can be summarized as follows.

- We propose a method for progressively building cross-modal relational graph networks by continuously learning object-object, word-word, and object-word co-occurrence statistics.
- We propose an aligned cross-modal representation learning method that learns semantic representations of objects and words through cross-entropy loss for self-identification of each entity and alignment loss for correspondence between objects and words.
- Finally, we validate our model with object-to-word mapping accuracy and zero-shot learning tasks. The experimental results show that the proposed methods significantly outperform the baselines and that each conceptual system is topologically aligned.

2 Related work

Cross-situational learning Various studies [4, 33, 36, 33] discuss XSL strategies based on distribution statistics in which words and referents occur simultaneously at multiple moments. There are...
works [8, 30] that examine how learners segment words and acquire their meanings through XSL. Fazly et al. [6] and Kádár et al. [13] propose computational models that learn word meanings by probabilistic associations between words and semantic elements. There are also studies using the XSL mechanism for robot learning. Taniguchi et al. [35] proposes a Bayesian generative model that can form multiple categories based on each sensory channel of a humanoid robot and can associate words with four sensory channels. Roesler et al. [32] introduces a Bayesian learning model for grounding synonymous object and action names using XSL. Recently, deep neural network models have been proposed to learn semantic knowledge [24] and representations [14] using images with corresponding descriptions and audio.

**Cross-modal representation learning** Castrejon et al. [3] introduces a new cross-modal scene dataset, and presents a method for training cross-modal scene representations to have a shared representation that is modal-agnostic. CLIP [29] demonstrates that a simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn image representations from large datasets of image-text pairs. ALIGN [10] consists of a simple dual-encoder architecture and learns to align visual and language representations of image and text pairs using contrastive loss. OSCAR [19] learns cross-modal representations using masked token loss and contrastive loss with image-text and object tags. The most relevant work to ours is [31], which proposes a way to align conceptual systems in an unsupervised manner using representation vectors trained with the GloVe [25] method. These methods are difficult to explicitly learn complex relationships between entities. We solve this problem by constructing the cross-modal relational graph networks through a cross-situational learning method.

### 3 Cross-modal relational graph networks

#### 3.1 Problem settings and preliminaries

In this section, we introduce a method to build cross-modal relational graph networks by continuously learning the co-occurrence statistics of objects and words. To investigate the proposed methods, we use image-level pseudo labels (i.e. categories of objects included in the image) and text descriptions for images in the Open Images Dataset V6 [18]. Formally, given an image \(I\) containing \(M\) objects, we use the set of objects \(B = \{b_m\}_{m=1}^M\), where \(b_m\) denotes the pseudo label of the object. Also, given a text description \(S\) corresponding to the image \(I\), we use the word sequence \(S = [s_n]_{n=1}^N\) of length \(N\), where \(s_n\) denotes the word tokenized by NLTK [2]. Our goal is to build cross-modal relational graph networks \(G(G^O, G^W, E^X)\) by continuously learning the relationships between objects and words from the input of \(B\) and \(S\). Here, \(G^O\) and \(G^W\) represent weighted relational graph networks for objects and words, respectively, and edge set \(E^X\) represents associations between the two graphs.

To this end, we generate multi-modal sequential inputs \([((B, Q_1))_{t=1}^T\] using local context words \(Q_t\) for each paired input \((B, S)\) to simulate the stream input and reflect the local context. To explain local context words \(Q_t\), we first define a padded word sequence \(\tilde{S} = [D; S] = [\tilde{s}_i]_{i=1}^T\) filled with dummy words \(D\) at the beginning of the sequence \(S\). The length of the dummy word sequence \(D\) is determined as \(H - 1\) for the size \(H\) of the local context window. Therefore, the length \(T\) of the padded word sequence \(\tilde{S}\) is determined as \(N + H - 1\). The local context \(Q_t = \{q_0\}_{i=1}^K\) refers to set of all words included in the local context window of \(\tilde{s}_i\) (i.e. \(H\) words after \(\tilde{s}_i\) including itself) in the padded word sequence \(\tilde{S}\). Finally, we get multi-modal sequential inputs \([((B, Q_1))_{t=1}^T = \{([b_1, b_2, ..., b_n], [q_1, q_2, ..., q_k])\}_{i=1}^T\) for each input pair \((B, S)\). In the following subsections, we first describe methods to form the undirected weighted graph networks \(G^O\) and \(G^W\) from these stream inputs, and then we describe how to build a cross-modal relational graph networks \(G(G^O, G^W, E^X)\) with edge set \(E^X\) using the proposed XSL method. Graph networks \(G\) is described as an undirected graph, and edge \((u_a, v_b)\) and edge \((v_b, u_a)\) connecting the two nodes \(u_a\) and \(v_b\) are equivalent.

#### 3.2 Object and word graph networks

In natural language, semantically similar words occur together more often than unrelated words. Based on this assumption, we incrementally build the relational graph networks \(G^W(V^W, E^W)\) of words, where \(V^W\) is the set of word nodes and \(E^W\) is the set of edges between the nodes. Each word node \(v_{w_i}\) has a variable \(d_{w_i}\), representing the cumulative number of inputs for the word \(w_i\). For each
local context $Q_t = \{q_1, q_2, \ldots, q_k\}$ in the input stream, if a new word $w_{\text{new}}$ not included in $V^W$ is observed, a new node $v_{w_{\text{new}}}$ is created, and $d_{w_{\text{new}}}$ is initialized to 0. Then $d_{q_t}$ is increased by 1 for every node $v^W_{q_t}$ corresponding to all words $q_t$ in $Q_t$.

Each edge $(v_{w_i}, v_{w_j})$ has the number of co-occurrence $c_{w_i, w_j}$ of the two words $w_i$ and $w_j$ as a variable. For each local context $Q_t = \{q_1, q_2, \ldots, q_k\}$ in the input stream, when a new node $v_{w_{\text{new}}}$ is created, edges $(v_{w_{\text{new}}}, v_{w_i})$ connecting that node $v_{w_{\text{new}}}$ and all other nodes $v_{w_i}$ are created, and $c_{w_{\text{new}}, w_i}$ is initialized to 0. Then, for every word pairs $(q_i, q_j)$ in $Q_t$, the co-occurrence count $c_{w_i, w_j}$ is increased by 1. By these processes, the connection strength (i.e. weight of edge) $e_{w_i, w_j}$ between two nodes $v_{w_i}$ and $v_{w_j}$ is defined as Equation 1. The edge weight $e_{w_i, w_j}$ has a value of 1 for the two words $w_i$ and $w_j$ that always appear together, and the lower the frequency of their co-occurrence, the closer to 0. In this way, the distribution and relationship of words are learned, and the undirected weighted graph networks $G^W$ is continuously formed.

Similarly in vision, semantically related objects appear together more often in a scene. Based on this assumption, we also build the relational graph networks $G^O(V^O, E^O)$ of objects in the same way as the relational graph networks formation of words. Here, $V^O$ is the set of object nodes and $E^O$ is the set of edges between the nodes. To form the relational graph networks for objects, we use the set of the objects $B$ in the input stream instead of the local context $Q_t$ of word sequence $S$. For each object set $B = \{b_1, b_2, \ldots, b_m\}$, if a new object $o_{\text{new}}$ not included in $V^O$ is observed, a new node $v_{o_{\text{new}}}$ is created, and $d_{o_{\text{new}}}$ and $c_{o_{\text{new}}, b_j}$ are initialized to 0. Then $d_{b_j}$ is increased by 1 for every node $v_{o_{\text{new}}}$ corresponding to every object in $B$, and the number of co-occurrences $v_{o_j}$ for edge $(v_{o_{\text{new}}}, v_{b_j})$ for every pair of objects $(b_j, b_j)$ in $B$ is also increased by 1. Finally, using the cumulative input counts $d_{o_j}$ and $d_{b_j}$ of the two object nodes $v_{o_j}$ and $v_{b_j}$ and their co-occurrence count $c_{o_j, b_j}$, the edge weight $e_{o_j, b_j}$ of the object nodes is defined as shown in Equation 1.

Following the formation of relational graph networks $G^W$ and $G^O$, the next subsection describes how to form relationships between objects and words with edge set $E^X$.

### 3.3 Cross-situational learning

XSL paradigm allows learners to gradually acquire the meaning of words through multiple exposures to the environment. We implement XSL for incremental learning of the relationships between objects and words using a cross-mapping approach. For the set $E^X$ of edges connecting all object nodes $v_{o_j}$ and word nodes $v_{w_i}$, each edge $(v_{o_i}, v_{w_j})$ has the cumulative number of co-occurrence $c_{o_i, w_j}$ as a variable for the object $o_i$ and word $w_j$. For each multi-modal input $(B, Q_t) = \{(b_1, b_2, \ldots, b_m); \{q_1, q_2, \ldots, q_k\}\}$, when a new object node $v_{o_{\text{new}}}$ is generated, edges $(v_{o_{\text{new}}}, v_{w_i})$ connecting $v_{o_{\text{new}}}$ and all word nodes $v_{w_i}$ are created and the co-occurrence counts $c_{o_{\text{new}}, w_i}$ are initialized 0. Likewise, when a new word node $v_{w_{\text{new}}}$ is generated, edges $(v_{o_j}, v_{w_{\text{new}}})$ connecting $v_{w_{\text{new}}}$ and all object nodes $v_{o_j}$ are created and $c_{o_j, w_{\text{new}}}$ are initialized 0. Then, for all object-word pairs $(b_j, q_j)$ in $(B, Q_t)$, $c_{o_j, q_j}$ is increased by 1. The connection strength $e_{o_i, w_j}$ between the object node $o_i$ and word node $w_j$ is formalized as Equation 2.

$$e_{o_i, w_j} = c_{o_i, w_j}^2 / (d_{o_i} \cdot d_{w_j}), \quad a_{o_i, w_j} = c_{o_i, w_j}^2 / (d_{o_i} \cdot d_{w_j})$$

In this equation, $a_{o_i, w_j}$ and $a_{o_i, w_j}$ are considered as the mapping probabilities in terms of objects and words, respectively. We define the connection strength $e_{o_i, w_j}$ as the product of these two probabilities. Experimental results in Section 6.1 show that this cross-mapping approach makes associations between objects and words more accurate.

Finally, the cross-modal relational graph networks $G$ is constructed, and we can obtain the object-to-word mapping probability distributions $p(w_j|o_i)$ as shown in Equation 3 by using the object-word connection strengths.

$$p(w_j|o_i) = e_{o_i, w_j} / \sum_{w_{\text{new}} \in V^W} e_{o_i, w_{\text{new}}}$$

In the next section, we describe a method for learning an aligned cross-modal representation from the cross-modal graph $G$ where objects and words are aligned.
4 Aligned cross-modal representations

We propose an aligned cross-modal representation learning method that learns the semantic representations of objects and words based on the constructed cross-modal relational graph networks. It allows entities of different modalities with the same conceptual meaning to have the same representation vector. Each object and word node has its own semantic representation vector \( \vec{r}_{oi} \) or \( \vec{r}_{wi} \), each initialized to a random vector. The vectors are trained with the cross-entropy loss for self-identification of each node and the alignment loss for the correspondence between objects and words.

4.1 Neighborhood aggregation

Although object and word co-occurrence statistics were extracted from the local context, a neighborhood aggregation method can be used to reflect the global context. The neighborhood aggregation function for cross-modal relational graph networks \( G \) is defined as follows [15].

\[
\vec{r}^{(l)}_{oi} = \vec{r}^{(l-1)}_{oi} + \delta \sum_{o_j \in N_{oi}} e_{o_i,o_j} \vec{r}^{(l-1)}_{o_j}, \quad \forall o_i \in V^O
\]

\[
\vec{r}^{(l)}_{wi} = \vec{r}^{(l-1)}_{wi} + \delta \sum_{w_j \in N_{wi}} e_{w_i,w_j} \vec{r}^{(l-1)}_{w_j}, \quad \forall w_i \in V^W
\]

Here, \( \delta \) is a constant between 0 and 1 to determine the propagation rate, and \( N_{oi} \) and \( N_{wi} \) is 1-hop neighbors of nodes \( o_i \) and \( w_i \), respectively. By performing this process on the representation vectors \( \vec{r}^{(l-1)}_{oi} \) and \( \vec{r}^{(l-1)}_{wi} \) of all nodes, the updated vectors \( \vec{r}^{(l)}_{oi} \) and \( \vec{r}^{(l)}_{wi} \) for the next layer \( l \) is obtained. For layer number \( l \), neighborhood aggregation is repeatedly performed from \( l = 1 \) to \( l = L \), and the base value \( \vec{r}^{(0)}_{oi} \) and \( \vec{r}^{(0)}_{wi} \) for layer 0 are the node’s own representation vectors \( \vec{r}_{oi} \) and \( \vec{r}_{wi} \).

As this method is repeatedly performed in multiple layers, information from more distant nodes can be aggregated. Finally, we can train the representation vectors \( \vec{r}^{(L)}_{oi} \) and \( \vec{r}^{(L)}_{wi} \) of all nodes using the aggregated vectors \( \vec{r}^{(L)}_{oi} \) and \( \vec{r}^{(L)}_{wi} \) of the final layer \( L \).

4.2 Learning aligned cross-modal representations

For each modality, we can train semantic representation vectors by inputting the aggregated vectors into a multi-layer perceptron (MLP) and performing self-supervised learning to classify the identity of each node. More specifically, we use an MLP with a single hidden layer \( h^O \) with no activation function for the object-relational graph networks \( G^O \). The dimension of the output layer of the MLP is set equal to the number of nodes in \( G^O \), and the softmax function is applied to the output to identify the input vector. We also use an MLP with a hidden layer \( h^W \) for word-relational graph networks \( G^W \). This method enables the semantic representation vectors to maintain their own identity while reflecting the information of neighboring nodes.

We add an alignment loss to the objective function \( L_{total} \) based on the connection strengths \( e_{o_i,w_j} \) of the cross-modal relational graph networks \( G \). The connection strength \( e_{o_i,w_j} \) between the object node
We used the Open Image Dataset V6 [18], which consisted of approximately 1.9 million images. We designed two tasks to investigate the effectiveness of the proposed model. The object-to-word mapping task described in Section 5.2 evaluates the performance of the proposed XSL method. Also, the zero-shot learning task in Section 5.3 validates the effectiveness of the aligned cross-modal representation learning method. In all experiments, the size of the local context window \( H \) was set to 10, and the propagation rate \( \delta \) and the number of layers \( L \), which are parameters for neighborhood aggregation, were set to 0.3 and 5, respectively. The value of \( \lambda_{align} \) for the alignment loss term was set to 1.6, and the size of the embedding vector of each object and word was set to 100 dimensions. We used the Adam optimizer [16] to minimize the objective function (5) for training the representation vectors, and the learning rate and weight decay are set to 0.001 and 1e-5, respectively. For experiments on hyperparameter settings, please refer to Appendix.

5 Experiments

We designed two tasks to investigate the effectiveness of the proposed model. The object-to-word mapping task described in Section 5.2 evaluates the performance of the proposed XSL method. Also, the zero-shot learning task in Section 5.3 validates the effectiveness of the aligned cross-modal representation learning method. In all experiments, the size of the local context window \( H \) was set to 10, and the propagation rate \( \delta \) and the number of layers \( L \), which are parameters for neighborhood aggregation, were set to 0.3 and 5, respectively. The value of \( \lambda_{align} \) for the alignment loss term was set to 1.6, and the size of the embedding vector of each object and word was set to 100 dimensions. We used the Adam optimizer [16] to minimize the objective function (5) for training the representation vectors, and the learning rate and weight decay are set to 0.001 and 1e-5, respectively. For experiments on hyperparameter settings, please refer to Appendix.

5.1 Datasets

We used the Open Image Dataset V6 [18], which consisted of approximately 1.9 million images containing 601 visual object categories and natural language descriptions. We used verified image-level labels for all images, and on average, each image contained 3.8 object categories. Each image description consists of an average of 38.3 tokenized words. For efficient experiments, we removed words that appear less than five times in all descriptions, so there are a total of 6,209 words in the vocabulary. The number of images without description is about 1.2 million. In this case, we only updated the object-relational graph networks using image-level pseudo labels. Since we assumed a stream input environment, all data samples are input to the proposed model only once, one by one.

5.2 Object-to-word mapping

We measured the mapping accuracy of objects and words for evaluation of the constructed cross-modal relational graph networks. Accuracy assessments were performed on 369 objects whose names consisted of only one word and were included in the word vocabulary. We did not limit the words that can map to objects to nouns, but used all words in the vocabulary: nouns, verbs, adjectives, adverbs, interrogatives, function words, etc. Therefore, it is not an easy task to find the exact mapping of one of the 6,209 words for each object. Mappings of the singular and plural with the same meaning (e.g., dog and dogs) were treated as true positives, and other cases such as synonyms were treated as incorrect mappings. We measured top-\( K \) accuracy, which evaluates the accuracy of the \( K \) most probable mappings predicted by the model. As a baseline, we investigated a probabilistic method [6, 13] ‘Prior’ which defined as \( e_{o_i, w_j} = p(o_i|w_j)/\sum_{w_z \in W} p(o_i|w_z) \) where \( p(o_i|w_j) = e_{o_i, w_j}/d_{o_i} \). We also performed ablation studies to verify the effectiveness of the cross-mapping approach described in Section 3.3. Methods ‘w/o \( a_o \)’ and ‘w/o \( a_w \)’ in Table I use \( e_{o_i, w_j} = a_{w_j}(o_i) \) and \( e_{o_i, w_j} = a_{o_i}(w_j) \), respectively, instead of \( e_{o_i, w_j} = a_{o_i}(w_j) \cdot a_{w_j}(o_i) \) in Equation (2). Furthermore, to examine the applicability of the
Table 1: Object-to-word mapping accuracies from top-1 to top-5. (unit : %)

| Pseudo label Method | Top-1   | Top-2   | Top-3   | Top-4   | Top-5   |
|---------------------|---------|---------|---------|---------|---------|
| Prior               | 37.13   | 50.68   | 57.18   | 60.98   | 64.23   |
| Human verified      |         |         |         |         |         |
| w/o a_o             | 45.53   | 62.87   | 71.54   | 75.07   | 78.05   |
| w/o a_c             | 70.19   | 79.13   | 82.11   | 84.28   | 85.91   |
| Ours                | 75.88   | 84.01   | 88.08   | 90.51   | 91.06   |

Table 2: Successful and unsuccessful cases of object-to-word mapping using human verified labels.

| Object    | Word (top-3) | Prob. (%) | Object    | Word (top-3) | Prob. (%) |
|-----------|--------------|-----------|-----------|--------------|-----------|
| porcupine | porcupine    | 50.51     | ambulance | ambulance    | 85.68     |
| hedgehog  | 21.66        |           | stopped   | 0.53         |           |
| porcupines| 19.07        |           |           |              |           |
| Successful cases |
| strawberry | strawberries | 69.31     | watch     | watch        | 47.86     |
|          | 29.79        |           | dial      | 19.31        |           |
|          | 0.17         |           | watches   | 11.71        |           |
|          |              |           |           |              |           |
| train     | pen          | 48.03     |          |              |           |
| track     | 19.28        |           |          |              |           |
| railroad  | 17.75        |           |          |              |           |
| Unsuccessful cases |
| blender   | mixer        | 30.35     | rat       | 85.27        |           |
| juicer    | 26.23        |           | rats      | 7.06         |           |
| grinder   | 22.85        |           | drinks    | 17.67        |           |
| grinder   |              |           |           |              |           |
| seafood   | fish         | 13.97     | bird      | 19.90        |           |
|          | prawns       | 10.65     | dog       | 6.56         |           |
|          | shells       | 8.30      | birds     | 4.61         |           |

proposed method to datasets without categorical labels of objects, we performed experiments using an object detector instead of human verified pseudo-labels. Here, we used YOLOv5 [11] pre-trained using the MS COCO dataset [20]. It can detect 80 categories of objects, and the mapping accuracy was measured for 65 of these categories with single-word object names.

Finally, the continuous learning performance of the proposed model was verified by measuring the mapping accuracy according to the training samples and vocabulary increase. Accuracy evaluation was performed every time 10k image samples were trained, and each time, the number of objects and word vocabularies learned by the model was counted.

5.3 Zero-shot learning

The zero-shot learning task was designed to show that conceptual systems are topologically aligned even for never co-occurred object-word pairs. We first set the cumulative number of co-occurrences $c_{o_i,w_j}$ to 0 for the $Z$ correct object-word pairs ($o_i$-$w_j$) in the cross-modal relational graph networks. We then trained the aligned cross-modal representations using the proposed method.

Zero-shot mapping accuracy was measured for 10 and 20 object-word pairs. For a given $Z$ object-word pair, there are $Z!$ possible one-to-one mappings. To find the best case among them, we selected the one with the largest sum of cosine similarities between the mapped representation vectors. In this process, the Hungarian algorithm [17] was used to reduce the computation time. The ‘w/o $L_{align}$’ method in Table 3 is an ablation study to verify the effect of alignment loss in the proposed method, and was performed by removing the term $L_{align}$ from the total loss $L_{total}$ in Equation 6.

As a baseline, we examined the mapping accuracy based on the alignment correlation (i.e., Spearman correlation) of the two conceptual systems [2]. Here, Spearman correlations were obtained for up to 1M potential mappings randomly selected to find the best mapping with an acceptable computational cost. We also measured random mapping accuracy for comparison. For statistical significance, we repeated each experiment 30 times and $Z$ correct object-word pairs were randomly selected for each trial. Finally, we qualitatively evaluated the proposed method by performing zero-shot learning on 100 correct object-word pairs and visualizing their similarity.
Figure 3: Object-to-word mapping accuracy with increasing training samples. The red and green lines indicate the accumulated size of object and word vocabulary, respectively. As learning progresses, the size of vocabulary increases, and the mapping accuracy also tends to increase continuously.

6 Results and discussion

6.1 Objects-to-words mapping

Table 1 shows the top-1 to top-5 object-to-word mapping accuracies after training all data. In all experiments, the accuracy of the proposed method (‘Ours’) was much higher than that of the baseline ‘Prior’. Ours method also showed higher accuracy than methods ‘w/o o’ and ‘w/o w’, verifying that the proposed cross-mapping approach is effective. Examples of successful top-3 object-to-word mappings are shown in Table 2. In general, the mapping probability of the same word as the object name is high, and the related object name or verb is also included in the top-3. For instance, pairs that are not accurately mapped, such as ‘train-track’, ‘watch-dial’, and ‘eagle-flying’ are also composed of highly related ones. Please refer to the Appendix for more examples.

Figure 3 shows the mapping accuracy as the number of learned samples increases. As the size of the vocabulary increases rapidly at the beginning of training, the mapping accuracy also increases steeply. Even after most of the vocabulary has been observed, the accuracy tends to steadily increase as the model is exposed to various situations. It indicates that the proposed XSL method works properly without forgetting in the aspect of continuous learning.

Limitations of evaluation methodology

We list unsuccessful cases of object-to-word mapping in Table 2. Since we treated the mapping for synonyms as false positives, the most unsuccessful cases were mappings with words that have almost the same meaning as objects, such as ‘blender-mixer’ and ‘mouse-rat’. When the scope of the object name is inclusive, the hyponym was mapped in some cases, such as ‘seafood-fish’ and ‘animal-bird’. There are other cases such as ‘drink-bottle’ and ‘desk-laptop’, but most of them were mapped to words that are highly related to objects.

6.2 Zero-shot learning

Table 3 shows the mapping accuracy of zero-shot learning task. For all tasks, the proposed method (‘Ours’) showed significantly higher mapping accuracy than ‘w/o L_align’ and other baselines. It confirms that objects and words with the same meaning have similar representation vectors in each conceptual system, even if object-word pairs are not directly learned. Also, there was no significant difference in the accuracy of the ‘Random’ method and the ‘w/o L_align’ method, indicating that alignment loss played a decisive role for topologically aligned representation learning.

The perplexity of self-supervised loss $L_s$ was measured as low as less than 1.05. Maintaining the identity of an entity’s representation vector is an important aspect when using it for downstream tasks. We used the neighborhood aggregation method to reflect the relationships of entities, and the alignment loss to make the representations of related objects and words similar. These methods negatively affect the semantic representation of entities to maintain their identity. Nevertheless, the learned representation retains its identity.
Table 3: Mapping accuracy (mean ± std) of the zero-shot learning task. Perplexity (mean ± std) for self-identification was measured for all objects and words.

| Number of pairs | Method              | Perplexity       | Mapping accuracy (%) |
|-----------------|---------------------|------------------|----------------------|
| 10              | Random              | N/A              | 13.00 ± 10.88        |
|                 | Spearman correlation| 1.03 ± 0.02      | 10.67 ± 10.15        |
|                 | w/o \( L_{align} \)| 1.00 ± 0.01      | 13.33 ± 11.35        |
|                 | Ours                | 1.03 ± 0.02      | 63.33 ± 18.07        |

| Number of pairs | Method              | Perplexity       | Mapping accuracy (%) |
|-----------------|---------------------|------------------|----------------------|
| 20              | Random              | N/A              | 6.17 ± 8.27          |
|                 | Spearman correlation| 1.03 ± 0.02      | 7.17 ± 6.91          |
|                 | w/o \( L_{align} \)| 1.01 ± 0.03      | 5.67 ± 5.83          |
|                 | Ours                | 1.03 ± 0.02      | 54.67 ± 15.70        |

Figure 4: **Left:** Similarity matrix visualized by performing the zero-shot learning task on 100 object-word pairs. For correct object-word pairs, objects are arranged on the vertical axis and words are arranged on the horizontal axis in order. High similarity values are distributed along the diagonal axis, which is the correct object-word mapping combination. **Right:** The visualization of the representation vectors of the learned object and word using t-SNE [22]. Objects and words with the same or similar meaning are distributed close to each other.

Additionally, we performed a zero-shot learning task on 100 object-word pairs and visualized their similarity in Figure 4 (Left). Although object-word co-occurrence information for 16.7% of the total visual object vocabulary was removed, high similarity values were generally shown for objects and words with the same meaning. This indicates that the constructed object-relational and word-relational graph networks are topologically aligned.

## 7 Conclusions

We propose a cross-modal alignment learning method that learns the association of objects and words online based on the co-occurrence statistics of vision-language data. The proposed model continuously constructs cross-modal relational graph networks using a cross-situational learning method. By continuously adding and modifying nodes and edges of the graph networks, continuous learning is possible without specifying the number of entities to be learned in advance.

Additionally, we propose an aligned cross-modal representation learning method that learns semantic representations of objects and words from the constructed cross-modal relational graph networks through self-supervision. By adding an alignment loss term to the objective function, objects and words with conceptually similar meanings have similar semantic representation vectors.

**Limitations and societal impact**  Our work has limitations, such as not being able to map the names of objects composed of two or more words, and not learning the relationship between words such as synonym and hyponym. Experiments in this study were performed using a vision-language dataset, but this does not mean that language is acquired only by audiovisual stimuli. Our approach can be extended to various modalities in the future works.
References

[1] Elika Bergelson and Daniel Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012.

[2] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[3] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2940–2949, 2016.

[4] Elizabeth M Clerkin and Linda B Smith. The everyday statistics of objects and their names: How word learning gets its start. In *CogSci*, pages 240–246, 2019.

[5] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018.

[6] Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063, 2010.

[7] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[8] Angela Xiaoxue He and Sudha Arunachalam. Word learning mechanisms. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(4):e1435, 2017.

[9] Jessica S Horst and Chris M Bird. Conceptual systems align to aid concept learning. *Nature Machine Intelligence*, 2(2):92–93, 2020.

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[11] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammmana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiachen Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, October 2020. URL [https://doi.org/10.5281/zenodo.4154370](https://doi.org/10.5281/zenodo.4154370).

[12] Peter W Jusczyk and Richard N Aslin. Infants’ detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23, 1995.

[13] Ákos Kádár, Afra Alishahi, and Grzegorz Chrupała. Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*, 55(3), 2015.

[14] Khazar Khorrami and Okko Räsänen. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning?–a computational investigation. *arXiv preprint arXiv:2109.14200*, 2021.

[15] Taehyeong Kim, Injune Hwang, Hyundo Lee, Hyunseo Kim, Won-Seok Choi, Joseph J Lim, and Byoung-Tak Zhang. Message passing adaptive resonance theory for online active semi-supervised learning. In *International Conference on Machine Learning*, pages 5519–5529. PMLR, 2021.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[21] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11645–11652, 2020.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[23] Jean M Mandler. On the origins of the conceptual system. American Psychologist, 62(8):741, 2007.

[24] Mitja Nikolaus and Abdellah Fourtassi. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In Workshop on Cognitive Modeling and Computational Linguistics, pages 200–210. Association for Computational Linguistics, 2021.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[26] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In European Conference on Computer Vision, pages 647–664. Springer, 2020.

[27] Willard Van Orman Quine. Word and object. MIT press, 2013.

[28] Paul C Quinn, Peter D Eimas, and Michael J Tarr. Perceptual categorization of cat and dog silhouettes by 3-to 4-month-old infants. Journal of experimental child psychology, 79(1):78–94, 2001.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

[30] Okko Räsänen and Heikki Rasilo. A joint model of word segmentation and meaning acquisition through cross-situational learning. Psychological review, 122(4):792, 2015.

[31] Brett D Roads and Bradley C Love. Learning as the unsupervised alignment of conceptual systems. Nature Machine Intelligence, 2(1):76–82, 2020.

[32] Oliver Roesler, Amir Aly, Tadahiro Taniguchi, and Yoshikatsu Hayashi. A probabilistic framework for comparing syntactic and semantic grounding of synonyms through cross-situational learning. In ICRA-2018 Workshop on "Representing a Complex World: Perception, Inference, and Learning for Joint Semantic, Geometric, and Physical Understanding", 2018.

[33] Sumarga H Suanda, Nassali Mugwanya, and Laura L Namy. Cross-situational statistical word learning in young children. Journal of experimental child psychology, 126:395–411, 2014.

[34] Daniel Swingley. Contributions of infant word learning to language development. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1536):3617–3632, 2009.

[35] Akira Taniguchi, Tadahiro Taniguchi, and Angelo Cangelosi. Cross-situational learning with bayesian generative models for multimodal category and word learning in robots. Frontiers in neurorobotics, 11:66, 2017.

[36] Haley A Vlach and Scott P Johnson. Memory constraints on infants’ cross-situational statistical learning. Cognition, 127(3):375–382, 2013.

[37] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5945–5954, 2017.

[38] Chen Yu and Linda B Smith. Rapid word learning under uncertainty via cross-situational statistics. Psychological science, 18(5):414–420, 2007.

[39] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. arXiv preprint arXiv:1805.03508, 2018.
A Appendix

A.1 Dataset

We used ‘Image labels’ of the Open Image Dataset V6 [18] with text descriptions from Localized Narratives [26]. Since the evaluation target of our experiments is the constructed cross-modal relational graph, we did not distinguish the split (i.e., train, validation, and test) of the provided dataset and used them all together. Thus, the dataset consisted of a total of 1,908,119 images containing 601 visual object categories, of which text descriptions for 669,490 images were obtained from the Localized Narratives. The number of images without descriptions is 1,238,629, and these samples were used only to update the object-relational graph.

We also performed experiments using a pre-trained object detector as pseudo labeler instead of human-verified labels to investigate the applicability of the proposed method. We used YOLOv5s [11] as an object detector trained on the COCO dataset [20]. The COCO dataset contains a total of 80 object categories for object detector training. We validated the proposed method for 80 categories of pseudo-labels obtained using the object detector from image samples of Open Image V6.

The download link and license for each dataset are as follows:

- Open Image Dataset V6: https://storage.googleapis.com/openimages/web/download.html (Apache License 2.0)
- Localized Narratives: https://google.github.io/localized-narratives/ (CC BY 4.0)
- COCO: https://cocodataset.org/ (CC BY 4.0)
### A.2 Additional experimental results

#### A.2.1 Object-to-word mapping

**Using human-verified labels**  
Table 4 and 5 show the object-to-word mapping results using human-verified pseudo-labels.

Table 4: Successful cases of object-to-word mapping using human-verified pseudo-labels. The mapping probability of the same word as the object name is generally high, and related object names or verbs are also included in the top-3.

| Object       | Word (top-3) | Prob. (%) | Object       | Word (top-3) | Prob. (%) | Object       | Word (top-3) | Prob. (%) |
|--------------|--------------|-----------|--------------|--------------|-----------|--------------|--------------|-----------|
| dog          | dog          | 77.47     | poster       | 71.62        |           | table        | 26.65        |           |
|              | dogs         | 11.26     | text         | 2.99         |           | chairs       | 6.63         |           |
|              | belt         | 3.91      | images       | 2.34         |           | tables       | 6.18         |           |
| coconut      | coconuts     | 75.68     | caterpillar  | 90.03        |           | sushi        | 96.33        |           |
|              | coconut      | 24.09     | caterpillars | 5.71         |           | delicious    | 1.18         |           |
|              | cocoa        | 0.13      | worm         | 2.74         |           | polaroid     | 0.35         |           |
| coat         | coat         | 22.93     | jeans        | 43.78        |           | pillow       | 43.14        |           |
|              | jacket       | 9.75      | standing     | 3.71         |           | beds         | 24.98        |           |
|              | mannequin    | 8.77      | walking      | 2.77         |           | pillow       | 14.65        |           |
| earrings     | earrings     | 87.31     | vase         | 83.72        |           | helicopter   | 86.20        |           |
|              | earring      | 5.18      | vases        | 6.00         |           | helicopters  | 6.27         |           |
|              | ear          | 4.75      | pot          | 2.18         |           | chopper      | 4.27         |           |
| swan         | swan         | 63.25     | broccoli     | 96.14        |           | teapot       | 65.17        |           |
|              | swans        | 35.40     | cauliflower  | 1.67         |           | kettle       | 29.36        |           |
|              | ducks        | 0.64      | cauliflowers | 1.14         |           | teapots      | 2.15         |           |
| seahorse     | seahorses    | 65.66     | dishwashes   | 66.92        |           | dinosaur     | 74.78        |           |
|              | northern     | 1.46      | mortar       | 62.38        |           | skeleton     | 10.05        |           |
|              |              |           | electronics  | 4.83         |           |              |              |           |
|              |              |           |              |              |           |              |              |           |
| candy        | candies      | 48.91     | cucumbers    | 58.42        |           | house        | 27.35        |           |
|              | candy        | 11.41     | cucumber     | 38.08        |           | porch        | 26.94        |           |
|              | jellies      | 7.07      | pickles      | 0.64         |           | balcony      | 7.53         |           |
| laptop       | laptop       | 67.09     | door         | 61.21        |           | skirt        | 80.75        |           |
|              | laptops      | 25.11     | doors        | 24.14        |           | skirts       | 4.85         |           |
|              | working      | 1.80      | entrance     | 2.62         |           | chalkboard   | 3.54         |           |
| chair        | chairs       | 25.66     | rocket       | 67.82        |           | football     | 70.96        |           |
|              | chair        | 23.19     | rockets      | 12.96        |           | ball         | 12.34        |           |
|              | tables       | 5.54      | launch       | 4.12         |           | kicking      | 4.13         |           |
| dragonfly    | dragonfly    | 89.73     | clock        | 80.47        |           | girl         | 29.44        |           |
|              | dragonflies  | 4.90      | clocks       | 9.28         |           | girls        | 5.88         |           |
|              | fly          | 3.21      | seconds      | 1.37         |           | woman        | 4.86         |           |
| bear         | bear         | 82.39     | sofa         | 51.19        |           | cheetah      | 52.34        |           |
|              | bears        | 8.63      | couch        | 19.13        |           | leopard      | 25.52        |           |
|              | panda        | 5.16      | sofas        | 10.16        |           | leopards     | 12.83        |           |
| toy          | toy          | 37.69     | snowmobile   | 58.31        |           | shower       | 76.30        |           |
|              | toys         | 37.27     | snowmobiles  | 40.29        |           | showers      | 8.36         |           |
|              | lego         | 8.49      | snow         | 0.36         |           | washroom     | 3.26         |           |
| rifle        | gun          | 51.58     | pig          | 57.43        |           | plants       | 20.56        |           |
|              | rifle        | 17.08     | pigs         | 34.42        |           | grass        | 8.29         |           |
|              | guns         | 9.94      | piglets      | 5.45         |           | trees        | 6.78         |           |
| wheelchair   | wheelchair   | 55.54     | tank         | 19.63        |           | accordions   | 6.95         |           |
|              | wheelchairs  | 41.58     | tanker       | 17.39        |           | harmonium    | 24.71        |           |
|              | wheel        | 2.16      | war          | 12.01        |           |              |              |           |
| backpack     | backpack     | 66.61     | suit         | 24.59        |           | tent         | 46.10        |           |
|              | backpacks    | 7.93      | blazer       | 12.98        |           | tents        | 28.55        |           |
|              | bag          | 5.37      | suits        | 10.87        |           | camping      | 19.20        |           |
| tree         | trees        | 25.85     | food         | 28.22        |           | cheese       | 93.57        |           |
|              | sky          | 9.35      | plate        | 12.08        |           | pizza        | 1.12         |           |
|              | the          | 3.98      | bowl         | 8.90         |           | stuffings    | 0.99         |           |
| lipstick     | lipstick     | 67.53     | house        | 28.64        |           | suitcase     | 48.13        |           |
|              | lips         | 19.76     | houses       | 10.60        |           | luggage      | 26.43        |           |
|              |              | 5.15      | windows      | 8.97         |           | suitcases    | 14.47        |           |
Table 5: Unsuccessful cases of object-to-word mapping using human-verified pseudo-labels. Although the mapping is not exact, words with similar or related meanings are mostly ranked.

| Object       | Word (top-3) | Prob. (%) | Object       | Word (top-3) | Prob. (%) | Object       | Word (top-3) | Prob. (%) |
|--------------|--------------|-----------|--------------|--------------|-----------|--------------|--------------|-----------|
| fax          | xerox        | 44.18     | platter      | plates       | 25.84     | mouse        | rat          | 85.27     |
|              | photocopies  | 29.17     |              | forks        | 17.79     |             | rats         | 7.06      |
|              | answering    | 23.67     |              | plate        | 9.72      |             | guinea       | 2.69      |
| person       | the          | 5.87      |              | deer         | 41.97     |             | fish         | 13.97     |
|              | a            | 5.57      |              | impala       | 18.28     |             | prawns       | 10.65     |
|              | and          | 4.87      |              | deers        | 15.81     |             | shells       | 8.30      |
| dessert      | cake         | 22.26     |              | suitcase     | 32.43     |             | jeans        | 16.67     |
|              | food         | 10.15     |              | handler      | 14.15     |             | pant         | 11.50     |
|              | plate        | 8.54      |              | leather      | 13.27     |             | standing     | 4.05      |
| canoe        | paddles      | 25.04     |              | spectacles   | 59.37     |             | bottle       | 17.67     |
|              | boat         | 18.03     |              | specs        | 8.49      |             | wine         | 13.03     |
|              | rowing       | 15.25     |              | spectacle    | 8.26      |             | bottles      | 10.98     |
| snack        | food         | 16.58     |              | cows         | 43.86     |             | thermos      | 84.39     |
|              | packet       | 8.10      |              | cow          | 37.39     |             | headscarf    | 13.25     |
|              | packets      | 7.39      |              | buffalo      | 4.24      |             | juices       | 0.67      |
| heater       | burning      | 42.45     |              | car          | 18.68     |             | purifier     | 38.88     |
|              | flame        | 24.17     |              | road         | 9.35      |             | fluid        | 28.44     |
|              | firewood     | 13.77     |              | cars         | 7.05      |             | headscarf    | 18.28     |
| countertop   | kitchen      | 25.88     |              | a            | 5.66      |             | mixer        | 30.35     |
|              | cupboards    | 12.79     |              | the          | 5.43      |             | blender      | 26.23     |
|              | sink         | 8.87      |              | and          | 4.71      |             | grinder      | 22.85     |
| houseplant   | pots         | 28.59     |              | laptop       | 9.72      |             | flowerpot    | 38.47     |
|              | pot          | 26.71     |              | computer     | 8.51      |             | pot          | 37.99     |
|              | potted       | 10.99     |              | monitor      | 7.71      |             | potted       | 8.80      |
| footwear     | shoes        | 8.24      |              | juice        | 18.17     |             | pistol       | 39.71     |
|              | the          | 5.05      |              | straw        | 16.25     |             | gun          | 24.58     |
|              | people       | 3.55      |              | cubes        | 14.51     |             | revolver     | 15.78     |
| shelf        | racks        | 25.06     |              | chicken      | 49.20     |             | bees         | 46.75     |
|              | shelves      | 23.67     |              | hens         | 24.64     |             | honeycomb    | 19.47     |
|              | books        | 18.60     |              | cock         | 11.30     |             | honey        | 10.17     |
| ladle        | use          | 45.19     |              | car          | 14.47     |             | weighing     | 79.95     |
|              | scrub        | 32.37     |              | vehicle      | 8.78      |             | weight       | 9.19      |
|              | droplet      | 14.31     |              | tyre         | 8.49      |             | balance      | 3.45      |
| furniture    | cosmetic     | 41.00     |              | boats        | 25.97     |             | shells       | 65.53     |
|              | ointment     | 26.60     |              | boat         | 22.40     |             | seashells    | 11.22     |
|              | tubes        | 13.43     |              | ship         | 20.65     |             | shell        | 5.61      |
| bomb         | greenhouse   | 0.00      |              | blades       | 48.22     |             | bird         | 19.90     |
|              | chisel       | 0.00      |              | tool         | 47.11     |             | dog          | 6.56      |
|              | generator    | 0.00      |              | tools        | 3.28      |             | birds        | 4.61      |
| furniture    | table        | 10.17     |              | nail         | 38.47     |             | hangers      | 70.69     |
|              | chair        | 8.06      |              | polish       | 35.55     |             | textile      | 6.99      |
|              | chairs       | 7.88      |              | nails        | 10.10     |             | wardrobe     | 3.73      |
| wok          | pan          | 29.46     |              | the          | 6.79      |             |             |           |
|              | cooking      | 20.87     |              | see          | 5.60      |             |             |           |
|              | pans         | 14.49     |              | in           | 5.50      |             |             |           |
Table 6: Successful cases of object-to-word mapping using the YOLOv5l object detector as a pseudo labeler. The mapping probability of the same word as the object name is generally high, and related object names or verbs are also included in the top-3.

| Object   | Word (top-3) | Prob. (%) |
|----------|--------------|-----------|
| bicycle  | bicycle      | 46.38     |
|          | bicycles     | 23.28     |
|          | riding       | 10.06     |
| airplane | aircraft     | 23.80     |
|          | airplane     | 19.25     |
|          | runway       | 7.90      |
| truck    | vehicles     | 16.23     |
|          | vehicle      | 15.83     |
|          | track        | 10.00     |
| bird     | bird         | 40.30     |
|          | birds        | 14.58     |
|          | branch       | 3.72      |
| horse    | horse        | 62.99     |
|          | horses       | 25.44     |
|          | cart         | 3.96      |
| elephant | elephant     | 58.63     |
|          | elephants    | 31.42     |
|          | rhinoceros   | 3.21      |
| backpack | bags         | 13.55     |
|          | backpack     | 12.62     |
|          | bag          | 8.57      |
| suitcase | luggage      | 28.17     |
|          | suitcase     | 21.90     |
|          | suitcases    | 10.58     |
| snowboard| snowboard    | 40.81     |
|          | snow         | 25.43     |
|          | snowboarding | 11.59     |
| surfboard| surfing      | 45.23     |
|          | surfboard    | 27.92     |
|          | surfboards   | 5.36      |
| fork     | fork         | 56.69     |
|          | forks        | 14.31     |
|          | spoons       | 5.28      |
| bowl     | bowl         | 33.45     |
|          | bowls        | 14.78     |
|          | food         | 8.83      |
| orange   | oranges      | 22.56     |
|          | fruits       | 20.21     |
|          | lemon        | 13.01     |
| pizza    | pizza        | 73.74     |
|          | pizzas       | 5.81      |
|          | food         | 5.23      |
| chair    | chairs       | 12.71     |
|          | chair        | 7.36      |
|          | sitting      | 5.75      |
| toilet   | toilet       | 54.72     |
|          | flush        | 9.93      |
|          | commode      | 7.87      |
| remote   | remote       | 52.08     |
|          | remotes      | 12.17     |
|          | joystick     | 5.79      |

| Object   | Word (top-3) | Prob. (%) |
|----------|--------------|-----------|
| car      | car          | 16.33     |
|          | cars         | 11.65     |
|          | vehicles     | 11.59     |
| motorcycle| bike        | 42.41     |
|          | bikes        | 15.45     |
|          | motorcycle   | 9.35      |
| train    | train        | 38.92     |
|          | track        | 19.61     |
|          | railway      | 18.48     |
| boat     | boat         | 23.51     |
|          | boats        | 23.43     |
|          | water        | 13.39     |
| dog      | dog          | 72.70     |
|          | dogs         | 11.88     |
|          | belt         | 4.69      |
| sheep    | sheep        | 52.74     |
|          | goats        | 9.49      |
|          | animals      | 9.41      |
| cow      | cows         | 23.61     |
|          | cow          | 19.08     |
|          | bull         | 13.75     |
| zebras   | zebra        | 60.72     |
|          | zebra        | 23.83     |
|          | tiger        | 9.95      |
| giraffe  | giraffe      | 55.38     |
|          | giraffes     | 30.60     |
| umbrella | umbrella     | 37.10     |
|          | umbrellas    | 35.24     |
|          | tents        | 9.29      |
| tie      | tie          | 13.90     |
| suit     | suit         | 11.22     |
| blazer   | blazer       | 5.76      |
| frisbee  | frisbee      | 53.29     |
|          | disc         | 9.33      |
|          | weights      | 2.74      |
| skis     | skiing       | 33.43     |
|          | ski          | 17.41     |
|          | skis         | 15.44     |
| parachute| parachute    | 28.16     |
|          | kite         | 18.86     |
|          | parachutes   | 14.90     |
| skating  | skateboard   | 41.10     |
|          | skateboard   | 37.74     |
|          | skate        | 8.15      |
| bottle   | bottle       | 26.99     |
|          | bottles      | 25.66     |
|          | table        | 3.08      |
| cup      | cup          | 8.57      |
|          | table        | 8.24      |
|          | glasses      | 6.80      |
| knife    | knife        | 50.50     |
|          | knives       | 13.11     |
|          | spoons       | 4.44      |
| spoon    | spoon        | 49.17     |
|          | spoons       | 9.46      |
|          | bowl         | 4.80      |
| bananas  | banana       | 57.04     |
|          | bananas      | 21.11     |
|          | artichokes   | 1.76      |
| apple    | apples       | 35.01     |
|          | apples       | 20.46     |
|          | apple        | 10.09     |
| broccoli | broccoli     | 56.21     |
| cauliflower| cauliflower | 4.76     |
|          | vegetables   | 3.48      |
| carrot   | carrot       | 29.52     |
|          | carrot       | 15.21     |
|          | vegetables   | 7.91      |
| doughnuts| doughnuts    | 18.32     |
|          | cookies      | 17.59     |
|          | donuts       | 12.93     |
| cake     | cake         | 58.80     |
|          | cupcakes     | 12.86     |
|          | cakes        | 5.69      |
| sofa     | sofa         | 48.10     |
|          | couch        | 21.80     |
|          | sofas        | 8.89      |
| bed      | bed          | 60.00     |
| pillows  | pillows      | 10.46     |
|          | beds         | 6.37      |
| laptop   | laptop       | 50.09     |
|          | laptops      | 17.73     |
|          | working      | 1.94      |
| mouse    | mouse        | 62.09     |
|          | keyboard     | 5.99      |
|          | mouses       | 5.12      |
| microwave| microwave    | 29.45     |
|          | microwave    | 29.45     |
|          | kitchen      | 3.98      |
Table 7: Unsuccessful cases of object-to-word mapping using the YOLOv5l object detector as a pseudo labeler. Although the mapping is not exact, words with similar or related meanings are mostly ranked.

| Object     | Word (top-3) | Prob. (%) | Object     | Word (top-3) | Prob. (%) | Object     | Word (top-3) | Prob. (%) |
|------------|--------------|-----------|------------|--------------|-----------|------------|--------------|-----------|
| person     | the          | 6.84      | bear       | animal       | 16.28     | bag        | 8.34         |           |
|            | a            | 5.18      | monkey     |              | 11.61     | handbag    |              |           |
|            | in           | 4.85      | polar      |              | 10.66     | people     | 5.37         |           |
| sandwich   | burger       | 27.05     | screen     |              | 18.55     | bags       | 4.10         |           |
|            | food         | 8.94      | tv         | television   | 16.89     |            |              |           |
|            | bread        | 8.69      | monitor    |              | 6.83      |            |              |           |
A.2.2 Zero-shot learning

Parameter search We searched the hyperparameters for training the aligned cross-modal representations in the space of \( \delta \in [0, 0.5] \), \( L \in [0, 6] \), and \( \lambda_{\text{align}} \in [0, 2] \), with the zero-shot learning task. They only affect the zero-shot learning performance marginally unless extreme values are used. When using too large values of \( \delta \) and \( L \) (i.e., \( \delta = 0.5 \) and \( L = 6 \)), the representation vector of each entity does not maintain its identity. Meanwhile, the alignment of the cross-modal conceptual systems is not guaranteed if we use a too small value of \( \lambda_{\text{align}} \) (i.e., \( \lambda_{\text{align}} = 0 \)). Table 8 and Table 9 show the empirical results for the hyperparameter search. In Table 9, we made some changes to the zero-shot learning task to reduce the influence of hyperparameter \( \lambda_{\text{align}} \) and highlight the impact of \( \delta \) and \( L \). Specifically, we set \( c_{a, w} \) and \( c_{o, w} \) to 0 for the object-word pairs \( o_i \sim w_j \) which are the evaluation targets of the zero-shot learning task, where \( o_i \) and \( w_j \) indicate all objects and words, respectively.

Baseline using Spearman correlation In the previous study \([31]\), the authors found that the alignment correlation (i.e., Spearman correlation) of two conceptual systems was positively correlated with the object-word mapping accuracy. However, given the \( Z \) object-word pairs, there are \( Z! \) possible one-to-one mappings. As an example, there are a total of 3,628,800 possible one-to-one mappings for 10 object-word pairs. Thus, examining the correlation of all possible mappings cases requires excessive computational cost. They also show that maximizing the alignment correlation does not guarantee the best mapping, especially when the number of pairs is small. These limitations indicate difficulties in aligning conceptual systems in an unsupervised manner. On the other hand, our method using aligned cross-modal representations overcomes these limitations. In the zero-shot learning task, the proposed method directly measures the similarity of each object-word pair based on the aligned cross-modal representations, and uses the Hungarian algorithm to find the optimal one-to-one mapping. As an example, there are a total of 3,628,800 possible one-to-one mappings.

Table 8: Mapping accuracy (mean ± std) of zero-shot learning task with various \( \delta \) and \( L \). The experiment was performed on 10 object-word pairs, and a value of 1.0 was used for \( \lambda_{\text{align}} \). **Boldface** indicates the top three and the **blue text** indicates the bottom three. The mean and standard deviation were drawn from the results of 20 trials.

| \( L \) | \( \delta = 0.1 \) | \( \delta = 0.3 \) | \( \delta = 0.5 \) |
|-------|-----------------|-----------------|-----------------|
|       | Perplexity      | Mapping accuracy (%) | Perplexity      | Mapping accuracy (%) | Perplexity      | Mapping accuracy (%) |
| 1     | 1.01 ± 0.01     | 69.50 ± 20.37    | 1.01 ± 0.01     | 75.30 ± 18.24    | 1.02 ± 0.02     | 75.50 ± 17.17    |
| 2     | 1.04 ± 0.06     | 68.50 ± 20.07    | 1.02 ± 0.03     | 65.50 ± 18.83    | 1.02 ± 0.04     | 67.00 ± 22.83    |
| 3     | 1.02 ± 0.02     | 65.00 ± 18.03    | 1.01 ± 0.01     | 74.50 ± 11.17    | 1.03 ± 0.05     | 77.50 ± 18.13    |
| 4     | 1.02 ± 0.02     | 68.00 ± 20.15    | 1.01 ± 0.01     | 74.50 ± 18.83    | 1.12 ± 0.17     | 62.00 ± 19.64    |
| 5     | 1.01 ± 0.03     | 70.00 ± 14.83    | 1.01 ± 0.01     | 71.00 ± 19.72    | 22.63 ± 20.82   | 76.50 ± 15.90    |
| 6     | 1.01 ± 0.01     | 68.50 ± 22.20    | 1.12 ± 0.20     | 71.00 ± 17.58    | N/A             | 70.00 ± 17.03    |

Table 9: Mapping accuracy (mean ± std) of zero-shot learning task for hyperparameter search with various \( \delta \) and \( L \). The experiment was performed on 10 object-word pairs, and a value of 1.0 was used for \( \lambda_{\text{align}} \). **Boldface** indicates the top three and the **blue text** indicates the bottom three. The mean and standard deviation were drawn from the results of 20 trials.

| \( L \) | \( \delta = 0.1 \) | \( \delta = 0.3 \) | \( \delta = 0.5 \) |
|-------|-----------------|-----------------|-----------------|
|       | Perplexity      | Mapping accuracy (%) | Perplexity      | Mapping accuracy (%) | Perplexity      | Mapping accuracy (%) |
| 1     | 1.02 ± 0.02     | 13.50 ± 11.52   | 1.02 ± 0.02     | 21.00 ± 13.75    | 1.02 ± 0.02     | 26.00 ± 16.85    |
| 2     | 1.02 ± 0.02     | 22.00 ± 12.08   | 1.01 ± 0.01     | 27.00 ± 17.91    | 1.02 ± 0.04     | 31.00 ± 15.46    |
| 3     | 1.02 ± 0.02     | 20.00 ± 13.78   | 1.01 ± 0.01     | 31.50 ± 13.52    | 1.02 ± 0.02     | 28.50 ± 17.68    |
| 4     | 1.02 ± 0.02     | 19.50 ± 11.17   | 1.01 ± 0.02     | 36.00 ± 22.89    | 1.11 ± 0.19     | 34.00 ± 22.45    |
| 5     | 1.02 ± 0.02     | 25.00 ± 16.88   | 1.03 ± 0.03     | 37.50 ± 18.13    | 19.52 ± 13.61   | 27.00 ± 14.18    |
| 6     | 1.01 ± 0.01     | 25.00 ± 12.04   | 1.05 ± 0.08     | 27.50 ± 18.94    | N/A             | 30.00 ± 15.17    |
Table 10: Mapping accuracy (mean ± std) of zero-shot learning task according to $\lambda_{align}$ value. For the hyperparameters $\delta$ and $L$, values of 0.3 and 5 were used, respectively. Perplexity (mean ± std) for self-identification was measured for all objects and words. We investigated the Spearman correlations for 1M potential mappings for 10 or 20 pairs and all 120 possible mappings for 5 pairs and reported the best mappings. For statistical significance, each experiment was repeated 30 times and correct object-word pairs were randomly selected for each trial.

| Number of pairs | Method          | $\lambda_{align}$ | Perplexity        | Mapping accuracy (%) |
|-----------------|-----------------|-------------------|-------------------|----------------------|
| 5               | Random          | N/A               | N/A               | 19.33 ± 19.56        |
|                 | Spearman correlation | 1.6          | 1.03 ± 0.02      | 23.14 ± 26.68        |
|                 | w/o $L_{align}$ | 0.0               | 1.00 ± 0.00      | 24.00 ± 22.53        |
|                 | Ours            | 0.2               | 1.01 ± 0.02      | 74.67 ± 27.26        |
|                 |                 | 0.4               | 1.01 ± 0.01      | 72.00 ± 27.09        |
|                 |                 | 0.6               | 1.02 ± 0.04      | 87.33 ± 22.58        |
|                 |                 | 0.8               | 1.02 ± 0.03      | 80.67 ± 27.53        |
|                 |                 | 1.0               | 1.03 ± 0.07      | 79.33 ± 30.39        |
|                 |                 | 1.2               | 1.03 ± 0.04      | 69.33 ± 30.95        |
|                 |                 | 1.4               | 1.03 ± 0.03      | 68.00 ± 28.09        |
|                 |                 | 1.6               | 1.03 ± 0.02      | 87.33 ± 21.96        |
|                 |                 | 1.8               | 1.04 ± 0.03      | 81.33 ± 24.60        |
|                 |                 | 2.0               | 1.04 ± 0.05      | 75.33 ± 28.62        |
| 10              | Random          | N/A               | N/A               | 13.00 ± 10.88        |
|                 | Spearman correlation | 1.6          | 1.03 ± 0.02      | 10.67 ± 10.15        |
|                 | w/o $L_{align}$ | 0.0               | 1.00 ± 0.01      | 13.33 ± 11.55        |
|                 | Ours            | 0.2               | 1.05 ± 0.15      | 62.33 ± 22.54        |
|                 |                 | 0.4               | 1.01 ± 0.02      | 67.67 ± 19.77        |
|                 |                 | 0.6               | 1.01 ± 0.01      | 61.67 ± 16.63        |
|                 |                 | 0.8               | 1.01 ± 0.01      | 61.67 ± 21.35        |
|                 |                 | 1.0               | 1.03 ± 0.07      | 69.67 ± 17.52        |
|                 |                 | 1.2               | 1.03 ± 0.02      | 67.33 ± 18.18        |
|                 |                 | 1.4               | 1.03 ± 0.03      | 71.33 ± 19.61        |
|                 |                 | 1.6               | 1.03 ± 0.02      | 63.33 ± 18.07        |
|                 |                 | 1.8               | 1.04 ± 0.05      | 67.67 ± 21.92        |
|                 |                 | 2.0               | 1.04 ± 0.03      | 65.00 ± 20.80        |
| 20              | Random          | N/A               | N/A               | 6.17 ± 8.27          |
|                 | Spearman correlation | 1.6          | 1.03 ± 0.02      | 7.17 ± 6.91          |
|                 | w/o $L_{align}$ | 0.0               | 1.01 ± 0.03      | 5.67 ± 5.83          |
|                 | Ours            | 0.2               | 1.04 ± 0.10      | 50.83 ± 10.75        |
|                 |                 | 0.4               | 1.01 ± 0.01      | 52.33 ± 15.07        |
|                 |                 | 0.6               | 1.01 ± 0.01      | 55.17 ± 11.02        |
|                 |                 | 0.8               | 1.01 ± 0.01      | 56.67 ± 12.27        |
|                 |                 | 1.0               | 1.02 ± 0.03      | 51.33 ± 12.99        |
|                 |                 | 1.2               | 1.03 ± 0.02      | 54.67 ± 13.51        |
|                 |                 | 1.4               | 1.03 ± 0.02      | 56.67 ± 13.54        |
|                 |                 | 1.6               | 1.03 ± 0.02      | 54.67 ± 15.70        |
|                 |                 | 1.8               | 1.03 ± 0.02      | 57.00 ± 11.72        |
|                 |                 | 2.0               | 1.04 ± 0.03      | 56.83 ± 11.71        |
**Computational cost**  Table 11 shows the run-time of our Python implementation on a 3.9 GHz CPU machine. Our method has a much shorter computation time than the ‘Spearman correlation’, and the increase in computation time is not significant as the number of pairs increases. On the other hand, the method using ‘Spearman correlation’ excessively increases the computation time as the number of object-word pairs increases.

Table 11: Comparison of average run-time for zero-shot learning task according to the number of object-word pairs. Spearman correlations were investigated for 1M potential mappings for 10 or 20 pairs and for all 120 possible mappings for 5 pairs. For statistical significance, each experiment was repeated 30 times and correct object-word pairs were randomly selected for each trial.

| Number of pairs | Method                | Computation time (ms) |
|-----------------|-----------------------|-----------------------|
| 5               | Spearman correlation  | 52.25                 |
|                 | Ours                  | 0.18                  |
| 10              | Spearman correlation  | 449887.29             |
|                 | Ours                  | 0.19                  |
| 20              | Spearman correlation  | 484357.88             |
|                 | Ours                  | 0.44                  |