2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Visual Noun Modifiers: The Problem of Binding Visual and Linguistic Cues*

Mohamadreza Faridghasemnia, Jennifer Renoux, Alessandro Saffiotti[1]

***Abstract*— In many robotic applications, especially those involving humans and the environment, linguistic and visual information must be processed jointly and bound together. Existing works either encode the image or the language into a subsymbolic space, like the CLIP model, or create a symbolic space of extracted information, like the object detection models. In this paper, we propose to describe images by nouns and modifiers and introduce a new embedded binding space where the linguistic and visual cues can effectively be bound. We investigate how state-of-the-art models perform in recognizing nouns and modifiers from images, and propose our method by introducing a dataset and CLIP-like recognition techniques based on transfer learning and metric learning. We show real-world experiments that demonstrate the practical applicability of our approach to robotics applications. Our results indicate that our method can surpass the state-of-the-art in recognizing nouns and modifiers from images. Interestingly, our method exhibits a language characteristic related to context sensitivity.**

I. Introduction

In many robotics applications, it is crucial to bind information from different modalities. As an example, a robot doing visual language grounding needs to bind linguistic and visual information. The *binding problem* originates in neuroscience [1] and concerns how the brain binds information from separately processed sensory input [2]. An example is binding color and motion of the same object, given that color and movement are known to be processed in different regions in the brain [3], [4]. The binding problem has also been addressed in cognitive psychology and AI [5], [6], [7], as how information from different modalities should be processed, represented, and bound together. This paper is about enabling a robot to bind linguistic and visual cues.

In AI systems, the binding problem is often addressed by representing information from different modalities in encoded spaces. For example, Contrastive Language–Image Pre-training (CLIP) encoders [8] are state-of-the-art in creating a shared binding space that represents information from language and vision into spaces with sub-symbolic characteristics. How to best preprocess and represent modalities' information in a shared space, ensuring that characteristics of symbolic and subsymbolic methods are preserved, is still an open problem. While subsymbolic encoders are compelling in expressing details, communication with humans and many existing methods for cognitive processes in AI, like reasoning or planning, use symbols.

In robots, a binding space between linguistic and visual cues of objects can be used to recognize visual cues of objects (e.g., object names and attributes) or ground sentences to objects. Visual cues recognition in existing works is often specific to object names [9], attributes [10], or their joint recognition [11], [12]. In language, one can relate these to head nouns and attributive adjectives in noun phrases that describe objects, as in "a small green apple".

Besides being an input modality, natural language is a symbolic system humans use to describe objects. Among different phrases in human language, noun phrases are among the most common types that humans use for describing visual information [13], [14], [15]. Noun phrases are grammatical structures composed of a head noun, and some modifiers that contribute to the meaning of the head noun.

This paper introduces the Visual Noun Modifiers (ViNoM) problem, that is, how an AI system can recognize nouns and modifiers from images. With respect to the state of the art that addresses the recognition of head nouns and attributive adjectives, the ViNoM problem also includes the recognition of attributive nouns, quantifiers, and propositional phrases. It allows a broader range of linguistic information to be identified from objects, inspired by how humans identify objects. To the best of our knowledge, our work is the first that proposes to bind nouns and modifiers to images.

This work also studies some characteristics of symbolic systems in the subsymbolic spaces. Rooted and inspired by characteristics of language, we argue that binding linguistic constructs to objects' images in AI systems has to be *decomposable* and *context sensitive*. By decomposable, we mean that when a system is trained on pairs of object images and their linguistic descriptions (made of multiple lexical units), it should also recognize each lexical unit individually. By context sensitive, we mean that by adding a linguistic context to a lexical unit the system should also recognize the contextualized lexical unit. Namely, due to relations between lexical units, the system can recognize a pack of lexical units, possibly even better than an individual lexical unit. These points are further discussed in Section IV.

This paper's key idea is to exploit visual noun modifiers in robotic applications that need a joint space for visual and linguistic cues. The main novel contributions are: (1) in contrast to prior work, we describe objects by nouns and modifiers, (2) we introduce the problem of visual noun modifiers (ViNoM) binding, (3) we create a dataset of visual

*This work has been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and has also been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016442 (AIPlan4EU).

[1]Center for Applied Autonomous Sensor Systems (AASS), Örebro University, Sweden. mohamadreza.farid@oru.se

noun modifiers,[1] (4) we implement a method, ViNoM$_s$, for binding linguistic cues to visual cues, which bridges the two modalities, (5) we argue for decomposability and context sensitivity in multi-modal binding, and (6) we evaluate our method against prior ones based on those qualities. The key moves in ViNoM$_s$ are to create a shared embedded binding space using pre-trained CLIP models [8], and to apply metric learning to improve recognition of noun modifiers.

The next section reviews related work, and Section III describes the ViNoM dataset. Section IV introduces the ViNoM problem and our method, by describing binding space, kernel function, and evaluation criteria. Results of experiments are reported in Section V, followed by a conclusion.

II. Related work

In robotics, many applications require finding the category of objects in an environment [16], [17], [18]. Several works extract different features of objects that a robot can perceive, such as category, color, shape, and size [19], [20], [21]. These works extract a limited number of objects' attributes and used an individual function for each attribute type. Recognizing a wide range of object attributes can improve a robot's perception of its surroundings, enabling it to acquire a more comprehensive knowledge of the objects in the environment. This has been used, e.g., in image retrieval [22], referential expression grounding [23], visual question answering [24], [25], image captioning [26] and object recognition [27], [28], [29], [30], [31]. Recent research focused on describing objects in finer grained detail, such as PACO [32] focused on describing objects by their parts and attributes. Some works [11], [12] focused on methods that can combine object names and attribute recognition methods together. Our work is peculiar in describing objects through nouns and modifiers, inspired by how humans describe objects in natural language.

Existing works on attribute detection with deep learning methods are based on various labeled datasets focusing on images of different concepts, such as birds dataset [33] or specialized datasets for other objects [34], [35], [36], [37], [38], [39]. Attribute annotation in free-domain datasets can be per-image [40], [41], or per-object, such as the Visual Genome attribute dataset [42]. The latter has subsequently been refined into GQA [43], and PhraseCut [44] datasets. A more refined version has been released as the VAW dataset [10]. Our gathered dataset is an extended version of the VAW dataset, providing additional annotations per object extracted from object descriptions and including more attribute types.

Various methods can be used for attribute recognition. If the number of recognized attributes is low, methods similar to object classification can be used: for example, Deep Attribute Network that ignores recognition of object category and solely focuses on attributes [45]. Some work, such as SCoNe [10], takes the image and the object category and returns a set of predicted attribute labels. We believe these methods fall short in a few aspects; they fail with out-of-vocabulary attributes, and they cannot recognize the most important attribute of an object, which is the object's category.

Some previous works used approaches similar to ours, using CLIP models [8] to recognize a large number of classes of object categories and attributes. For example, Duet [12] performs zero-shot recognition of attributes for an image and a prompt, and OvarNet [11] demonstrated that the CLIP model can outperform previous works in recognition of out-of-vocabulary attributes. While our method shares similarities with these works, we took different objectives in learning the kernel function and focused on a different problem. While other works aim to enhance the classification within the range of VAW label set, our objective is to enhance similarities and quality of binding space by reducing the gap between information modalities, and recognizing a wider range of visual cues.

III. A Visual-Noun-Modifier Dataset

Existing datasets for object attributes (such as VAW) lack labels encompassing all modifier types. This section describes our automated procedure for creating a dataset with objects' images and noun and modifiers labels, by curating existing attribute datasets. Note that our objective in this paper is to find nouns and modifiers used to describe objects. Modifiers include lexical units that serve as attributive nouns, attribute adjectives, and quantifiers. We also store propositional phrases in our dataset for further studies but did not include them in this paper's reports. We aim to have a large open-domain dataset in the number of objects and annotations per instance to create a more generalized binding space. In this work, we used the VAW dataset [10] as the basis of our dataset and enriched it with other modifiers extracted from Visual Genome (VG) region descriptions [42]. Every entry of the VAW dataset (region$_{\text{VAW}}$) has an image name (corresponds to VG images, let's call it *image*), region attributes, object name, and region bounding box. Our approach is to find the correspondence of each VAW entry with VG region descriptions and populate VAW entries based on corresponding VG region description (region$_{\text{VG}}$).

For every entry in VAW, we find the closest VG region description based on L-2 distance. Note that for some entries in VAW, the closest region description does not overlap with the VAW entry, which we filter out using a threshold over the L-2 distance. For any selected region description, nouns, and modifiers (adjectives, propositional phrases, and quantitive determiners) are extracted using a dependency tree and Part-Of-Speech tags computed by the Spacy library (spacy.io). We also applied VG aliases to maintain the label's consistency.

The outcome dataset consists of 2260 (1952 with aliasing) nouns and modifiers, from which 685 are modifiers, distributed over 260895 objects in 72274 images. Each entry of our dataset corresponds to an entry in VAW, added with region description, labels extracted from region descriptions, and L_2-Norm. While the average number of labels per object in the VAW dataset is 2.51, our dataset averages 3.62.

Synthesizing noun phrases: A noun phrase is created for each region from all extracted nouns and modifiers. We take

[1]Dataset available on request from the authors.

the object name as the *headNoun*, quantitative determiners as *quantifier*, and all the rest as *modifiers*, and define the synthesized noun phrase (*NM-Phrase*) as the concatenation of them. As an example, for an apple image that has attribute annotations of ["big", "green"], and has two region descriptions, "It is one fresh apple" and "The picture of a green gala apple", the *NM-Phrase* becomes "one big fresh green gala apple".

IV. Visual Noun Modifiers

This section defines a formal binding space for recognition tasks (Sec. IV-A), followed by the enhancement of noun modifier recognition using a kernel function (Sec. IV-B). We also specify evaluation criteria of binding space's quality (Sections IV-C and IV-D) and the parameters of our experiments (Sec. IV-E).

A. Binding space

A binding space between language and vision is an encoded space where both modalities' information can be represented and bound together via symbols or subsymbols. This contrasts with, e.g., CLIP [8] where binding is only based on subsymbols. Language is traditionally represented by symbols, and binding with symbols requires a model to map images to symbols (e.g., image classification or caption generation). However, these models have limitations in capturing important cues. Classifiers rely on fixed and limited label sets when categorizing images. With generative models, the generated labels lack controllability, namely, it cannot be guaranteed that information about attributes (e.g., color) can be recognized. This is due to generation and recognition being different types of problems.

Subsymbolic spaces can represent nuances of information without predefined label-sets [12], [11], and both language and visual information can be encoded into them. We define two subsymbolic spaces for representing linguistic and visual information from modalities. Consider a set of images (img_i) with their corresponding linguistic description ($\mathscr{L}$), as:

$$\begin{aligned} D_{lang} &= \{\mathscr{L}_1, \mathscr{L}_2, ..., \mathscr{L}_n\} \\ D_{vis} &= \{img_1, img_2, ..., img_n, \} \end{aligned} \tag{1}$$

where elements of D_{vis} correspond to elements in D_{lang}. Assuming a *Encoder* function similar to CLIP [8], that takes the modality input (language or image) and represents them as a vector of numbers into similar spaces, e.g., with the same dimensionality. Then, we can define language and vision spaces ($\mathcal{S}_L$ and $\mathcal{S}_V$, respectively) via the set of encoded representations from $\mathrm{D_{lang}}$ and $\mathrm{D_{vis}}$ as:

$$\begin{aligned} \mathcal{S}_L &= \{x_{language} = Encoder(\mathscr{L}), \mathscr{L} \in D_{lang}\} \\ \mathcal{S}_V &= \{x_{image} = Encoder(image), image \in D_{vis}\} \end{aligned} \tag{2}$$

Note that points in each space are represented as vectors of real numbers. Then, given a point (encoded image) and a set of candidate points (encoded language), one can bind the image to languages by:

$$B = \underset{x \in \mathcal{S}_L}{\operatorname{argmin}}\, Distance(x_{image}, x) \tag{3}$$

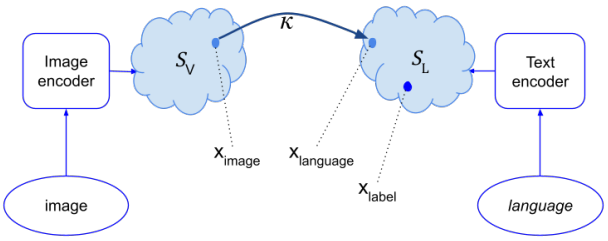

Fig. 1. Overall schema of our method. $\mathcal{S}_V$ and $\mathcal{S}_L$ are constructed from images and descriptions in $\mathrm{ViNoM_{dataset}}$. Zero-shot recognition relies on x_{image} and x_{label}. Decomposability relies on x_{language} and x_{label}, and visual noun modifier recognition relies on $\mathcal{K}(x_{\mathrm{image}})$ and x_{label}.

where B stands for the binding between the image x_{image} and the most similar point (x) in the language space ($\mathcal{S}_L$), with respect to the *Distance* function (here, Cosine distance) between them.

B. Kernel function

In subsymbolic binding spaces generated by large encoders, there is a geometric gap between language and vision spaces, which affects Zero-Shot (ZS) recognition performance. Liang et al. [46] studied this gap and noticed that encoded spaces of large Contrastive learning models are similar but separate. In this work, we revisited metric learning to reduce this gap and improve the recognition performance. Metric learning, also known as similarity learning, is a method aimed at minimizing distance functions [47], [48]. Our approach is to apply metric learning with a kernel function that can minimize the angular distance (*Distance* in Eq. 3) between images (x_{image}) and their corresponding noun modifiers (x_{language}) to improve their recognition. This is shown in Fig. 1, where a kernel function $\mathcal{K}$ learns to minimize the distance between $\mathcal{K}(x_{\mathrm{image}})$ and x_{language}. A kernel is a function that learns to represent data in a new space with particular objectives. We used a fully connected neural network as our kernel function. We set the metric learning regime for our neural network to minimize cosine distance (the same distance function used in Eq. 3):

$$loss_{cosine} = -\frac{1}{N}\sum_{i=1}^{N}(||x^i_{language}||_2 . ||\mathcal{K}(x^i_{image})||_2) \tag{4}$$

where x_{image} is the image embedding, x_{language} is the embedding of the noun phrase (*NM-Phrase*), and $\mathcal{K}$ is the kernel function. Note that cosine loss minimizes the distance between one pair of an image and its description; hence, all the descriptions of an image should be merged into one (*NM-Phrase*).

Binding Noun modifiers in the encoded space: Recognition can be done in the binding space with the help of similarity metrics, as in Eq. 3. For example, an image and a set of descriptions can be represented in their corresponding spaces, and the best description is the closest one to the image representation, as shown by Radford et al. [8].

In order to bind nouns and modifiers to an image representation, our approach is to train the kernel function $\mathcal{K}$

to match the image representation (x_{image} in Fig. 1) to its corresponding encoded *NM-Phrase* (x_{language} in Fig. 1). Since the synthesized noun phrase comprises nouns and modifiers, it is possible to use Eq. 3 to identify nouns and modifiers individually from the kernelized image $\mathcal{K}(x_{\text{image}})$. To this aim, one can use the same encoder that is used for encoding *NM-Phrase* and encode a set of individual nouns and modifiers (x_{label} in Fig. 1) that can be sorted and bound based on their similarity with the encoded image.

In this approach, the performance of recognizing nouns and modifiers (constituents of the noun phrase) depends on the similarity between the kernelized image and each lexical unit in the synthesized expression (x_{label} in Fig. 1). Namely, while the kernel $\mathcal{K}$ learns to minimize the distance between a synthesized expression and an image, the performance of recognition depends on how well it captures the relation between the image and the words that constitute the expression. We call this property "decomposability" of language in the encoded space.

C. Decomposability in the encoded space

Recalling Sec. I, we define an entity in a space as decomposable if one can recognize the constituents of an entity given that one recognizes the whole entities, for example, recognizing the lexical units within a sentence. To describe the decomposability of binding spaces, let us assume a very simple and trivial symbolic binding space of language to lexical units, given a sentence "small green apple" and its lexical units (as "modifier-1", "modifier-2", and "head-noun"). Now, let us consider an encoder that can represent the whole sentence and also each lexical unit into an embedded space, shown in Fig. 2. In this case, decomposability refers to how well the embeddings of each lexical unit can be recognized from the embedding of the whole sentence, similar to decomposability in symbolic language. Figure 2 shows decomposability when lexical units ($x_{\text{label}-\text{i}}$ in Fig. 2) that constitute a language have to be recognized from the whole sentence (x_{language}) in the embedded space. Decomposability is relevant to robotic applications that rely on models like CLIP for fine-grained recognition tasks. Models like CLIP are trained on pairs of images and their full descriptions, and decomposability indicates how well the model can relate the image to the constituents of the descriptions. In other words, decomposability shows the maximum score that such methods can achieve when lexical units of descriptions have to be recognized. Thus, decomposability can indicate the maximum achievable performance: we shall consider this quality in our empirical evaluation below.

D. Context sensitivity in the encoded space

Human use of nouns and modifiers shows that these are syntactically and semantically related. Semantically, both nouns and modifiers may contribute to each other's meanings. Consider the context of the modifier to be the object name. In "a thin laptop" and "a thin pillow", thin is a modifier in different contexts, where it describes a property of the head noun, laptop or pillow. But the norms of laptops

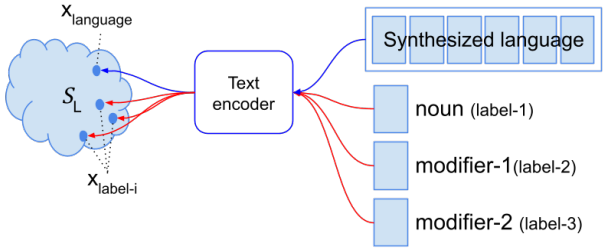

Fig. 2. Decomposability in an embedded space. Labels are nouns and modifiers, and Synthesized language is a composition of them.

and pillows describe the meaning of thin: the context of modifiers affecting their meaning. Syntactically, humans' choice of modifiers depends on the object type and situation. Even though many studies (e.g., [49]) argue that this choice concerns situation, others (e.g., [50]) found that the choice of attribute also depends on the object type, by investigating the object-attribute co-occurrence. For example, while it is common to refer to apples by color, referring to oranges by color is less common. In other words, the first argument suggests the importance of visual context, and the second argument suggests the importance of linguistic context.

We stress the importance of linguistic context and argue that these mutual dependencies can enable a system to recognize nouns and modifiers jointly better than recognizing them individually. Conventionally, most works (e.g., [19]) use different recognition functions and create separate binding spaces for nouns and modifiers; we argue that this separation misses the mutual relation between head nouns and modifiers across modalities. In human-robot interaction, it may be natural for a user to make grounding more explicit by providing extra descriptions in their referential expression. In such cases, context sensitivity becomes an important quality, as it indicates how the performance of grounding changes when the user adds more descriptions to its referential expression. Given the importance of context sensitivity in multi-modal binding, we shall consider this quality in our empirical evaluation below.

E. Training parameters

This section outlines the design choices in preprocessing and training. $\text{ViNoM}_{\text{dataset}}$ is processed by removing labels (head nouns or modifiers) that appeared less than 50 times in the dataset, ignoring region descriptions that their L_2-Norm were above 100, and removing samples with no label. Our kernel function is a network with two middle layers of size 4096, trained with cosine loss and RMS-prop optimizer for 300 epochs, at 0.1 learning rate and with a batch size of 40000. Except for the test of decomposability (described in its respective section), data is randomly split for 70%, 20%, and 10% for train, test, and validation, respectively. Experiments are done on a machine with two Nvidia 3090 GPUs, Core-i9 10900X CPU, and 64GB of RAM.

V. Results

The overall setup of our ViNoM$_s$ method is shown in Fig. 1. To apply transfer learning, we used pre-trained OpenAI CLIP models [8] as our language and vision encoders. These models are used for creating the $\mathcal{S}_L$ and $\mathcal{S}_V$ spaces, representing information from modalities into subsymbolic spaces. To enhance the performance, we add kernel function $\mathcal{K}$ and train it over samples of $\mathrm{ViNoM}_{\mathrm{dataset}}$ while keeping the pre-trained encoders frozen. Our evaluation is in three steps. In the first step, the quality of the binding space ($\mathcal{S}_L$) for decomposability is evaluated, and the most suitable OpenAI language encoder is selected [8]. In the subsequent steps, we evaluate context sensitivity, noun modifier recognizability, and its applicability in real-world scenarios.

Evaluating multi-label problems is not straightforward, particularly in nouns and modifiers recognition, where data is inherently labeled partially. Regarding evaluation metrics, we followed the metrics suggested for the noisy large-scale datasets by Veit et al. (also followed by VAW) [10], [51], and reported the weighted Mean Average Precision, Recall, and F1 of the top K prediction as MAP@K, MAR@K, and MAF1@K, respectively.

A. Decomposibality

In the decomposability test, we encode *NM-Phrase* and each of its lexical units and use Eq. 3 to recognize the lexical units (nouns and modifiers) in the encoded space (binding x_{language} to x_{label} in Fig. 2). This test evaluates the limits of embedded-based recognition models. The decomposability performance of each OpenAI CLIP language encoder [8] is evaluated over a small randomly selected sample (10k in our experiments, expediting the experiments) with MAF1@K, K=5 score. We found the worst model, ViT-B/16, achieved 48.54, and the best model, RN50X4, scored 52.31. We also observed all RN language models performed better than the ViT models. We further evaluated the decomposability of RN50X4 by enlarging the test size to 30% of data (around 75k samples), randomly selected, and tested decomposability over different Ks, reported in Fig. 3.

While the average label per object is 3.6 in our dataset, at K=4, recall reaches 84% as shown in Fig. 3, which shows this space represents our dataset distribution and can be utilized for recognition within our dataset. However, results show some limitations of embedded-based recognition; in our case, we found the maximum precision at K=1 (79.5%) and the maximum recall at K=10 (91.5%). This comes from the fact that long *NM-Phrase* generates an encoding distanced from its lexical units' encoding. This limitation should mainly concern densely-labeled multi-label recognition problems.

Based on the results of decomposability, we chose RN50X4 as the pre-trained encoders and trained the kernel function for the rest of the results. We trained the kernel using specifications and hyperparameters mentioned in Sec. IV, and observed that our model could minimize the loss function to -0.8 over the validation set.

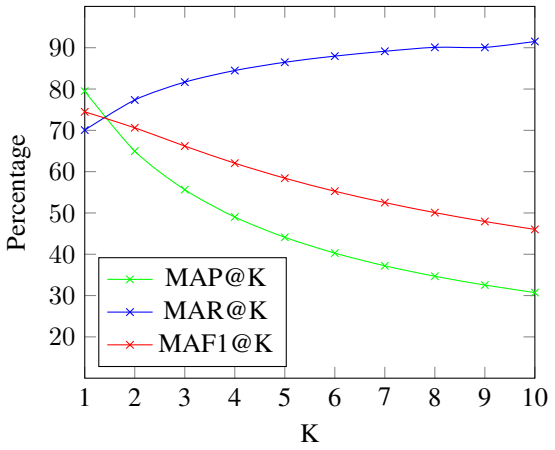

Fig. 3. Decomposability of RN50X4 language model at K=1 to K=10.

TABLE I
Context-sensitivity at k=1 and k=5.

MAF1@k	k=1		k=5	
Modifier's recognition	CLIP	ViNoM$_s$	CLIP	ViNoM$_s$
withoutContext	19.35	30.52	22.71	31.08
withContext	15.00	36.50	19.12	34.45

B. Context sensitivity

The context sensitivity is evaluated in the second step with the pre-trained RN50X4 model and the trained $\mathcal{K}$. In recognition of modifiers, we defined the context as the head noun. We recalled a modifier with the context if the label x_{label} is computed from the concatenation of the head noun and the modifier. Alternatively, when labels (x_{label}) are calculated solely from modifiers, we recall the recognition of modifiers is without context.

The result of context sensitivity is shown in Table I. Results show that the original CLIP image space poorly reflects the context sensitivity. For example, the CLIP model can bind the image of a red apple to the label "red" better than "red apple". We suspect this behavior is because the space created by the CLIP image model ($\mathcal{S}_V$) lacks decomposability, in opposition to its language model.

As presented in Table I, our method (ViNoM$_s$) demonstrates improved context sensitivity, with slightly enhanced recognition results when the context is incorporated. Although results do not improve for some classes, adding context improves the overall performance of recognition of modifiers, indicating that the mutual relationship between nouns and modifiers is captured. We expected this behavior since there is an occasional relation between head nouns and modifiers. We also point out that in the ZS CLIP model adding context dropped the binding performance.

C. Recognizability of nouns and modifiers from image

In the third step, our method is evaluated for recognizing nouns and modifiers, from the encoded images. As suggested by Fig. 4, our method shows a better performance with respect to the ZS RN50X4 model (in K=1 to K=10). Given

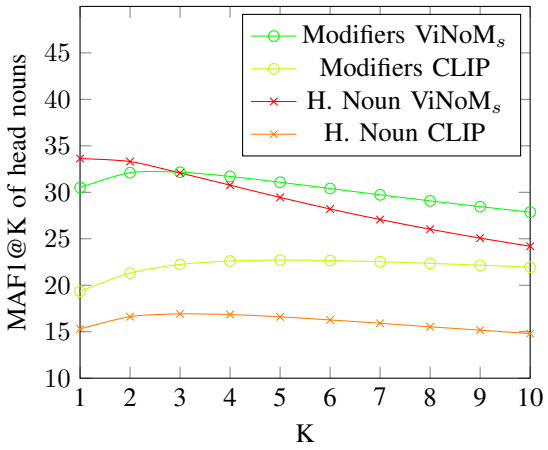

Fig. 4. Recognizability of nouns and modifiers at K=1 to K=10.

TABLE II

RECOGNIZABILITY OF HEAD NOUN AT K=1 AND K=5.

	K=1		K=5	
Head noun's recognition	CLIP	ViNoM$_s$	CLIP	ViNoM$_s$
MAP@K	17.55	38.43	12.44	22.64
MAR@K	13.57	29.90	24.92	42.14
MAF1@K	15.30	33.63	16.60	29.45

the average modifier per object of 2.61, our methods could achieve the maximum performance at K=3, which compared with CLIP (achieved at K=5), indicates that our method performs better in learning the data distribution.

In head noun recognition, each object has only one true label corresponding to the object name. Our method achieves the maximum MAF1@K at K=1, while CLIP space achieves its maximum score at K=3, indicating that our method is more accurate in its top prediction. Moreover, MAP@K and MAR@K of object name recognition for K=1 and K=5 are reported in table II. We observe that our method's precision and recall outperform the original CLIP space, with more significant improvement at K=1.

D. Real world experiment

To test our approach in a real-world scenario, we set up a question-answering scenario with SoftBank's Pepper robot and objects within the PEIS home environment [52], depicted in Fig. 5. Our experiments included 30 objects of types mugs, books, laptops, boxes, utensils, and toys. During verbal interaction between Pepper and users, our method enabled Pepper to ground users' referential expressions and answer user questions about objects' attributes. Users could ask two types of questions: about a specific attribute value or the top three attributes recognized by Pepper.

The scene image is captured by Pepper camera, and objects are cropped using the YOLOv8n model [53]. These object images are encoded using CLIP RN50X4 and the $\mathcal{K}$ function. The provided expression is encoded with the CLIP language encoder and bound to the closest encoded image for grounding. For recognition, the labels in the dataset are encoded, and sorted based on their similarity with the

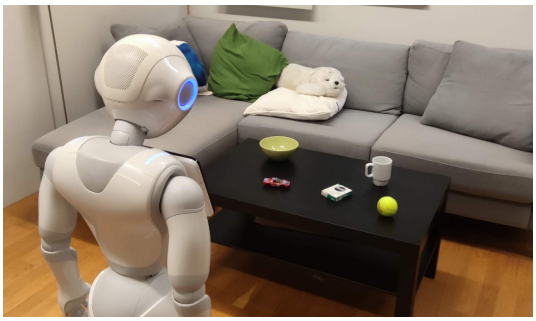

Fig. 5. Setup of objects' QA with Pepper in PEIS home (A video is available at: https://youtu.be/d3txLqHJbWM).

encoded image. We deployed Pepper's Audio Speech Recognition (ASR) and connected the functions through ROS. In a QA session for 30 objects with successful referential expression grounding, a user asked Pepper about the object's name and the top three predicted attributes. We evaluated answers with and without using the kernel function. Our evaluation revealed that the top-1 recognized object name was correct 56% of the time, regardless of whether the kernel function was used. We suspect this is because the test samples were limited to reflect the difference in accuracy on object names. Moreover, with the kernel function, 70% of the top-3 predicted attributes were correct, compared to 44% without it. While both models performed equally well in recognizing object names, our method performed better with the object attributes.

VI. CONCLUSION

Robotic systems, especially those interacting with humans, need to handle both visual and linguistic information, and must ensure the proper binding between them. In this paper, we have singled out an important aspect of this problem which has been under-estimated in the literature so far: the structuring of this combined information in terms of nouns and modifiers. We have defined a shared binding space where linguistic information can be bound to objects' images in these terms. Our initial results, based on offline data as well as online interaction through a Pepper robot, suggest that the ViNoM$_s$ method outperforms current methods, and exhibits properties like decomposability and context sensitivity that are usually afforded by symbolic models.

The work presented in this paper is a first step in the use of linguistic phrase structures for fine-grained binding in robotics, and it has limitations to be addressed in future work. First, using attributes inevitably leads to datasets with contradictory and missing labels, since the same object may be assigned different labels in different images, and not all relevant labels may be used. Moreover, while we focused this investigation on modifiers that are recognizable from the object's images, other types should be considered as well, especially propositional phrases. Finally, it would be interesting to investigate the compositionality of meaning, as the reverse of decomposability, as a further research direction.

REFERENCES

[1] M. Herzog, *Binding Problem*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 388–391. [Online]. Available: https://doi.org/10.1007/978-3-540-29678-2_626

[2] R. Velik, "From simple receptors to complex multimodal percepts: A first global picture on the mechanisms involved in perceptual binding," *Frontiers in psychology*, vol. 3, p. 259, 2012.

[3] A. Bartels and S. Zeki, "The temporal order of binding visual attributes," *Vision research*, vol. 46, no. 14, pp. 2280–2286, 2006.

[4] H. Schneider, "Causal cognitive architecture 3: a solution to the binding problem," *Cognitive Systems Research*, vol. 72, pp. 88–115, 2022.

[5] A. Revonsuo, "Binding and the phenomenal unity of consciousness," *Consciousness and cognition*, vol. 8, no. 2, pp. 173–185, 1999.

[6] K. Greff, S. Van Steenkiste, and J. Schmidhuber, "On the binding problem in artificial neural networks," *arXiv preprint arXiv:2012.05208*, 2020.

[7] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[10] K. Pham, K. Kafle, Z. Lin, Z. Ding, S. Cohen, Q. Tran, and A. Shrivastava, "Learning to predict visual attributes in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 018–13 028.

[11] K. Chen, X. Jiang, Y. Hu, X. Tang, Y. Gao, J. Chen, and W. Xie, "Ovarnet: Towards open-vocabulary object attribute recognition," in *CVPR*, 2023.

[12] Z. Chen, Y. Huang, J. Chen, Y. Geng, W. Zhang, Y. Fang, J. Z. Pan, W. Song, and H. Chen, "Duet: Cross-modal semantic grounding for contrastive zero-shot learning," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[13] E. Kreiss, F. Fang, N. D. Goodman, and C. Potts, "Concadia: Towards image-based text generation with a purpose," in *EMNLP*. Association for Computational Linguistics, 2022, pp. 4667–4684.

[14] P. Rodrigues and S. Kübler, "Part of speech tagging bilingual speech transcripts with intrasentential model switching," in *2013 AAAI Spring Symposium Series*, 2013.

[15] Y. Goyal, A. Mohapatra, D. Parikh, and D. Batra, "Towards transparent AI systems: Interpreting visual question answering models," *arXiv preprint arXiv:1608.08974*, 2016.

[16] S. Ekvall, D. Kragic, and P. Jensfelt, "Object detection and mapping for service robot tasks," *Robotica*, vol. 25, no. 2, pp. 175–187, 2007.

[17] L. Lamanna, L. Serafini, M. Faridghasemnia, A. Saffiotti, A. Saetti, A. Gerevini, and P. Traverso, "Planning for learning object properties," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 12 005–12 013. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26416

[18] L. Lamanna, M. Faridghasemnia, A. Gerevini, A. Saetti, A. Saffiotti, L. Serafini, and P. Traverso, "Learning to act for perceiving in partially unknown environments," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 5485–5493. [Online]. Available: https://doi.org/10.24963/ijcai.2023/609

[19] D. Nyga, S. Roy, R. Paul, D. Park, M. Pomarlan, M. Beetz, and N. Roy, "Grounding robot plans from natural language instructions with incomplete world knowledge," in *Conference on robot learning*. PMLR, 2018, pp. 714–723.

[20] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, and G. Bartels, "Knowrob 2.0 – a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *International Conference on Robotics and Automation (ICRA)*, 2018. [Online]. Available: https://ai.uni-bremen.de/papers/beetz18knowrob.pdf

[21] M. Faridghasemnia, D. Nardi, and A. Saffiotti, "Towards abstract relational learning in human robot interaction," *arXiv preprint arXiv:2011.10364*, 2020.

[22] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[23] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[24] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[25] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, 2017.

[26] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[27] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1778–1785.

[28] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 951–958.

[29] Q. Meng and S. Shin'ichi, "Adinet: Attribute driven incremental network for retinal image classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4033–4042.

[30] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, "Attribute prototype network for zero-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 969–21 980, 2020.

[31] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1200–1209, 2020.

[32] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, *et al.*, "Paco: Parts and attributes of common objects," *arXiv preprint arXiv:2301.01795*, 2023.

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[34] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[35] Y. Zhang, P. Zhang, C. Yuan, and Z. Wang, "Texture and shape biased two-stream networks for clothing classification and attribute recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 538–13 547.

[36] S. Zhang, Z. Song, X. Cao, H. Zhang, and J. Zhou, "Task-aware attention model for clothing attribute prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1051–1064, 2019.

[37] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1667–1675.

[38] Y. Zhong, J. Sullivan, and H. Li, "Face attribute prediction using off-the-shelf cnn features," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–7.

[39] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

[40] G. Patterson and J. Hays, "Coco attributes: Attributes for people, animals, and objects," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 85–100.

[41] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, 2010.

[42] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[43] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.

[44] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, "Phrasecut: Language-based image segmentation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 216–10 225.

[45] S. Banik, M. Lauri, and S. Frintrop, "Multi-label object attribute classification using a convolutional neural network," *arXiv preprint arXiv:1811.04309*, 2018.

[46] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *NeurIPS*, 2022.

[47] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[48] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.

[49] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 69–85.

[50] A. C. Berg, T. L. Berg, H. Daumé, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3562–3569.

[51] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 839–847.

[52] A. Saffiotti, M. Broxvall, M. Gritti, K. LeBlanc, R. Lundh, J. Rashid, B. Seo, and Y.-J. Cho, "The PEIS-ecology project: vision and results," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008, pp. 2329–2335.

[53] Ultralytics, "YOLOv8: A real-time object detection and image segmentation model," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics