PATRON: Perspective-Aware Multitask Model for Referring Expression Grounding Using Embodied Multimodal Cues

Md Mofijul Islam, Alexi Gladstone, Tariq Iqbal

School of Engineering and Applied Science, University of Virginia
{mi8uu,abg4br,tiqbal}@virginia.edu

Abstract

Humans naturally use referring expressions with verbal utterances and nonverbal gestures to refer to objects and events. As these referring expressions can be interpreted differently from the speaker's or the observer's perspective, people effectively decide on the perspective in comprehending the expressions. However, existing models do not explicitly learn *perspective grounding*, which often causes the models to perform poorly in understanding embodied referring expressions. To make it exacerbate, these models are often trained on datasets collected in non-embodied settings without nonverbal gestures and curated from an exocentric perspective. To address these issues, in this paper, we present a perspective-aware multitask learning model, called PATRON, for relation and object grounding tasks in embodied settings by utilizing verbal utterances and nonverbal cues. In PATRON, we have developed a guided fusion approach, where a perspective grounding task guides the relation and object grounding task. Through this approach, PATRON learns disentangled task-specific and task-guidance representations, where task-guidance representations guide the extraction of salient multimodal features to ground the relation and object accurately. Furthermore, we have curated a synthetic dataset of embodied referring expressions with multimodal cues, called CAESAR-PRO. The experimental results suggest that PATRON outperforms the evaluated state-of-the-art visual-language models. Additionally, the results indicate that learning to ground perspective helps machine learning models to improve the performance of the relation and object grounding task. Furthermore, the insights from the extensive experimental results and the proposed dataset will enable researchers to evaluate visual-language models' effectiveness in understanding referring expressions in other embodied settings.

Introduction

Humans naturally use multimodal cues, such as verbal utterances and non-verbal signals (gazes and pointing gestures), to refer to objects and events, known as referring expressions (McNeill 2012; Arbib, Liebal, and Pika 2008; Liszkowski et al. 2006, 2004; Tomasello 2010; Tang et al. 2020; Stacy et al. 2020; Kratzer et al. 2020). In prior work, understanding referring expressions has been generally modeled as grounding relations and objects in visual scenes using verbal utterances, which is known as referring expression comprehension (REF) (Yang, Li, and Yu 2019; Yu et al. 2016; Kamath et al. 2021; Akula et al. 2021). These models are often trained in non-embodied settings, where the visual scenes contain objects but disregard human nonverbal signals. Consequently, these models cannot generalize well in comprehending real-world human interactions.

Several recent works have attempted to address the task of comprehending referring expressions by incorporating nonverbal gestures with verbal utterances in embodied settings (known as embodied referring expression comprehension (E-REF)) (Chen et al. 2021; Schauerte and Fink 2010). However, some crucial issues remain unaddressed in these recent works. Particularly, most embodied referring expression datasets only capture human interactions from an observer perspective with exo-centric views. People innately use an understanding of perspective, which can be observed in how humans interchangeably use perspectives from the speaker and the observer when referring to objects during interactions. For example, a person can refer to an object as "the red lamp to the left of the black hat" from the speaker's perspective or "the red lamp to the right of the black hat" from the observer's perspective (Fig. 1). Thus, understanding perspectives can help a model to ground relations and objects. However, the existing datasets do not contain data from other perspectives (e.g., speaker, observer, neutral) and visual views (e.g., exo, ego, top) to train such a model.

Recent works studied REF and E-REF by designing two separate tasks: a relation grounding task (Goyal et al. 2020; Dai, Zhang, and Lin 2017; Zhuang et al. 2017; Zhang et al. 2017) and an object grounding task (Chen et al. 2021; Achlioptas et al. 2020; Lee et al. 2022; Kazemzadeh et al. 2014). In a non-embodied setting, the relation grounding task is defined as determining whether a verbal utterance appropriately describes the spatial relationships between objects in a visual scene. In an embodied setting, this relation grounding task is defined as determining whether a verbal utterance and nonverbal signals (gazes and pointing gestures) refer to the same object. The object grounding task aims to identify a referred object using a verbal utterance and nonverbal gestures. These tasks have many use-cases in real-world interactions. For example, if a person verbally describes an object but nonverbally points to another object, an AI-driven agent can identify these incoherent multimodal

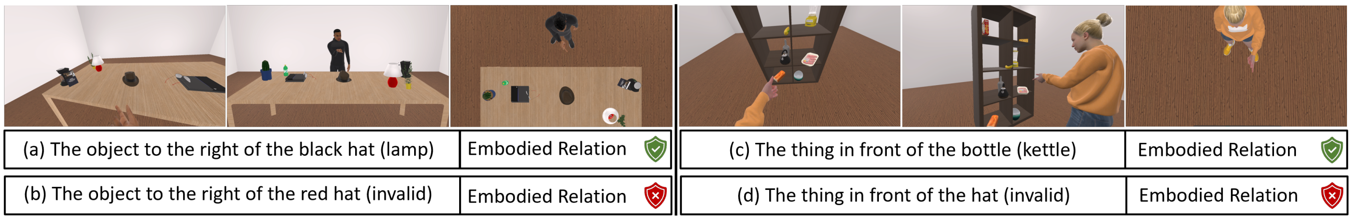
(a) The object to the right of the black hat (lamp)
Embodied Relation
(b) The object to the right of the red hat (invalid)
Embodied Relation
(c) The thing in front of the bottle (kettle)
Embodied Relation
(d) The thing in front of the hat (invalid)
Embodied Relation

Figure 1: Comprehending embodied referring expressions requires an understanding of the perspective, i.e., whether an object is verbally described from the speaker's or observer's perspective. In these scenarios, nonverbal signals (gaze and pointing gesture) can complement verbal utterance to ground an object (a & c). However, sometimes people verbally describe an object and point to or gaze at another object (b & d). Thus, it is also crucial to ground relation for comprehending referring expressions.

cues, using the relation grounding task, and request clarification. In another case, if a person points to an object and asks, "what is the object to the right of the black hat?", then the AI agent can use the object grounding task to identify the referred object (presented in Fig. 1). Thus, training models on these two related tasks (relations and objects grounding) and the previously mentioned perspective grounding task can enable achieve seamless human-AI interactions (HAI).

To address these challenges, we have developed a novel perspective-aware multitask model, PATRON, for the relation and object grounding task using multimodal cues. In PATRON, we have designed two cooperative tasks, one for the perspective grounding (auxiliary task) and another for the relation and object grounding (target task). In the auxiliary task module, PATRON learns disentangled representations (task-specific and task-guidance) to learn perspective grounding. In the target task module, PATRON uses our proposed guided fusion approach that utilizes task-guidance representations from the auxiliary task as prior information to extract guided multimodal representations. PATRON uses a self-attention-based fusion approach to extract supplementary target task-specific representations. Finally, PATRON fuses task-guided and target task-specific disentangled representations to learn relation and object grounding.

Additionally, to overcome the shortcomings of the existing datasets, we have developed a dataset, called CAESAR-PRO, to train and evaluate models for comprehending embodied referring expressions. In CAESAR-PRO, each embodied referring expression is captured from three visual views (ego, exo, and top), and the verbal utterances are generated from three perspectives: speaker, observer, and neutral. We have evaluated the performance of PATRON and state-of-the-art visual-language models by applying on the CAESAR-PRO dataset for perspective and relation-object grounding tasks. Our extensive experimental analysis suggests that perspective learning can improve the performance of visual-language models, including PATRON, for the relation-object grounding task. Moreover, our proposed perspective-aware guided fusion approach helps PATRON to outperform all the evaluated models by achieving the highest accuracy of 74.13% and 81.15% in relation-object and perspective grounding tasks, respectively. Moreover, our ablation study indicates that disentangling multitask representations can help extract salient multimodal features and significantly improve the performance of the relation-object grounding task. Our proposed perspective-aware E-REF model, the dataset, and the insights from our studies open new research directions in HAI.

Related Work

Embodied Referring Expression Comprehension: Several datasets and models have been developed for REF (Mao et al. 2016; Liu et al. 2019; Viethen and Dale 2008). For example, Kazemzadeh et al. (2014) developed a dataset of referring expressions to ground objects in photographs of natural scenes. Lee et al. (2022) curated another dataset to develop model for comprehending referring expressions through visual question-answering. One of the crucial limitations of these datasets is that the data samples are curated in non-embodied settings, where human presence and nonverbal gestures are not considered in referring expressions. As a result, the models trained on these datasets cannot perceive the nonverbal cues that are often necessary for E-REF.

Recently, a few datasets have been developed for E-REF with verbal utterances and nonverbal gestures. For example, Chen et al. (2021) developed a dataset of embodied referring expressions where a human refers to an object using a verbal utterance and nonverbal gestures. However, one of the crucial limitations of this dataset is that the data samples are captured solely from an exocentric perspective. Thus, models trained on these datasets will be biased towards an exocentric perspective and will not comprehend embodied referring expressions from different perspectives.

Multimodal Representation Learning: Several multimodal representation learning models have been proposed for various tasks, such as human activity recognition (Islam and Iqbal 2020, 2021; Samyoun* et al. 2022; Islam, Yasar, and Iqbal 2022; Feichtenhofer et al. 2019), motion prediction (Yasar*, Islam*, and Iqbal 2022; Yasar and Iqbal 2021), visual-question answering (Lu et al. 2019; Li et al. 2019), and referring expression comprehension (Yu et al. 2016; Mao et al. 2016). Existing models for REF predominately use similar cross-attention or self-attention methods to fuse multimodal representations (Goyal et al. 2020; Lee et al. 2022). However, due to lack of perspective diversity and nonverbal gestures, these models do not explicitly learn the perspective taking and the understanding of human nonverbal interaction necessary to comprehend REF.

Multitask Learning: Several multitask models have been designed to learn multiple tasks (Ruder 2017; Hashimoto et al. 2016; Zhang and Yang 2017; Guo et al. 2018; Gagné 2019; Zhou et al. 2020a). The aim of designing these models is two-fold: maximizing shared representations among the tasks and compressing the size of models by maximizing shared learnable parameters across tasks (Ruder 2017; Xu et al. 2018; Zhou et al. 2020b; Achille et al. 2019; Zamir et al. 2018). Moreover, since tasks in these models are occasionally independent and competitive, training approaches can determine which tasks should be learned together (Guo, Lee, and Ulbricht 2020; Jang, Gu, and Poole 2016), further improving model performance. For example, Standley et al. (2020) uses cooperative and competitive task relationships to optimize the shared representations and improve the performance of multiple tasks. A few works use cooperative relationships among tasks to fuse multimodal representations. For example, Islam and Iqbal (2022) uses activity-group information (auxiliary task) to fuse representations for activity recognition (target task), where the classes of the activity recognition task can be directly mapped to the classes of the activity-group recognition task. However, if there is no direct mapping of classes between the tasks, then using the auxiliary task's representation to fuse representations for target tasks can degrade the overall tasks' performance.

PATRON: Perspective-aware Multitask Model

In PATRON, we have designed two tasks: an auxiliary task (perspective grounding) and a target task (relations and objects grounding). We combine relation and object grounding task in a single task (relation-object grounding), where the models identify the referred object if the verbal and nonverbal cues refer to the same object; otherwise, it will report a failed condition. In PATRON, the auxiliary task learns disentangled representations, auxiliary task-specific and task-guidance, where task-guidance representations are used to guide the target task to extract complementary representations. PATRON also learns disentangled representations for target tasks, task-guided and target task-specific, where task-guided representations are learned using task-guidance representations from the auxiliary task. In the following subsections, we present different modules of PATRON.

Unimodal Feature Encoders

PATRON uses modality-specific encoders to encode data from visual and verbal modalities. Visual modalities capture nonverbal gestures in three image views (X_{ego}, X_{exo}, and X_{top}). Verbal utterances (X_{verbal}) refer to an object from a perspective (ego, exo, and neutral). As different modalities have different feature characteristics, PATRON uses separate encoders to encode visual and verbal modalities. This architecture design enables PATRON to utilize state-of-the-art models (F_m) to extract salient unimodal representations (E_m). In our implementation of PATRON, we use ResNet and DistilBERT to extract unimodal representations:

$$E_m = F_m(X_m) \quad , \quad m \in (ego, exo, top, verbal) \quad (1)$$

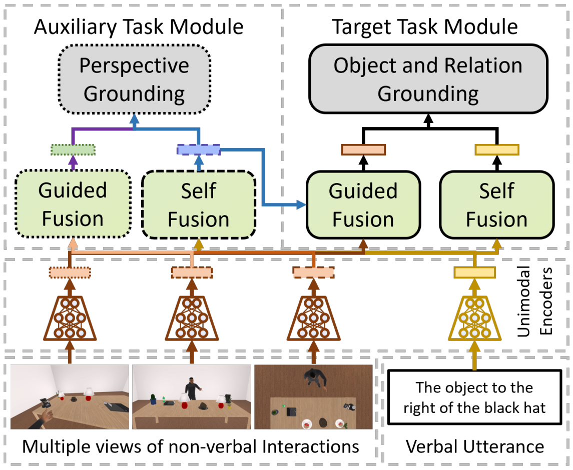

Figure 2: PATRON: Perspective-aware Multitask Learning Model. PATRON learns disentangled representations (i.e., auxiliary task-specific and task-guidance representations) for the auxiliary task (perspective grounding) and disentangled representations (i.e., task-guided and target task-specific) for the target task (relation and object grounding). Here, the proposed guided fusion approach extracts the task-guided representations using the task-guidance representations as prior information from the auxiliary task.

Here, $E_m \in \mathbb{R}^{(B \times D^m)}$, B is the batch size, and D^m is the representation dimension of modality m.

Auxiliary Task Module

In PATRON, the auxiliary task module extracts task-specific and task-guidance disentangled representations from unimodal representations $E_u = (E_{ego}, E_{exo}, E_{top}, E_{verbal})$ (u indicates for unimodal). These disentangled representations are used together to learn perspective grounding, whereas task-guidance representations are also used to guide the target task module to extract perspective-aware complementary representations for relations and objects grounding.

Auxiliary Task-Specific Representation Learning: In PATRON, we have designed a guided fusion approach to fuse unimodal representations. In the auxiliary task module, PATRON uses verbal representation (E_{verbal}) as queries to fuse visual modalities ($E_{visual} = (E_{ego}, E_{exo}, E_{top})$) and produce task-specific representations. At first, PATRON projects (E_{verbal}) to produce queries ($Q = E_{verbal}W^Q$) and projects E_{visual} to produce key ($K = E_{visual}W^K$) and value ($V = E_{visual}W^V$) representations. Here, W^Q, W^K, and W^V are learnable parameters. Finally, queries are used to extract multimodal representations from keys and values in the following way:

$$E^{'} = \sigma\left(\frac{QK^T}{\sqrt{D^u}}\right)V \quad (2)$$

$$E^{aux}_{task\_specific} = W^o E^{'} \quad (3)$$

D^u is the unimodal representation dimension and W^o is a learnable parameter. As we also use guided fusion approach in the target task module, we can summarize this as,

$$E^{aux}_{task\_specific} = Guided\_Fusion(Query, E_u) \tag{4}$$

Task-Guidance Representation Learning: PATRON uses self-attention approaches to fuse unimodal representations and extract task-guidance representations ($E^{aux}_{task\_guidance}$), which is disentangled from $E^{aux}_{task\_specific}$:

$$E^{aux}_{task\_guidance} = Self\_Attn(E_u) = \sum_{m \in M} \alpha_m E_m \tag{5}$$

$$\alpha_m = \frac{exp(\beta_m)}{\sum\limits_{m \in M} exp(\beta_m)} \quad , \quad m \in M \tag{6}$$

$$\beta_m = (W^{aux})^T E_m \quad , \quad m \in M \tag{7}$$

Here, M is the modality list (ego, exo, top, verbal), W^{aux} is a learnable parameter, and α_m is the attention weights.

Perspective Grounding Task: PATRON fuses disentangled representations ($[E^{aux}_{task\_specific}; E^{aux}_{task\_guidance}]$) using a self-attention approach ($Self\_Attn$: Eq. 5) to learn perspective. PATRON uses $E^{aux}_{task\_guidance}$ to learn perspective for ensuring that it contains perspective-aware information, which PATRON uses in the target task module:

$$E^{aux}_{fused} = Self\_Attn([E^{aux}_{task\_specific}; E^{aux}_{task\_guidance}]) \tag{8}$$

$$y_P = F_{Perspective}(E^{aux}_{fused}) \tag{9}$$

Here, $F_{perspective}$ is a multi-layer perceptron (MLP).

Target Task Module

In PATRON, the auxiliary task module (perspective grounding) guides the target task module to extract salient multimodal representations for grounding relations and objects. PATRON uses Guided and Self Fusion modules to extract representations for target task learning.

Task-Guided Representation Learning: PATRON uses our Guided Fusion approach (Section: Task-Specific Representation Learning and Eq.4), to fuse unimodal representations (E_u). In the target task module, the guided fusion approach aims to extract perspective-aware multimodal representations that can be used for grounding relations and objects. PATRON utilizes the guidance representations ($E^{aux}_{task\_guidance}$) from the auxiliary task module as prior information to extract salient multimodal representations: $E^{target}_{guided} = Guided\_Fusion(E^{aux}_{task\_guidance}, E_u)$

Target Task-Specific Representation Learning: Although a guided fusion approach helps PATRON to extract perspective-aware representations (E^{target}_{guided}), verbal and visual modalities can provide additional information to E^{target}_{guided}. PATRON uses $Self\_Attn$ (described in Section Task-Guidance Representation Learning and Eq. 5) to extract supplementary representations for relation-object grounding: $E^{target}_{task\_specific} = Self\_Attn(E_u)$.

Relation-Object Grounding Task: PATRON grounds relations and objects together. PATRON identifies the target object that is referred to by multimodal cues - verbal utterances and nonverbal gestures (gazes and pointing gestures). If the verbal utterance and nonverbal gestures refer to two different objects, then the model should identify these inconsistencies (invalid embodied referring relations) and should not ground any objects. To accomplish this, PATRON fuses guided representations (E^{target}_{guided}) and target task-specific representations ($E^{target}_{task\_specific}$) through a self-attention approach ($Self\_Attn$: Eq. 5):

$$E^{target}_{fused} = Self\_Attn([E^{target}_{task\_guided}; E^{target}_{task\_specific}]) \tag{10}$$

$$y_{OR} = F_{OR}(E^{target}_{fused}) \tag{11}$$

Here, F_{OR} is a MLP to learn target task.

Multitask Learning

We use a multitask learning loss to train PATRON for jointly learning auxiliary (perspective grounding) and target tasks (relations and objects grounding). We use cross-entropy to calculate the loss for auxiliary task ($L_P(y_P, \hat{y}_P) = \frac{1}{B}\sum_{i=1}^{B} y_{(P,i)} \log \hat{y}_{(P,i)}$) and target task ($L_{RO}(y_{RO}, \hat{y}_{RO}) = \frac{1}{B}\sum_{i=1}^{B} y_{(RO,i)} \log \hat{y}_{(RO,i)}$). These losses are used to calculate the multitask loss ($L_{multitask} = \gamma_p L_P + \gamma_{RO} L_{RO}$), where, γ_p and γ_{OR} are task loss weights. L_P helps to learn perspective-aware representations for grounding the perspective. L_P is also used to learn disentangled representation ($E^{aux}_{task\_guidance}$) for guiding the target task to learn perspective-aware multimodal representations.

CAESAR-PRO Dataset

We have used an embodied simulator, CAESAR (Islam, Gladstone, and Iqbal 2022) to develop a dataset of embodied referral expression. CAESAR allows to automatically generate datasets and synthesizing human gaze and gestures from multiple perspectives (ego, exo, and top). Moreover, CAESAR can generate contrastive situations where the person verbally and nonverbally referring two different objects.

We have developed an additional embodied environment in CAESAR, called a shelf environment, where various objects are located on a shelf (Fig. 1-right), whereas the original CAESAR simulator contains only a table-top environment (Fig. 1-left). These two environments allow us to generate diverse data samples with more spatial relations, such as above and below, enabling the model to understand spatial relations in three dimensions. Moreover, due to the locations of cameras in the shelf environment, the observer's point of view differs from the table-top environment (Fig. 1), where the observer is always placed in front of the speaker.

Dataset Generation

To accomplish a realistic and sufficiently variable synthesis of human gaze and pointing gestures, we have used the CAESAR simulator, which uses a gesture synthesis algorithm on real-world data collected using a motion capture system. CAESAR uses inverse kinematics applied to

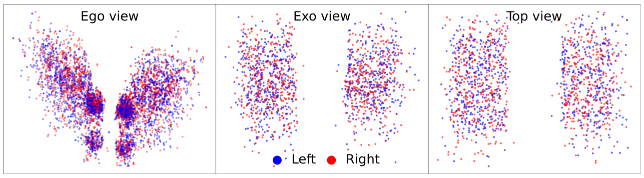

Figure 3: A visualization of referred object locations from different views in the table-top environment is presented here. These locations indicate that the CAESAR-PRO dataset has little to no bias toward object locations in visual scenes and is evenly distributed for a given *left* and *right* spatial relations in verbal utterances.

both the chest and head of the human to generate gestures. To construct verbal referring expressions, we used several templates from CAESAR. We have also included additional spatial relations, such as the above and below relation, which allows for generating more diverse data samples and training models to learn 3D spatial relations. The general structure of these templates: <referred object location><referred object properties><spatial relation>< reference object location><reference object properties>. We have varied this template structure and created eight unique sub-templates. Additionally, we have varied the object names, colors, sizes, locations, and spatial relations to generate diverse verbal expressions to identify one of up to ten objects in a scene from multiple perspectives. These templates are further described in the supplementary document.

Dataset Analyses

The CAESAR-PRO dataset consists of $128,100$ samples, with a train $(79,431)$, validation $(24,597)$, and test split $(24,072)$. The CAESAR-PRO dataset is composed of $229,036$ images, which are mixed with different verbal expressions from multiple perspectives. Images are rendered at a resolution of (480×320) pixels, and an object library of 61 objects is used. We randomly sampled 10 objects which are used as referred objects. Each data sample consists of RGB images, skeletal images, depth map images, and object segmentation mask images for three camera views and a verbal utterance. We also generated task labels for perspective, relation, and object grounding tasks.

We aim to generate a dataset that is not biased to object locations and spatial relations. For example, the term "on the left" always refers to objects on the left side of the scene from the observer's perspective would cause trained models to bias towards only using verbal cues, resulting in a model not being aware of the perspective-taking necessary to ground real-world referring expressions with verbal and nonverbal signals. In Fig. 3, it is evident that the object locations in CAESAR-PRO (table-top environment) are not tied to the spatial relations in the verbal utterances. These diverse data ensure that models use nonverbal cues to ground objects rather than solely relying on verbal cues. We present additional dataset analysis in the supplementary document.

Experimental Setup

We have compared the performance of PATRON by comparing its performance against the following models for the perspective and relation-object grounding tasks: MuMu (Islam and Iqbal 2022), VisualBERT (Li et al. 2019), CLIP (Radford et al. 2021), Dual-Encoder (similar to Zhai et al. (2022)), and Late-Fusion (similar to Islam and Iqbal (2020)). The evaluated models produce visual-language (VL) representations from a single visual image and verbal utterance. We extend these models to process multiple visual views (ego, exo, and top) and a verbal utterance. For VisualBERT, we extract visual representations of multiple views using ResNet-101 and pass these representations with a verbal utterance to learn VL representations. For the CLIP and Dual-Encoder models, we pair the verbal utterance to each visual view and pass each visual-verbal pair through the model to extract VL representations, which are later concatenated. For the Late-Fusion and MuMu models, we extract visual and verbal representations using ResNet-101 and DistillBERT (Sanh et al. 2019), respectively. Late Fusion model fuses these representations using the Multi-Head Self-Attention approach (Vaswani et al. 2017), whereas MuMu uses a guided fusion approach.

We have evaluated these models' performances by applying on the CAESAR-PRO dataset. As some classes contain more data samples than others, we used macro-accuracy metrics to evaluate perspective, object, and relation grounding tasks. We trained each model for eight epochs with a learning rate set to $1e^{-5}$ in a distributed cluster environment with eight A100 GPUs in each cluster node. We train all the models using Pytorch-lightning environment with a fixed seed to ensure reproducibility. For more implementation details, please check the supplementary materials.

Results and Discussion

Comparison of Multitask Learning Approaches

We evaluated the performance of PATRON and other models by applying on the CAESAR-PRO dataset in single and multitask learning settings. In these experiments, a model takes multiple visual views (ego, exo, and top) and a verbal utterance from multiple perspectives (speaker, observer, and neutral) to learn two tasks: (i) perspective grounding task and (ii) relation-object grounding task. In the multitask model, we chose either perspective or relation-object grounding task as the auxiliary task (the first task in the model architecture) and another task as the target task (the second task in the model architecture). For example, in Task Order I, we chose perspective grounding as the auxiliary task and relation-object grounding as the target task. In Task Order II, we chose relation-object grounding as the auxiliary task and perspective grounding as the target task. We have also evaluated state-of-the-art visual-language models in single-task learning settings, where we trained perspective and relation-object grounding tasks using two separate models. We did not evaluate MuMu and PATRON in single-task learning settings, as these models are designed for multitask learning. We present the results of single task and multitask models in Table 1 (a) & (b), respectively.

(a) Single task models trained separately

Models	Perspective (Pers.)	Relation-Object (RO)
Late Fusion	74.90	65.50
Dual Encoder	**77.87**	54.47
CLIP	71.23	65.26
VisualBERT	77.20	**66.53**

(b) Multitask models with different task order in model

Models	Task Order I		Task Order II	
	Pers →	RO	RO →	Pers
Late Fusion	72.30	61.80	65.40	75.12
Dual Encoder	75.67	64.99	43.66	75.77
CLIP	74.52	68.14	56.82	73.02
VisualBERT	74.52	65.90	62.15	69.44
MuMu	73.65	67.48	63.22	75.27
PATRON	**79.85**	**74.13**	**67.63**	**81.15**

Table 1: Top-1 macro accuracy of various models of perspective and relation-object grounding tasks.

Results: The results in Table 1 suggest that PATRON outperforms all the single and multitask models for grounding perspective and relation-object tasks by achieving 81.15% and 74.13% in macro-accuracy, respectively. Among the other visual-language multitask models, CLIP and Dual Encoder achieve the next highest accuracy for relation-object and perspective grounding tasks by achieving 68.14% and 75.77%, respectively. However, among the single task models, VisualBERT and Dual Encoder achieve the next highest accuracy for relation-object and perspective grounding tasks by achieving 66.53% and 77.87%, respectively.

Discussion: The results in Table 1 indicate that for both Task Orders (I & II), the performance of PATRON improves compared to the single and multitask models. Although MuMu uses a guided fusion approach and outperforms single task models for relation-object grounding, it fails to outperform PATRON. However, when considering the task order, some multitask models show improved results compared to their single task models. For example, when Task Order I was considered, the CLIP model showed better accuracy than its single task counterpart for both grounding tasks. Similarly, for Task Order II, the CLIP model showed improved performance for the perspective grounding task; however, the performance degrades for the relation-object grounding task compared to the single task model. One can also observe performance degradation of several models in some multitask settings compared to single task settings. For example, the accuracy of the perspective grounding task degrades for the Dual Encoder and VisualBERT models, whereas the accuracy of the relation-object grounding task degrades for the Late Fusion and the VisualBERT models.

The reasoning behind the performance degradation of the multitask models compared to their single-task counterparts is that the baseline models try to learn a shared representation for all tasks in the multitask setting. As multiple tasks compete to maximize their task-specific representations, a shared representation can discard salient representations of individual tasks. On the other hand, PATRON extracts task-specific and task-guidance disentangled representations. In this process, PATRON uses the task-guidance representations to guide other tasks using our proposed guided fusion approach to extract salient multimodal representations. In the same way, PATRON also learns to extract disentangle representations for the target task and trains these tasks cooperatively, whereas most of the other models train these tasks independently. These findings indicate that a multitask model can improve the tasks' performance if the model can disentangle visual-language representations while training the model in a cooperative learning setting, where one task can guide the learning of other tasks.

Models	Non-Embodied		Embodied		
	V	V+NH	V+G	V+P	V+G+P
BERT	26.44	-	-	-	-
Late Fusion	-	**56.33**	54.91	55.30	61.80
Dual Encoder	-	51.53	53.51	56.93	64.99
CLIP	-	52.63	57.38	60.67	68.14
VisualBERT	-	54.45	58.87	57.05	65.90
PATRON	-	54.24	**65.24**	**66.65**	**74.13**

Table 2: Impact of nonverbal signals (gaze and pointing gesture) on the performance (Top-1 macro accuracy) of the multitask models in the relation and object grounding task. The results suggest that nonverbal signals improve the performance of the models. (V: Verbal, NH: Visual without Human, G: Gaze, P: Pointing Gesture).

Impact of Nonverbal Gestures

We aim to investigate how nonverbal cues impact the performance of the models in the relation-object grounding task. We have conducted this analysis in different settings by varying nonverbal gestures: two non-embodied settings (only verbal (no visual), verbal + visual (scenes without human)), and three embodied settings (verbal + gaze, verbal + pointing gesture, and verbal + gaze + pointing gesture). We trained the models in a multitask learning setting (auxiliary task: prospective grounding, target task: relation-object grounding). We used visual scenes captured from multiple views (ego, exo, and top) and multiple verbal perspectives (speaker, observer, and neutral) to train the models. Table 2 shows the top-1 macro accuracy of the target task.

Results and Discussion: The results in Table 2 suggest that PATRON outperforms all the baseline models in all the evaluated settings for the target task (achieving the highest accuracy of 74.13%). The results also indicate that PATRON achieves the highest accuracy when both gaze and pointing gestures were used, compared to when only gaze or only pointing gestures were used in the embodied setting, and only verbal + visual (scenes without humans) were used in the non-embodied setting. Similarly, other baseline models' performances were also improved when nonverbal cues were used compared to the same model trained with a partial set of nonverbal cues or without any nonverbal cues. Additionally, when only verbal utterances were used, without

Models	Training Perspectives			
	Speaker	Observer	Neutral	All
Late Fusion	60.42	53.71	60.53	**61.80**
Dual Encoder	59.43	45.23	57.95	**64.99**
CLIP	62.36	58.04	60.99	**68.14**
VisualBERT	55.71	43.46	49.68	**65.90**
PATRON	60.36	47.23	57.85	**74.13**

Table 3: Top-1 macro accuracy of the multitask learning models when trained on data samples from single and multiple verbal perspectives and tested on data samples from multiple visual and verbal perspectives.

visual scene (i.e., BERT model), the model achieved only 26.44% accuracy. As the dataset contains verbal expressions that can be interpreted differently from different perspectives, nonverbal gestures can help the models disambiguate and accurately perform the relation-object grounding task. These findings suggest that using nonverbal gestures can improve a model's performance in comprehending E-REF.

Importance of Multi-Perspectives

Here, we investigate how varying verbal perspectives (speaker, observer, and neutral) can impact the performance of the models. We trained PATRON and baseline models on the CAESAR-PRO dataset by varying the verbal perspectives while utilizing all the visual views (ego, exo, and top). During testing, we used all the verbal perspectives and visual views. These models are trained in a multiple-task learning setting (auxiliary task: prospective grounding, target task: relation-object grounding). We have reported the top-1 macro accuracy of the target task in Table 3.

Results and Discussion: The results in Table 3 suggest that all the models demonstrated the highest performance in comprehending E-REF when the models were trained utilizing the data with all the perspectives. For example, training PATRON on multiple perspectives improves the performance of relation-object grounding tasks (achieved 74.13% accuracy) compared to training the same model only on a single perspective. Baseline models also gain similar performance improvement when training the models with data from multiple perspectives. These findings indicate that training models on data samples from multi-perspective can help the models to comprehend E-REF more accurately.

Ablation Study and Significance Analysis

We have conducted ablation studies to evaluate whether our proposed disentangle representation-based guided fusion approach can significantly improve the performance of the relation-object grounding task. We evaluated PATRON and the baseline models on our CAESAR-PRO dataset in the multitask setting (auxiliary task: prospective grounding, target task: relation-object grounding). These models disentangle representations for auxiliary task (task-specific and task-guidance) and target task (task-guided and task-specific). We have conducted a significance analysis ($\alpha = 0.05$) by evaluating these models five times with different parameters initialization (Following Dror, Shlomov, and Reichart (2019)). The results are presented in Table 4.

Models	Auxiliary Task	Target Task	Guided Fusion	Acc.	Std. Dev.	Significant Over [§]
M1	✗	✗	✗	61.38	0.97	None
M2: MuMu	✗	✗	✓	64.21	2.27	M1
M3	✗	✓	✓	64.25	1.07	M1
M4	✓	✗	✓	70.38	0.72	M1-3
PATRON	✓	✓	✓	**74.09**	**0.56**	**M1-4**

Table 4: The results (Top-1 macro accuracy) of the ablation study, where various components of the model are evaluated on the relation-object grounding task. The results of five runs with different initial parameters are presented. ✓and ✗ denote whether a task learns disentangled representations or not, respectively. [§] Significance analysis at level $\alpha = 0.05$.

Results and Discussion: The results in Table 4 suggest that the models with guided fusion can improve the performance of relation-object grounding tasks compared to the model that does not use guided fusion. For example, MuMu (M2) can improve the performance of relation-object grounding by 2.83% compared to a model which does not use guide fusion (M1). Additionally, the models can significantly improve performance if they can disentangle the representation for auxiliary and target tasks (e.g., M3, M4, and PATRON) compared to the models that cannot (e.g., M1). For example, PATRON improves the performance of relation-object grounding tasks by 12.71% by disentangling multiple task representations and using these representations in the guided fusion approach compared to M1. The reasoning behind this significant performance improvement is that learning disentangled representations allows these models to learn task-specific and task-guidance salient representations, which can be used to guide other tasks. On the other hand, models learning non-disentangle representations need to use the same representations for task learning and guiding other tasks. Consequently, the shared representations neglect task-specific salient representation for learning generalized representations for all tasks and degrade the task performance.

Conclusion

We developed a perspective-aware multitask learning model, PATRON, for comprehending referring expressions in embodied settings. We also curated a dataset of embodied referring expressions, CAESAR-PRO, to develop and evaluate learning models. Our extensive experimental results suggest that our perspective-aware guided fusion approach can extract salient multimodal representations for relation and object grounding. Additionally, the results provide valuable insights into developing robust learning models, such as the effects of the order of the tasks, non-verbal cues, verbal perspective, and disentanglement of representations. Finally, we believe the proposed perspective-aware learning model and dataset will be useful in other embodied tasks, such as embodied question answering, embodied navigation, and conversational human-AI interactions.

References

Achille, A.; Lam, M.; Tewari, R.; Ravichandran, A.; Maji, S.; Fowlkes, C. C.; Soatto, S.; and Perona, P. 2019. Task2Vec: Task Embedding for Meta-Learning. In *IEEE International Conference on Computer Vision (ICCV)*.

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. J. 2020. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *16th European Conference on Computer Vision (ECCV)*.

Akula, A.; Jampani, V.; Changpinyo, S.; and Zhu, S.-C. 2021. Robust visual reasoning via language guided neural module networks. *Advances in Neural Information Processing Systems*, 34.

Arbib, M. A.; Liebal, K.; and Pika, S. 2008. Primate vocalization, gesture, and the evolution of human language. *Current anthropology*, 49(6): 1053–1076.

Chen, Y.; Li, Q.; Kong, D.; Kei, Y. L.; Zhu, S.-C.; Gao, T.; Zhu, Y.; and Huang, S. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1385–1395.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 3076–3086.

Dror, R.; Shlomov, S.; and Reichart, R. 2019. Deep dominance-how to properly compare deep neural models. In *ACL*, 2773–2785.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gagné, C. 2019. A Principled Approach for Learning Task Similarity in Multitask Learning. In *IJCAI*.

Goyal, A.; Yang, K.; Yang, D.; and Deng, J. 2020. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *Advances in Neural Information Processing Systems*, 33: 10514–10525.

Guo, M.; Haque, A.; Huang, D.-A.; Yeung, S.; and Fei-Fei, L. 2018. Dynamic task prioritization for multitask learning. In *European Conference on Computer Vision (ECCV)*, 270–287.

Guo, P.; Lee, C.-Y.; and Ulbricht, D. 2020. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, 3854–3863. PMLR.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Islam, M. M.; Gladstone, A.; and Iqbal, T. 2022. CAESAR: A Multimodal Simulator for Generating Embodied Relationship Grounding Dataset. In *NeurIPS*.

Islam, M. M.; and Iqbal, T. 2020. HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10285–10292.

Islam, M. M.; and Iqbal, T. 2021. Multi-GAT: A Graphical Attention-based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. In *IEEE Robotics and Automation Letters (RA-L)*.

Islam, M. M.; and Iqbal, T. 2022. MuMu: Cooperative Multitask Learning-Based Guided Multimodal Fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 1043–1051.

Islam, M. M.; Yasar, M. S.; and Iqbal, T. 2022. MAVEN: A Memory Augmented Recurrent Approach for Multimodal Fusion. In *IEEE Transaction on Multimedia*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. Doha, Qatar: Association for Computational Linguistics.

Kratzer, P.; Bihlmaier, S.; Midlagajni, N. B.; Prakash, R.; Toussaint, M.; and Mainprice, J. 2020. Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters*, 6(2): 367–373.

Lee, J. H.; Kerzel, M.; Ahrens, K.; Weber, C.; and Wermter, S. 2022. What is Right for Me is Not Yet Right for You: A Dataset for Grounding Relative Directions via Multi-Task Learning. *arXiv preprint arXiv:2205.02671*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *Advances in Neural Information Processing Systems*.

Liszkowski, U.; Carpenter, M.; Henning, A.; Striano, T.; and Tomasello, M. 2004. Twelve-month-olds point to share attention and interest. *Developmental science*, 7(3): 297–307.

Liszkowski, U.; Carpenter, M.; Striano, T.; and Tomasello, M. 2006. 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, 7(2): 173–187.

Liu, R.; Liu, C.; Bai, Y.; and Yuille, A. L. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4185–4194.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20.

McNeill, D. 2012. *How language began: Gesture and speech in human evolution*. Cambridge University Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Samyoun*, S.; Islam*, M. M.; Iqbal, T.; and Stankovic, J. 2022. M3Sense: Affect-Agnostic Multitask Representation Learning using Multimodal Wearable Sensors. In *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schauerte, B.; and Fink, G. A. 2010. Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450304146.

Stacy, S.; Zhao, Q.; Zhao, M.; Kleiman-Weiner, M.; and Gao, T. 2020. Intuitive Signaling Through an" Imagined We'". In *CogSci*.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *ICML*, 9120–9132. PMLR.

Tang, N.; Stacy, S.; Zhao, M.; Marquez, G.; and Gao, T. 2020. Bootstrapping an Imagined We for Cooperation. In *CogSci*.

Tomasello, M. 2010. *Origins of human communication*. MIT press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 5999–6009.

Viethen, J.; and Dale, R. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, 59–67.

Xu, P.; Madotto, A.; Wu, C.-S.; Park, J. H.; and Fung, P. 2018. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. ACL.

Yang, S.; Li, G.; and Yu, Y. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4145–4154.

Yasar, M. S.; and Iqbal, T. 2021. A Scalable Approach to Predict Multi-Agent Motion for Human-Robot Collaboration. In *IEEE Robotics and Automation Letters (RA-L)*.

Yasar*, M. S.; Islam*, M. M.; and Iqbal, T. 2022. IMPRINT: Interactional Dynamics-aware Motion Prediction in Teams using Multimodal Context. In *ACM Transactions on Human-Robot Interaction (under-review)*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, 69–85. Cham: Springer International Publishing.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling Task Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18123–18133.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5532–5540.

Zhang, Y.; and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhou, F.; Shui, C.; Abbasi, M.; Robitaille, L.-É.; Wang, B.; and Gagné, C. 2020a. Task Similarity Estimation Through Adversarial Multitask Neural Network. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhou, L.; Cui, Z.; Xu, C.; Zhang, Z.; Wang, C.; Zhang, T.; and Yang, J. 2020b. Pattern-Structure Diffusion for Multi-Task Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuang, B.; Liu, L.; Shen, C.; and Reid, I. 2017. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE international conference on computer vision*, 589–598.