Which One? Leveraging Context Between Objects and Multiple Views for Language Grounding

Anonymous ACL submission

Abstract

When connecting objects and their language referents in an embodied 3D environment, it is important to note that: (1) an object can be better characterized by leveraging comparative information between itself and other objects, and (2) an object's appearance can vary with camera position. As such, we present the Multi-view Approach to Grounding in Context (MAGiC) model, which selects an object referent based on language that distinguishes between two similar objects. By pragmatically reasoning over both objects and across multiple views of those objects, MAGiC improves over the state-of-the-art model on the SNARE object reference task with a relative error reduction of 12.9% (representing an absolute improvement of 2.7%). Ablation studies show that reasoning jointly over object referent candidates and multiple views of each object both contribute to improved accuracy.

1 Introduction

To distinguish a "thin handled mug" between two mugs, we must contextually reason about the object with the *relatively thinner* handle. Such *grounded language* can connect to machine representations of the world (Harnad, 1990). Considering pragmatic context (Potts, 2022; Fried et al., 2022) in grounded natural language can assist applications in vision and robotics (Tellex et al., 2020; Krishna et al., 2017; Lu et al., 2019; Li et al., 2022; Desai and Johnson, 2021). Additionally, object features like mug handles may be occluded from certain viewpoints, requiring multiple views or 3D information (Huang et al., 2022; Wang et al., 2021b).

In the real world, language use is situated in a 3D environment and must consider a rich context of alternatives. However, for tasks like object disambiguation, some models score referring expression compatibility with visual observations of an object in isolation (Thomason et al., 2021; Corona et al., 2022), while broader methods for aligning vision and language representations often consider only static images of objects and scenes (Radford et al., 2021; Kim et al., 2021). Such work can miss language information which can contain comparative information in language and embodied visual information from multiple viewpoints.

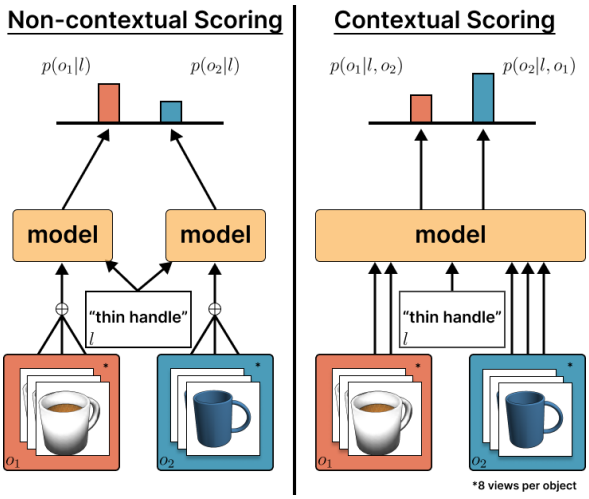

Figure 1: *Left:* Previous methods for identifying object referents of language expressions in the SNARE benchmark consider target and distractor objects independently and pool multiple views before grounding. *Right:* By contrast, MAGiC jointly reasons over target and distractor objects and their views from different angles to identify the correct referent with higher accuracy than the previous state-of-the-art model.

We introduce Multi-view Approach to Grounding in Context (MAGiC) [1]. MAGiC jointly reasons over candidate referent objects *and* considers each object from multiple possible vantage points (Figure 1). We evaluate MAGiC via the **S**hape**N**et **A**nnotated with **R**eferring **E**xpressions (SNARE) benchmark (Thomason et al., 2021). In SNARE, candidate objects are always of the same high-level category, such as *chair* or *mug*, and lan-

[1] Code for MAGiC will be released after anonymous review

guage references uniquely identify one target referent object in contrast to the distractor object of the same category. Embodied agents operating in real-world environments analogously need to disambiguate between similar objects, such as mugs in a kitchen, parts on a conveyor belt, or rocks on the seafloor. By reasoning about both objects and their views, MAGiC achieves a relative error reduction of 12.9% (improved accuracy by 2.7%). Our contributions include:

- MAGiC, a transformer-based model that reasons over multiple 2D-image views of 3D objects and implicitly considers the relative differences between objects;
- state-of-the-art SNARE accuracy;[2]
- ablation studies that show both multi-object and multi-view inputs are needed for the MAGiC accuracy gains; and
- analysis showing MAGiC outperforms previous methods even as fewer object viewpoints are available.

2 Background and Related Work

Embodied agents increasingly operating alongside humans must understand the relationships between natural language and the objects they reference. To best capture these relationships, our method synthesizes the comparative context afforded by reasoning over multiple objects and considering each one in multiple views.

2.1 Object Referent Identification

Object referent identification selects specific object referents given natural language descriptors. Several datasets are prominent in 3D object referent identification. ShapeGlot (Achlioptas et al., 2019) focuses on chairs and lamps, training models to distinguish target objects using shape-based descriptions. PartGlot (Koo et al., 2022) employs a reference task for implicit learning of point cloud part-segmentation. SNARE (Thomason et al., 2021) uses the ShapeNetSem dataset, featuring 262 object categories, while ShapeTalk (Achlioptas et al., 2023) introduces 29 object classes for learning grounded point cloud representations. We utilize the SNARE dataset, leveraging extensive object variety to highlight generalizability.

Previous SNARE task methods scored objects individually (Thomason et al., 2021; Corona et al., 2022). We discuss the limitations of these methods by considering two specific principles in pragmatics (Potts, 2022). The first is the consideration of contrastive object sets in reference games (Andreas and Klein, 2016). Another similarly relevant pragmatics principle relevant to our work is the consideration of alternatives (Fried et al., 2022). These principles suggest the importance of utilizing comparative information between presented objects when completing SNARE or a similar task.

MAGiC employs language grounding to capture object distinctions in the SNARE task. Our core insight is joint reasoning over both objects, diverging from methods that independently score reference-referent and reference-distractor pairs (Thomason et al., 2021; Corona et al., 2022).

2.2 3D Language Grounding

In the domain of grounding language to visual representations, significant progress has been made in 2D (Sadhu et al., 2019; Yu et al., 2016; Plummer et al., 2015; Wang et al., 2021a). This research can be extended to work in three dimensions, incorporating more information such as the relative positions and views of multiple objects. There are many common 3D object representations such as point clouds (Qi et al., 2017; Guo et al., 2020), meshes (Lin et al., 2021; Bouritsas et al., 2019), voxels (Yagubbayli et al., 2021), and neural radiance fields (Mildenhall et al., 2020; Yu et al., 2021).

Applications of language and 3D representations include resolving spatial reference for language localizing objects in a 3D scene (Zhang et al., 2017; Huang et al., 2018, 2022). Language guidance can also inform real-world tasks in 3D such as vision-and-language navigation (Gu et al., 2022) or robot instruction following (Shridhar et al., 2020, 2022). In all these tasks, grounded language understanding of objects from different viewpoints is necessary.

The necessity of this 3D, rotational understanding is more prominent in 3D object referent identification tasks such as SNARE (Thomason et al., 2021) and ShapeGlot (Achlioptas et al., 2019). While the model may be presented with explicit 3D object representations to provide rotational information in other identification tasks, SNARE provides multiple 2D views of the referent and distractor objects. The previous SoTA methods on SNARE have all aggregated these views before generating a score for each object. However, in keeping with Grice's maxim of quantity (Grice, 1975), the MAGiC transformer attends over all the views of both objects, in contrast to previous meth-

[2] https://github.com/snaredataset/snare#leaderboard

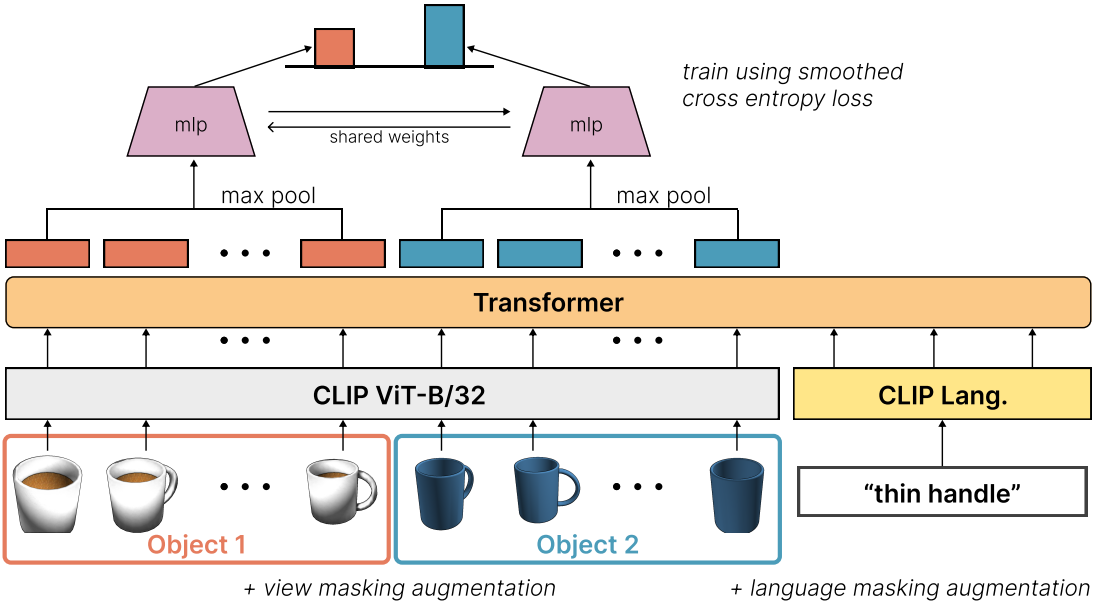

Figure 2: **Model Architecture.** MAGiC consists of a multi-view transformer that attends to CLIP language embeddings for the description and CLIP image embeddings across multiple views for both objects. This transformer allows our model to contextually reason across views about both objects at the same time with respect to a language description. We do not use any positional encodings, and MAGiC is invariant to the input order of images and objects. Unlike previous methods for SNARE, we pool information from object views only after updating their representations with respect to the language referring expression. We apply view masking and language masking augmentations to regularize the model during training.

ods attempting SNARE that performed early fusion on view representations.

3 Object Reference Task

We define an object reference task where, given one or more visual views of candidate objects and a natural language description, the in an object reference task is to select the referent representing the content of the description. Formally, a model must use a given language description l to predict a target object o^l that is aligned with the language description from among a set of m objects $O = \{o^l, o^{c_1}, o^{c_2}, ..., o^{c_{m-1}}\}$. Besides object o^l, there are $m-1$ distractor objects o^{c_i} that contribute to the context in which an model needs to reason about. For each object o, the model is able to perceive n views for each object $o_1, ..., o_n$. These objects are unordered, and we do not assume access to the relative positions between each view. The goal of the task is to learn a classifier function $f(O, l) \rightarrow [0, 1]^m$ such that a higher probability is assigned to the target object.

Previous approaches (Koo et al., 2022; Achlioptas et al., 2023; Thomason et al., 2021; Corona et al., 2022) learn f for single objects, then each object $o \in O$ is scored separately using a single-object classifier $s(o, l) \rightarrow [0, 1]$. While classifying only individual objects simplifies the implementation, it limits the model's ability to comparatively reason about objects in context. Also, previous image-based methods for object reference tasks (Thomason et al., 2021; Corona et al., 2022) aggregate each object's n views without reasoning about each view's relationship to the language description. To overcome these limitations, a model needs to address two key challenges: 1) reasoning about the contextual relationships between objects, and 2) reasoning about multiple views of each object in relation to the language description. We propose MAGiC, a transformer-based architecture that enables joint reasoning about object-specific and view-specific contextual dependencies for 3D language grounding.

4 MAGiC

We introduce Multi-view Approach to Grounding in Context (MAGiC) for language grounding of 3D objects (Figure 2). In contrast to previous work that individually score each object, MAGiC considers both the language and the objects, along with their views, simultaneously.

The design of our model is guided by principles

Model	Considers Both Objects	Lang Attends to Ind. Views	VALIDATION ACC. Visual	Blind	All	TEST ACC. Visual	Blind	All
Human (U)	✓	✓	94.0	90.6	92.3	93.4	88.9	91.2
ViLBERT	✗	✓	89.5	76.6	83.1	80.2	73.0	76.6
MATCH	✗	✗	89.2(0.9)	75.2(0.7)	82.2(0.4)	83.9(0.5)	68.7(0.9)	76.5(0.5)
LAGOR	✗	✗	89.8(0.4)	75.3(0.7)	82.6(0.4)	84.3(0.4)	69.4(0.5)	77.0(0.5)
VLG	✗	∼	91.2(0.4)	78.4(0.7)	84.9(0.4)	86.0	71.7	79.0
MAGiC	✓	✓	**92.1(0.4)**	**81.3(0.9)**	**86.8(0.5)**	**87.7**	**75.4**	**81.7**

Table 1: **SNARE Benchmark Performance.** Mean accuracy ± standard deviation over 10 seeds for existing SNARE approaches, whether those approaches reason over objects jointly, and whether they perform language grounding over individual object views versus pooled representations. Note: ∼ indicates that VLG enables language grounding to LegoFormer (Yagubbayli et al., 2021) features of object views, but not RGB views. We find that MAGiC outperforms all other models on SNARE and is statistically significantly better than VLG, the previous state-of-the-art approach, under a Welch's unpaired two-tailed t-test with a $p < 0.001$.

in 3D language grounding and pragmatics. Grice's maxim of quantity reinforces the need for leveraging all available information necessary for solving a given task. In SNARE, a model should consider information about comparative differences between two objects to identify the correct referent of the language expression. To enable a model to more effectively ground language to these visual dissimilarities, we focus on two key elements of the model formulation: (1) **object context**, which involves jointly reasoning over both objects and the referring expression, and (2) **multi-view context**, where multiple views of the object representation are explicitly utilized throughout the model without aggregating their representation as a preprocessing step. We adopt this paradigm in 3D language grounding and design our model to concurrently process features from *both* objects and the referring expression to leverage context-dependent information.

With the common-ground context established by considering both objects and the referring expression, our model can leverage context-dependent information effectively. More concretely, consider the scenario of the model being asked to choose between two chairs given the referring expression "the tall, skinny chair". The model can exploit context-dependent information, such as using the descriptor "tall" to reason over both objects comparatively to ascertain which is taller. Additionally, by incorporating features from multiple views of the object, our model benefits from the additional 3D perspective, ensuring that important object information, even if initially rotated out of view, is captured and utilized.

A transformer architecture is well-suited for context-based 3D language grounding due to its wide receptive field and low inductive bias (Vaswani et al., 2017). Unlike CNN-based architectures that have a spatial locality bias, transformers have a wide receptive field that includes *all* input features after just one transformer layer. This architecture enables our model to attend to all inputs and effectively leverage both object and multi-view context for 3D language grounding. Moreover, the low inductive bias of transformers makes the design choice suitable for 3D language grounding, as the transformers are particularly good at handling multiple modalities (Xu et al., 2023).

4.1 Model Architecture

Given a target object o^l, a single distractor object o^c, and the language description l, MAGiC employs a transformer-based architecture to learn a classifier $f([o^l, o^c], l)$. We conjecture that our architecture will effectively learn contextual relationships between views and objects. Our approach focuses on the use of images from each view to represent an object, without relying on additional depth or camera information. Thus, each object o has n views that represent the object. Unlike previous work that used additional features, such as voxel-based information (Corona et al., 2022) or point cloud information (Huang et al., 2022; Achlioptas et al., 2019), we demonstrate the effectiveness of using image-based views alone for 3D language grounding. Thus, our model is agnostic to specific orderings of views for an object.

For each view, we utilize a CLIP-ViT (Radford et al., 2021) image encoder g to obtain view-specific visual embeddings $v_i = g(o_i)$. Similarly, a CLIP language encoder h is employed to encode the given language description l, generating a sequence of token embeddings $[e_d^1, ..., e_d^k] = h(l)$. Similar to the previous state-of-the-art model

(Corona et al., 2022), we use the token-level text embeddings from CLIP rather than the CLIP's end-of-token feature that SNARE's baselines use (Thomason et al., 2021). To distinguish between image-view embeddings and language embeddings, we add a learned token-type embedding to each token to indicate whether it is an image-view embedding or a language embedding (Kim et al., 2021). To ensure permutation invariance between objects, we do not add a token embedding to distinguish whether a view belongs to the first or second object. To remain agnostic to view orderings, we deliberately exclude positional encodings from all views.

Using these representations for the objects and language, we construct a sequence $r = [v_0^l, ..., v_n^l, v_0^c, ...v_n^c, e_l^1, ..., e_l^k]$, which is then passed as input to the transformer encoder t:

$$[w_0^l, ..., w_n^l, w_0^c, ...w_n^c, q_l^1, ..., q_l^k] = t(r),$$

where w is a contextualized representation for an object's view, and q are output representations for the language input. The resulting contextualized representations capture the interplay between views and the language input.

The object-specific output representations $w_0, ..., w_n$ from the transformer t for an object o are aggregated using max pooling, yielding a single aggregate embedding u representing object o. This aggregate embedding captures the contextual relationships between multiple views of the object in consideration. A classifier MLP $s(u)$ takes the contextualized embeddings for an object o as input and generates a score s indicating the likelihood of the object being the target.

Given a target object o^l that is aligned with a language description l and a single distractor object o^c, we apply a sigmoid to the scores for each object to compute the probabilities $p(o^l|l, o^c)$ and $p(o^c|l, o^l)$ of the target and distractor objects, respectively.

4.2 Attention Masking Augmentation

Humans often adapt and rely on a subset of views or language cues when faced with challenging circumstances or limited information. This observation motivates the exploration of masking techniques in language grounding tasks, aiming to enhance model performance by selectively blocking out certain inputs and encouraging the model to focus on the most relevant information.

We incorporate attention masking augmentations into our model, specifically targeting the transformer's attention weights for both the view and language inputs (Girdhar and Grauman, 2021; Vaswani et al., 2017; Cho et al., 2022). This masking strategy encourages the model to develop a better understanding of multi-view contextual relationships and effectively capture the essential aspects for accurate predictions.

For view masking, we introduce a 10% probability of masking out each individual view during training. This process promotes view invariance as well as the ability to generalize to unseen viewpoints. Similarly, for language masking, we apply a 20% probability of masking out each word in the input language description. By randomly masking a portion of the word and image embeddings, we encourage the model to learn more robust vision and language representations that are capable of handling missing or incomplete information.

	VALIDATION ACC.		
Model	Visual	Blind	All
MATCH	90.6(0.5)	77.0(0.7)	83.9(0.4)
+ obj. context	90.5(0.5)	76.8(0.6)	83.7(0.3)
MAGiC	**92.1(0.4)**	**81.3(0.9)**	**86.8(0.5)**
- obj. context	91.1(0.5)	79.4(1.1)	85.3(0.5)
- mv. context	91.0(0.6)	79.5(0.8)	85.3(0.4)
- both contexts	90.5(0.6)	78.2(1.2)	84.4(0.6)

Table 2: **Context Ablations.** We investigate the importance of multi-view context (mv. context) and object context (obj. context). On the validation set, we report 10 averaged seeds and the standard deviation on ablations of both contexts for MATCH and MAGiC. We note that the MATCH performance is different from Table 1 as these are our replications of MATCH results as opposed to the paper (Thomason et al., 2021) report. We find that if we remove one type of context or both, performance is degraded for MAGiC.

5 Evaluation

We evaluate the effectiveness of our method on the SNARE (Thomason et al., 2021) benchmark, a language grounding task that draws from a subset of items in the ShapeNetSem (Chang et al., 2015; Savva et al., 2015) dataset, specifically those included in the ACRONYM (Eppner et al., 2020) robot grasping dataset. The SNARE benchmark adversarially selects similar target and distractor objects to challenge 3D language grounding approaches. In the object reference task, the model is presented with a natural language description l and must correctly identify the target object o^l from a set of $m = 2$ objects $O = \{o^l, o^{c_1}\}$. Each object o

in the benchmark is accompanied by $n = 8$ image views, capturing the object from different perspectives at 45-degree intervals. As both target and distractor objects are from the same ShapeNetSem category, the SNARE benchmark aims to evaluate a model's contextual reasoning abilities.

The SNARE benchmark encompasses two types of object descriptions: **visual** and **blindfolded**. Visual descriptions are generated by annotators who are guided to include the object's name, shape, and color. These visual descriptions aim to capture a comprehensive understanding of the object, providing relevant visual cues to guide the grounding process (e.g., "the red mug"). On the other hand, blind descriptions predominantly focus on the object's shape and specific distinguishing attributes, intentionally omitting color and other visual characteristics that might aid identification (e.g., "the one with a tapered lip").

The SNARE benchmark is split into training, validation, and test splits. The train/validation/test sets are split over (207 / 7 / 48) ShapeNetSem object categories, containing (6,153 / 371 / 1,357) unique object instances and (39,104 / 2,304 / 8,751) object pairings, each accompanied by a referring expression. The validation and test sets include unseen object categories that were not encountered during the model training phase, thus evaluating the generalizability and robustness of different methods.

5.1 Models

In our evaluation, we compare the performance of MAGiC against several baselines, including the previous state-of-the-art (SOTA. We describe these baselines below:

Human accuracy serves as an upper bound for performance. These results are provided from SNARE (Thomason et al., 2021). Human performance is determined by evaluating whether the annotators can unanimously identify the corresponding object based on the provided natural language description.

MATCH (Thomason et al., 2021) uses CLIP-ViT to encode the views of each object. These encoded views are then max-pooled and concatenated to the language description embedding. Then, an MLP is trained to assign scores to each object independently based on the concatenated representation.

ViLBERT (Lu et al., 2019; Thomason et al., 2021) uses 14 views as opposed to the standard 8 views in SNARE. These images are tiled into a single image based on the camera view. ViLBERT then attends to the bounding boxes of each view to provide an image representation that is used in a MATCH model instead of the CLIP-ViT encoder.

LAGOR (Thomason et al., 2021) (Language Grounding through Object Rotation) builds upon the MATCH model. LAGOR introduces additional regularization through a view prediction loss on each view. The model is presented with only two random views of each object, and it scores each view individually for language grounding in addition to view prediction.

VLG (Corona et al., 2022) (Voxel-informed Language Grounding) uses a pretrained LegoFormer (Yagubbayli et al., 2021) model for image-to-voxel map prediction. VLG employs a factorized representation of the predicted voxel map, CLIP image embeddings, and CLIP language embeddings to score an object. By incorporating voxel-based information, the VLG baseline serves as a strong comparison against our model, which suggests an alternative pragmatic approach.

5.2 Training Details

We train MAGiC on the SNARE dataset using a smoothed binary cross-entropy loss. We adopt a similar training strategy as VLG. We train our model for 75 epochs using the AdamW optimizer. The learning rate is set to 1e-3, and we incorporate a linear learning rate warmup for the first 10,000 steps of training. Our model uses 3 transformer encoder layers, 8 attention heads, and a hidden size of 256 for a total of 3.6 million trainable parameters. We train our models with a batch size of 64.

6 Results

In this section, we present the test set performance of our model and compare it with the previous state-of-the-art models. Additionally, we report the average performance and standard deviation of our model and various ablations on the validation set, calculated over 10 different seeds.

6.1 MAGiC improves over SOTA

Table 1 presents the performance comparison of models on the SNARE benchmark. MAGiC outperforms all other models with a 2.7% absolute accu-

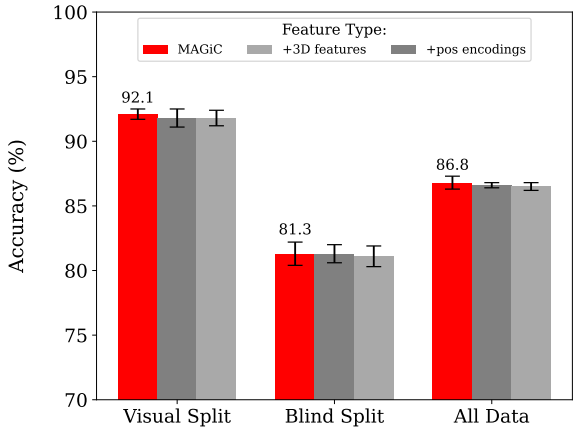

Figure 3: **Explicit 3D Features.** We find that adding 3D structural information to MAGiC does not improve accuracy.

racy improvement on the test set over VLG. In the blindfolded split, MAGiC has a 3.7% performance increase over VLG. Across the entire validation set, MAGiC is statistically significantly better in grounding accuracy than VLG with a $p < 0.001$ under a Welch's unpaired two-tailed t-test.

MATCH aggregates the CLIP embeddings using max pooling, removing its ability to effectively reason over the 3D structure of an object. VLG explicitly uses 3D features and improves 2.5% on grounding accuracy compared to MATCH. MAGiC however is able to improve performance by 5.2% over MATCH. These results suggest that our model does not explicitly require additional 3D structure like VLG.

Though VLG also uses a transformer-based architecture, VLG uses max pooling to aggregate image features before it is input into the transformer model. In the blindfolded subset, ViLBERT previously had the top performance of 73.0%, likely beating VLG since it used 14 views instead of 8 views. Although ViLBERT reasons explicitly over multi-view context rather than pooling view information like VLG and MATCH, MAGiC improves over ViLBERT by 2.4% on the blindfolded set using fewer views. This performance difference implies that by leveraging CLIP image features for each view independently, MAGiC demonstrates the ability to capture and reason about multi-view context effectively.

We believe our performance gain can also be attributed to capturing object and multi-view context. In the next subsection, we present ablations to further demonstrate this result.

6.2 Ablation Study:

We present several ablations performed on the SNARE validation split. We first investigate the precise contributions of **object** and **view** context to our method's improvement on the benchmark as shown in Figure 2. We also examine the effect of additional 3D information and varying the number of views on our method.

Context improves validation accuracy. In Table 2, we find that using context improves validation accuracy on SNARE, implying that MAGiC can capture and utilize contextual dependencies, showcasing its advantage over MLP-based architectures. To assess the significance of object context in our model, we added object context to a MATCH model and removed it from MAGiC. We find that adding object context to MATCH does not help improve performance. In contrast, removing object context from MAGiC decreases grounding accuracy by 1.5%. MAGiC without object context is similar to the ViLBERT-based MATCH model in Table 1, as both only use multi-view context. These two models have a noticeable 2.3% difference in grounding accuracy, though some of this difference could be attributed to ViLBERT's weaker representational capacity for language grounding compared to CLIP. These results suggest that MAGiC is able to effectively leverage object context.

To understand the importance of multi-view context, we remove multi-view context and only reason over object context. MAGiC without multi-view context is conceptually similar to MATCH with object context, but we find a 1.6% difference in validation performance between the models. MATCH's lower performance with multi-view context implies that MAGiC can contextually reason between objects better than MLP-based architectures.

Most notably, we find that MAGiC without multi-view and object context has 0.7% higher overall validation accuracy than MATCH, which is reasonably within error bounds. The validation performance of VLG in Table 1 also performs similarly to MAGiC without both types of context. The similarity in their performance indicates that the difference between MATCH, VLG, and MAGiC comes from MAGiC's ability to reason contextually between both views and objects.

MAGiC does not require 3D information. In Figure 3, we investigate whether 3D information is necessary to comparatively ground two objects by conducting two experiments that introduce 3D

structure explicitly: via positional encodings and explicitly adding 3D features.

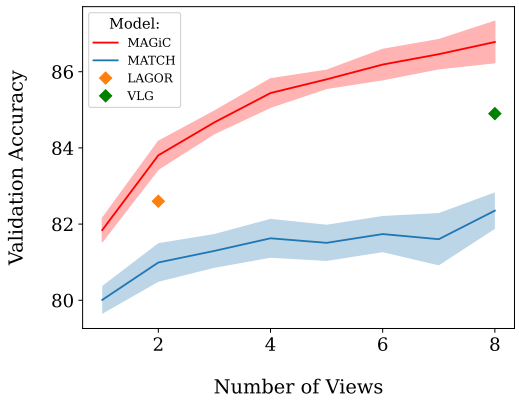

Figure 4: **Fewer Views Impact on Performance.** We report results on the validation set on the impact of fewer views on performance. We find that MAGiC outperforms MATCH, LAGOR, and VLG, achieving greater accuracy with fewer views.

Positional Encodings. Given that our model does not impose any specific ordering for the views, we rely on our model to learn the 3D structure of objects implicitly from unordered 2D-image views. To investigate the maximum potential of an image-based 3D object grounding model, we experiment with enforcing canonical image view orderings and incorporating learnable positional encodings for the views. While MAGiC handles unordered input views without relying on knowledge about camera rotations, we specifically enforce a canonical ordering scheme based on 45-degree rotations for the inputs and add learnable positional encodings for each view. If consistent view orderings and positional encodings help in learning 3D structure, we would expect improved performance. However, our findings in Figure 3 indicate that enforcing order and using positional encodings do not result in performance changes, implying that MAGiC can capture view-specific contextual relationships without explicit positional information.

3D Features. The performance gains of VLG over MATCH in Table 1 can be attributed to the addition of explicit 3D information. To assess whether our model can benefit from explicit 3D information, we investigate the impact of incorporating supplementary, view-specific 3D features into the transformer input. We use features precomputed using the Point-E (Nichol et al., 2022) transformer for each object view and language description. Point-E is a language-conditioned point cloud diffusion transformer that captures both 3D and language information through a reconstruction task, so we believe it will effectively capture relevant 3D information. View masking augmentations are applied as necessary. Also, we add token-type embeddings so the model can distinguish between the 2D image features and the 3D features. We find that the explicit inclusion of 3D features does not improve accuracy.

These results further reinforce the importance of grounding fine-grained object differences over the use of 3D information in improving comparative language grounding (as posited in prior works).

MAGiC is more robust to fewer views. Stronger performance by MAGiC on view-limited experiments compared to the previous SOTA demonstrates MAGiC's ability to handle limited visual information in language grounding tasks. By retraining MAGiC and MATCH on a reduced number of views as shown in Figure 4, we can assess a model's ability to effectively leverage limited visual information and still accurately understand and interpret natural language descriptions. We find that on the validation set, MAGiC achieves higher accuracy with fewer views compared to other models. For instance, with only 4 views, MAGiC achieves an accuracy of 85.4%, surpassing VLG, which attains 84.9% accuracy with 8 views. This suggests that MAGiC can more efficiently leverage available information from fewer views. Our findings contribute to a deeper understanding of the significance of exploiting multiple views in language-grounding tasks.

7 Discussion

In this work, we present MAGiC, which demonstrates significant improvements in language grounding accuracy on an object reference task by reasoning jointly over objects and their multi-view contexts when scoring their compatibility with referring expressions. We find that comparatively reasoning over multiple objects is central to capturing contextual relationships that enhance the model's ability to ground object descriptions, with added multi-view context also contributing to better language-to-object grounding. The experimental results from the SNARE object identification benchmark highlight the effectiveness of MAGiC, which outperforms all methods on both the validation and test sets.

8 Limitations

MAGiC heavily relies on having access to multiple views of objects. While using multiple views allows for capturing richer context and improving performance, it also requires obtaining and processing multiple images for each object, which may not always be feasible or practical in certain scenarios. Future work could consider actively selecting views that promote the most information gain. Additionally, our experiments focus on a single distractor object. We provide preliminary multiple distractor experiments in the appendix to showcase the practicality of MAGiC in the real world, which is provides a foundation for future work on comparatively reasoning over multiple objects.

Additionally, MAGiC uses CLIP embeddings for encoding visual information. While CLIP provides powerful pre-trained image and text encoders, its representations may not fully capture the intricacies and characteristics of 3D objects. This limitation could potentially impact the model's ability to discriminate between visually similar objects or capture fine-grained details crucial for accurate language grounding.

9 Potential Negative Societal Impact

MAGiC was designed to ground language to 3D household objects. However, MAGiC has direct potential uses for sensitive applications such as face identification and surveillance. For instance, law enforcement agencies may use MAGiC with vague witness testimony to discern a suspect given two sets of mugshots with multiple views. In these high-stakes applications, our model could generate harmful and discriminatory identifications that would further negatively impact historically minoritized peoples. Furthermore, our model uses a CLIP-backbone, and previous literature has shown that CLIP reinforces malignant sexist and racist stereotypes (Hundt et al., 2022) and exhibits gender bias (Wang et al., 2022; Agarwal et al., 2021) which are part of broader patterns of marginalization in society. Vision-and-language models have also been shown to compound gender biases that exist separately in language and vision (Srinivasan and Bisk, 2021). Therefore, these models must account for the ways in which language and perception reflect social norms.

References

Panos Achlioptas, Judy Fan, Robert X. D. Hawkins, Noah D. Goodman, and Leonidas J. Guibas. 2019. ShapeGlot: Learning Language for Shape Differentiation. *International Conference on Computer Vision (ICCV)*.

Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. 2023. ShapeTalk: A Language Dataset and Framework for 3D Shape Edits and Deformations. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *arXiv Preprint*.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. *ArXiv*, abs/1604.00562.

Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. 2019. Neural 3d Morphable Models: Spiral Convolutional Networks for 3d Shape Representation Learning and Generation. *International Conference on Computer Vision (ICCV)*.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. 2022. Cross-Attention of Disentangled Modalities for 3D Human Mesh Recovery with Transformers. *European Conference on Computer Vision (ECCV)*.

Rodolfo Corona, Shizhan Zhu, Dan Klein, and Trevor Darrell. 2022. Voxel-informed Language Grounding. *Association for Computational Linguistics (ACL)*.

Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Clemens Eppner, Arsalan Mousavian, and Dieter Fox. 2020. ACRONYM: A Large-Scale Grasp Dataset Based on Simulation. *International Conference on Robotics and Automation (ICRA)*.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in Grounded Language Learning: Phenomena, Tasks, and Modeling Approaches. *arXiv Preprint*.

Rohit Girdhar and Kristen Grauman. 2021. Anticipative Video Transformer. *International Conference on Computer Vision (ICCV)*.

Herbert P Grice. 1975. Logic and Conversation. *Speech acts*.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. 2022. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. *Association for Computational Linguistics (ACL)*.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep Learning for 3D Point Clouds: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *International Conference on Multimedia (MM)*.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. 2022. Multi-View Transformer for 3D Visual Grounding. *Conference Computer Vision and Pattern Recognition (CVPR)*.

Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. 2018. Cooperative Holistic Scene Understanding: Unifying 3d Object, Layout, and Camera Pose Estimation. *Conference on Neural Information Processing Systems (NeurIPS)*.

Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots Enact Malignant Stereotypes. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language Transformer Without Convolution or Region Supervision. *International Conference on Machine Learning (ICML)*.

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J Guibas, and Minhyuk Sung. 2022. PartGlot: Learning shape part segmentation from language reference games. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded Language-Image Pre-training. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. Mesh Graphormer. *International Conference on Computer Vision (ICCV)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Conference on Neural Information Processing Systems (NeurIPS)*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *European Conference on Computer Vision (ECCV)*.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv Preprint*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Conference on Computer Vision (ICCV)*.

Christopher Potts. 2022. Pragmatics. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. Pointnet: Deep Learning on Point Sets for 3d Classification and Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning (ICML)*.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2019. Zero-Shot Grounding of Objects From Natural Language Queries. *International Conference on Computer Vision (ICCV)*.

Manolis Savva, Angel X Chang, and Pat Hanrahan. 2015. Semantically-enriched 3D Models for Common-sense Knowledge. *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. *Conference on Robot Learning (CoRL)*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tejas Srinivasan and Yonatan Bisk. 2021. Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems*.

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. 2021. Language Grounding with 3D Objects. *Conference on Robot Learning (CoRL)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Conference on Neural Information Processing Systems (NeurIPS)*.

Junyang Wang, Yi Zhang, and Jitao Sang. 2022. FairCLIP: Social Bias Elimination based on Attribute Prototype Learning and Representation Neutralization. *arXiv Preprint*.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2021a. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue Wang, Vitor Campanholo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. 2021b. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. *ArXiv*, abs/2110.06922.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Farid Yagubbayli, Alessio Tonioni, and Federico Tombari. 2021. LegoFormer: Transformers for Block-by-Block Multi-view 3D Reconstruction.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. *Conference on Computer Vision and Pattern Recognition (CPVR)*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. *European Conference on Computer Vision (ECCV)*.

Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. 2017. DeepContext: Context-encoding Neural Pathways for 3d Holistic Scene Understanding. *International Conference on Computer Vision (ICCV)*.

A Appendix

In this supplementary section, we describe additional experiments, ablations, and results related to our work.

A.1 Masking Ablations

As discussed in Section 4, in order to improve the robustness and generalization capabilities of MAGiC, we employ masking augmentations on both the language embeddings and the view embeddings as regularization for our model. Specifically, we applied random masking to a certain percentage of the language and view embeddings during training, analyzing the impact of different masking percentages as depicted in Figure 5. Through hyperparameter tuning on the validation set, we determined that a 20% language masking and a 10% view masking yield language grounding accuracy improvements. We ran each model for 10 seeds. However, we also noticed that excessive regularization can have a detrimental effect on accuracy, highlighting the need for a balanced application of masking augmentations.

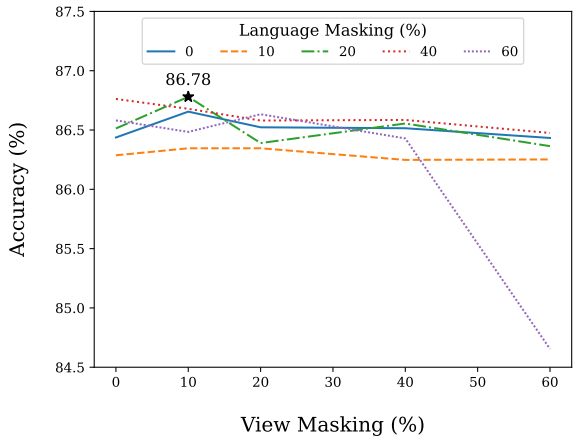

Figure 5: **View and language masking.** We show the impact of different attention masking percentages for the view and language tokens that are input into MAGiC. Each variant is trained for 10 seeds. We find that 10% view masking and 20% language masking achieved the highest validation set accuracy.

A.2 Contrastive Loss

We investigated additional regularization by using CLIP-like contrastive losses on the output representations. Losses that are similar in spirit have been used in face recognition and clustering research (Schroff et al., 2015) as well as multi-modal sentiment analysis research (Hazarika et al., 2020). At a high-level, we implement a contrastive loss that motivates the embedded target-object image features to be similar to the embedded object description language features. Our model does not have any supervision on the output language representations, and thus, we hypothesized that a contrastive loss would have led to a more structured embedding space. Additionally, we expected that the additional supervision from the contrastive loss on the output embeddings from the language inputs would help improve grounding accuracy. However, we did not find any improvements on MAGiC's accuracy on the validation set as shown in Figure 6. These findings indicate that the transformer model was already able to contrastively structure the embedding space given access to both objects and the language description such that the added contrastive loss was not further advantageous towards that goal.

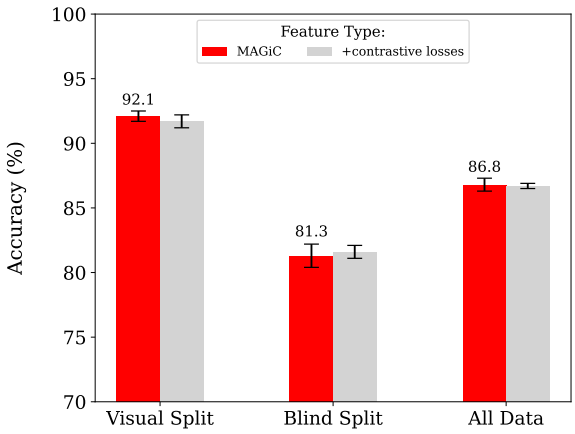

Figure 6: **Contrastive Loss.** We train MAGiC on 10 seeds on the validation set with and without contrastive losses. We show that there is no noticeable impact on MAGiC's validation accuracy when a contrastive loss is added during training.

A.3 Additional Discussion

We would also like to note that our model outperformed another model on the SNARE leaderboard called LOCKET. However, we were unable to find any code or paper publicly associated with LOCKET at the time of submission, and omitted it from Table 1.

The code for MAGiC will be made public after anonymity restrictions are lifted.

A.4 Multiple Distractor Experiment

While the SNARE benchmark presents the model with one target and one distractor object, we

demonstrate MAGiC's ability to generalize to multiple distractors, as may be the case in a more realistic use case. SNARE provides adversarially-selected pairs with language annotations. To have multiple distractor objects, we randomly select an object in the same train/val set. There is no guarantee for new distractor objects will be in the same category of as the initial two objects since the language might not differentiate additional objects of the same category.

Our results in Figure 7 show that while overall performance decreases, MAGiC generally retains its strong performance over an architecture without object context. MAGiC without object context is similar to the MLP-based MATCH model, as they score each object individually. We find that reasoning over all objects generally outperforms scoring the objects individually. We note that performance clearly degrades as more objects are added, and we show a line depicting random chance to show that our model has generally high performance. Due to SNARE being a dataset for loading 2 objects at a time, implementation constraints limited us from scaling up these experiments efficiently. Thus each variant is trained only on 1 seed, which makes it clear that this result becomes noisy as more distractor objects are added. We also note that for MAGiC, for additional distractor objects, MAGiC is trained and evaluated on the same number of distractor objects.

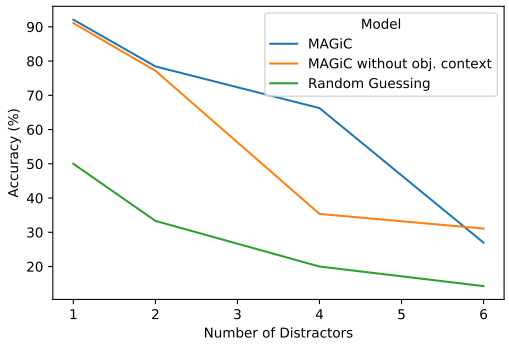

Figure 7: **Number of distractors.** We show the impact of different numbers of distractors on the performance of MAGiC and MAGiC without context. Each variant is trained for 1 seed. We find that MAGiC