Imagining Grounded Conceptual Representations from Perceptual Information in Situated Guessing Games

Alessandro Suglia[1], **Antonio Vergari**[2], **Ioannis Konstas**[1], **Yonatan Bisk**[3], **Emanuele Bastianelli**[1], **Andrea Vanzo**[1], and **Oliver Lemon**[1]
[1]Heriot-Watt University, Edinburgh, UK
[2]University of California, Los Angeles, USA
[3]Carnegie Mellon University, Pittsburgh, USA
[1]{as247,i.konstas,a.vanzo,e.bastianelli,o.lemon}@hw.ac.uk
[2]aver@cs.ucla.edu, [3]ybisk@cs.cmu.edu

Abstract

In visual guessing games, a Guesser has to identify a target object in a scene by asking questions to an Oracle. An effective strategy for the players is to learn conceptual representations of objects that are both discriminative and expressive enough to ask questions and guess correctly. However, as shown by Suglia et al. (2020), existing models fail to learn truly multi-modal representations, relying instead on gold category labels for objects in the scene both at training and inference time. This provides an unnatural performance advantage when categories at inference time match those at training time, and it causes models to fail in more realistic "zero-shot" scenarios where out-of-domain object categories are involved. To overcome this issue, we introduce a novel "imagination" module based on Regularized Auto-Encoders, that learns context-aware and category-aware latent embeddings without relying on category labels at inference time. Our imagination module outperforms state-of-the-art competitors by 8.26% gameplay accuracy in the CompGuessWhat?! zero-shot scenario (Suglia et al., 2020), and it improves the Oracle and Guesser accuracy by 2.08% and 12.86% in the *GuessWhat?!* benchmark, when no gold categories are available at inference time. The imagination module also boosts reasoning about object properties and attributes.

1 Introduction

Humans do not learn conceptual representations from language alone, but from a wide range of situational information (Beinborn et al., 2018; Bisk et al., 2020) as highlighted also by property-listing experiments (McRae et al., 2005). When humans experience the concept of "boat", they *simulate* a new representation by reactivating and aggregating *multi-modal* representations that reside in their memory and are associated with the concept of "boat" (e.g., what a boat looks like, the action of sailing, etc) (Barsalou, 2008). This simulation process is called *perceptual simulation*. Therefore, it is no wonder that recent trends in learning conceptual representations adopt multi-modal and holistic approaches (Bruni et al., 2014) wherein abstract distributional lexical representations (Landauer and Dumais, 1997; Laurence and Margolis, 1999) learned from text corpora are augmented or refined with *perceptual information* for concrete and context-aware representations built from visual (Kiela et al., 2018; Lazaridou et al., 2015), olfactory (Kiela et al., 2015), or auditory (Kiela and Clark, 2015) modalities.

Language games between AI agents, inspired by Wittgenstein's Language Games among humans (Wittgenstein et al., 1953), are an excellent test bed for such approaches since concepts are expected to emerge when agents are required to communicate to solve specific tasks in specific environments. *GuessWhat?!* (De Vries et al., 2017) is a prototypical language game of this kind: a Guesser has to identify a target object in a scene represented as an image by asking questions to an Oracle. Learning to ground pixels of the scene into object representations that are relevant for the object category they belong to (*category-aware*), but are also particularized for the specific scene (*context-aware*), is fundamental for the Guesser to effectively converse with the Oracle and vice-versa.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1090–1102
Barcelona, Spain (Online), December 8-13, 2020

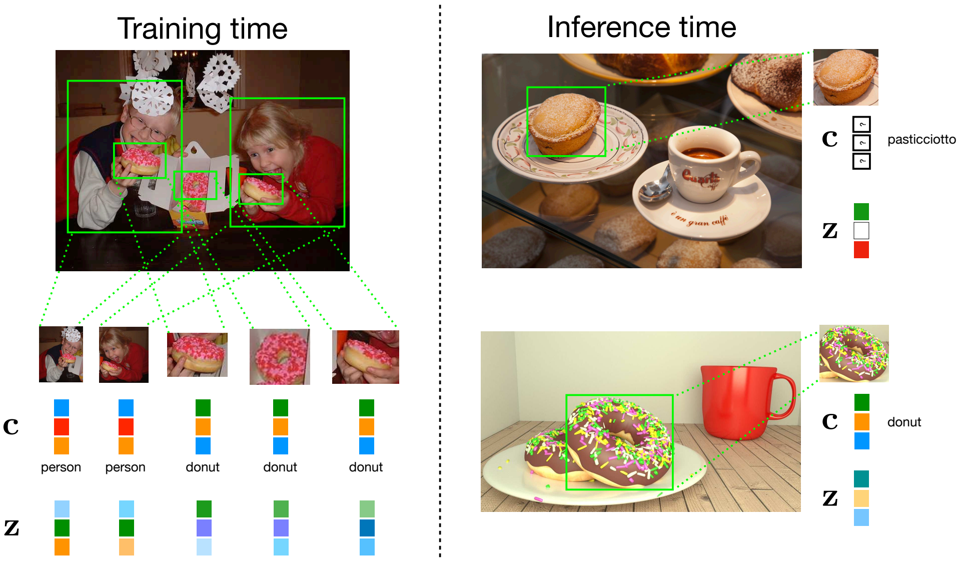


Figure 1: Common approaches to visual grounding such as De Vries et al. (2017) and Zhuang et al. (2018) rely on gold category labels at test time, thereby failing to ground novel objects from categories not seen during training (e.g., a "pasticciotto", top right) or to properly encode known categories but with unseen visual features (like a "frosted donut", bottom right) since they employ *category embeddings* $\mathbf{c}$ from a predefined set that are fixed for each object. Instead, embeddings $\mathbf{z}$ learned by our imagination module can be flexibly category-aware allowing them to generalize to unseen categories.

We consider a model truly *multi-modal* if it always uses all the modalities to make decisions. However, existing approaches (De Vries et al., 2017; Shekhar et al., 2019) rely instead on *gold category labels* that are assumed to be available also at inference time, thus making these models depend on this modality and discarding the others. This not only poses an unnatural performance advantage for players in controlled benchmark scenarios like the *GuessWhat?!* game when categories at inference time match those at training time, but causes them to fail in more realistic zero-shot scenarios (Suglia et al., 2020) where players are required to generalize to out-of-domain object categories. For example, consider an agent that during training has only seen glazed donuts, associated with the fixed "donut" category embedding (cf. Figure 1). At inference time, the model cannot ground visual representations for objects belonging to the "pasticciotto" (an Italian pastry) category, since such a category was not in its repertoire. Similarly, it will likely represent frosted donuts with a generic "donut" embedding, despite the perceptual differences among different types of donut.

In this paper, we tackle the above limitations by introducing a novel *imagination module* based on Regularized Auto-encoders (Ghosh et al., 2019), which are able to derive *imagination embeddings* directly from perceptual information in the form of the object crop. Our formulation of the reconstruction loss allows the model to learn *context-aware* and *category-aware* imagination embeddings. Thus, removing the need for gold category labels at inference time and greatly improving zero-shot generalization. Section 4.2 integrates our imagination component into the Oracle model of De Vries et al. (2017) and the Guesser model of Shekhar et al. (2019). We show that the new imagination models are state-of-the-art in the recently introduced *CompGuessWhat?!* benchmark (Suglia et al., 2020) outperforming current models by 8.26%. It also improves the Oracle's and Guesser's accuracy (by 2.08% and 12.86%, respectively) in the standard *GuessWhat?!* when no gold category labels are available. Lastly, we show that imagining latent object representations greatly helps to reason about object visual properties (i.e., color, shape, etc.), qualifying our module as a generic *perceptual simulation* component alà Barsalou (2008).

2 Background: Guessing Games and Concept Representations

GuessWhat?! is an instance of a multi-word guessing game (Steels, 2015). Every game involves two players: an *Oracle* and a *Guesser* conversing about a *scene* $\mathcal{S}$ (a natural image). A scene $\mathcal{S}$ can be abstracted into a collection of *objects* $\mathcal{O}$, each of which is associated with a category $c_i \in \mathcal{C}, i = \{1, \ldots, K\}$. The aim of the Guesser is to identify a *target object* $o^* \in \mathcal{O}$ by asking questions about $\mathcal{S}$ to the Oracle. The gameplay of *GuessWhat?!* thus comprises three tasks: i) *question generation* where the Guesser inquires about an object in the scene $\mathcal{S}$ given the dialogue generated so far; ii) *answer prediction*, where the Oracle answers $a \in \mathcal{A} = \{\mathsf{Yes}, \mathsf{No}, \mathsf{N/A}\}$ given the scene $\mathcal{S}$, question and the target object o^*; and iii) *target prediction* where the Guesser selects a candidate object with the highest relevance score $r(o_i)$.

Several architectural variants have been proposed to tackle *GuessWhat?!* (cf. Section 5 for some related works). In this work we adopt the recent GDSE model (Shekhar et al., 2019), which learns a visually grounded dialogue state used to learn both question generation and target object prediction. As shown below, GDSE does not deliver the desired multi-modality needed, therefore we extend it with our Imagination component to obtain more effective multi-modal object representations.

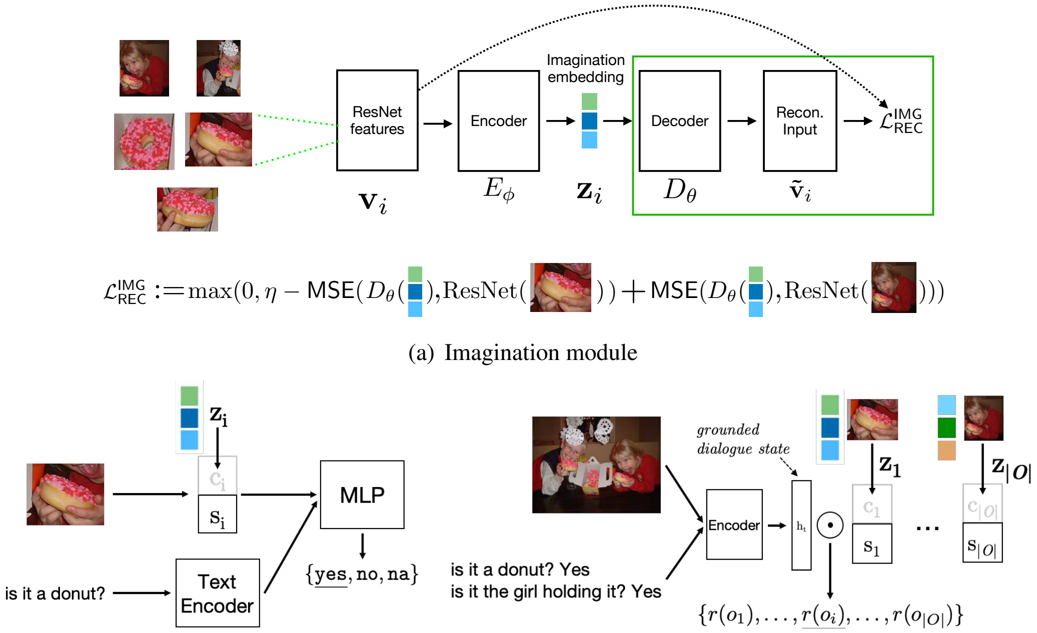
For successful gameplay, both the Guesser and Oracle must build representations of the scene that contain specific perceptual information of objects (object-aware), are relevant for the object category they belong to (category-aware), and are specialized to the scene in which the game is played (context-aware). As the scene $\mathcal{S}$ is an image, it is natural to associate each object $o_i \in \mathcal{O}$ with a *perceptual embedding*, i.e., a vector $\mathbf{v}_i \in \mathbb{R}^{d_\mathcal{O}}$ extracted from the penultimate layer of a pretrained vision model (e.g. ResNet-152 (Shekhar et al., 2019)) based on their bounding box.[1]

However, these representations are not sufficient as they are neither *context-aware* nor *category-aware*, i.e., they ignore other objects in the scene and do not leverage their category information. GDSE and other recent approaches (De Vries et al., 2017; Shekhar et al., 2019; Zhuang et al., 2018; Shukla et al., 2019) coped with the second issue by introducing *category embeddings* as $d_\mathcal{C}$-dimensional continuous representations $\mathbf{c}_k \in \mathbb{R}^{d_\mathcal{C}}$ for $k = 1, \ldots, K$. Once learned, a category embedding $\mathbf{c}$ is then concatenated to an 8-dimensional feature vector $\mathbf{s}_i$ derived from the object bounding box (cf. De Vries et al. (2017)). While these embeddings partially solve category-awareness, they are *not object-aware*. For instance, the embedding for the object category "apple" will be the same regardless of a particular object to be a red or green apple, i.e., most likely a centroid representation of the objects seen only during training. Moreover, if during training we only see red apples, at inference time, we will likely fail to detect green apples as belonging to the same category (Figure 2(a)). These issues have gone unnoticed since category embeddings usually boost performances on the original *GuessWhat?!* task, given that gold category labels are also available at inference time. However, this boost is illusory: models relying on this symbolic information to be always available are not learning to exploit all modalities. In fact, a 20% drop in the Guesser accuracy if gold category labels are not provided has been reported in Zhuang et al. (2018) for *GuessWhat?!* and analogous poor results in more realistic benchmarks measuring zero-shot generalization such as *CompGuessWhat?!* (Suglia et al., 2020).

3 Imagination Module: Learning Context- and Category-aware Object Representations

To overcome the limitations of GDSE and competitors and realize a form of *perceptual simulation* in a learning system, we introduce a generic component—named the *imagination module*—which learns latent concept representations that are both context- and category-aware, without relying on category labels at inference time. Our imagination model can be understood in the context of representation learning via deep generative models (Bengio et al., 2013) which has been popularized by variational autoencoders (VAEs) (Kingma and Welling, 2013; Kingma et al., 2014), and GANs (Goodfellow et al., 2014). Specifically, we substantially extend the recently introduced regularized autoencoders (RAEs) framework (Ghosh et al., 2019). RAEs are simplified VAEs where stochasticity in the encoder and decoder is dropped in favor of more stable training and more informative embedding learning. In fact, RAEs do not suffer from several issues known to affect VAEs, such as poor convergence and the possibility of learning embeddings that are independent of the input images (cf. Ghosh et al. (2019) for a detailed discussion). More crucially for our purposes, RAEs do not have to compromise the informativeness of the learned embeddings with a fixed a-priori structure in the latent space that enables simple

[1] Bounding boxes are assumed to be given, e.g. by using object recognition as a pre-processing step (Anderson et al., 2018).



(a) Imagination module

(b) Oracle model

(c) Guesser model

Figure 2: Imagination-based Representation Learning: Given the perceptual information $\mathbf{v}_i$ of object o_i, we learn an imagination embedding $\mathbf{z}_i$ generated by Encoder E_ϕ. The latent code is optimized to reconstruct the original visual representation $\mathbf{v}_i$ (the "donut" ResNet encoding) via the reconstruction loss $\mathcal{L}^{\text{IMG}}_{\text{REC}}$ using the Decoder D_θ. Figures 2(b) and 2(c) show how the imagination embedding $\mathbf{z}$ replaces the category embedding $\mathbf{c}$ in the Oracle model from De Vries et al. (2017) and Guesser model from Shekhar et al. (2019) respectively, and is concatenated to the spatial information $\mathbf{s}_i$.

sampling (e.g., an isotropic Gaussian prior). VAEs which need to have such a fixed prior, instead, are deemed to learn embeddings that are less informative w.r.t. objects, categories, and context information.

Module architecture. Figure 2(a) summarizes our imagination module. Its aim is to distill a context and category-aware embedding $\mathbf{z}_i \in \mathbb{R}^{dz}$ per object o_i in scene $\mathcal{S}$. To this end, we adopt an encoder E_ϕ parameterized by ϕ that maps a perceptual embedding $\mathbf{v}_i$ of object o_i to its imagined counterpart $\mathbf{z}_i$, i.e., $E_\phi(\mathbf{v}_i) = \mathbf{z}_i$. A decoder D_θ realizes the inverse mapping $\tilde{\mathbf{v}}_i = D_\theta(\mathbf{z}_i)$, with $\tilde{\mathbf{v}}_i \in \mathbb{R}^{d\mathcal{O}}$ being also called the reconstruction of the input $\mathbf{v}_i$. As in RAEs, our per-object loss $\mathcal{L}_{\text{IMG}}$ comprises a reconstruction loss ($\mathcal{L}_{\text{REC}}$), weighting how good the reconstructions of D_θ are w.r.t. the encoded representations by E_ϕ, and a regularization term ($\mathcal{L}_{\text{REG}}$) enhancing generalization by smoothing the decoder D_θ. This leads to the following composite loss:

$$\mathcal{L}_{\text{IMG}} = \mathcal{L}_{\text{REC}} + \alpha \mathcal{L}_{\text{REG}}, \tag{1}$$

where α is an hyperparameter controlling regularization.[2] As in L_2-RAE (Ghosh et al., 2019), the regularization component is defined as $\mathcal{L}_{\text{REG}} := ||\mathbf{z}_i|| + ||\theta||_2$: the first term bounds the latent embedding space learned by E_ϕ easing optimization; the second enforces smoothing over D_θ improving generalization over regions of the latent space that are unseen during training.

Differently from RAEs, we devise a specific reconstruction loss tailored to learn contextual and category-aware representations. In conventional RAEs, in fact, the reconstruction loss is defined as the Mean Squared Error (MSE) representing the distance between $\mathbf{v}_i$ and its reconstruction $\tilde{\mathbf{v}}_i$, so that $\mathcal{L}^{\text{RAE}}_{\text{REC}} := \text{MSE}(\mathbf{v}_i, \tilde{\mathbf{v}}_i)$. This loss is purely unsupervised and as such agnostic to object categories or to the scene context. To our aims, we define a custom *imagination reconstruction loss* $\mathcal{L}^{\text{IMG}}_{\text{REC}}$ as an instance of a max-margin triplet-loss (Wang et al., 2014; Schroff et al., 2015), as follows. Let c_i be the category

[2] Ghosh et al. (2019) use two different hyperparameters for the two terms in $\mathcal{L}_{\text{REG}}$ Optimizing them independently had no evident benefit in our experiments, hence we simply treat them as a single regularizer together.

of object o_i with perceptual embedding $\mathbf{v}_i$ in scene $\mathcal{S}$ and let $\mathcal{O}_{\neg c_i} = \{o_j \mid o_j \in \mathcal{O} \wedge c_j \neq c_i\}$ be the set of all objects in $\mathcal{S}$ belonging to a different category than c_i. Our per-object $\mathcal{L}^{\mathsf{IMG}}_{\mathsf{REC}}$ term is defined as:

$$\mathcal{L}^{\mathsf{IMG}}_{\mathsf{REC}} := \max(0, \eta - \mathsf{MSE}(\mathbf{v}_i, D_\theta(\mathbf{z}_i)) + \mathsf{MSE}(\mathbf{v}_j, D_\theta(\mathbf{z}_i))), \tag{2}$$

where η is the minimum margin between two components: *i*) the distance between the perceptual embedding $\mathbf{v}_i$ and its reconstruction $D_\theta(\mathbf{z}_i)$, and *ii*) the distance between the perceptual embedding $\mathbf{v}_j$ of a randomly sampled object $o_j \in \mathcal{O}_{\neg c_i}$ and the reconstruction $D_\theta(\mathbf{z}_i)$. By doing so, we enforce each object representation to be *representative of its category given a specific context* by locally contrasting it to another object of a different category in the same scene. Note that this is strikingly different from previous approaches employing a max-margin loss (Elliott and Kádár, 2017; Kiros et al., 2018) where "negative" objects are arbitrarily sampled from other scenes in the same batch.

Imagining at inference time. Differently from the category embeddings $\mathbf{c}$ employed by all previous work, our imagination embeddings $\mathbf{z}$ *do not depend on gold category labels at inference time*, while still being context-aware and category-aware. In fact, once parameters ϕ have been learned, the encoder E_ϕ contains all the information needed to distill embeddings $\mathbf{z}$ independently of $\mathcal{L}_{\mathsf{IMG}}$, which is necessary *only* at training time. We consider *imagination* the ability of the model of generating latent representations on-the-fly. Therefore, for both Guesser and Oracle models we consider an object representation for object o_i that replaces $\mathbf{c}_i$ with $\mathbf{z}_i$ and concatenates it with its spatial information $\mathbf{s}_i$ (see Figures 2(b) and 2(c) and Appendix A.1 for details). By doing so, we consider every gameplay situated in a reference scene as an experience where our imagination module is able to derive a latent conceptual representation simply by "looking" at objects, realizing a *perceptual simulator* (Barsalou, 2008). We plan to investigate how to combine label-dependent category embeddings $\mathbf{c}$ with our imagination embeddings $\mathbf{z}$, similarly to how some VAE variants tackle semi-supervised classification scenarios (Kingma et al., 2014).

4 Experimental Investigation

To assess the impact of using the imagination embeddings against the category embeddings, we use two evaluation benchmarks: *GuessWhat?!* and *CompGuessWhat?!*. More information about the training procedure can be found in Appendix A.2.

4.1 *GuessWhat?!* Evaluation

In this experiment, we evaluate the accuracy of the Oracle in answering questions and the accuracy of the Guesser in selecting the target object. We consider as both training and evaluation data all the gold dialogues (and questions) that have been labeled as successful in the dataset (De Vries et al., 2017). We want to highlight that in this evaluation phase, the models using label-aware object encodings have gold information both at training and test time. This is true both for the Oracle and Guesser models. However, this does not hold for all other models using the imagination component.

4.1.1 Experimental Setup

Oracle task. We evaluate the imagination-based Oracle and compare it to several combinations of the following baselines with and without category embeddings from De Vries et al. (2017): 1) MAJORITY: majority classifier; 2) QUESTION: uses only the question; 3) IMAGE: uses only the image representation; 4) CROP: uses only the crop representation of the target object.

Guesser task. Similarly, we compare the GDSE model using imagination embeddings (GDSE+IMAGINATION) with the following *label-aware* baselines: 1) text-only baselines using LSTM encoder (LSTM) and Hierarchical Recurrent Encoder-Decoder architecture (Serban et al., 2017) (HRED) as well as their corresponding multi-modal models LSTM+IMAGE and HRED+IMAGE; 2) PARALLELATTENTION (Zhuang et al., 2018) and GDSE (Shekhar et al., 2019). We also compare with variants of the above that do not use any category embeddings or gold category labels (*-NOCAT), as well as models with predicted category labels (*-PREDCAT).[3]

[3] We train an object classifier using as input the ResNet-101 features generated for the object crop. It achieves 65% accuracy evaluated on *all* objects in the *GuessWhat?!* test set.

	Perceptual Information					Categorical Information	
Model (DV-Ques+Spatial)	Location	Shape	Color	Texture	Size	Super Category	Object
+ Crop	66.86%	69.08%	67.25%	68.30%	**65.09%**	88.94%	80.48%
+ Category	67.48%	68.42%	61.83%	**70.08%**	60.14%	**97.09%**	**88.82%**
+ Category + Crop	65.27%	60.34%	59.14%	65.76%	59.08%	96.19%	86.32%
+ Imagination	**68.62%**	**69.08%**	**67.64%**	69.86%	62.65%	90.05%	82.32%

Table 2: Oracle accuracy grouped by question type for the best Oracle model with category information (DV-Ques+Spatial) and for multi-modal variants using either perceptual or categorical information.

4.1.2 Results

Oracle task. In Table 1, we divide configurations into *category-aware* (De Vries et al., 2017) and *multi-modal.* The model reference for several other publications on *GuessWhat?!* is a *category-aware* model Question+Spatial+Category. However, by relying on symbolic information in the form of category labels, it is inevitably not truly multimodal anymore because the heavy-lifting is done by these embeddings. As shown in the results, other multi-modal models such as Question+Spatial+Crop and Question+Crop, are not able to learn effective representations to bridge the gap between category-aware and category-free models. On the other hand, the proposed imagination model is able to reduce this gap without relying on gold information as input. Indeed, we are able to learn category-aware and context-aware latent codes by using category information only in our loss function.

	Model	Val	Test
BASE	Majority	53.80%	49.10%
	Ques	58.30%	58.80%
	Img	53.30%	53.30%
	Crop	57.30%	57.00%
W/ CAT	DV-Ques+Cat	74.20%	74.30%
	DV-Ques+Crop+Cat	75.60%	75.30%
	DV-Ques+Spatial+Cat	**78.90%**	**78.50%**
	DV-Ques+Spatial+Crop+Cat	78.30%	77.90%
	DV-Ques+Spatial+Img+Cat	76.80%	76.50%
MM	DV-Ques+Crop	70.90%	70.80%
	DV-Ques+Img	59.80%	60.20%
	DV-Ques+Spatial	68.80%	68.70%
	DV-Ques+Spatial+Crop	74.00%	73.80%
	DV-Ques+Spatial+Crop+Img	72.30%	72.10%
	Imagination	**75.78%**	**75.88%**

Table 1: Oracle results on gold questions: we compare the Imagination Oracle model to models from De Vries et al. (2017) (DV-*). We group them into models relying on gold category labels (**W/ CAT**) and models that only use multi-modal perceptual information (**MM**).

We investigate this argument further by using a rule-based question classifier (Shekhar et al., 2019) to partition the test questions according to their type. Table 2 summarizes this analysis; we include models considered truly multi-modal and the best Oracle model Question+Spatial+Category. The latter can answer with high accuracy questions about specific object instances (e.g., "is it the dog?") or super-categories (e.g., "is it an animal?") since it is using category embeddings as input. However, when it comes to answering questions about perceptual properties of the target object, it loses some accuracy points because the perceptual information is missing from the category embedding representing a centroid of typical instances seen at training time only. On the other hand, the Imagination model is able to bring improvements of 1.34%, 5.81%, and 2.52% for location, color, and shape questions, respectively. On questions related to perceptual information, models using crop information seem to be on par with the Imagination model. However, our model is able to obtain an improvement over +CROP of 1.84% in object questions and of 1.11% on super category questions solely by relying on the imagination embeddings.

Guesser task. Table 3 compares several category-aware and multi-modal models; ParallelAttention and GDSE-SL are the two best performing configurations. However, when ParallelAttention does not have access to category information (ParallelAttention-nocat) its performance drops by 3.7% (also noted by Zhuang et al. (2018)). We confirmed the same behavior for GDSE-SL as well (GDSE-SL-nocat), noticing a more significant drop in performance of 16.95% which is in line with the simpler LSTM+Image model. On the other hand, GDSE-SL with our imagination component (GDSE-SL+Imagination), performs comparably with the category-aware model and better then all

	Gameplay	*Attribute Prediction*				*Zero-shot Gameplay*		
	ACCURACY	**A-F1**	**S-F1**	**AS-F1**	**L-F1**	**ND-ACC**	**OD-ACC**	**GROLLA**
RANDOM	15.81%	15.1	0.1	7.8	2.8	16.9%	18.6%	13.3
DEVRIES-SL	41.5%	46.8	39.1	48.5	42.7	31.3%	28.4%	38.5
DEVRIES-RL	53.5%	45.2	38.9	47.2	42.5	43.9%	38.7%	46.2
GDSE-SL	49.1%	**59.9**	47.6	**60.1**	48.3	29.8%	22.3%	43.0
GDSE-CL	**59.8**%	59.5	47.6	59.8	48.1	43.4%	29.8%	50.1
GDSE-SL+IMAGINATION	43.82%	56.23	47.37	57.2	**51.73**	39.19%	39.90%	45.50
GDSE-CL+IMAGINATION	51.98%	57.59	*47.6*	58.31	50.42	**46.56**%	**46.96**%	**50.74**

Table 4: Results for the *CompGuessWhat?!* benchmark (Suglia et al., 2020). We assess model quality in terms of *gameplay* accuracy, *attribute prediction* quality, measured in terms of F1 for the *abstract* (A-F1), *situated* (S-F1), *abstract+situated* (AS-F1) and *location* (L-F1) prediction scenario, as well as *zero-shot learning gameplay*. GROLLA is a macro-average of the individual scores.

multi-modal models. Therefore we argue that it is possible to learn object representations that, given a representation for the current dialogue state, allow for discriminating the target object among other candidates *without* relying on symbolic information.

4.2 *CompGuessWhat?!* Evaluation

CompGuessWhat?! is a benchmark proposed to assess the quality of models' representations and out-of-domain generalization. It includes the following tasks: a) *in-domain gameplay accuracy*, – selecting the target object with model generated dialogues as input, b) *attribute prediction task* – assessing the ability of the dialogue representation to recover target object attributes, and c) *zero-shot gameplay accuracy* – selecting the target object among objects belonging to categories never seen by the model during training. In contrast to *GuessWhat?!*, the attribute prediction and zero-shot tasks give us more insights about the quality of the learned representations and the model's generalization ability.

	MODEL	**VAL**	**TEST**
	HUMAN	90.80%	90.80%
	RANDOM	17.10%	17.10%
CATEGORY	LSTM	62.10%	61.30%
	HRED	61.80%	61.00%
	LSTM+IMAGE	61.50%	60.50%
	HRED+IMAGE	61.60%	60.40%
	PARALLELATTENTION	63.80%	**63.40%**
	GDSE-SL	63.14%	62.96%
	GDSE-SL-PREDCAT	52.08%	51.00%
MM	LSTM+IMAGE-NOCAT	50.10%	48.60%
	PARALLELATTENTION-NOCAT	55.70%	**59.70%**
	GDSE-SL-NOCAT	46.11%	46.01%
	GDSE-SL-IMAGINATION	**59.54%**	58.90%

Table 3: Guesser accuracy on successful gold dialogues: we compare GDSE-SL-IMAGINATION with i) models that are truly multi-modal (**MM**) and ii) use category information (**CATEGORY**).

4.2.1 Experimental Setup

We compare imagination-based models with baselines used in Suglia et al. (2020): 1) RANDOM: randomly selects an object; 2) DEVRIES-SL: presented in De Vries et al. (2017) trained using Supervised Learning; 3) DEVRIES-RL: DEVRIES-SL with Questioner fine-tuned using Reinforcement Learning (Strub et al., 2017); and where 4) GDSE-SL and 5) GDSE-CL are the same as used in Section 4.1.

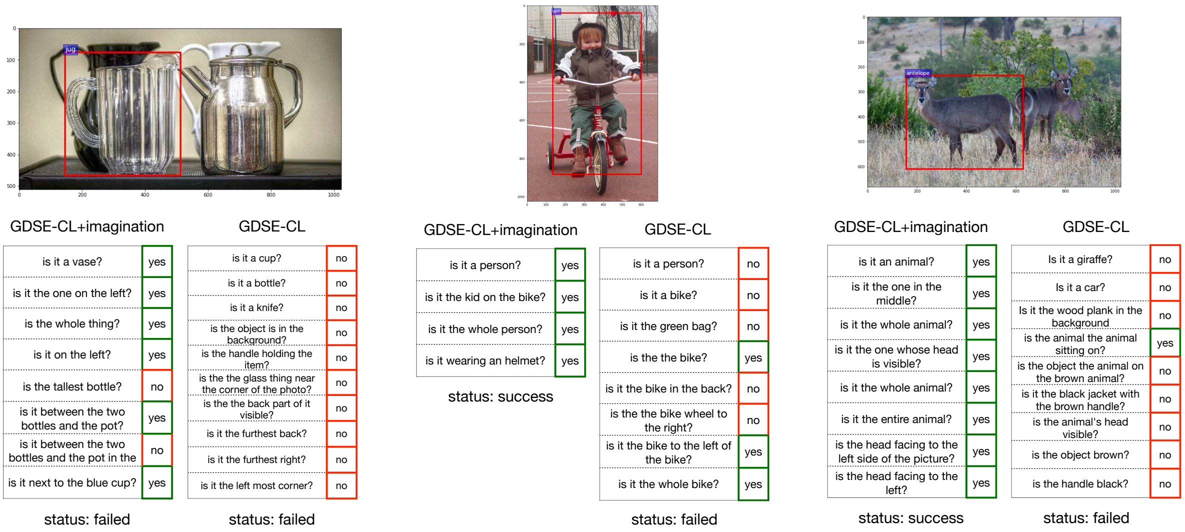
4.2.2 Results

In-domain gameplay. Table 4 presents the results on the *CompGuessWhat?!* benchmark. Models are tasked to play the game by generating up to 10 questions and corresponding answers. Firstly, we note that the results for GDSE-CL+IMAGINATION—the collaborative version of the model with Imagination—is still in the same ballpark of more complex models, such as DEVRIES-RL that is using category embeddings as input. At the same time, we notice that overall both imagination models perform worse than the GDSE-* models. We impute this drop to the introduction of additional loss terms that probably have changed the training dynamic of a cumbersome modulo-n multi-task training (Shekhar et al., 2019). This downside calls for a more principled way of handling tasks of different complexity (i.e., question generation and target prediction) in a multi-task learning system; we leave this for future work.

Attribute prediction. Table 4 reports the attribute prediction task results. In this scenario, we underline the fact that the dialogue state representation generated by the Guesser model is used to recover several types of attributes associated with the target object. In this work, we use the same dialogue state representation as used by Shekhar et al. (2019) and only focus on improving the object representations using the imagination component. Indeed, the best imagination model GDSE-SL+IMAGINATION is in line with GDSE-SL, currently the best model in terms of attribute prediction. In particular, even though the dialogue state representation is only indirectly affected by the imagination embeddings (via a dot-product operation to score the candidate objects), we can still see an improvement in terms of F1 for *Location* attributes (L-F1) and similar performance for *Situated* attribute prediction (S-F1). Both can be considered, to some extent, a result of better situated object representations.

Zero-shot gameplay. As underlined in Section 3, the imagination module's main strength is to be able to distill imagination embeddings from perceptual information only, without relying on externally provided category labels. The zero-shot gameplay scenario from *CompGuessWhat?!* (Table 4) sheds some light on the ability of the model to generalize to out-of-distribution examples. In the out-of-domain gameplay scenario where candidate objects belonging to categories never seen before are present, both imagination-based models GDSE-SL+IMAGINATION and GDSE-CL+IMAGINATION outperform the previous best performing system DEVRIES-RL by 1.2% and 8.26%, respectively in terms of OD accuracy (OD-ACC). By analyzing their output, we notice that the best imagination model achieves higher accuracy by learning a better gameplay strategy involving half the amount of location questions generated by DEVRIES-RL (39.68% vs 75.84%; see Appendix A.3 for more details). A further improvement in the near-domain scenario (ND-ACC) confirms the effectiveness of the imagination component to generate category embeddings for objects on-the-fly using only perceptual information.

Out-of-domain error analysis. Lastly, we report an error analysis comprising 50 dialogues selected at random from out-of-domain games (for more details refer to Appendix A.3). First, we manually annotated the Oracle answers and partitioned them according to their type using the same question classifier used for the Oracle Task (Section 4.1.2). 83% of super-category questions (from a total of 80) were correctly answered by the model and 63.36% color related questions (from a total of 88) were correctly answered. For instance, as shown in Figure 3, GDSE-CL is not able to answer correctly the question "is it a person?" because it does not have category information for the label "girl" but only for the label



GDSE-CL+imagination	
is it a vase?	yes
is it the one on the left?	yes
is the whole thing?	yes
is it on the left?	yes
is the tallest bottle?	no
is it between the two bottles and the pot?	yes
is it between the two bottles and the pot in the	no
is it next to the blue cup?	yes

status: failed

GDSE-CL	
is it a cup?	no
is it a bottle?	no
is it a knife?	no
is the object is in the background?	no
is the handle holding the item?	no
is the the glass thing near the corner of the photo?	no
is the the back part of it visible?	no
is it the furthest back?	no
is it the furthest right?	no
is it the left most corner?	no

status: failed

GDSE-CL+imagination	
is it a person?	yes
is it the kid on the bike?	yes
is it the whole person?	yes
is it wearing an helmet?	yes

status: success

GDSE-CL	
is it a person?	no
is it a bike?	no
is it the green bag?	no
is the the bike?	yes
is it the bike in the back?	no
is the the bike wheel to the right?	no
is it the bike to the left of the bike?	yes
is it the whole bike?	yes

status: failed

GDSE-CL+imagination	
is it an animal?	yes
is it the one in the middle?	yes
is it the whole animal?	yes
is it the one whose head is visible?	yes
is it the whole animal?	yes
is it the entire animal?	yes
is the head facing to the left side of the picture?	yes
is the head facing to the left?	yes

status: success

GDSE-CL	
Is it a giraffe?	no
Is it a car?	no
Is it the wood plank in the background	no
is the animal the animal sitting on?	yes
is the object the animal on the brown animal?	no
is it the black jacket with the brown handle?	no
is the animal's head visible?	no
is the object brown?	no
is the handle black?	no

status: failed

Figure 3: Qualitative examples in the zero-shot gameplay scenario: the categories 'girl' and 'antelope' are not present in MSCOCO and therefore cannot be encoded by the GDSE-CL model. On the other hand, the imagination model is able to *distill* imagination embeddings by using the crop features only (for the sake of presentation quality we remove consecutive repeated questions).

"person". On the other hand, GDSE-CL+IMAGINATION is able to a) categorize the object as a member of the super-category "person", and b) correctly ground the expression "kid on the bike" to the target object. The same behavior can be observed when the "antelope" is the target object. Antelopes are not part of the MSCOCO classes, and therefore have not been seen by the model during training. First, the model refers to it as "animal", hence the Oracle is able to correctly answer the question even though "antelope" was never involved in the training. Secondly, we found that the number of No answers for GDSE-CL is considerably higher (88.06%) than GDSE-CL+IMAGINATION (51.02%), validating our hypothesis that the Oracle does not know how to deal with unseen instances. Finally, in the imagination dialogue of the first example, even though the generated question/answers were probably referring to the correct object, the Guesser model is eventually unable to guess correctly. More work is required to better fuse the language modality and the object representations to improve its performance.

5 Related Work

Concerning unsupervised learning of concept representations, Bruni et al. (2014) first learn modality-specific representations and then fuse them into a unified representation for each concept. However, they rely on hand-crafted bags of visual features, making the approach laborious to extend to new domains and games. Kiela et al. (2018) cope with this issue by relying on CNN models to extract latent features from images for instances of specific objects. Lazaridou et al. (2015) use a margin loss but in the context of maximizing the similarity between the visual representation of a noun phrase and its corresponding text representation. Similarly, Collell et al. (2017) learn a mapping between the ResNet features and the word embeddings of a concept. As discussed in Section 2, unlike our imagination embeddings, these purely-perceptual representations are neither category-aware nor context-aware. Silberer et al. (2016) present a multi-modal model that uses a denoising auto-encoder framework. Unlike us, they do not use perceptual information as input but rely on an attribute-based representation derived from an additional attribute predictor. However, they do use a reconstruction loss (cross-entropy loss for attribute prediction) and an auxiliary category loss during training. Their training scheme is more complex as they first separately train the AE for each modality and then fuse them, which we avoid by adopting a single end-to-end architecture. Ebert and Pavlick (2019) used VAEs to learn grounded representations for lexical concepts. However, as discussed in Section 3, VAEs are not as well suited as RAEs to representation learning for our imagination module. In the context of guessing games, all the previous approaches rely on categories embeddings (De Vries et al., 2017; Shekhar et al., 2019; Strub et al., 2017; Zhuang et al., 2018; Shukla et al., 2019) (see Section 2). Our imagination component can be flexibly integrated in any of them by replacing the category embeddings with imagination embeddings.

6 Conclusions

We argued that existing models for learning grounded conceptual representations fail to learn compositional and generalizable multi-modal representations, relying instead on the use of category labels for every object in the scene both at training and inference time (De Vries et al., 2017). To address this, we introduced a novel "imagination" module based on Regularized Auto-Encoders, that learns a context-aware and category-aware latent embedding for every object directly from its image crop, without using category labels. We showed state-of-the-art performance in the CompGuessWhat?! zero-shot scenario (Suglia et al., 2020), outperforming current models by 8.26% in gameplay accuracy while performing comparably on the other tasks to models which use category labels at training time. The imagination-based model also shows improvements of 2.08% and 12.86% in Oracle and Guesser accuracy. Finally, we conducted an extensive error analysis and showed that imagination embeddings help to reason about object visual properties and attributes. For future work, we plan to 1) integrate category labels at training time in a more principled way following advances in semi-supervised learning (Kingma et al., 2014); 2) improve the multi-task learning procedure presented in (Shekhar et al., 2019) to optimize at the same time multiple tasks of different complexities.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Thirty-First AAAI Conference on Artificial Intelligence*.

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. 2013. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8609–8613. IEEE.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Dylan Ebert and Ellie Pavlick. 2019. Using grounded word representations to study theories of lexical concepts. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 160–169, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. 2019. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

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

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Stephen Laurence and Eric Margolis. 1999. Concepts and cognitive science. *Concepts: core readings*, 3:81.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should i ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2016. Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2284–2297.

Luc Steels. 2015. *The Talking Heads experiment: Origins of words and meanings*, volume 1. Language Science Press.

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courvile, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2765–2771.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. CompGuessWhat?!: A multi-task evaluation framework for grounded language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7625–7641, Online, July. Association for Computational Linguistics.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393.

Ludwig Wittgenstein, Gertrude Elizabeth Margaret Anscombe, and Rush Rhees. 1953. *Philosophische Untersuchungen.(Philosophical investigations)*. Basil Blackwell.

Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261.

A Appendix

A.1 Model details

As described in Section 3 of the main paper, we extend both the Oracle and Guesser model with an imagination component. For both roles, we keep the same model structure for the imagination component. In this paper we implement E_ϕ as a 2-layer feed-forward neural network with ReLU (Dahl et al., 2013) activation function. We acknowledge that many other implementations are possible in this case and we leave more complex designs for future work. Given the latent code $\mathbf{z}_i$ generated by the function E_ϕ, we use a decoder D_θ to generate the reconstructed perceptual input (*imagined*) of the object o_i, $D_\theta(\mathbf{z}_i) = \tilde{\mathbf{v}}_i$. As common practice, we define the decoder D_θ as symmetric to the architecture of the encoder E_ϕ. For the category embeddings size d_c, as in (Shekhar et al., 2019), we use 256 and 512 for the Oracle and Guesser respectively. For the imagination component, we run a grid search involving several parameters for the latent code $\mathbf{z}$ such as $(16, 32, 64, 128, 256, 512)$. For both roles, we choose 512 because it was the value that lead to the highest accuracy on the validation set. We also experimented with several values for the coefficient α of the regularization term $\mathcal{L}_{\text{REG}}$: (1e-3, 1e-5, 1e-6, 1e-7). For the Oracle the best value resulted to be $1e-7$, while $1e-5$ for the Guesser. When training the imagination component with the object category loss, due to the class imbalance, we apply loss weighting. We compute the class weights using the method reported in (King and Zeng, 2001). For the margin value η we opted for 1.0 after experimenting with a less effective dynamic margin that would change depending on the distance between the concepts in the WordNet hierarchy.

A.2 Training details

For both roles, we train the models using the Adam optimizer (Kingma and Ba, 2014). For the Oracle and Guesser training we use 0.0001 as learning rate. In both cases, we use the original *GuessWhat?!* validation set to select the best model that is used in the evaluation on the test set. As described in (Shekhar et al., 2019), we use a modulo-n training procedure to jointly optimize both the Guesser and Questioner. In our experimental evaluation we run a grid search of several values of n such as $3, 5, 7$. We selected 5 as the best performing value on the validation set. For a fair comparison with all the GDSE model variants trained with Supervised Learning and Collaborative Learning, we made the same architectural choices and hyperparameters values. Please refer to the original codebase implementation available on GitHub [4]. Another point of difference is in the Collaborative Learning fine-tuning phase for the Guesser model. During this phase, only the Questioner and Guesser models are fine-tuned whereas the Oracle model is fixed (Shekhar et al., 2019) therefore, we decided to use the best performing Oracle so that the Guesser model is not negatively affected by a less performing Oracle and also to be comparable with the original implementation.

A.3 Error analysis

In order to provide a more fine-grained evaluation of the generated dialogues, we adapt the quality evaluation script presented by Suglia et al. (2020) and extend it with additional metrics. First of all, it relies on a rule-based question classifier that classifies a given question in one of seven classes: 1) super-category (e.g., "person", "utensil", etc.), 2) inanimate object (e.g., "car", "oven", etc.), 3) animate object (e.g., "dog", "cat", etc.), 3) "color", 4) "size", 5) "texture", 6) "shape" and "location". The question classifier is useful to evaluate the dialogue strategy learned by the models. In particular, we look at two types of turn transitions: 1) super-category → object/attr, it measures how many times a question with an affirmative answer from the Oracle related to a super-category is followed by either an object or attribute question (where "attribute" represents the set {color, size, texture, shape and location}; 2) object → attr, it measures how many times a question with an affirmative answer from the Oracle related to an object is followed by either an object or attribute question. We compute the *lexical diversity* as the type/token ratio among all games, *question diversity* and the percentage of games with repeated questions. We also evaluate the percentage of dialogue turns involving location questions. Table 5 and 6 show the results of these analysis for the models `GDSE-CL` and `GDSE-CL+imagination` analyzed in this paper.

[4] `https://github.com/shekharRavi/Beyond-Task-Success-NAACL2019`

Using the above-mentioned question classifier, we completed an error analysis trying to understand the quality of the generated gameplay in a zero-shot scenario from the point of view of the answers prediction performance and the guesser accuracy. In particular, we randomly sampled a pool of 50 reference games from the out-of-domain zero-shot scenario and we manually annotated whether a given answer generated by the Oracle model was correct or not. Table 7 shows the results of the manual annotation step. The model confirms high performance in answering questions about super-category information demonstrating that it is able to correctly categories objects in macro-categories even though is has not seen them before.

A.3.1 Zero-shot gameplay quality

Model	Lexical diversity	Question diversity	% games repeated questions	Super-cat -> obj/attr	Object -> attribute	% turns location questions	Vocab. size	Accuracy
DeVries-RL	0.13	1.77	99.48	97.39	98.70	78.07	702.00	43.92%
GDSE-CL	0.17	13.74	66.75	93.62	66.27	31.23	1260	43.42%
GDSE-CL + Imagination	0.10	8.56	91.80	93.15	60.72	39.90	808	46.70%

Table 5: Comparison between the quality of gameplay in the near-domain zero-shot scenario between GDSE-CL and GDSE-CL with imagination. Number of total turns 10.

Model	Lexical diversity	Question diversity	% games repeated questions	Super-cat -> obj/attr	Object -> attribute	% turns location questions	Vocab. size	Accuracy
DeVries-RL	0.24	2.96	98.49	91.26	98.57	75.84	1275	38.73%
GDSE-CL	0.14	7.86	66.32	91.67	72.33	26.03	1002	29.83%
GDSE-CL + Imagination	0.10	8.57	89.19	94.82	58.51	39.68	814	46.93%

Table 6: Comparison between the quality of gameplay in out-of-domain zero-shot scenario between GDSE-CL and GDSE-CL with imagination. Number of total turns 10.

Question type	Accuracy	Count
Inanimate object	65.48%	168
Animate object	53.33%	15
Super category	83.33%	60
Location	78.86%	175
Size	100.00%	1
Color	58.33%	24
Parts	100.00%	2

Question type	Accuracy	Count
Inanimate object	81.71%	164
Animate object	70.00%	10
Super category	67.61%	71
Location	72.97%	148
Size	100%	1
Color	63.64%	88
Parts	71.43%	7

Table 7: Error analysis results completed on the Out-of-domain zero-shot scenario for the model `GDSE-CL+Imagination` (on the left) and `GDSE-CL` (on the right).