# Resilience through Scene Context in Visual Referring Expression Generation

**Simeon Junker** and **Sina Zarrieß**
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{simeon.junker,sina.zarriess}@uni-bielefeld.de

## Abstract

Scene context is well known to facilitate humans' perception of visible objects. In this paper, we investigate the role of context in Referring Expression Generation (REG) for objects in images, where existing research has often focused on distractor contexts that exert pressure on the generator. We take a new perspective on scene context in REG and hypothesize that contextual information can be conceived of as a resource that makes REG models more resilient and facilitates the generation of object descriptions, and object types in particular. We train and test Transformer-based REG models with target representations that have been artificially obscured with noise to varying degrees. We evaluate how properties of the models' visual context affect their processing and performance. Our results show that even simple scene contexts make models surprisingly resilient to perturbations, to the extent that they can identify referent types even when visual information about the target is completely missing.[1]

## 1 Introduction

Objects do not appear randomly in the world that surrounds us, but they occur in predictable spatial, semantic, or functional configurations and relations to their environment. Research on human perception shows that we "see the world in scenes" (Bar, 2004), and that prior experience and knowledge of the world helps us to efficiently process visual stimuli. Even with an extremely short glimpse at an image, humans remember essential semantic aspects of the scene and object arrangement (Oliva and Torralba, 2006). This rapid scene understanding allows us to handle the complexity of the visual world and to recognize objects in context, e.g., when they are not fully visible (Võ, 2021).

Today's systems for Vision and Language (V&L) commonly process visual inputs that represent "real-world" scenes (e.g. Lin et al. 2014; Antol et al. 2015; Krishna et al. 2016; Das et al. 2017) which, to some extent, exhibit the regularities that human perception is known to be exploiting. Yet, it is not clear *how* current V&L systems process context and whether they rely on strategies of scene understanding similar to humans. In this paper, we aim to investigate this question for Referring Expression Generation (REG, Dale and Reiter 1995; Mao et al. 2016), a controlled set-up that is well established in NLG research, by testing how scene context supports reference generation for objects that are difficult to recognize.


| | | |
|---|---|---|
| noise 0.0 | $\text{TRF}_{tgt}$ | red van (A) |
| | $\text{TRF}_{vis}$ | red truck (A) |
| | $\text{TRF}_{sym}$ | red truck (A) |
| noise 1.0 | $\text{TRF}_{tgt}$ | left elephant (F) |
| | $\text{TRF}_{vis}$ | white truck (A) |
| | $\text{TRF}_{sym}$ | car on left (A) |

Figure 1: Example from RefCOCO (displayed with noise level 0.5) with generated expressions and human judgments. Visual or symbolic scene context allows to identify even fully occluded targets (noise 1.0).

Whereas classical REG algorithms mostly build on pre-defined symbolic representations (Krahmer and van Deemter, 2012), neural generation models in *visual* REG have to extract object properties from low-level visual representations (i.e., photographs) of the target and its context (Schüz et al., 2023). This even applies to properties as fundamental as the *type* of an object, i.e. how it is *named*

[1] Code, models and data for this project are available at: https://github.com/clause-bielefeld/REG-Scene-Context



*Proceedings of the 17th International Natural Language Generation Conference*, pages 344–357
September 23–27, 2024. 

in the expression. Under ideal conditions, determining a referent's type and properties can be regarded as a relatively simple task, but it becomes non-trivial in the presence of imperfect visual information, occlusion or noise. Here, in light of previous findings on human scene understanding (cf. Section 2), scene context can be expected to be of great support. However, to date, little is known as to how processes of scene understanding and object type identification interact in REG.

In this work, we hypothesize that visual scene context makes REG models more *resilient*, i.e., it allows them to recalibrate predictions that were based on imperfect target representations. To test this, we use a novel and highly controllable experimental setup for REG: we train and test different Transformer-based model architectures with target representations that have been artificially obscured with varying degrees of noise (cf. Figure 1), simulating scenarios that are common in the real world but insufficiently represented in current REG datasets. We provide the models with different context representations and compare their performance on common quality metrics and a focused human evaluation of their ability to determine referent types. Our results show that context makes models surprisingly resilient to perturbations in target representations, to the extent that they can identify referent types even when information about the objects themselves is completely missing. We believe that these results open up new perspectives on how information about the structure and content of surrounding scenes facilitate the description of objects in REG and related tasks.

## 2 Background

**Human scene understanding** Research on human vision and perception emphasizes the fact that scenes are not mere collections of objects (Võ, 2021). When humans *view* a scene, they do not simply recognize the objects in it, but *understand* it as a coherent whole. Oliva and Torralba (2006) observe that humans perceive the so-called gist of a scene rapidly and even when local information is missing (e.g. blurred). Other experiments indicate that contextual information can facilitate the recognition of visible objects across different tasks (Oliva and Torralba, 2007; Divvala et al., 2009; Galleguillos and Belongie, 2010; Parikh et al., 2012), and that incongruent context can also be misleading (Zhang et al., 2020; Gupta et al., 2022) demonstrating that the human vision exploits learned knowledge about regularities of the visual word for visual processing (Biederman, 1972; Bar, 2004; Greene, 2013; Pereira and Castelhano, 2014; Sadeghi et al., 2015; Võ, 2021).

**Scenes, objects, and image captioning** Much research on V&L is concerned with modeling the generation and understanding of image descriptions, e.g. in image captioning (Xu et al. 2015; Anderson et al. 2018; Cornia et al. 2020, among many others). Yet, many captioning tasks focus on rather object-centric descriptions that mention objects and their spatial relationships (Cafagna et al., 2021). A common representation of scene context in image captioning is scene graphs (Yang et al., 2023), which are usually modeled via spatial relations between bounding boxes of objects. Cafagna et al. 2023 propose a new task and dataset that foregrounds scene-level instead of object-centric descriptions. Another perspective on scene knowledge in captioning models is coming from work that focuses on probing them with perturbed or systematically varied images: Yin and Ordonez (2017) find that captioning with extremely reduced inputs of labeled object layouts performs surprisingly well. Related to this, Nikolaus et al. (2019) show that image captioning models often rely on regularities in object occurrences, to the extent that they fail to generalize to new combinations of objects. Their solution is to generate unseen combinations and challenge models on these. Our goal in this work is complementary: we aim to understand how exactly generation models may be able to leverage regular scene knowledge and patterns of object co-occurrence, and how this may facilitate the handling of imperfect visual information.

**REG and scene context** REG is concerned with the generation of descriptions that distinguish a particular object in a given visual context, cf. Krahmer and van Deemter 2012. In past years, REG research has largely transitioned from symbolic settings to *visual REG*, focusing on referring expressions for objects in photographs (Kazemzadeh et al., 2014; Mao et al., 2016). Recent models usually build on image captioning models but are adapted to generate more pragmatically informative expressions, using e.g. training objectives (Mao et al., 2016), comprehension modules (Luo and Shakhnarovich, 2017), reinforcement agents (Yu et al., 2017) or decoding strategies (Schüz and Zarrieß, 2021).

Visual REG models usually process different

forms of context information. Whereas some models encode differences in appearance between targets and surrounding objects (Yu et al., 2016, 2017; Tanaka et al., 2019; Kim et al., 2020; Liu et al., 2020), others use representations of the global image (Mao et al., 2016; Luo and Shakhnarovich, 2017; Zarrieß and Schlangen, 2018; Panagiaris et al., 2020, 2021), both commonly supplemented with the relative position and size of the target in the image. On a conceptual level, however, recent work in visual REG generally follows the traditional paradigm by Dale and Reiter 1995, i.e. context is mainly considered in terms of so-called distractor or competitor objects, that are similar to the target and must therefore be excluded by naming differences (Schüz et al. 2023, but see Ilinykh and Dobnik 2023 for context influences in object naming). In this view, context "exerts pressure", as the speaker needs to reason about which attributes and words make the expression unambiguous (Cohn-Gordon et al., 2018; Schüz and Zarrieß, 2021). In this paper, we investigate how contextual information can be conceived as a resource that makes the generation of descriptions easier rather than harder.

**Research gap** Little is known about how visual REG models internally exploit their context representations and in what way context exactly enhances the generation of expressions. A key difference to symbolic REG is that in visual REG failures in scene and object understanding due to e.g. imperfect visual input can lead to semantic errors, cf. Schüz et al. (2023). This is especially evident for the *type* of objects: this attribute had a privileged role in early works (Dale and Reiter, 1995) as it is essential as the head of referential noun phrases. In visual REG, referents must first be correctly identified to *name* them appropriately (Zarrieß and Schlangen, 2017; Silberer et al., 2020a,b; Ilinykh and Dobnik, 2023), which is challenging in cases of deficient input, e.g. small or partially occluded objects (Yao and Fei-Fei, 2010). In this paper, we aim to close this gap and investigate how visual context information helps REG models to be more resilient to deficits in their target inputs.

## 3 Experimental Set-Up

### 3.1 Outline and Research Hypotheses

The main idea of this work is to train and test standard REG models on visual target representations occluded with varying amounts of noise, to investigate how different combinations of target and context can compensate for this perturbation. For this, we draw on existing model architectures, and evaluate the trained models using both out-of-the-box quality metrics and more fine-grained human evaluation capturing the validity of assigned referent type labels, given the challenges of type identification in visual REG discussed in the previous section. The evaluation results are also supported by supplementary analyses.

Generally, we expect that automatic metrics and human evaluation scores will drop for increasing amounts of target noise. However, we also hypothesize that visual context makes models more resilient, i.e., for the same amount of noise, models supplied with context outperform variants with only target information. While we expect this general effect across all conditions, it should be more pronounced as the amount of occlusion increases.

### 3.2 Models

We set up two transformer-based REG models: TRF is a transformer model trained from scratch on REG data, CC builds upon a pre-trained language model. We define variants of both models using a) different combinations of target and context representations as the respective model inputs, and b) the amount of target noise during training and inference. Implementation and training details for our models can be found in appendix B.

Target representations include the visual contents of the target bounding box ($V_t$) and its location and size relative to the global image ($Loc_t$). As context representations, we use the embedding of the global image with the target masked out ($V_c$). We also experiment with symbolic representations about what kinds of objects the surrounding scene is composed of (*scene summaries*, $S_c$). Incorporating symbolic scene features renders the task a multimodal fusion problem, i.e. the model has to align information from low-level visual and location information and symbolic scene summaries. Models processing only target information are indicated with the subscript $tgt$, whereas models processing $V_c$ and $S_c$ context information are indexed with $vis$ and $sym$, respectively.

To test our systems for perturbed target representations, we randomly replace a fixed proportion of the pixels in the bounding box with random noise during both training and inference. With this, we simulate cases of occlusion or other visual disturbances, which are common in real-world scenarios but rarely found in RefCOCO objects. We

opted for pixel-wise occlusion for controllability reasons: Masking continuous sections would arguably be more akin to real-world occlusion by other objects, but could raise further questions, for example whether the parts masked out are important for determining the target class. All systems are trained and tested with three noise settings: $0.0$ as our baseline setting, where no pixels are perturbed; $0.5$, where $50\%$ of the pixels are replaced with noise; and $1.0$, where the entire content of the target bounding box is occluded, i.e. no visual target information is available, similar in spirit to the *Context-Obj* condition in Ilinykh and Dobnik (2023). Importantly, models are trained separately for noise levels, i.e. a model evaluated for noise 0.5 is trained with the same noise level.

**REG Transformer (TRF)** We train a standard transformer architecture from scratch, which allows to carefully control and probe the effects of different target and context information. We use the model from Schüz and Zarrieß (2023), which is based on an existing implementation for image captioning.[2] The model builds on ResNet (He et al., 2015) encodings for targets and context, which are passed on to an encoder/decoder transformer in the style of Vaswani et al. (2017), and is largely comparable to the system in Panagiaris et al. (2021), but without self-critical sequence training and layer-wise connections between encoder and decoder. Unlike e.g. Mao et al. (2016), we train the model using Cross Entropy Loss.

We compare three variants of this model, which take as input concatenated feature vectors comprised of the representations described above. TRF$_{tgt}$ receives only target information, i.e. an input vector $[V_t; Loc_t]$. TRF$_{vis}$ additionally receives visual context representations, namely $[V_t; Loc_t; V_c]$. TRF$_{sym}$ takes symbolic scene summaries as context, i.e. $[V_t; Loc_t; S_c]$.

For both $V_t$ and $V_c$, the respective parts of the image are scaled to $224 \times 224$ resolution (keeping the original ratio and masking out the padding) and encoded with ResNet-152 (He et al., 2015), resulting in 196 features ($14 \times 14$) with hidden size 512 for both target and context. $Loc_t$ is a vector of length 5 with the corner coordinates of the target bounding box and its area relative to the whole image, projected to the model's hidden size.

The scene summary input for TRF$_{sym}$ consists of 134 features, representing the relative area each of the object or stuff categories in COCO occupies in the visual context. $S_c$ features are based on 2D panoptic segmentation maps (cf. Section 3.3): We mask out the target bounding box and calculate the number of pixels assigned to each COCO category in the remaining image, then normalize the number of pixels assigned to each class by the total number of pixels. In TRF$_{sym}$, we add a further layer with jointly trained embeddings for all object and stuff types. In the model's forward pass, we concatenate all 134 embeddings, weighted by the respective coverage in the input image.

**Fine-tuned GPT-2 (CC)** We adapt the *ClipCap* model in Mokady et al. (2021) to the REG task. The authors use a simple MLP-based mapping network to construct fixed-size prefixes for GPT-2 (Radford et al., 2019) from CLIP encodings (Radford et al., 2021), and fine-tune both the mapping network and the language model for the image captioning task. To the best of our knowledge, this is the first model tested for REG which utilizes a pre-trained language model.

As for the TRF model, we compare different variants of this base architecture. First, in CC$_{tgt}$, GPT-2 prefixes are constructed as $[V_t; Loc_t]$, where $V_t$ is computed like the CLIP prefix in the original paper (but for the contents of the target bounding box) and $Loc_t$ is the location features described above, projected into a single prefix token. In CC$_{vis}$, prefixes contain visual context representations, i.e. $[V_t; V_c; Loc_t]$. Here, $V_c$ is computed like $V_t$, but with a separate mapping network and with the global image (minus the target) as the visual input. Finally, CC$_{sym}$ includes symbolic scene summaries, i.e. $[V_t; S_c; Loc_t]$. Similar to the visual inputs, we use a mapping network to project the features before concatenation.

### 3.3 Data

We use RefCOCO and RefCOCO+ (Kazemzadeh et al., 2014) for training and evaluation. Both contain bounding boxes and expressions for the same objects in MSCOCO images (Lin et al., 2014), but while the location attributes *left* and *right* are highly frequent in RefCOCO, they have been excluded in RefCOCO+. The datasets contain separate *testA* and *testB* splits (1.9k and 1.8k items), where *testA* only contains humans as referents and *testB* all other object classes (but not humans). To construct scene summaries ($S_c$) and analyze attention allocation patterns, we use annotations for panoptic

[2] https://github.com/saahiluppal/catr

segmentation (Kirillov et al., 2018), i.e. dense pixel-level segmentation masks for *thing* and *stuff* classes in MSCOCO images (Caesar et al., 2016).

### 3.4 Evaluation

**Generation Quality / N-Gram Metrics** To estimate the general generation capabilities of our models we rely on BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2014) as established metrics for automatic evaluation. As target occlusion involves random processes, we repeat inference ten times for all settings and average the results.

**Referent Type Assignment / Human Evaluation** To test whether our models succeed in assigning valid types to referents, we collect human judgments for generated expressions for a subset of 200 items from the RefCOCO *testB* split, which is restricted to non-human referents. Unlike for the automatic metrics, we use the results of a single inference run for each system. The annotators were instructed to rate only those parts of the expressions that refer to the type of the referential target. For example, "the black dog" should be rated as correct if the target is of the type dog, but is actually white. All items should be assigned exactly one of the following categories:

- **Adequate / A**: The generated expression contains a valid type description for the referent.
- **Misaligned / M**: Type designators do not apply to the intended target, but to other objects (partially) captured by the bounding box.
- **Omission / O**: Omission of the target type, e.g. description via non-type attributes, pronominalization or general nouns such as "thing".
- **False / F**: Type designations that do not apply to the intended target or other objects captured by the bounding box.

Previous research has shown considerable variation in object naming (Silberer et al. 2020a,b, among others). Therefore, for the *A* category, type descriptions do not have to match the ground truth annotations, but different labels can be considered adequate if they represent valid descriptions of the target type. For example, *dog*, *pet* and *animal* would be considered equally correct for depicted dogs. Subsequent to the human evaluation, we investigate correlations between the evaluation results and further properties of the visual context.

**Attention Allocation** We also examine how our $\text{TRF}_{vis}$ model allocates attention over different parts of the input as a result of different noise levels during training. First, we follow Schüz and Zarrieß (2023) in measuring the attention directed to the target and its context in both the encoder and decoder. For this, we compute $\alpha_t$, $\alpha_l$ and $\alpha_c$ as the cumulative attention weights directed to $V_t$, $Loc_t$ and $V_c$, respectively, normalized such that $\alpha_t+\alpha_l+\alpha_c = 1$. We report the difference of attention directed to target and context, calculated as $\Delta_{t,c} = (\alpha_t+\alpha_l)-\alpha_c$, i.e. $0 < \Delta_{t,c} \leq 1$ if there is relative focus on the target, $-1 \leq \Delta_{t,c} < 0$ if there is relative focus on the context, and $\Delta_{t,c} = 0$ when both are weighted equally. Second, we measure the model attention allocated to different classes of objects in the visual context, using the panoptic segmentation data described in Section 3.3. Here, we first interpolate the model attention map to fit the original dimensions of the image and retrieve the respective segmentation masks. For each category $x \in X$, we then compute the cumulative attention weight $\alpha_x$ by computing the sum of pixels attributed to this category, weighted by the model attention scores over the image and normalized such that $\sum_{x \in X} \alpha_x = 1$. We report $\alpha_{x=tgt}$, i.e. attention allocated to areas of the visual context assigned *the same category as the referential target*.

## 4 Results

### 4.1 Automatic Quality Metrics

Table 1 shows the results of the automatic evaluation of our systems on the testA and testB splits in RefCOCO and RefCOCO+. Interestingly, the simpler TRF model outperforms CC, although the latter builds on pre-trained CLIP and GPT-2 which are known to be effective for image captioning (Mokady et al., 2021). Possible reasons for this can be seen in structural differences between bounding box contents and full images as used in the CLIP pre-training, or in higher compression when constructing the GPT prefixes. Without target occlusion, model variants with access to visual context generally achieve the highest scores for both architectures ($\text{TRF}_{vis}$ and $\text{CC}_{vis}$, although $\text{CC}_{sym}$ exceeds the latter on testB+).

As expected, scores consistently drop with increasing target noise. However, this is mitigated if context is available: For both TRF and CC, variants incorporating visual context are substantially more robust against target noise, even if target rep-

| | noise | testA Bl$_1$ | testA Bl$_2$ | testA CDr | testB Bl$_1$ | testB Bl$_2$ | testB CDr | testA+ Bl$_1$ | testA+ Bl$_2$ | testA+ CDr | testB+ Bl$_1$ | testB+ Bl$_2$ | testB+ CDr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TRF$_{tgt}$ | | 0.55 | 0.35 | 0.86 | 0.57 | 0.35 | 1.28 | 0.49 | 0.31 | 0.77 | 0.36 | 0.19 | 0.68 |
| TRF$_{vis}$ | 0.0 | 0.58 | 0.39 | 0.93 | 0.61 | 0.39 | 1.36 | 0.50 | 0.32 | 0.83 | 0.37 | 0.20 | 0.73 |
| TRF$_{sym}$ | | 0.54 | 0.34 | 0.84 | 0.57 | 0.35 | 1.27 | 0.46 | 0.29 | 0.78 | 0.37 | 0.19 | 0.72 |
| TRF$_{tgt}$ | | 0.49 | 0.32 | 0.73 | 0.52 | 0.32 | 1.06 | 0.42 | 0.27 | 0.64 | 0.29 | 0.14 | 0.53 |
| TRF$_{vis}$ | 0.5 | 0.53 | 0.35 | 0.81 | 0.56 | 0.36 | 1.24 | 0.43 | 0.26 | 0.67 | 0.34 | 0.18 | 0.62 |
| TRF$_{sym}$ | | 0.53 | 0.35 | 0.81 | 0.57 | 0.35 | 1.28 | 0.45 | 0.29 | 0.71 | 0.36 | 0.19 | 0.68 |
| TRF$_{tgt}$ | | 0.35 | 0.17 | 0.34 | 0.30 | 0.14 | 0.20 | 0.29 | 0.15 | 0.20 | 0.07 | 0.01 | 0.04 |
| TRF$_{vis}$ | 1.0 | 0.46 | 0.29 | 0.60 | 0.55 | 0.36 | 1.14 | 0.32 | 0.17 | 0.34 | 0.29 | 0.14 | 0.47 |
| TRF$_{sym}$ | | 0.42 | 0.24 | 0.51 | 0.53 | 0.33 | 1.12 | 0.31 | 0.15 | 0.31 | 0.30 | 0.14 | 0.48 |
| CC$_{tgt}$ | | 0.48 | 0.30 | 0.70 | 0.47 | 0.28 | 0.88 | 0.42 | 0.27 | 0.70 | 0.29 | 0.14 | 0.53 |
| CC$_{vis}$ | 0.0 | 0.57 | 0.38 | 0.92 | 0.58 | 0.37 | 1.25 | 0.45 | 0.29 | 0.77 | 0.33 | 0.18 | 0.62 |
| CC$_{sym}$ | | 0.45 | 0.28 | 0.66 | 0.56 | 0.36 | 1.22 | 0.44 | 0.28 | 0.73 | 0.37 | 0.20 | 0.70 |
| CC$_{tgt}$ | | 0.38 | 0.21 | 0.48 | 0.36 | 0.20 | 0.51 | 0.40 | 0.25 | 0.64 | 0.27 | 0.14 | 0.47 |
| CC$_{vis}$ | 0.5 | 0.51 | 0.32 | 0.75 | 0.50 | 0.31 | 0.97 | 0.41 | 0.26 | 0.68 | 0.30 | 0.16 | 0.55 |
| CC$_{sym}$ | | 0.44 | 0.27 | 0.61 | 0.57 | 0.36 | 1.17 | 0.35 | 0.21 | 0.46 | 0.33 | 0.17 | 0.57 |
| CC$_{tgt}$ | | 0.35 | 0.16 | 0.37 | 0.29 | 0.12 | 0.16 | 0.27 | 0.14 | 0.20 | 0.10 | 0.02 | 0.06 |
| CC$_{vis}$ | 1.0 | 0.40 | 0.23 | 0.46 | 0.38 | 0.21 | 0.46 | 0.29 | 0.15 | 0.30 | 0.20 | 0.09 | 0.27 |
| CC$_{sym}$ | | 0.42 | 0.25 | 0.52 | 0.55 | 0.34 | 1.17 | 0.31 | 0.16 | 0.32 | 0.32 | 0.16 | 0.53 |

Table 1: BLEU$_1$, BLEU$_2$ and CIDEr scores on RefCOCO testA and testB for all TRF and CC variants. Systems indicated with *tgt* can only access target information, *vis* and *sym* models are supplied with visual context and symbolic *scene summaries*, respectively. Generally, context information leads to improved results, especially for high noise settings.

resentations are entirely occluded, cf. Figure 2. For example, for RefCOCO testB, CIDEr drops to 0.20 for TRF$_{tgt}$ with noise 1.0 but TRF$_{vis}$ achieves scores as high as 1.14, indicating that visual context combined with location features provides valuable information for describing (occluded) targets. Generally, TRF$_{vis}$ appears to be more effective at exploiting the visual context, e.g. CC$_{vis}$ with noise 1.0 drastically underperforms with CIDEr 0.46 on testB. Although CC$_{tgt}$ is still outperformed (CIDEr 0.16), this suggests problems for extracting relevant information from the visual context.

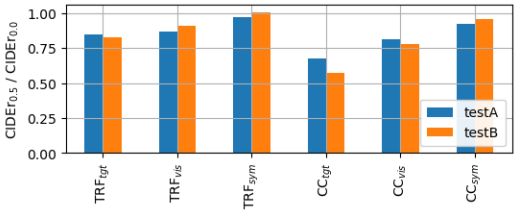
Similar patterns emerge when replacing visual context with symbolic *scene summaries*: For both TRF and CC, model variants incorporating symbolic context features outperform their target-only counterparts in most cases, highlighting the potential of object co-occurrence information for making predictions robust to noise. For example, TRF$_{sym}$ achieves CIDEr 1.12 for noise 1.0 in testB, comparable to TRF$_{vis}$. CC$_{sym}$ even outperforms CC$_{vis}$ for high noise settings (and all settings on testB+). On testB, CC$_{sym}$ scores are almost constant across noise levels, suggesting that the model is strongly relying on the scene summary information.

Interestingly, we see considerable differences between testA and testB: For both RefCOCO and RefCOCO+, target-only variants suffer less from occlusion on the testA splits (containing references to humans), but context is more effective on testB (containing references to other objects). We hypothesize that models without meaningful visual input but access to location and size information can often *guess right* on the frequent human classes in testA, but struggle with the higher variation in testB. Conversely, while human referents appear in a wide range of environments, other objects in testB rather tend to occur in specific surroundings, making context information more informative regarding their identity.

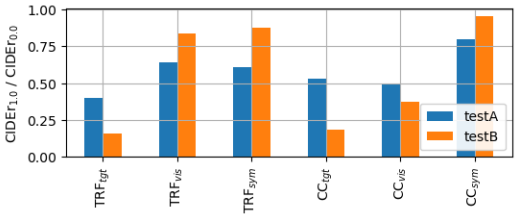
### 4.2 Target Identification

Human judgments were collected from 6 expert annotators, including the first author. Every system was evaluated independently by three annotators, with a Fleiss' Kappa of 0.85, indicating *almost perfect* agreement (Landis and Koch, 1977). The final judgments are determined by majority vote.

The human evaluation results for the 200-item subset of RefCOCO testB are shown in Table 2. Generally, we see similar patterns as in the BLEU and CIDEr scores discussed previously: Ratios of ***A**dequate* descriptions drop if noise ratios increase, while ***F**alse* ratios increase at the same time. For




(a) CIDEr for noise 0.5, relative to noise 0.0




(b) CIDEr for noise 1.0, relative to noise 0.0

Figure 2: Relative CIDEr scores with respect to noise 0.0 for RefCOCO testA and testB. For both TRF and CC, model variants with access to context are more robust against noise, especially for testB.

| | noise | % A | % F | % O | % M |
|---|---|---|---|---|---|
| $\text{TRF}_{tgt}$ | 0.0 | 84.0 | 10.5 | 5.0 | 0.5 |
| $\text{TRF}_{vis}$ | | 81.0 | 11.5 | 5.5 | 2.0 |
| $\text{TRF}_{sym}$ | | 89.0 | 7.0 | 3.5 | 0.5 |
| $\text{TRF}_{tgt}$ | 0.5 | 66.5 | 28.0 | 4.0 | 1.5 |
| $\text{TRF}_{vis}$ | | 70.5 | 18.5 | 7.0 | 4.0 |
| $\text{TRF}_{sym}$ | | 81.5 | 14.5 | 2.5 | 1.5 |
| $\text{TRF}_{tgt}$ | 1.0 | 1.5 | 75.5 | 19.5 | 3.5 |
| $\text{TRF}_{vis}$ | | 66.0 | 26.5 | 4.0 | 3.5 |
| $\text{TRF}_{sym}$ | | 68.0 | 22.0 | 1.5 | 8.5 |
| $\text{CC}_{tgt}$ | 0.0 | 46.0 | 46.5 | 7.0 | 0.5 |
| $\text{CC}_{vis}$ | | 75.5 | 21.5 | 3.0 | 0.0 |
| $\text{CC}_{sym}$ | | 70.5 | 17.5 | 5.5 | 6.5 |
| $\text{CC}_{tgt}$ | 0.5 | 23.0 | 61.0 | 13.0 | 3.0 |
| $\text{CC}_{vis}$ | | 55.5 | 35.5 | 6.5 | 2.5 |
| $\text{CC}_{sym}$ | | 69.0 | 19.5 | 2.5 | 9.0 |
| $\text{CC}_{tgt}$ | 1.0 | 0.5 | 84.5 | 11.0 | 4.0 |
| $\text{CC}_{vis}$ | | 19.5 | 68.5 | 9.0 | 3.0 |
| $\text{CC}_{sym}$ | | 70.5 | 16.0 | 4.5 | 9.0 |
| *human* | 0.0 | 90.5 | 2.5 | 6.0 | 1.0 |

Table 2: Ratios of **A**dequate, **F**alse, **O**mitted and **M**isaligned type descriptions (human annotation for 200 items from RefCOCO testB). Generally, contextual information leads to more adequate type descriptions, even if target representations are entirely occluded.

***M**isalignments* and ***O**missions*, higher noise generally leads to higher rates than the baseline setting. $\text{TRF}_{sym}$ and $\text{CC}_{sym}$ show particularly high *M* rates for high noise settings, suggesting that the models often select object types that appear in the image, but not as the referent. In the vast majority of cases, TRF variants outperform their CC counterparts. Again, the systems show large differences in exploiting visual context: Whereas $\text{CC}_{vis}$ assigns *adequate* types in almost 20% of all cases for noise 1.0 (as compared to 0.5% without context information), $\text{TRF}_{vis}$ scores an impressive 66%.

Interestingly, symbolic scene summaries appear to be more effective for identification than visual context features: In most cases, models taking $S_c$ as input generate more adequate descriptions and fewer false descriptions and omissions than corresponding variants with visual context. For $\text{TRF}_{sym}$, this even extends to cases without target occlusion, unlike for BLEU and CIDEr (cf. Section 4.1). Surprisingly, $\text{CC}_{sym}$ achieves very similar *A* scores across all noise settings, narrowly exceeding $\text{TRF}_{sym}$ with noise 1.0. In line with the diminished influence of target occlusion observed for CIDEr and BLEU on testB, this indicates heavy reliance on symbolic scene representations (irrespective of the availability of visual target information), possibly due to problems with fusing symbolic (scene) and visual (target) information, a process that has received much attention in e.g. Visual Question Answering (Zhang et al., 2019; Lu et al., 2023).

### 4.3 How do models exploit scene context?

So far, our results indicate that the scene context of referential targets greatly improves the resilience of REG models, to the extent that correct predictions are possible to a surprising rate even if target information is missing. Here, we aim to analyze how exactly contextual information is exploited by the models. As discussed in Section 2, previous research indicates that regularities of object co-occurrence and scene properties facilitate e.g. object recognition in context. However, qualitative inspection indicates that for high noise, our systems often *copy* from context, i.e. predict referent types that are also present in the surrounding scene, given that many classes of objects tend to appear in groups. To investigate this, we (a) perform statistical tests to check whether similar objects in context support identification performance and (b) analyze the attention distribution for $\text{TRF}_{vis}$ to see how the respective context objects are weighted by the model.

**Statistical analysis: Target categories in context** We hypothesize that recalibration through context is more effective when the target class is also present in the scene. To test this, we conduct

| | noise | corr. | p |
|---|---|---|---|
| $\text{TRF}_{tgt}$ | 0.0 | 0.128 | – |
| $\text{TRF}_{vis}$ | | 0.109 | – |
| $\text{TRF}_{sym}$ | | 0.154 | $< 0.05$ |
| $\text{TRF}_{tgt}$ | 0.5 | 0.071 | – |
| $\text{TRF}_{vis}$ | | 0.186 | $< 0.01$ |
| $\text{TRF}_{sym}$ | | 0.157 | $< 0.05$ |
| $\text{TRF}_{tgt}$ | 1.0 | 0.046 | – |
| $\text{TRF}_{vis}$ | | 0.321 | $< 0.001$ |
| $\text{TRF}_{sym}$ | | 0.277 | $< 0.001$ |
| $\text{CC}_{tgt}$ | 0.0 | 0.156 | $< 0.05$ |
| $\text{CC}_{vis}$ | | 0.142 | $< 0.05$ |
| $\text{CC}_{sym}$ | | 0.353 | $< 0.001$ |
| $\text{CC}_{tgt}$ | 0.5 | 0.049 | – |
| $\text{CC}_{vis}$ | | 0.145 | $< 0.05$ |
| $\text{CC}_{sym}$ | | 0.249 | $< 0.001$ |
| $\text{CC}_{tgt}$ | 1.0 | 0.045 | – |
| $\text{CC}_{vis}$ | | 0.136 | – |
| $\text{CC}_{sym}$ | | 0.246 | $< 0.001$ |

Table 3: Correlation between identification accuracy and relative coverage of the target class in context. For most model variants with access to context, higher prevalence of the target class in the visual context leads to significantly higher scores in human evaluation.

a correlation analysis between identification accuracy and the relative coverage of the target class in the context. For this, we again rely on panoptic segmentation annotations (cf. Section 3.3) to compute the proportion of pixels of the same class as the referential target, normalized by the total size of the context. We binarize the human evaluation scores (*True* if rated as *A*, else *False*) and compute the Point-biserial correlation coefficient between the relative coverage of the target class in context and the identification accuracy. The results are shown in Table 3. In almost all systems including visual or symbolic context representations, a higher prevalence of the target class in the visual context leads to significantly higher scores in human evaluation ($p < 0.05$ or higher significance for all systems except $\text{TRF}_{vis}$ / noise 0.0 and $\text{CC}_{vis}$ / noise 1.0), i.e. systems can easier compensate a lack of visual target information if the context contains similar objects. For TRF variants, the correlation is increasing with higher noise ratios, whereas it is more stable for CC. Interestingly, without access to context, both $\text{CC}_{tgt}$ and $\text{TRF}_{tgt}$ show weak correlation for the noise 0.0 setting (albeit only the former is significant), indicating the possibility of more general biases in the data.

| | noise | Encoder $\Delta_{t,c}$ | Encoder $\alpha_{x=tgt}$ | Decoder $\Delta_{t,c}$ | Decoder $\alpha_{x=tgt}$ |
|---|---|---|---|---|---|
| $TRF_{vis}$ | 0.0 | 0.07 | 36.70 | 0.25 | 26.94 |
| $TRF_{vis}$ | 0.5 | -0.30 | 35.27 | -0.06 | 40.56 |
| $TRF_{vis}$ | 1.0 | -0.17 | 35.63 | -0.12 | 43.66 |

Table 4: Attention allocation scores for $\text{TRF}_{vis}$, averaged over RefCOCO testB. $\Delta_{t,c}$ is the attention ratio between target and context, $\alpha_{x=tgt}$ is the % of context attention directed to instances of the target class.

**Model attention to target category in context** In Table 4, we report the results of our attention analysis for $\text{TRF}_{vis}$ (cf. Section 3.4), averaged over all items in RefCOCO testB. For the target/context deltas $\Delta_{t,c}$, we expect that context is weighted more (i.e., scores are decreasing) as noise levels increase. Surprisingly, in the encoder, context is attended most in the 0.5 noise setting. Decoder attention, however, follows our expected pattern. Similarly, as shown by the $\alpha_{x=tgt}$ scores in Table 4, target noise does not seem to have a consistent effect on encoder attention to context objects sharing the target category. For the decoder, however, we see a notable increase: Whereas the baseline model assigns an average of 26.94 % of its attention mass on context objects with the target class, this is significantly increased for higher noise settings (40.56 % and 43.66 %), suggesting that the TRF model learns to exploit the occurrence of similar objects in target and context as a common property of scenes in RefCOCO.

### 4.4 Qualitative Examples and Error Analysis

Figure 3 shows expressions generated by all TRF variants and human identification judgments for three examples from RefCOCO.[3] We identify both *recognition errors*, where visual representations are incorrectly categorized, and *inference errors*, where contextual information is misinterpreted.

Examples of recognition errors can be seen in Figure 3a, where $\text{TRF}_{tgt}$ predicts incorrect but visually related object types for noise 0.5 (*horse*) and mostly unrelated types for noise 1.0 (*man*). Here, both symbolic and visual context allow for robust predictions across noise levels. This is different in Example 3b: While similar problems can be seen for $\text{TRF}_{tgt}$ (*monitor* instead of *microwave* for noise 0.5), symbolic context leads to inference errors, i.e.

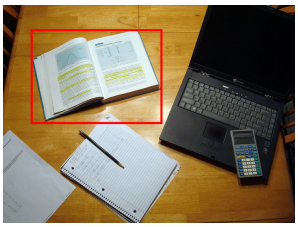
[3] For brevity, we present only expressions generated by TRF. For CC we observe similar patterns, the expressions can be found in Appendix E.

| | | |
|---|---|---|
| noise 0.0 | $TRF_{tgt}$ | cow (A) |
| | $TRF_{vis}$ | left cow (A) |
| | $TRF_{sym}$ | cow on left (A) |
| noise 0.5 | $TRF_{tgt}$ | white horse (F) |
| | $TRF_{vis}$ | cow on left (A) |
| | $TRF_{sym}$ | cow (A) |
| noise 1.0 | $TRF_{tgt}$ | man (F) |
| | $TRF_{vis}$ | left cow (A) |
| | $TRF_{sym}$ | cow on left (A) |

(a) Recognition errors for $TRF_{tgt}$ with target noise, mitigated by context.

| | | |
|---|---|---|
| noise 0.0 | $TRF_{tgt}$ | top left micro (A) |
| | $TRF_{vis}$ | top left microwave (A) |
| | $TRF_{sym}$ | top left microwave (A) |
| noise 0.5 | $TRF_{tgt}$ | left monitor (F) |
| | $TRF_{vis}$ | top microwave (A) |
| | $TRF_{sym}$ | top oven (F) |
| noise 1.0 | $TRF_{tgt}$ | top left donut (F) |
| | $TRF_{vis}$ | top microwave (A) |
| | $TRF_{sym}$ | stove top (F) |

(b) $TRF_{sym}$ predictions are incorrect, but congruent with the scene.



| | | |
|---|---|---|
| noise 0.0 | $TRF_{tgt}$ | top book (A) |
| | $TRF_{vis}$ | top book (A) |
| | $TRF_{sym}$ | paper on top (A) |
| noise 0.5 | $TRF_{tgt}$ | white book (A) |
| | $TRF_{vis}$ | top laptop (F) |
| | $TRF_{sym}$ | open book (A) |
| noise 1.0 | $TRF_{tgt}$ | top left (O) |
| | $TRF_{vis}$ | left laptop (F) |
| | $TRF_{sym}$ | laptop on left (F) |

(c) Copying errors (*laptop*) for $TRF_{vis}$ and $TRF_{sym}$.

Figure 3: Examples from RefCOCO with generated expressions and human judgments (targets are marked red).

$TRF_{sym}$ predicts incorrect object types that however fit into the general scene surrounding the target (*oven* and *stove top* as examples for kitchen appliances). Finally, in Example 3c we see evidence for the copying strategy discussed in Section 4.3: With increasing noise, both $TRF_{vis}$ and $TRF_{sym}$ incorrectly predict *laptop* as an object class present in the surrounding scene.

## 5 Discussion and Conclusion

Our findings show that contextual information about the surroundings of referents makes REG models more resilient against perturbations in visual target representations. Even if no target information is present at all, context allows REG models to maintain good results in automatic quality metrics and to identify referent types with high accuracy, as shown in the human evaluation results. This holds for different kinds of context: While especially the $TRF_{vis}$ model is able to leverage scene information from ResNet encodings of image contents outside the target bounding box, the same applies to symbolic scene representations, as included in $TRF_{sym}$ and $CC_{sym}$. This adds another perspective to basic assumptions of the REG paradigm, where context information is considered important mainly to ensure that references can be resolved without ambiguity. Here, we show, that it is also a valuable source for further communicative goals, i.e. the *truthfulness* of generated expressions.

Interestingly, while related studies on human perception emphasize the importance of e.g. learned co-occurrence patterns between objects, our subsequent analysis rather points to implicitly learned copying strategies that appear to be highly effective for the relatively regular RefCOCO data. While this can also be seen as exploiting scene patterns, it is fundamentally different from the ways in which scene information is interpreted by humans (cf. Section 2). Therefore, we see an urgent need for data more representative of real-world scenarios to further investigate the impact of scene context on multimodal language generation.

Overall, our results indicate that the influence of visual context in REG is more multifaceted than reflected in previous studies. Importantly, this study only provides an initial spotlight, as research in related fields suggests that there are other and more complex ways in which visual scene context may facilitate reference production. With this in mind, we strongly advocate further research into scene context at the interface of perceptual psychology and V&L generation.

**Risks and Ethical Considerations** We do not believe that there are significant risks associated with this work, as we consider the generation of general expressions for generic objects in freely available datasets with limited scale. When selecting samples for human evaluation, we refrain from descriptions of people (that could potentially be perceived as hurtful). No ethics review was required. Our data does not contain any protected information and is fully anonymized.

**Supplementary Materials Availability Statement:**

- RefCOCO and RefCOCO+ annotations and the RefCOCO API for computing BLEU and

CIDEr scores are available on GitHub[4]

- COCO images and panoptic segmentation annotations are available at https://cocodataset.org/
- Source code for the TRF base model are available on GitHub[5]
- Source code for the CC base model are available on GitHub[6]
- Our own code and data are available on GitHub[7]

## Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project B02.

## 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Irving Biederman. 1972. Perceiving real-world scenes. *Science*, 177(4043):77–80.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2016. Coco-stuff: Thing and stuff classes in context.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language models 'see' when they see scenes.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2023. HL dataset: Visually-grounded description of scenes, actions and rationales. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 293–312, Prague, Czechia. Association for Computational Linguistics.

[4]https://github.com/lichengunc/refer
[5]https://github.com/saahiluppal/catr
[6]https://github.com/rmokady/CLIP_prefix_caption
[7]https://github.com/clause-bielefeld/REG-Scene-Context

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert. 2009. An empirical study of context in object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Carolina Galleguillos and Serge Belongie. 2010. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722.

Michelle R. Greene. 2013. Statistics of high-level scene context. *Frontiers in Psychology*, 4.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Nikolai Ilinykh and Simon Dobnik. 2023. Context matters: evaluation of target and context features on variation of object naming. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 12–24, Ingolstadt, Germany. Association for Computational Lingustics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1952–1962, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2018. Panoptic segmentation.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29:5244–5258.

Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. 2023. The multimodal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, 9:e1400.

R. Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134.

Junhua Mao, J. Huang, A. Toshev, Oana-Maria Camburu, A. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

Aude Oliva and Antonio Torralba. 2006. Chapter 2 building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, pages 23–36. Elsevier.

Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2020. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech & Language*, 68:101184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Devi Parikh, C. Lawrence Zitnick, and Tsuhan Chen. 2012. Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1978–1991.

Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Effie J. Pereira and Monica S. Castelhano. 2014. Peripheral guidance in scenes: The interaction of scene context and object content. *Journal of Experimental Psychology: Human Perception and Performance*, 40(5):2056–2072.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Zahra Sadeghi, James L. McClelland, and Paul Hoffman. 2015. You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, 76:52–61.

Simeon Schüz and Sina Zarrieß. 2021. Decoupling pragmatics: Discriminative decoding for referring expression generation. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, Gothenburg, Sweden. Association for Computational Linguistics.

Simeon Schüz and Sina Zarrieß. 2023. Keeping an eye on context: Attention allocation over input partitions in referring expression generation. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 20–27, Prague, Czech Republic. Association for Computational Linguistics.

Simeon Schüz, Albert Gatt, and Sina Zarrieß. 2023. Rethinking symbolic and visual context in referring expression generation. *Frontiers in Artificial Intelligence*, 6.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020a. Object naming in language and vision: A survey and a new dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5792–5801, Marseille, France. European Language Resources Association.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020b. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

M. Tanaka, Takayuki Itamochi, K. Narioka, Ikuro Sato, Y. Ushiku, and T. Harada. 2019. Generating easy-to-understand referring expressions for target identifications. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5793–5802.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation.

Melissa Le-Hoa Võ. 2021. The meaning and structure of scenes. *Vision Research*, 181:10–20.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. 37:2048–2057.

Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Songfang Huang, Fei Huang, Zhangzikang Li, and Yu Zhang. 2023. Transforming visual scene graphs to image captions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12427–12440, Toronto, Canada. Association for Computational Linguistics.

Bangpeng Yao and Li Fei-Fei. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.

Xuwang Yin and Vicente Ordonez. 2017. Obj2Text: Generating visually descriptive language from object layouts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 177–187, Copenhagen, Denmark. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Sina Zarrieß and David Schlangen. 2017. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 243–254, Vancouver, Canada. Association for Computational Linguistics.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.

Dongxiang Zhang, Rui Cao, and Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion*, 52:268–280.

Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. 2020. Putting visual object recognition in context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12982–12991.

## A Limitations

We identify the following limitations in our study:

First, in both training and evaluation, we do not consider pragmatic informativeness as a core criterion for the REG task. We train our models using Cross Entropy Loss and do not test whether the generated expressions unambiguously describe the referential target, instead focusing on semantic adequacy as an important prerequisite for the generation of successful referential expressions. However, we acknowledge that a comprehensive view

| | noise | RefCOCO epochs | RefCOCO CIDEr (val) | RefCOCO+ epochs | RefCOCO+ CIDEr (val) |
|---|---|---|---|---|---|
| $TRF_{tgt}$ | 0.0 | 8 | 1.074 | 7 | 0.803 |
| $TRF_{vis}$ | 0.0 | 6 | 1.156 | 7 | 0.828 |
| $TRF_{sym}$ | 0.0 | 8 | 1.075 | 5 | 0.794 |
| $TRF_{tgt}$ | 0.5 | 11 | 0.936 | 4 | 0.647 |
| $TRF_{vis}$ | 0.5 | 9 | 1.035 | 11 | 0.697 |
| $TRF_{sym}$ | 0.5 | 14 | 1.032 | 10 | 0.74 |
| $TRF_{tgt}$ | 1.0 | 5 | 0.302 | 3 | 0.173 |
| $TRF_{vis}$ | 1.0 | 6 | 0.869 | 5 | 0.449 |
| $TRF_{sym}$ | 1.0 | 12 | 0.818 | 5 | 0.45 |
| $CG_{tgt}$ | 0.0 | 7 | 0.824 | 4 | 0.673 |
| $CG_{vis}$ | 0.0 | 4 | 1.103 | 5 | 0.754 |
| $CG_{sym}$ | 0.0 | 8 | 0.908 | 8 | 0.756 |
| $CG_{tgt}$ | 0.5 | 8 | 0.554 | 14 | 0.603 |
| $CG_{vis}$ | 0.5 | 10 | 0.894 | 5 | 0.679 |
| $CG_{sym}$ | 0.5 | 11 | 0.89 | 11 | 0.553 |
| $CG_{tgt}$ | 1.0 | 2 | 0.294 | 4 | 0.174 |
| $CG_{vis}$ | 1.0 | 7 | 0.526 | 11 | 0.334 |
| $CG_{sym}$ | 1.0 | 9 | 0.823 | 8 | 0.45 |

Table 5: Training information for all TRF and CC variants. CIDEr scores are computed for the val splits in RefCOCO / RefCOCO+.

would require the consideration of both semantic and pragmatic aspects.

Also, we do not consider recent developments such as multimodal LLMs, although the high diversity of their training data would contribute an interesting aspect to this study. Here, we selected our models with a focus on both modifiability and transparent processing.

Finally, additional vision and language datasets such as VisualGenome (Krishna et al., 2016) would have made the results more representative. However, due to time and space constraints, we leave this for future research.

## B Model implementation and training

For the hyperparameters of our models, we largely followed Panagiaris et al. (2021) (TRF) and Mokady et al. (2021) (CC). During inference, we relied on greedy decoding.

The TRF model has 3 encoder and 3 decoder layers with 8 attention heads, hidden dimension and feedforward dimension of 512, and was trained with an initial learning rate of 0.0001 for the transformer encoder and decoder, and 0.00001 for the pre-trained ResNet-152 backbone. Our TRF models have approximately 103,000,000 parameters.

For our CC model, we kept the settings defined by Mokady et al. (2021). From the two models proposed in this work, we used the variant where a simple MLP is used as a mapping network and the GPT-2 language model is fine-tuned during training. However, we have different prefix sizes than in the original paper: For $CC_{tgt}$, we have a prefix size of 11, i.e. 10 for the visual target representation and 1 for the target location information. For $CC_{vis}$ and $CC_{sym}$, our prefix size is 21, with additional 10 tokens for the context. The model was trained using a learning rate of 0.00001. $CC_{vis}$ has approximately 338,000,000, $CC_{sym}$ has 337,000,000 and $CC_{tgt}$ has 307,000,000 parameters.

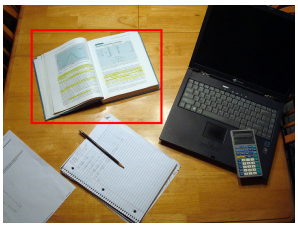
We trained our models on an Nvidia RTX A40. Both RefCOCO and RefCOCO+ contain approximately 42k items for training. The number of training epochs per system and the final CIDEr scores over the validation sets are displayed in Table 5. We trained all our models for a maximum of 15 epochs, with early stopping if no new maximum for CIDEr over the validation set has been achieved for three consecutive epochs. Per epoch, the compute time was approximately 2.30 h for all systems.

## C Scientific Artifacts

In our work, we mainly used scientific artifacts in the form of existing model implementations, all of which are cited or referenced in Section 3. The model implementations were published under permissive licences, i.e. *MIT* (TRF) and *Apache 2.0* (CC). We publish our modifications to the model

| | | |
|---|---|---|
| noise 0.0 | $CC_{tgt}$ | left bird (F) |
| | $CC_{vis}$ | white cow (A) |
| | $CC_{sym}$ | cow on left (A) |
| noise 0.5 | $CC_{tgt}$ | left giraffe (F) |
| | $CC_{vis}$ | left cow (A) |
| | $CC_{sym}$ | cow on left (A) |
| noise 1.0 | $CC_{tgt}$ | left guy (F) |
| | $CC_{vis}$ | cow on left (A) |
| | $CC_{sym}$ | cow on left (A) |

| | | |
|---|---|---|
| noise 0.0 | $CC_{tgt}$ | left one (O) |
| | $CC_{vis}$ | top microwave (A) |
| | $CC_{sym}$ | left stove (F) |
| noise 0.5 | $CC_{tgt}$ | left clock (F) |
| | $CC_{vis}$ | left microwave (A) |
| | $CC_{sym}$ | stove top (F) |
| noise 1.0 | $CC_{tgt}$ | top left donut (F) |
| | $CC_{vis}$ | left umbrella (F) |
| | $CC_{sym}$ | top left stove (F) |



| | | |
|---|---|---|
| noise 0.0 | $CC_{tgt}$ | left monitor (F) |
| | $CC_{vis}$ | book on left (A) |
| | $CC_{sym}$ | left laptop (F) |
| noise 0.5 | $CC_{tgt}$ | left monitor (F) |
| | $CC_{vis}$ | keyboard on left (F) |
| | $CC_{sym}$ | left laptop (F) |
| noise 1.0 | $CC_{tgt}$ | top left donut (F) |
| | $CC_{vis}$ | left laptop (F) |
| | $CC_{sym}$ | left laptop (F) |

Figure 4: Examples from RefCOCO with expressions generated by CC variants and human judgments (targets are marked red).

implementations using the same licences, and our other code and data using permissive licences.

Apart from this, we relied on scikit-learn (version 1.2.0, Pedregosa et al. 2011) for our statistic analysis and the RefCOCO API (Kazemzadeh et al., 2014; Yu et al., 2016)[8] for computing BLEU and CIDEr scores.

## D Human Evaluation

We conducted a human evaluation in which the adequacy of assigned referent types in English referring expressions was assessed. The annotation guidelines are published in our code repository.

Our annotators were undergrad student assistants from linguistics and computational linguistics, which were paid by the hour according to the applicable pay scale. The annotators were informed about the intended use of their produced data. Along with our code, we publish the fully anonymized raw and aggregated results of the human evaluation.

## E Qualitative Examples for CC

In Section 4.4 we presented expressions generated by all TRF variants and discussed different types of errors in the model outputs. CC responses for the same examples are shown in Figure 4. In general, we observe similar patterns as for TRF, but with some additional errors (especially for $CC_{tgt}$).

[8] https://github.com/lichengunc/refer