Human-Like Distractor Response in Vision-Language Model

Xiaonan Xu
University of Cologne, Germany
xux0@smail.uni-koeln.de

Haoshuo Chen
Nokia Bell Labs, USA

Abstract

Previous studies exploring the human-like capabilities of machine-learning models have primarily focused on pure language models. Limited attention has been given to investigating whether models exhibit human-like behavior when performing tasks that require the integration of visual and language information. In this study, we investigate the impact of tags of semantic, phonological, and bilingual features on the visual question-answering task performance of an unsupervised model. Our findings reveal its similarities with the influence of distractors in the picture-naming task (known as the picture-word-interference paradigm) observed in human experiments: 1) Semantically-related tags have a more negative effect on task performance compared to unrelated tags, indicating a more robust competition between visual and tag information which are semantically closer to each other when generating an answer. 2) Even presenting a partial section (wordpiece) of the originally detected tag significantly improves task performance, with the portion that plays a lesser role in determining the overall meaning of the original tag leading to a more pronounced improvement. 3) Tags in two languages that refer to the same meaning exhibit a symmetrical-like effect on performance in balanced bilingual models. Datasets and code of this project are released at https://github.com/NLPbelllabs/PWI

1 Introduction

Machine learning models possess a broad range of reasoning abilities like humans. Comparing these models to human capabilities can aid in understanding the decision-making process underlying their predictions and ultimately enhance their accuracy. Numerous studies exploring human-like behavior in models primarily focus on language processing alone, revealing that, even though these models may encounter difficulties in certain specialized areas of language, they can attain significant human-like capabilities across diverse linguistic domains (Ettinger, 2020; Rogers et al., 2021). In this study, we expand the scope of the investigation to the field of vision and language, an area that, to the best of our knowledge, has been relatively unexplored from the perspective of the language community (Dobreva and Keller, 2021; Cao et al., 2020a).

In cognitive psychological research on human language production, the paradigm of picture-word interference (PWI) plays a crucial role in understanding how humans access the appropriate words in their mental processes (Bürki et al., 2020; van Maanen et al., 2009; Lupker, 1979). In a typical PWI task, participants are presented with a picture and asked to name the object (target word, e.g., *dog*) depicted in the picture accurately and quickly. Concurrently, a linguistically related distractor word (e.g. *cat* is semantically related) is presented superimposed on the picture. Compared to an unrelated distractor (e.g. *cap*), a related distractor can either interfere with or facilitate the process by which humans produce the correct response. The PWI is employed to simulate the cognitive process of selecting the most appropriate representation of a word (target) from multiple possibilities stored in long-term memory.

This selection process involved in PWI can be likened to a visual question answering (VQA) task, where a single answer is assigned the highest probability of being the correct response among a set of choices. To achieve successful performance in the task, a vision-language pre-training (VLP) model relies on the inclusion of additional tags, which are accurately detected by an object detector. These tags enhance the visual information extracted from the image, which is particularly important for the unsupervised model which is pre-trained with unaligned text and image corpora (Li et al., 2020). Compared to the baseline cases where no tags are included, the inclusion of tags that possess certain

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 174–185
November 1–4, 2023.

features may enhance task performance, similar to the facilitation effect in the PWI paradigm. Conversely, the presence of tags with some other features may negatively impact task performance, akin to the interference effect in the PWI paradigm.

From this perspective, we examine how tags with different features affect the VQA task performance of unsupervised VisualBERT model (Li et al., 2019, 2020). Our findings reveal similarities between the effects of tags and the influence of distractors observed in PWI experiments with human participants: A) Semantically-related tags have a more pronounced negative impact on task performance compared to unrelated tags. B) A partial section (wordpiece) of the detected tags improves task performance. C) Tags in two languages that refer to the same meaning exhibit a symmetrical-like effect on a balanced bilingual model which is fine-tuned in both languages. Additionally, our results indicate that: D) When visual and tag information are semantically closer, there is a heightened competition between them to be chosen as the final answer. E) The portion of the tag that has a lesser role in determining its overall meaning contributes to a more substantial improvement in task performance.

We outline three main contributions of our study: Firstly, we extend the scope of investigation on human-like intelligence beyond pure language models to include models that integrate visual and language information. Secondly, our findings indicate that the model's performance demonstrates some degree of similarity to human cognitive abilities under various distractor conditions. Thirdly, our study highlights the impact of tag quality on the effectiveness of the model and underscores the need for careful attention to this aspect during the model design phase.

2 Related Work

Object Tag Previous studies have demonstrated that incorporating object tags can improve performance in various vision-language tasks, including VQA (Fang et al., 2021; Cho et al., 2021; Zhang et al., 2021b; Wang et al., 2020), image captioning (Hu et al., 2021; Zhang et al., 2021a; Hu et al., 2020) and visual commonsense reasoning (Lin et al., 2019). However, there is still a lack of comprehensive understanding regarding the influence of tags on the task. In this study, we explore the impact of tags with different features on the performance of unsupervised VisualBERT.

Human-level Intelligence Numerous studies have been conducted to examine the linguistic capabilities of pre-trained transformer-based language models, exploring their resemblance to human abilities in various aspects such as syntactic knowledge (Linzen et al., 2016; Gulordava et al., 2018), semantic knowledge (Ettinger, 2020; Kementchedjhieva et al., 2021; Misra et al., 2020), and the integration of semantic and syntactic information (Xu and Chen, 2022). However, there is still limited knowledge about the linguistic capabilities of VLP models in relation to human behavior (Dobreva and Keller, 2021; Cao et al., 2020a). Our study seeks to expand the current understanding by investigating whether the impact of tags with various features on the VQA task is comparable to the effect of distractors in the PWI task.

3 PWI and Prediction

3.1 Semantic Relatedness

Inspiration PWI research has suggested that a distractor that shares a closer semantic relationship with the target word tends to have a more substantial interference effect on lexical selection. For instance, many studies (Vigliocco et al., 2004; Vieth et al., 2014; Aristei and Rahman, 2013; Rose et al., 2019; cf. Hutson and Damian, 2014) demonstrate that the time taken to name a target word (e.g., *dog*) is longer when presented with a distractor from the same category (e.g. *cat*), compared to an unrelated distractor (e.g. *cap*).

The interference effect caused by a semantically-related distractor can also be observed in pre-trained language models. Misra et al. (2020) investigated BERT's (Devlin et al., 2018) sensitivity to lexical cues and observed that when a target word is masked within a sentence, e.g., *bacon* in *pork/meteorite. she cooked up some eggs, [MASK], and toast*, the probability of [MASK] being predicted as *bacon* was lower when a semantically related prime word *pork* was present compared to an unrelated word *meteorite*. This finding indicates that a semantically related prime word acts as a negative distractor, causing interference effect in certain situations (see also Kassner and Schütze, 2019).

Prediction Building upon these findings, we hypothesize that the presence of semantically-related tags, i.e., from the same semantic category, leads to a more pronounced negative effect on VQA task performance compared to unrelated ones.

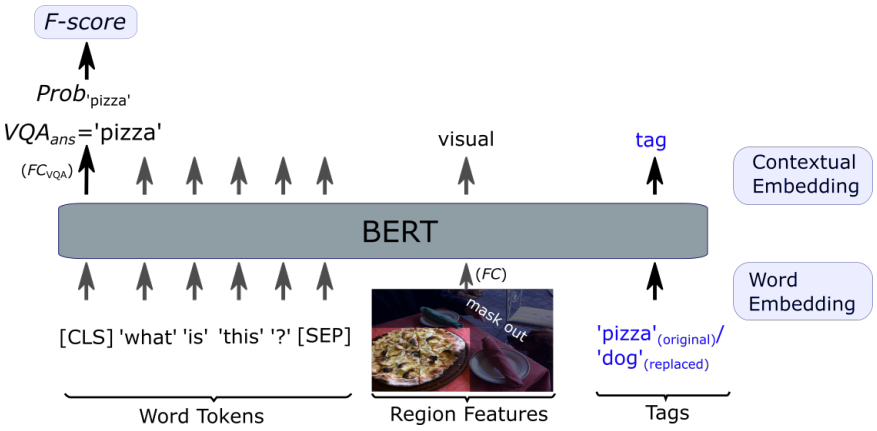

Figure 1: Illustration of unsupervised VisualBERT architecture in the fine-tuning and for inference. (FC: fully connected layer)

3.2 Phonological Relatedness

Inspiration In contrast to the interference effect from semantically-related distractors, phonologically-related ones contribute positively in the PWI task performance (Meyer and Schriefers, 1991; Schriefers, 1999; Ayora et al., 2011). For example, distractors sharing either the first syllable (e.g., *ha-vik*) or the second syllable (e.g., *zo-mer*) to disyllabic target words (e.g., *hamer* in Dutch) show a facilitation effect compared to unrelated distractors (Meyer and Schriefers, 1991). Moreover, the facilitation effect is stronger when the first syllable is shared compared to the second one[1] (Meyer and Schriefers, 1991; Schriefers, 1999).

Prediction We use the originally detected tags consisting of two wordpieces and substitute these tags with a single wordpiece. The replaced tag (a single wordpiece) shares one wordpiece with the original tag, which is in line with the methodology employed in PWI research. We predict that a) both the first and second wordpieces will yield a positive effect on the VQA task in comparison to the cases in which no tags are present, and b) the positive effect will be more prominent with the first wordpiece compared to the second one.

3.3 Bilingual Relatedness

Inspiration PWI research has suggested that the effect size of distractors varies to different degrees between balanced bilinguals, who are highly proficient in two languages, and dominant bilinguals, who exhibit higher proficiency in one language compared to the other. For example, for balanced bilinguals, no significant difference in effect size is observed concerning the languages in which the distractors are presented (Costa et al., 1999; Guo and Peng, 2006). This indicates a symmetrical pattern for balanced bilinguals, i.e., the effect size is consistent regardless of the language used for the distractors. Similarly, research on multilingual language models also suggests that they are capable of aligning word meaning across languages (Cao et al., 2020b; K et al., 2019; Wang et al., 2019; Schuster et al., 2019). However, such a symmetrical-like pattern is not found for dominant bilinguals. For example, different degrees of effect size were observed between semantically-related distractors in two languages (e.g., *valley* in English vs. *dal* meaning 'valley' in Dutch) when dominant bilinguals naming pictures in their second language with lower proficiency (e.g., *mountain* in English) (Hermans et al., 1998; Altarriba and Mathis, 1997).

Prediction We predict a symmetrical-like effect, i.e., the replaced tags in two languages referring to the same meaning will result in highly similar task performance for a balanced model that is fine-tuned in two languages. This pattern, however, is not expected for a dominant model that is exclusively fine-tuned in a single language.

4 Methods

4.1 VQA Model

Semantic and Phonological Relatedness Following Li et al. (2020), the monolingual English VQA model VQA_{PWI} is pre-trained using "mask-and-predict" objective with unaligned data: shuffled text segments (Conceptual Captions (Sharma et al.,

[1] A stronger effect means that the facilitation effect is consistently present across a wider range of time intervals between the onset of the picture and the presentation of the distractor.

2018) and BookCorpus[2]) and images (Microsoft COCO (Chen et al., 2015)). We use the Adam optimizer (Kingma and Ba, 2014) with a linear-decayed learning-rate schedule and pre-train the model for 10 epochs with a batch size of 144. In each batch, part of the text or part of the image regions is masked and the model is trained to predict the masked words or the image regions. We use the image region features and associated tags from LXMERT (Tan and Bansal, 2019), which are extracted and detected using Faster R-CNN (Anderson et al., 2018). These tags are appended as words to the visual input and the mask-and-predict objective is also applied to the tags.

Fine-tuning for the VQA downstream task (Goyal et al., 2017) is conducted using training questions that are related to the images taken from Microsoft COCO (Chen et al., 2015). To avoid intervention from factors such as unrelated image features, we only retained the region features that correspond to the correct answers as the visual input. The remaining region features were masked out by setting the unrelated feature vectors to zero. As shown in Figure 1, the input is composed of a question, an image feature that is detected as the *pizza* object, and the correctly detected tag *pizza*. To enhance the model's understanding, the replaced tags are the correct answers from other instances.

Bilingual Relatedness We include the German Wikipedia text corpus[3] in the pre-training of all three bilingual VQA models: VQA$_{EN\&DE}$, VQA$_{EN}$, and VQA$_{DE}$. In order to align the English and German text embeddings with the visual representations, we use the *googletrans*[4] tool to translate the object tags of 50% of the images into German and a multilingual tokenizer[5] to support both languages during pre-training. During the fine-tuning phase for the VQA task, VQA$_{EN}$ only uses the original tags in English detected in the images, while VQA$_{DE}$ only utilizes the translated tags in German. For VQA$_{EN\&DE}$, tags applied during fine-tuning are either in English or German for each image, which is expected to provide a better alignment between both languages.

[2] https://github.com/jackroos/VL-BERT/blob/master/data/PREPARE_DATA.md

[3] https://github.com/t-systems-on-site-services-gmbh/german-wikipedia-text-corpus

[4] https://pypi.org/project/googletrans/

[5] https://huggingface.co/bert-base-multilingual-uncased

4.2 Instance Selection

To maximize the effect of replaced tags and simplify the result analysis, we collect specific instances from the training dataset that meet the following two criteria: 1) the model outputs the detected tag with the highest probability as the final answer, 2) at least one region feature from the input image is detected as the answer object. Figure 1 provides an example instance, where the question *What is this?* is asked about an image COCO_train2014_000000074253[6] that contains a pizza object and the word *pizza* is detected as a tag. The VQA Annotations[6] correctly label *pizza* as the answer, and it is also identified as the output with the highest probability. The probability of *pizza* may decrease and an incorrect response might be generated as the final answer if no tag is present or a different tag such as *dog* is given.

4.3 Measures

Accuracy There exists a number of valid instances that identify one tag t_o^i in N original tags $t_o^0, ..., t_o^N$ as their correct answer. We replace each original tag t_o^i with N different new tags $t_r^0, ..., t_r^N$. A set V_i of instances that have t_o^i as their correct answer is collected with a number of m_i in total. The function $C_{VQA}()$ is used to count the number of instances where t_o^i remains selected as the correct answer after it is replaced by t_r^i. Note that t_o^i is the correct answer for all instances in V_i and the accuracy value with the original tags is 100%. This value will decrease with replaced tags or without any tags. A lower accuracy value indicates a greater impact of tag replacement on image-text alignment.

$$Accuracy = \sum_{i=1}^{N} C_{VQA}(t_r^i, V_i) / \sum_{i=1}^{N} m_i \quad (1)$$

$\mathbb{F}$-score and Similarity We use $\mathbb{F}$-score to further examine the degree of change in the probability of a correct answer caused by a replaced tag for each instance in the experiment on semantic- and bilingual-relatedness. For each valid instance v, where the original tag t_o is the labeled correct answer, we define surprisal $\mathbb{S}$ as the negative logarithm of the probability of the model outputting the correct answer ($VQA_{ans} = t_o$). The surprisal $\mathbb{S}o(v)$ (Eq. 2) is computed when the original tag t_o is used as the answer. When a replaced tag t_r is present, the corresponding surprisal $\mathbb{S}r(v)$ is calculated using Eq. 3. The $\mathbb{F}$-score (Eq. 4) measures

[6] https://visualqa.org/download.html

		Accuracy				Estimate (intercept)	
		$\mathbf{T_{orig}}$	$\mathbf{T_{none}}$	$\mathbf{T_{diff}}$	$\mathbf{T_{same}}$	Sim_{word}	$Sim_{contextual}$
food	*Acc*	100	79.1	66.0	58.8	-1.083***(2.103)	-1.8813***(2.982)
	example	*banana*	/	*cow*	*mango*		
animal	*Acc*	100	88.9	71.3	56.6	-0.6924***(1.177)	-3.011***(3.014)
	example	*dog*	/	*squash*	*cat*		

Table 1: (left) *Accuracy* (%) of the VQA task and examples for semantically-related conditions for the list food and the list animal. (right) Results of the linear mixed model, where cosine similarity (Sim_{word}/$Sim_{contextual}$) is treated as a fixed effect together with the intercepts of items in each list as random effects, using the formula: $\mathbb{F}$-score $\sim$ cosine similarity + (1|item), ***: $p < .001$.

the difference between $\mathbb{S}r(v)$ and $\mathbb{S}o(v)$, indicating the extent of the impact of t_r compared to t_o on the probability of a correct answer. If the replacement of t_o with t_r leads to a greater decrease in the probability, then the $\mathbb{F}$-score increases[7].

$$\mathbb{S}_o(v) = -log_e(prob(VQA_{ans} = t_o|(v, t_o))) \quad (2)$$

$$\mathbb{S}_r(v) = -log_e(prob(VQA_{ans} = t_o|(v, t_r))) \quad (3)$$

$$\mathbb{F}(v) = \mathbb{S}_r(v) - \mathbb{S}_o(v) \quad (4)$$

The $\mathbb{F}$-scores are analyzed in relation to the semantic distance between the replaced tags and the original ones. The vector similarity Sim between t_o and t_r is calculated using cosine similarity $cos()$. We define two types of similarities, word (non-contextual) similarity Sim_{word} and contextual similarity $Sim_{contextual}$, as:

$$Sim_{word}(v) = cos(e_o(t_o), e_o(t_r)) \quad (5)$$

$$Sim_{contextual}(v) = cos(e_1(t_o, v), e_1(t_r, v)) \quad (6)$$

The function $e_o()$ in Eq. 5 returns the token embeddings of t_o and t_r at the 0^{th} layer of the model, while Eq. 6 uses $e_1()$ to generate contextualized embeddings of t_o and t_r from the last (12^{th}) hidden layer for instance v.

5 Experiments

A comprehensive list of tags used in the three experiments can be found in Table 4 in the Appendix.

5.1 Semantic Relatedness

Setup To investigate whether the effect of semantically-related features on task performance can apply to various categories, we conduct the experiment using two word lists representing animals and food, respectively. The VQA performance is tested using different types of tags, including originally detected tags (T_{orig}), semantically-related tags from the same category (T_{same}), unrelated tags from a different category (T_{diff}) (using food words for animal images and vice versa), and without any tags (T_{none}). Examples of the types are shown in Table 1. Two additional types are introduced to expand the spectrum of semantic similarity alongside the existing tags: tags of hypernyms (T_{hype}) (Kuipers et al., 2006) (using *food* for the list food and *animal* for the list animal) and pseudowords (T_{pseu}) which are made up of random letters and have no meaning in English. To avoid any influence from wordpiece segmentation, the tags cannot be divided into subwords within the word embedding.

Result The task accuracy presented in Table 1(left) shows the same trend for both lists: T_{orig} > T_{none} > T_{diff} > T_{same}. This finding strongly supports the prediction that the tags from the same semantic category negatively impact task performance more significantly compared to unrelated ones.

To determine if the interference effect is linked to the semantic similarity between the original and replaced ones, we plot the $\mathbb{F}$-score against Sim_{word}/$Sim_{contextual}$ for all instances with the tags T_{hype}, T_{pseu}, T_{diff}, and T_{same}. Figure 2a illustrates that the tags with lower Sim_{word}/$Sim_{contextual}$ tend to have a more scattered distribution towards higher $\mathbb{F}$-scores for the list animal. A consistent tendency can be found for the list of food in Figure 4 in the Appendix. We apply a Gaussian distribution to fit the Sim for each $\mathbb{F}$-score and plot the mean of these fits against the $\mathbb{F}$-scores for the list food and the list animal in Figure 2b. Importantly, we evaluate the statistical significance of the relationship between $\mathbb{F}$-scores and Sim_{word}/$Sim_{contextual}$, and found significant

[7]The concept of surprisal and $\mathbb{F}$-score have similarities with that defined in Misra et al. (2020), where surprisal is based on the probability of a masked token instead.

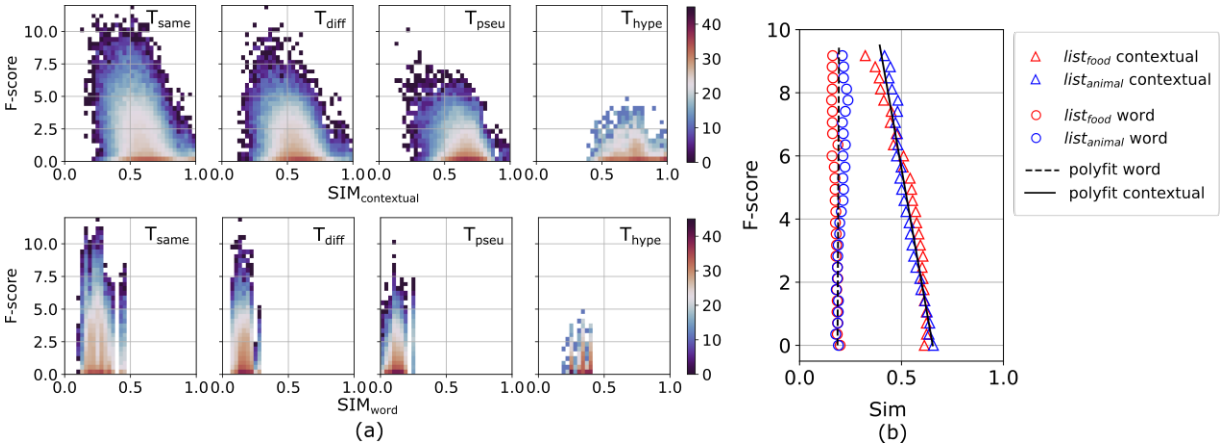

Figure 2: (a) 2D histogram displaying the $\mathbb{F}$-scores by $Sim_{contextual}$ (upper) and by Sim_{word} (lower) for the list animal; (b) $\mathbb{F}$-score versus Sim_{word}/$Sim_{contextual}$ from Gaussian fitting the 2D histogram for the list animal in (a) and for the list food in Figure 4 in Appendix.

coefficients in a linear mixed model with Sim as an explanatory variable and the $\mathbb{F}$-score as a dependent variable, as shown in Table 1(right). The statistical findings and the evident negative correlation between the $\mathbb{F}$-scores and $Sim_{contextual}$ in Figure 2b, collectively indicate that semantically closer tags result in a less pronounced decrease in the probability correct answers.

5.2 Phonological Relatedness

Setup To investigate whether the impact of phonologically-related features on VQA task performance can be applied to various word categories, we find instances with tags of three different types of word compounds: open compound words (C_{open}), which are composed of two words written separately with a space like *tennis player*, closed compound words (C_{closed}), which are composed of two words written together as a single word like *bathtub*, and non-compound words (C_{non}), which can not be divided into two words, such as *buoy*. We expect the predicted positive effect to be consistent for all three types of words. All the words can be split into two wordpieces using the tokenizer from the BERT base model (uncased)[8]. The first and second parts are labeled as T_{1st} and T_{2nd}, respectively. Table 2 shows examples of these tags and T_{none} refers to the cases where no tags are used.

Result Table 2 demonstrates that the task accuracy for the three types of words consistently follows this order: $T_{1st} > T_{2nd} > T_{none}$. This result supports the prediction that both wordpieces as tags contribute to improved task performance

[8] https://huggingface.co/bert-base-uncased

list	info	T_{orig}	T_{1st}	T_{2nd}	T_{none}
C_{open}	example	*tennis player*	*tennis*	*player*	/
	Acc	100	89.9	82.7	68.9
C_{closed}	example	*bathtub*	*bath*	*tub*	/
	Acc	100	91.2	88.4	77.3
C_{non}	example	*buoy*	*buo*	*y*	/
	Acc	100	96.3	95.9	69.1

Table 2: Examples and *Accuracy* (%) of VQA task for each phonologically-related condition within the three groups C_{open}, C_{closed} and C_{non}.

compared to the cases without tags, and that the first wordpiece has a stronger effect than the second wordpiece.

Interestingly, the result reveals a consistent trend in the accuracy values of T_{1st} and T_{2nd}: both exhibit the order of $C_{non} > C_{closed} > C_{open}$. In contrast to the groups C_{closed} and C_{open}, T_{1st}/T_{2nd} in the group C_{non} are wordpieces that do not have independent meaning. Their higher accuracy (96.3 for T_{1st} and 95.9 for T_{2nd}) compared to the T_{1st} and T_{2nd} for both C_{closed} (91.2 and 88.4, respectively) and C_{open} (89.9 and 82.7, respectively) suggests the need to consider linguistic factors in the observed trend, which will be discussed in the Discussion section.

5.3 Bilingual Relatedness

Setup The dominant bilingual models VQA_{EN} and VQA_{DE} are fine-tuned with tags in English and German, respectively. Their performance is tested with original tags in the corresponding language (T_{EN} for VQA_{EN} and T_{DE} for VQA_{DE}), and with replaced tags translated into the other language (T_{EN} for VQA_{DE} and T_{DE} for VQA_{EN}). For the balanced bilingual model $VQA_{EN\&DE}$ which is fine-tuned with an equal distribution of 50% En-

model	info	Accuracy					Estimate (intercept)	
		T_{EN}	T_{DE}	T_{none}	$T_{diff\text{-}DE}$	$T_{diff\text{-}EN}$	Sim_{word}	$Sim_{contextual}$
$VQA_{EN\&DE}$	*Acc*	100	96.6	67.3	61.6	61.0	-4.512***(1.866)	-4.913***(4.970)
VQA_{EN}	*Acc*	100	77.3	50.7	43.4	37.4	-6.265***(3.299)	-9.884***(8.895)
VQA_{DE}	*Acc*	90.3	100	44.4	39.5	40.6	-10.308***(4.089)	-7.465***(7.603)
	example	*car*	*Wagen*	/	*Hund*	*dog*		

Table 3: (left) *Accuracy* (%) of VQA task and examples for the bilingually-related conditions. (right) Results of the linear mixed model, where cosine similarity (Sim_{word}/$Sim_{contextual}$) is treated as a fixed effect together with the intercepts of items in each list as random effects, using the formula: $\mathbb{F}$-score $\sim$ cosine similarity + (1|item), ***: $p < .001$.

glish tags and 50% German tags, all the original tags are translated into the other language. We additionally add the types $T_{diff\text{-}DE}$ and $T_{diff\text{-}EN}$, which present the tags in a different category in German and in English, respectively. The type T_{none} refers to the cases without tags. Table 3 showcases one example for each type.

Result The results in Table 3 show that the tags referring to the same meaning as the original tags (T_{EN}/T_{DE}) achieve superior task performance compared to unrelated tags ($T_{diff\text{-}EN}$/$T_{diff\text{-}DE}$) and cases without tags (T_{none}). Importantly, $VQA_{EN\&DE}$ shows smaller accuracy differences between a) T_{EN} and T_{DE} (3.4) and b) $T_{diff\text{-}EN}$ and $T_{diff\text{-}DE}$ (0.6) than that for VQA_{EN} (22.7 and 5, respectively) and VQA_{DE} (9.7 and 1.1, respectively). This supports the prediction that tags in two languages referring to the same meaning result in similar task performance for the balanced model compared to the dominant model. This symmetrical-like effect is also supported by the 2-D histogram plot of the corresponding $\mathbb{F}$-scores of $T_{diff\text{-}DE}$ and $T_{diff\text{-}EN}$, see Figure 3(a, top) where balanced model $VQA_{EN\&DE}$ shows a clear diagonal distribution. Similarly to the result in the experiment on semantic relatedness, the significant negative coefficients of Sim_{word}/$Sim_{contextual}$ against $\mathbb{F}$-scores in Table 3(right) and the clear negative correlation between $\mathbb{F}$-score and $Sim_{contextual}$ in Figure 3(b) further reinforce the suggestion that the semantically closer tags result in a smaller decrease in the probability of correct answers.

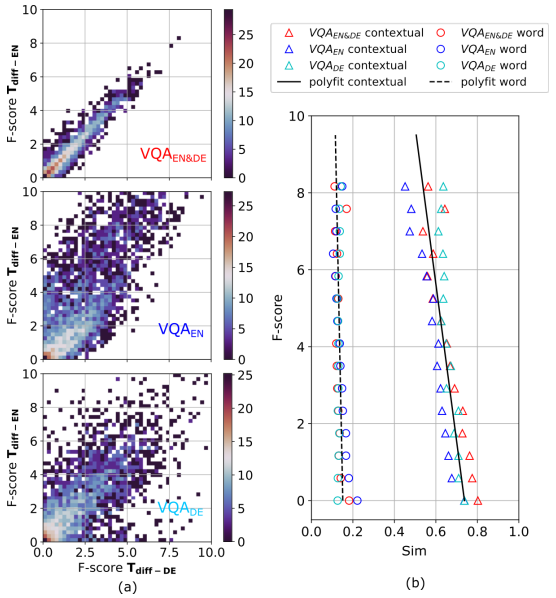

Figure 3: (a) 2D histogram displaying the $\mathbb{F}$-scores by $T_{diff\text{-}DE}$ and $T_{diff\text{-}EN}$ for model $VAQ_{EN\&DE}$ (top), VQA_{EN} (middle) and VQA_{DE} (bottom). (b) $\mathbb{F}$-scores versus Sim_{word}/$Sim_{contextual}$ from Gaussian fitting the 2D histogram in (a).

6 Discussion

Our study reveals that the influence of the tested tags on VQA task performance parallels the effects of distractors observed in PWI experiments on human participants. We will discuss the experimental results from the following perspectives.

Competition between semantically closer visual and tag information hinders performance. On the one hand, similar to the human performance in the PWI task, the experimental results on semantic relatedness (Table 1(left)) demonstrate that replacing tags within the same category leads to a lower accuracy compared to the unrelated tags of a different category. On the other hand, both the experiment on semantic (Table 1(right)) and bilingual relatedness (Table 3(right)) reveal a negative correlation between semantic similarity (Sim) and $\mathbb{F}$-scores, suggesting that semantically closer tags cause a smaller change in the probability of correct answers. The two metrics, despite their differences, are not contradictory. In the PWI paradigm, the processing of an image activates the representation of the target and related concepts (Bürki et al., 2020). Likewise, the results indicate that the image feature in the VQA task activates information about the correct answer and closely related objects. These related objects are more likely to be chosen as the final answer compared to unrelated ones. When a semantically-related tag is present, it tends to increase the likelihood of the model choosing itself as the final, but incorrect answer, thus leading to

reduced accuracy. In this case, the probability of the correct answer remains high due to the closer similarity between visual and tag information, resulting in a low $\mathbb{F}$-score. In contrast, an unrelated tag does not receive significant activation from the image feature, making it less likely to be chosen as the final answer, even when provided as a tag. As the visual and tag information diverges from each other, the probability of the correct answer is expected to be lower, resulting in a higher $\mathbb{F}$-score, compared to the cases in which the visual and tag information align closely.

Wordpiece partially representing visual objects enhances performance. In PWI experiments, sharing syllables between distractors and target words facilitates the picture-naming process. Consistent with this finding, the experiment on phonological relatedness (Table 2) shows that task performance is significantly improved when either T_{1st} or T_{2nd} are included as input. These results suggest that presenting a single wordpiece, regardless of its semantic meaning, aids the model in selecting the final correct answer, subsequently enhancing task performance.

Wordpiece determining overall word meaning may also compete with visual information. We observed that the first wordpieces tend to achieve better performance compared to the second wordpieces (Table 2). We attribute this to the greater weight placed on the second wordpieces in determining the overall word meaning, particularly evident in the cases of C_{open} like *tennis player* and C_{closed} like *bathtub*. Typically, the first wordpiece serves as a modifier or specifier, while the second wordpiece carries the central meaning of the words, representing the main object being referred to. For example, the first wordpiece *tennis* acts as a modifier in the compound *tennis player*, while the second wordpiece *player* represents the main object. The second wordpieces are more closely associated with the original tags, which can potentially lead to competition between visual and tag information (as discussed in the previous part) and result in worse task performance.

Compounding ability of wordpiece affects task performance. The result of the experiment on phonological relatedness shows a specific order of accuracy values: $C_{non} > C_{closed} > C_{open}$ for both T_{1st} and T_{2nd}. This pattern can be attributed to the varying degrees of compounding ability between three word groups. In the case of C_{non}, the wordpieces have limited possibilities to form meaningful words, as exemplified by *buo* in *buoy*. The association between the image feature representing a buoy and the wordpiece *buo* significantly narrows down the available options for the model, strongly indicating the correct answer as *buoy*. In contrast, a single wordpiece or lexeme in C_{open} offers more opportunities to create related words, such as *tennis* in *tennis player* can be a standalone word, or form words like *tennis racket*, *tennis coach*, *tennis ball*, etc. This offers the model a broader range of options to choose from, considering the image feature representing a tennis player and the tag *tennis*. Thus, the model is likely to make more accurate predictions when the answer options are limited (C_{non}) compared to when there is a wider range of options (C_{closed} and C_{open}), leading to the observed performance order of $C_{non} > C_{closed} > C_{open}$ for both T_{1st} and T_{2nd}.

7 Conclusion

Cognitive psychological research on PWI demonstrates that different distractors have varying effects on picture-naming tasks performed by human participants. We replace the tags detected in images with new words possessing semantic, phonological, and bilingual characteristics relative to the original tags, and examine their impact on the VQA task performance. Our findings indicate that the influence of these tags on task performance parallels the effects of distractors in PWI experiments on human participants.

Taking the task performance in cases where no tags are present as the baseline, we found that A) semantically-related tags have a greater negative impact on task performance compared to unrelated ones, suggesting that when visual and tag information is semantically closer to each other, they compete more strongly to be selected as the final answer. B) Presenting even a portion (wordpiece) of the original tag improves task performance significantly. The portion that plays a lesser role in determining its original tag's overall meaning leads to a more significant improvement. C) Tags in two languages referring to the same meaning exhibit a symmetrical-like effect on performance in balanced bilingual models which is fine-tuned in both languages. However, Similar behavior is not observed in dominant bilingual models that are fine-tuned in only one language.

For future work, we will explore additional prob-

ing measures, such as attention probing, to gain a deeper understanding of the internal behavior of tags in the VQA task. We will also investigate whether the observed effects of tags in this study can be generalized to other VLP models.

8 Limitations

We acknowledge several limitations in our study. First, the replaced tags used in our experiments are originally detected tags from other instances, which ensures the models' understanding of the replaced tags but restricts the inclusion of additional tag types. One type that warrants further investigation is the use of synonyms as tags, such as using *puppy* as a replacement for *dog*. It would also be valuable to examine the symmetry-like effect in the experiment on bilingual relatedness by using semantically related tags, e.g., from the same category.

Second, even though a majority of the questions used in our study focus on object identification such as *what is this?*, there is a small number of questions that involve additional object entities. For instance, *What is floating near the bird?* and *What is in the water?*. *Bird* and *water* in the questions may act as distractors and potentially affect the model's performance. Despite this potential impact, we chose to retain these questions in the study due to their limited quantity.

Ethics Statement

In adherence to their license agreements, we use publicly available resources in our experiments. The datasets have undergone complete anonymization, ensuring that they do not include any personal information regarding the caption annotators or any data that could expose the identities of the subjects photographed.

References

Jeanette Altarriba and Katherine M Mathis. 1997. Conceptual and lexical development in second language acquisition. *Journal of memory and language*, 36(4):550–568.

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.

Sabrina Aristei and Rasha Abdel Rahman. 2013. Semantic interference in language production is due to graded similarity, not response relevance. *Acta Psychologica*, 144(3):571–582.

Pauline Ayora, Francesca Peressotti, F-Xavier Alario, Claudio Mulatti, Patrick Pluchino, Remo Job, and Roberto Dell'Acqua. 2011. What phonological facilitation tells about semantic interference: A dual-task study. *Frontiers in psychology*, 2:Article 57.

Audrey Bürki, Shereen Elbuy, Sylvain Madec, and Shravan Vasishth. 2020. What did we learn from forty years of research on semantic interference? a bayesian meta-analysis. *Journal of Memory and Language*, 114:104125.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020a. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020b. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Albert Costa, Michele Miozzo, and Alfonso Caramazza. 1999. Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and language*, 41(3):365–397.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Radina Dobreva and Frank Keller. 2021. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Taomei Guo and Danling Peng. 2006. Event-related potential evidence for parallel activation of two languages in bilingual speech production. *NeuroReport*, 17(17):1757–1760.

Daan Hermans, Theo Bongaerts, Kees De Bot, and Robert Schreuder. 1998. Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: language and cognition*, 1(3):213–229.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. VIVO: Surpassing human performance in novel object captioning with visual vocabulary pre-training.

James Hutson and Markus F Damian. 2014. Semantic gradients in picture-word interference tasks: Is the size of interference effects affected by the degree of semantic overlap? *Frontiers in Psychology*, 5:Article 872.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: An empirical study. *arXiv preprint arXiv:1912.07840*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised mary because he? implicit causality bias and its interaction with explicit cues in LMs. *arXiv preprint arXiv:2106.01060*.

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

Jan-Rouke Kuipers, Wido La Heij, and Albert Costa. 2006. A further look at semantic context effects in language production: The role of response congruency. *Language and Cognitive Processes*, 21(7-8):892–919.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2020. Unsupervised vision-and-language pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*.

Jingxiang Lin, Unnat Jain, and Alexander Schwing. 2019. TAB-VCR: Tags and attributes based visual commonsense reasoning baselines. *Advances in Neural Information Processing Systems*, 32.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Stephen J Lupker. 1979. The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, 7(6):485–495.

Antje S Meyer and Herbert Schriefers. 1991. Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(6):1146–1160.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. *arXiv preprint arXiv:2010.03010*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Benjamin Rose, Sabrina Aristei, Alissa Melinger, and Rasha Abdel Rahman. 2019. The closer they are, the more they interfere: Semantic similarity of word distractors increases competition in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(4):753–763.

Herbert Schriefers. 1999. Phonological facilitation in the production of two-word utterances. *European Journal of Cognitive Psychology*, 11(1):17–50.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2556–2565.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Leendert van Maanen, Hedderik van Rijn, and Jelmer P Borst. 2009. Stroop and picture—word interference are two sides of the same coin. *Psychonomic Bulletin & Review*, 16(6):987–999.

HE Vieth, KL McMahon, and GI de Zubicaray. 2014. Feature overlap slows lexical selection: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology*, 67(12):2325–2339.

Gabriella Vigliocco, David P Vinson, William Lewis, and Merrill F Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4):422–488.

Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. MiniVLM: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. *arXiv preprint arXiv:1909.06775*.

Xiaonan Xu and Haoshuo Chen. 2022. Who did what to whom? Language models and humans respond diversely to features affecting argument hierarchy construction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 254–265.

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. 2021a. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3008.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

A Appendix

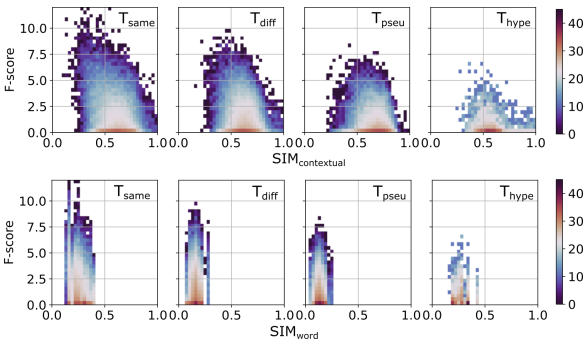

Figure 4: 2D histogram displaying the $\mathbb{F}$-scores by $Sim_{contextual}$ (upper) and by Sim_{word} (lower) for the list food.

list	T_{orig}/T_{same}	T_{diff}	T_{pseu}	T_{hype}
food	pepper(65), 'cabbage'(52), 'tomato'(2028), 'banana'(6903), 'apple'(2925), 'pasta'(338), 'bread'(1092), 'cheese'(378), 'egg'(468), 'chocolate'(715), 'pancakes'(52), 'sandwich'(3796), 'pizza'(18109), 'fries'(754)	'cat', 'dog', 'sheep', 'deer', 'cow', 'horse', 'zebra', 'elephant', 'goose', 'goat', 'bear', 'panda', 'pigeon', 'butterfly'	'san', 'lan', 'ren', 'fen', 'jia', 'cho', 'jon', 'nan', 'gan', 'kam', 'yan', 'abe', 'dia', 'pia'	'food'
animal	'dog'(11778), 'cat'(9386), 'cow'(5408), 'lamb'(117), 'sheep'(7488), 'deer'(91), 'zebra'(5057), 'horse'(8567), 'goat'(221), 'goose'(39), 'elephant'(6773), 'panda'(130), 'bear'(3471), 'butterfly'(26)	'pepper', 'cabbage', 'tomato', 'banana', 'apple', 'pasta', 'bread', 'cheese', 'egg', 'chocolate', 'pancakes', 'sandwich', 'pizza', 'fries'	'san', 'lan', 'ren', 'fen', 'jia', 'cho', 'jon', 'nan', 'gan', 'kam', 'yan', 'abe', 'dia', 'pia'	'animal'

list	T_{orig}
C_{open}	'stop sign'(100), 'knee pads'(18), 'tennis court'(16), 'train station'(75), 'tennis player'(6), 'toilet brush'(21), 'french fries'(104), 'christmas tree'(45), 'living room'(3100), 'teddy bear'(2893), 'shower curtain'(52), 'polar bear'(252), 'hot dog'(4005), 'bow tie'(204), 'soccer ball'(247), 'traffic light'(210), 'parking meter'(1144), 'teddy bears'(230), 'tank top'(480), 'home plate'(52), 'stuffed animals'(448), 'cutting board'(261), 'ski pole'(682), 'palm trees'(99), 'skate park'(1394), 'palm tree'(152), 'water bottle'(80), 'tennis ball'(41), 'cell phone'(42), 'trash can'(1496), 'toilet paper'(1457), 'wine glass'(98), 'stuffed animal'(228), 'fire truck'(1357), 'ski poles'(378), 'parking lot'(512), 'power lines'(198), 'shower head'(70), 'coffee maker'(1387), 'steering wheel'(75), 'ski lift'(3120), 'mouse pad'(81), 'hot dogs'(2132), 'traffic lights'(166), 'bus stop'(170)
C_{closed}	'dugout'(12), 'bathtub'(38), 'flamingo'(20), 'sailboat'(60), 'racket'(114), 'wetsuit'(854), 'skateboard'(5355), 'blender'(520), 'surfboard'(7007), 'snowboard'(2616), 'goggles'(780), 'wristband'(323), 'cupcake'(40), 'toaster'(230), 'surfboards'(240), 'blueberries'(325), 'stroller'(260), 'headband'(783), 'hoodie'(180), 'steeple'(324), 'toothbrush'(6623), 'pineapple'(39), 'pickle'(1560), 'headphones'(387), 'frosting'(276), 'tablecloth'(47), 'crosswalk'(384), 'snowsuit'(49), 'pepperoni'(6222), 'pickles'(104), 'blueberry'(159), 'sticker'(54), 'planter'(56), 'watermark'(59), 'spatula'(183), 'beanie'(63), 'wallpaper'(207), 'earring'(71), 'strawberries'(1725), 'spinach'(11324)
C_{non}	'urinal'(57), 'tarmac'(52), 'visor'(480), 'donut'(4620), 'donuts'(7339), 'bib'(294), 'kayak'(186), 'tarp'(65), 'buoy'(320)

model	T_{EN}	T_{DE}	$T_{diff\text{-}EN}$	$T_{diff\text{-}DE}$
$VQA_{EN\&DE}$ & VQA_{EN}	'car'(510), 'train'(649), 'airplane'(48), 'boat'(152), 'television'(57), 'clock'(357), 'phone'(780), 'camera'(105), 'dog'(535), 'cow'(484), 'tree'(567), 'mountain'(35)	'wagen', 'zug', 'flugzeug', 'boot', 'fernsehen', 'uhr', 'telefon', 'kamera', 'hund', 'kuh', 'baum', 'berg'	'cat', 'dog', 'sheep', 'deer', 'cow', 'horse', 'zebra', 'elephant', 'goose', 'goat', 'bear', 'panda', 'pigeon', 'butterfly'	'hund', 'kuh', 'ente', 'adler', 'mais', 'stein', 'blatt', 'zeitschrift', 'wagen', 'zug', 'wein', 'haar'
VQA_{DE}	'car', 'train', 'airplane', 'television', 'clock', 'phone', 'camera', 'dog', 'cow', 'tree', 'mountain'	'wagen'(483), 'zug'(447), 'flugzeug'(2), 'fernsehen'(4), 'uhr'(236), 'telefon'(429), 'kamera'(69), 'hund'(778), 'kuh'(337), 'baum'(447), 'berg'(95)	'dog', 'cow', 'duck', 'corn', 'stone', 'leaf', 'magazine', 'car', 'train', 'wine', 'hair'	'hund', 'kuh', 'ente', 'mais', 'stein', 'blatt', 'zeitschrift', 'wagen', 'zug', 'wein', 'haar'

Table 4: Lists of tags used in the semantically-related (top), phonologically-related (middle), and bilingually-related (bottom) experiments. The frequencies of the original tags are shown in parentheses. In the phonologically-related experiment, the wordpieces (examples can be found in Table 2 in the main text) from the original tags are omitted.