HL Dataset: Visually-grounded Description of Scenes, Actions and Rationales

Michele Cafagna[1] **Kees van Deemter**[2] **Albert Gatt**[1,2]

[1]University of Malta, Institute of Linguistics and Language Technology
[2]Universiteit Utrecht, Information and Computing Sciences
michele.cafagna@um.edu.mt
{a.gatt, c.j.vandeemter}@uu.nl

Abstract

Current captioning datasets focus on object-centric captions, describing the visible objects in the image, e.g. "people eating food in a park". Although these datasets are useful to evaluate the ability of Vision & Language models to recognize and describe visual content, they do not support controlled experiments involving model testing or fine-tuning, with more high-level captions, which humans find easy and natural to produce. For example, people often describe images based on the type of scene they depict ('people at a holiday resort') and the actions they perform ('people having a picnic'). Such descriptions draw on personal experience and commonsense assumptions. We present the High-Level Dataset [1]; a dataset extending 14997 images from the COCO dataset, aligned with a new set of 134,973 human annotated (high-level) captions collected along three axes: *scenes*, *actions* and *rationales*. We further extend this dataset with confidence scores collected from an independent set of readers, as well as a set of narrative captions generated synthetically, by combining each of the three axes. We describe this dataset and analyse it extensively. We also present baseline results for the High-Level Captioning task.

1 Introduction

Conceptual grounding broadly refers to the idea that symbols (e.g. language) are grounded in perception (Barsalou et al., 2008). Perceptually grounded communication is made possible by the fact that perceptual experiences are largely shared. However, individual experience can also license subjective inferences which inform not just what we express through language, but also what we choose to assume and leave unexpressed (Bisk et al., 2020).

Among the many modalities available in the perceptual spectrum, visual grounding has always been of primary interest as it provides a relatively straightforward way to link linguistic expressions to physical objects. Consistent with this claim, a glance at many widely used datasets and models in image captioning reveals a bias towards 'object-centric' descriptions, whereby models are trained on image-text pairs where the text consists of explicit mentions of objects visible in the scene. However, experience and perception also motivate other, non-object-centric ways of talking about the world, for example, when we talk about scenes, or when we describe actions or their underlying rationales. While such 'high-level' descriptions are also perceptually grounded, they incorporate world knowledge and subjective experience.

For example, the object-centric description in Table 1 certainly describes the visual content, though it is based mainly on the recognition of objects in the scene. By contrast, the three high-level captions (*scene, action, rationale*, from the HL-Dataset described below), provide three different perspectives of the scene among the many possible ones, which are triggered by expectations and assumptions based on subjective experience and world knowledge.

In this work, we tackle the issue of grounding high-level linguistic descriptions in the visual modality, proposing the High-Level (HL) Dataset: a resource for Vision and Language (V&L) modeling which aligns existing object-centric captions with human-collected high-level descriptions of images along three different axes: *scenes, actions* and *rationales*. The high-level captions capture the human interpretation of the scene which are complementary to object-centric captions used in current V&L datasets, e.g. in COCO (Lin et al., 2014). We take a step further, and we collect *confidence scores* from independent annotators, which serve to shed

[1] huggingface.co/datasets/michelecafagna26/hl
github.com/michelecafagna26/HL-dataset

Proceedings of the 16th International Natural Language Generation Conference, pages 293–312
September 11–15, 2023. ©2023 Association for Computational Linguistics

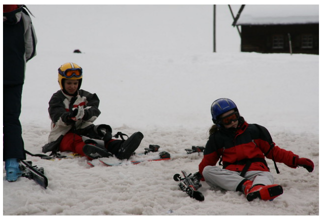
Image	**Axis**	**Caption**
	scene	the picture is shot in a ski resort
	action	they are just relaxing after a round of skiing
	rationale	they want to have a good time together
	object-centric (COCO)	a woman and a boy sitting in the snow outside of a cabin.

Table 1: Example of High-Level captions. It is shown one of the three captions available for the three axes collected: *scene, action, rationale*, combined with the object-centric captions from COCO.

light on the extent to which the high-level captions in the dataset correspond to widely-shared assumptions, or to idiosyncratic interpretations. Finally, we consider the task of generating captions that incorporate these different axes, yielding a more narrative-like description of images. Our contributions are:

- We present and release the HL Dataset, a new V&L resource, grounding high-level captions in images along three different axes and aligned with existing object-centric captions;
- We describe the collection protocol and provide an in-depth analysis of the data;
- We present baselines for the High-Level Captioning task and describe further potential uses for our data.

2 Related work

Hodosh et al. (2013), in their influential work, argue that image captioning is mostly interested in 'conceptual descriptions', which focus on what is actually in the image and differ from the so-called non-visual descriptions, which provide additional background information. This line of thought has been broadly followed in the field, resulting in datasets emphasizing object-centric content in V&L tasks involving text generation, like image captioning (Lin et al., 2014; Sharma et al., 2018; Agrawal et al., 2019) and visual question answering (Antol et al., 2015; Zhu et al., 2016).

For instance, in the instructions used to collect COCO (Lin et al., 2014), the annotators are explicitly asked to mention entities visible in the image. This is beneficial to enhance cross-modal interactions: Zhang et al. (2021) show that improving the visual backbone on object recognition tasks, improves the performance of visio-linguistic models in downstream tasks. Li et al. (2020) show that using object labels to bridge the two modalities improves grounding capabilities of V&L models. Object-centricity is also a feature of widely-used web-scraped datasets: in the Conceptual Captions dataset for instance, Sharma et al. (2018) filtered out all captions which did not overlap with object labels automatically identified by a computer vision model in the corresponding image.

Some efforts have been made to understand how low-level concepts improve generalization capabilities and connect to high-level concepts. Object-centric captions help to improve the generalization over unseen objects (Hu et al., 2021) and play a role in the model understanding of abstract concepts (Cafagna et al., 2022; Wang et al., 2022b). In our work, we are interested in the relations between what Hodosh et al. (2013) refer to as 'conceptual' and 'non-visual' descriptions, which we re-frame as a distinction between low-level (object-centric) and high-level descriptions in multimodal learning. We release a novel dataset to foster research in this direction.

Motivation for the present work is also provided by recent research exploring the visual correlates of inferences, temporal and causal relationships (e.g., Park et al., 2020), which also have implications for generation. In visual storytelling, for instance, a model has to understand actions and interactions among the visually depicted entities (Huang et al., 2016; Hu et al., 2020; Lukin et al., 2018; Hong et al., 2023). Identifying actions is a prerequisite for predicting their motivations or rationales as well as explaining automatically generated descriptions of images (Hendricks et al., 2018). Actions and intention are paramount to performing commonsense and temporal reasoning on visual inputs. Along these lines, Park et al. (2020) creates dynamic stories on top of static images, where the task is to predict priors and subsequent actions and rationales. Our work is similar in spirit, as we align high-level descriptions of *actions* and *rationales* with low-level descriptions of static images.

Some work has also been done to test multimodal model grounding capabilities from a more linguistic perspective. Parcalabescu et al. (2022) build a benchmark to test models on a variety of linguistic phenomena, like spatial relations, counting, existence, etc. Pezzelle et al. (2020) assess the integration of complementary information of V&L models across modalities, while Thrush et al. (2022) test multimodal models on compositional reasoning. In this context, the HL Dataset proposed here can offer another benchmark for V&L models' understanding of high-level descriptions of images. Such descriptions are licensed by the entities depicted in the visual modality and the relationships between them but they do not mention them explicitly.

3 Data

In this section, we describe the protocol used to collect annotations for *scenes, actions* and *rationales* and the subsequent collection of confidence scores through crowdsourcing. Differently from previous works, such as COCO, where human annotators are instructed to be objective and to mention only the objects clearly visible in the picture, we elicit high-level concepts in the form of captions by encouraging the annotators to rely on their subjective interpretation of the image.

3.1 Data collection

The task of collecting high-level descriptions is by nature hard to define and requires a clear and careful formulation, therefore we run a pilot study with the double goal of collecting feedback and fine-tuning the task instructions. Full details of the pilot are reported in Appendix D.

Procedure The participants are shown an image containing at least one human subject and three questions regarding three aspects or axes: *scene*, *actions* and *rationales* i,e. *Where is the picture taken?*; *What is the subject doing?*; and *Why is the subject doing it?* We explicitly ask the participants to rely on their personal interpretation of the scene and add examples and suggestions in the instructions to further guide the annotators. Moreover, differently from other VQA datasets like (Antol et al., 2015) and (Zhu et al., 2016), where each question can refer to different entities in the image, we systematically ask the same three questions about the same subject for each image. See Appendix D for the full instructions and Appendix C for details regarding the annotations costs.

Images As mentioned in Section 1 the COCO dataset has a very explicit object-centric orientation, therefore it provides a good starting point to select images, such that we can couple object-centric and high-level captions in a resource-lean approach. Moreover, the alignment of object-centric and high-level captions permits an investigation of the relationship between them.

We randomly select 14,997 images from the COCO 2014 train-val split. In order to answer questions related to *actions* and *rationales* we need to ensure the presence of a (human) subject in the image. Therefore, we leverage the entity annotation provided in COCO to select images containing at least one person.

The whole annotation is conducted on Amazon Mechanical Turk (AMT). We split the workload into batches in order to ease the monitoring of the quality of the data collected. Each image is annotated by three different annotators, therefore we collect three annotations per axis.

3.2 Confidence Scores

The high-level descriptions are collected by asking the participants to interpret the scene leveraging their personal experience. The element of subjectivity leads us to expect some variation in the resulting descriptions, especially where annotators need to infer actions and rationales. In order to distinguish what can confidently be considered widely-shared, or 'commonsense' descriptions, from more idiosyncratic interpretations, we conduct a separate study where we crowd-source *confidence scores* for each high-level caption. We ask independent participants to score the likelihood of a high-level description given the image and the corresponding question on a Likert scale from 1 to 5. For a detailed example of the form see Figure 23 in Appendix D.

Agreement-based worker selection The confidence scores are collected following the same protocol used to collect the high-level descriptions. Using the data from our pilot study, which was carried out with participants who had been thoroughly briefed on the task, we ran a preliminary qualification task where we employed an *automatic worker selection method* to hire qualified annotators from the crowd-sourcing platform.

Let's consider the participants of the pilot as gold annotators (as they were trained on the task) and their annotations as reference annotations. The inter-annotator agreement computed on the reference annotations can be considered the gold inter-annotator agreement α_{gold} of the task.

We run the qualification task using the same set of items used in the pilot, then for each worker w we re-compute the inter-annotator agreement (Hayes and Krippendorff, 2007), combining the workers and the reference annotations, obtaining α_w. We compute an agreement ratio

$$r = \frac{\alpha_w}{\alpha_{gold}} \tag{1}$$

Then, we select the worker w if $r > t$, where t is a threshold empirically set to 0.5. This is equivalent to choosing workers such that their contribution does not negatively affect α_{gold} by a factor greater than t. In other words, the workers are selected if they are relatively compliant with the gold annotators.

4 Dataset Analysis

In this section, we analyse the captions collected in the High-Level Dataset. To provide insights on the kind of captions collected, we analyse the distribution of the captions across different axes, also comparing them with the object-centric COCO captions[2]. Furthermore, we perform a grammatical error analysis, which we report in Appendix A.1.

4.1 High-Level descriptions

We collected 3 annotations per axis over a set of 14,997 images for a total of 134,973 captions. An example of high-level descriptions aligned with the original object-centric caption from COCO is shown in Table 1. We expect to observe shorter texts in the high-level captions as annotators were not giving highly descriptive details typical of object-centric captions. This is visible in Figure 1, which shows that the length of the high-level captions is roughly half of the object-centric COCO captions. Though shorter, they have a comparable number of unique tokens over all the axes (as reported in Table 2); this suggests that the high-level captions are not repetitive and contain a fair amount of lexical variability. A more detailed comparison of the statistics is reported in Table 2.

Data	# Tok	Avg Len	# Uniq	# Cap
actions	271168	6.02	7326	44991
scenes	233232	5.18	4157	44991
rationales	306396	6.81	8301	44991
HL (tot)	810796	6.00	12296	134973
COCO	857218	11.42	13300	75019

Table 2: HL dataset caption statistics compared the COCO captions (object-centric) for the shared set of images. We report the number of tokens (# Tok), average length (Len), number of unique tokens (# Uniq), and number of captions (# Cap).

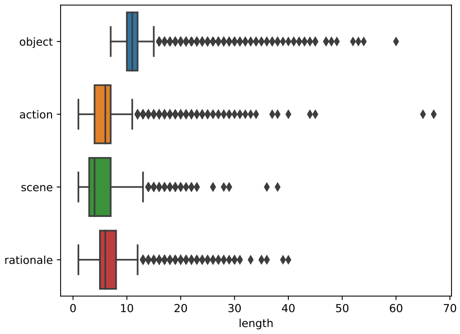


Figure 1: Caption length of the HL captions divided per axis (action, scene, rationale) in comparison to the object-centric COCO captions (object).

Moreover, as already mentioned, the COCO captions are object-centric, that is, these captions are collected to objectively represent the visual content. Although this is convenient in recognition-oriented tasks, they lack the situational knowledge required to contextualize scenes; knowledge that is instead an essential part of the cognitive processes underlying the grounding of language in vision. Indeed, as shown in Figure 2, the most frequent lemmas in the original COCO captions for the images used in the HL Dataset denote mostly objects visible in the picture. The high-level captions represent the same visual content with the addition of situational knowledge coming from the three axes, and this is also visible in different lexico-semantic choices in the texts. For example, Figure 3 shows the most frequent lemmas found in the *scene* axis. Because we align them to the same images, the dataset gives us a clean way to explore the relationship between objects and high-level axes.

Disentangling the content across the axes Asking the same three questions about the same subject for each image allows us to consistently compare the content of our captions across three well-defined axes. We analyse the most frequent nouns

[2] The analysis is performed by using Spacy v.3 pipeline for English using the `en_core_web_md` model to analyse the part of speech of the texts.

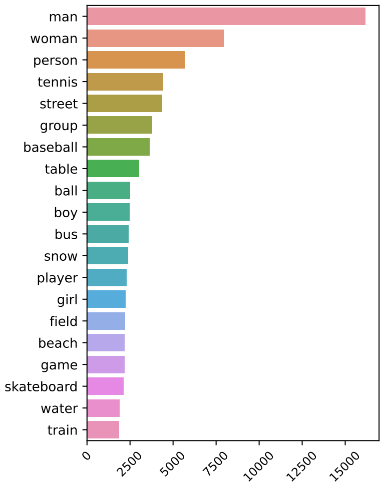


Figure 2: The most frequent nouns in the COCO captions of the shared set of images with the HL dataset. The majority of the terms correspond to physical objects visible in the image.

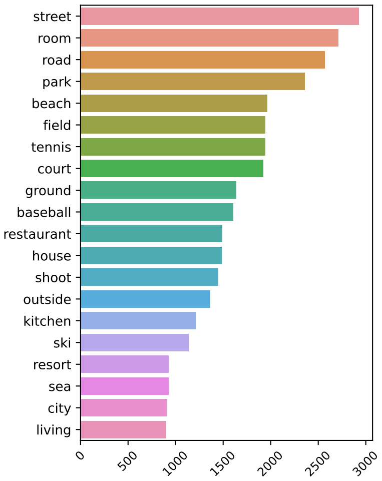


Figure 3: The most frequent lemmas of the captions in the *scene* axis of the HL dataset.

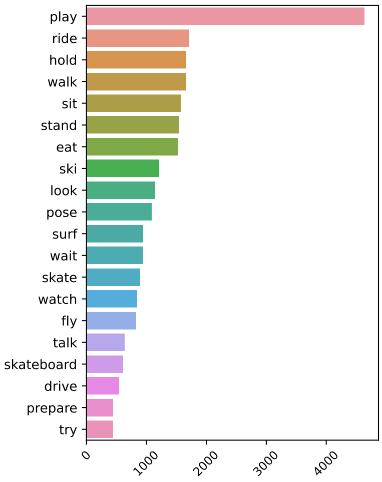


Figure 4: The most frequent verb lemmas of the captions in the *action* axis of the HL dataset.

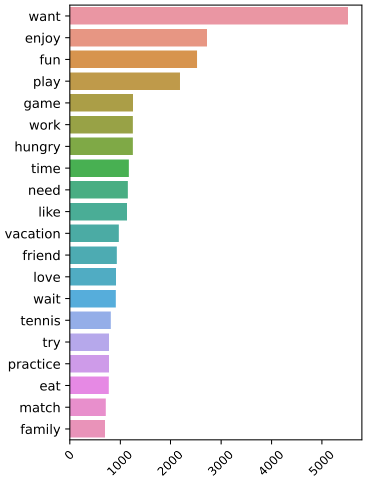


Figure 5: The most frequent noun and verb lemmas of the captions in the *rationale* axis of the HL dataset.

in the *scene* axis in order to characterize the kind of scenes mentioned in the captions collected. The top most frequent scenes include *street, room* and *road*. These are scene types which can encompass a very broad variety of objects. However, we can also identify scenes for which a narrower range of objects would be diagnostic, for example those related to sport activities like *baseball, tennis, ski, ground* and *court*, or domestic environments like *house, kitchen* and *living* (referring to 'living rooms'). For a more complete view see Figure 3 where we report the top 20 most frequent scenes in the HL dataset.

Similarly, we can characterize also the *action* and the *rationale* axes. We identify the *action* distribution by analysing the verbs contained in the captions. In Figure 4 we observe that the most frequent actions are related to sports activities, consistently with what was observed in the *scene* axis distribution. The most frequent verbs are *play, ski, surf, skateboard*, but we can also find generic actions like *hold, walk, sit* and *eat*.

In the *rationale* axis we analyse both nouns and verbs. In this axis we expect to observe more subjectivity and content variability, with more lemmas denoting intents, mental states and events, including psych verbs. Our hypothesis is that the annotators leverage their personal experience to infer these answers to a greater extent than they do for scene descriptions.

The majority of the rationales express intentions; in fact, *want* is by far the most frequent term in the lemmas distribution. As observed with the other two axes, terms related to sports activities are more frequent (*play, game, tennis, practice*), but also related to leisure (*enjoy, fun, vacation, love, family*) along with generic activities (*work, wait, try, eat*). For more details see Figure 5.

The systematic disentanglement of the content along three axes can serve as a filter to identify or analyse sub-samples of the data with specific characteristics. For instance, as observed so far, we can confidently say that sports-related activities are predominant in the dataset.

Connecting high- and low-level concepts One of the main goals of this resource is to enable the discovery of connections between high- and low-level captions, that are, descriptions of the same images at different levels of abstraction. By construction, the alignment provided by the HL Dataset allows us to identify concrete objects in images which provide 'support' to infer high-level concepts such as scenes, actions and rationales.

We dive deeper into our analysis and study the connection between high-level concepts related to scene, action and rationale, to low-level objects present in the aligned COCO captions. We ask: 'What are the most informative objects for a high-level concept (e.g. *enjoy*) found in a specific axis (e.g *rationale*)?'

We leverage the Point-wise Mutual Information (PMI) (Church and Hanks, 1990) to find the most informative objects linked to a high-level concept. This is helpful to discover connections between concepts across different levels of abstraction but also gives clues on the content distributions within the axes. We filter out object mentions which have a frequency less than 100 in the low-level captions. This leaves 475 object-denoting lemmas. Then, we compute the PMI between content words in the high-level captions and all these lemmas. For example, Figure 6 shows the nouns in the object-centric captions which have the strongest PMI with the verb 'enjoy' in the rationale axis.

We can observe that high-level captions can express different nuances of the same abstract concept. To take another example, *love* (in Figure 7) can refer to the love between an animal and its owner, between two partners (e.g. *wedding*) or the love for sports (e.g. *skate, snowboard*). In the same way, as shown in Figure 6 a general concept like *enjoy* can be characterized by object-level concepts leaning toward a specific nuance of meaning, like sports activities (e.g. *kite, snowboarder, skier*) or places (e.g. *sandy shore, ocean, lake*). More examples are provided in Appendix A.2.

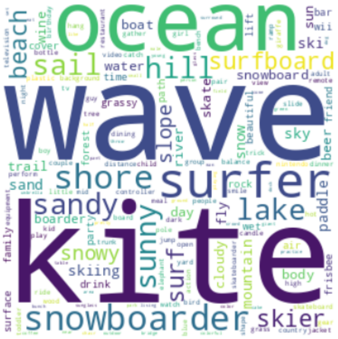


Figure 6: Most informative objects for the word *enjoy* in the *rationale* axis. Font size is proportional to PMI.

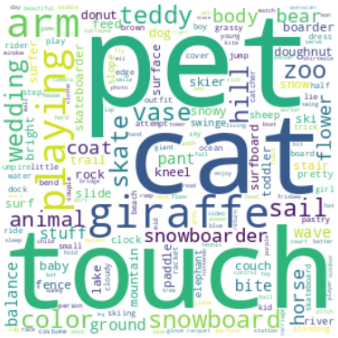


Figure 7: Most informative objects for the word *love* in the *rationale* axis. Font size is proportional to PMI.

4.2 Confidence scores analysis

Our confidence scores are similar in spirit to the *self-confidence* scores collected in the VQA dataset (Antol et al., 2015). However, they differ insofar as our scores are not self-reported by the authors of the captions, but collected from independent annotators. The inclusion of an external judgment plays an important role in determining the reliability of interpretation operated by the annotators in the caption collection and therefore, in shedding light on the extent to which an annotator's interpretation of a scene relies on 'shared' or 'commonsense' knowledge, or is entirely idiosyncratic.

We observe an average confidence score of 4.47 on a Likert scale from 1 to 5 (with a standard deviation of 0.78 and a median of 5) over all the axes. This suggests that, overall, according to independent judges, our high-level captions succeeded in capturing shared or 'commonsense' high-level interpretations of the scene.

Furthermore, the confidence scores provide an

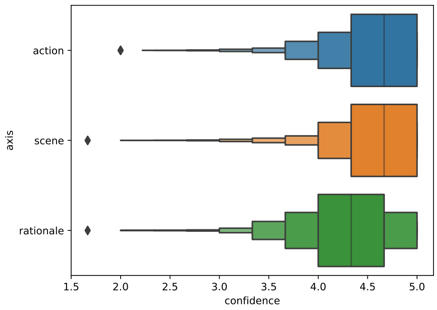
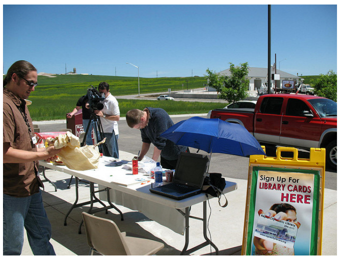


Figure 8: Axis-wise confidence score distribution of the high-level captions.



Idx	Scene caption	Confidence
1	in the restaurant	1
2	in the entrance of the library	1
3	the picture is taken outside a library	3

Figure 9: Example of a 'hard' sample in the HL dataset where the scene captions have low confidence scores.

additional perspective under which our data can be characterized: by performing an axis-wise analysis of the confidence scores distribution (see Figure 8), we observe that the *scene* and *action* captions feature the highest overall confidence, while the *rationale* axis lags behind by a small margin. We expect such differences, since determining the rationale of an action depicted in a static image is challenging, in particular, because annotators can leverage significant visual cues, but have no access either to temporal information or the subject's stated intentions. Therefore, they need to resort to their own priors and expectations which can also lead to idiosyncratic interpretations which independent judges – as in our confidence score analysis – would find relatively unlikely.

One important use of confidence scores is to provide a measure of uncertainty of the data, which can be used, for instance, to identify hard samples; an example is shown in Figure 9. The scene is hard to interpret even for humans and the scene captions display more variability and have low confidence scores. A detailed analysis of lexical and semantic variability in the presence of high-confidence scores is reported in Appendix A.3.

Model	Axis	Cider	SBLEU	Rouge-L
GIT	action	110.63	15.21	30.43
	rationale	42.58	5.90	18.57
	scene	103.00	24.67	33.92
BLIP	action	123.07	17.16	32.16
	rationale	46.11	6.21	19.74
	scene	116.70	26.46	35.30
ClipCap	action	**176.54**	**27.37**	**39.15**
	rationale	**78.04**	**11.71**	**25.76**
	scene	**145.93**	**36.73**	**42.83**

Table 3: Automatic metrics for baselines (GIT, BLIP, and ClipCap) fine-tuned along the three axes (*scene, action*, and *rationales*) of the HL dataset. The results are the average of 5 evaluation runs, by keeping the same decoding strategy and parameters for all the models.

5 Baselines and results

In this section, we show how the dataset can be used to finetune models to generate high-level, aspect-specific descriptions, e.g. image-to-scene or image-to-action. Below, in Section 6, we also describe a data augmentation and generation experiment, to merge the three axes into more 'narrative-like' descriptions of images.

We provide baselines for this task by fine-tuning three models, namely GIT (Wang et al., 2022a), BLIP (Li et al., 2022), and ClipCap (Mokady et al., 2021) on each separate axis. All the baselines were trained for a maximum of 10 epochs using a learning rate of $5e{-}5$, Adam optimizer, and half-precision (`fp16`).

Table 3 displays automatic evaluation results for the three models, on each axis. The first observation is that ClipCap outperforms by far the other models in each separate axis. Differently from the other models, which are natively multimodal, ClipCap leverages a LLM to generate captions, conditioning the text generation on a prefix representing the visual information, which is obtained by a mapping network trained to generate the prefix from CLIP's (Radford et al., 2021) image embeddings.

A second observation, consistent with the analysis presented in earlier sections, is that on all metrics, models fine-tuned to generate rationale-based descriptions receive lower scores. We hypothesise that this is due in part to the greater variability in this axis, and to its inherent difficulty, as reflected in lower confidence scores. Future work could leverage these scores as additional signal in fine-tuning models on captions that require more inference, compared to more descriptive ones.

6 Data augmentation and narrative generation

We now describe how we extend the dataset to combine the three axes to compose a short 'narrative', which describes the scene, action and rationale in tandem. We call this new dataset HL Narratives. To do this, we leverage the individual axes and synthesise this part of the data using a pre-trained language model. Since scenes, actions, and rationales were elicited individually in a visually grounded and controlled setting, a synthesised version of the three individual captions should also be true of the image to the same extent (modulo the variations in confidence that we observe).

6.1 Data generation process

We frame the synthesis of narrative captions as a paraphrasing task. We follow a human-in-the-loop approach consisting of three stages: (i) we manually annotate a small sample of gold data; (ii) we fine-tune a large pre-trained language model (LPLM); (iii) we use the fine-tuned model to generate a sample of data, which is manually corrected and then (iv) added to the gold annotations before fine-tuning again. This procedure allows us to use only a few iterations to annotate quickly a considerable amount of data because the model improves the quality of the generated data, making manual correction progressively easier.

We use a version of T5 (Raffel et al., 2020) already fine-tuned on paraphrase generation[3] as LPLM data generator. We initialise the process with manually paraphrased annotations for 50 images ($3 \times 50 = 150$), fine-tune the model for 2 epochs, and generate 150 captions for another 50 images, which are manually corrected and added to the original 150. The model is then fine-tuned for a further two epochs. In each iteration, we reserve 10% as validation data. After two epochs, we observe that the validation loss does not improve further. Finally, in the last iteration, we use all gold data to fine-tune the model and generate synthetic high-level captions for the whole HL dataset, obtaining 14,997 synthetic captions for training and 1499 for testing. In addition to the T5 paraphrase model, we also experimented with LLaMA (Touvron et al., 2023) in a few-shot setting; however, we find that T5 outperforms LLAMA in this task. See Appendix B for full details.

Model	SacreBLEU	ROUGE-L	Cider
GIT (PRE)	1.23	11.91	18.88
BLIP (PRE)	3.47	15.21	24.15
ClipClap (PRE)	8.72	19.45	40.47
GIT (FT)	11.11	**27.61**	75.78
BLIP (FT)	**11.70**	26.17	**79.39**
ClipCap (FT)	8.15	24.53	63.91

Table 4: Results of the narrative generation task, averaged over 5 runs using the same decoding parameters for all models. PRE: pretrained models; FT: finetuned on the synthetic data.

6.2 Results

We build three baselines by fine-tuning the same three large pre-trained models used in Section 5: GIT, BLIP, and ClipCap on our synthetic narrative captions. We fine-tune for 3 epochs with batch size 8, learning rate $5e^{-5}$, and Adam optimizer with weight decay (Loshchilov and Hutter, 2017). We test on our gold human-annotated data. As shown in Table 4, where we report results for automatic metrics, overall the models achieve worse results than in the aspect-specific caption generation task (reported in Table 3). This further highlights the difficulty of generating narrative captions of this kind for models trained on object-centric captions.

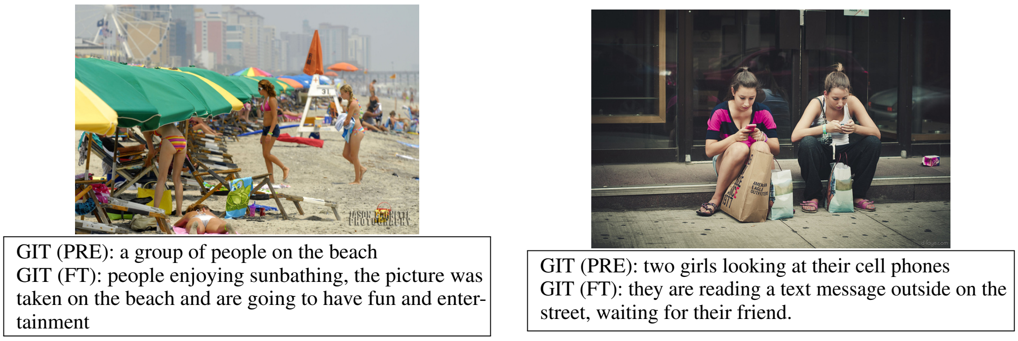
Notably, the best-performing model in the aspect-specific caption generation task, namely ClipCap, is the worst in the narrative caption generation, though by a small margin (Table 4). This suggests that although a conditioned LLM can greatly adapt to generate high-level descriptions of specific aspects of the scene, it struggles in generating comprehensive high-level descriptions involving multiple high-level aspects of the scene. Ultimately, this suggests that the multimodal representations learned by multimodal models are more robust and effective in generating natural captions than conditioned unimodal models such as ClipCap.

However, the exposure to a small amount of synthetic high-level captions is sufficient to drive the models' generated text toward more narrative-like outputs. See Appendix F for more examples from all models. Further progress can be done in this direction, for example by incorporating confidence scores during finetuning.

7 Further uses of the HL Dataset

We envision a wide set of use cases and tasks enabled by the HL Dataset.

[3] Details about the T5 fine-tuned on paraphrase generation are available at https://huggingface.co/Vamsi/T5_Paraphrase_Paws.



GIT (PRE): a group of people on the beach
GIT (FT): people enjoying sunbathing, the picture was taken on the beach and are going to have fun and entertainment

GIT (PRE): two girls looking at their cell phones
GIT (FT): they are reading a text message outside on the street, waiting for their friend.

Figure 10: Comparison between the object-centric captions generated by GIT pre-trained (PRE) and the high-level caption generated by the fine-tuned (FT) model. The generated high-level caption embeds high-level information regarding action, rationale, and scene, depicted in the visual content.

V&L generative tasks Our captions support image captioning generation tasks which encompass a broader range of visually grounded linguistic descriptions than the highly object-centric, 'conceptual' descriptions which dominate the captioning literature Hodosh et al. (2013). Moreover, the decomposition along three axes can be exploited to compose narratives of the image, as in image paragraph generation (Wang et al., 2019) and visual storytelling (Huang et al., 2016; Hu et al., 2020). They can be used in combination with the question each axis corresponds to, in order to generate micro-dialog scenarios.

We would also argue that the high-level captions are also more natural and human-like, since they were collected without enforcing any restriction on the content to be described. Given that the images are also aligned with object-centric captions, it is possible to envisage a scenario in which a model is trained to generate high-level captions, which are 'explained' or justified with reference to low-level, object-centric properties (see Hendricks et al., 2016, 2018, for some work in this direction). In this way, the dataset can be leveraged to provide captions and explanations. Furthermore, the confidence scores serve for the identification of hard samples in the data, both for evaluation purposes and to provide additional training signals, as recently shown by Ouyang et al. (2022).

Multimodal Grounding HL Dataset is also a useful resource to benchmark the grounding capabilities of large pre-trained V&L models. Along these lines, Cafagna et al. (2021) study the capability of V&L models to understand scene descriptions in zero-shot settings, finding that only large-scale pre-trained V&L models have enough generalization capabilities to handle unseen high-level scene descriptions. Cafagna et al. (2022) analyse the impact of exposure to high-level scene descriptions on multimodal representations in models pre-trained on object-centric captions. They show that exposure to high-level concepts mainly affects the model's attentional resource allocation over the visual input, even though the low-level concepts learned during pre-training provide enough signal to support and easily adapt to scene descriptions during fine-tuning. This is also supported by Wang et al. (2022b) who find that low-level concepts are needed to learn higher-level concepts, though this does not hold in the other direction.

8 Conclusions

In this paper, we introduced the High-Level (HL) Dataset. We extended 14,997 images from the popular COCO dataset with 134,973 human-annotated high-level descriptions systematically collected over three axes: *scene*, *action*, and *rationale*. We aligned high-level captions with object-centric captions and we provided human-collected confidence scores to measure the degree of commonsense expressed in the high-level captions. We also provided baseline results on generating captions for individual axes, as well as synthesised narrative captions by combining these three high-level axes of description.

Differently from current V&L captioning datasets, the high-level captions capture the human interpretation of the scene allowing for inference and expectations. We discussed how they can be used also in combination with low-level captions to improve research in visual commonsense reasoning and multimodal grounding of visual concepts into linguistic expressions and for generative tasks, hoping to foster future research in this direction.

Ethical Considerations

The data collection received ethical approval from the University of Malta Research Ethics Committee. This data is intended to be used for training, fine-tuning, and performing experimental evaluations of machine learning models. The dataset from which the images were originally sourced is a widely-studied, publicly available resource. As far as we are aware, the data does not contain harmful or offensive content. However, we acknowledge that any biases in the collection of images and/or captions in the original dataset will also be present in the HL Dataset.

Supplementary Materials Availability Statement:

The HL Dataset is publicly released on GitHub[4] and Huggingface[5]. The syntetic HL Narratives Dataset described in Section 6, is publicly released on Huggingface[6]. All the baselines described in Section 5 and 6 are available on Huggingface[7].

Acknowledgements

Contribution from the ITN project NL4XAI (*Natural Language for Explainable AI*). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Lawrence W Barsalou et al. 2008. Grounded cognition. *Annual review of psychology*, 59(1):617–645.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language models 'see' when they see scenes. *arXiv preprint arXiv:2109.07301*.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2022. Understanding cross-modal interactions in V&L models that generate scene descriptions. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 56–72, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating Visual Explanations. In *Proceedings of the 2016 European Conference on Computer Vision (ECCV'16)*, Amsterdam. ArXiv: 1603.08507.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

[4] `github.com/michelecafagna26/HL-dataset`

[5] `huggingface.co/datasets/michelecafagna26/hl`

[6] `https://huggingface.co/datasets/michelecafagna26/hl-narratives`

[7] `https://huggingface.co/michelecafagna26`

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976.

Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. Vivo: Visual vocabulary pre-training for novel object captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1575–1583.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 *European conference on computer vision*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32, New Orleans, Louisiana. Association for Computational Linguistics.

Kiwan Maeng, Alexei Colin, and Brandon Lucia. 2017. Alpaca: Intermittent execution without checkpoints. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):1–30.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be different to be better! a benchmark to leverage the complementarity of language and vision. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2751–2767.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 940–946. AAAI Press.

Zhecan Wang, Haoxuan You, Yicheng He, Wenhao Li, Kai-Wei Chang, and Shih-Fu Chang. 2022b. Understanding ME? multimodal evaluation for fine-grained visual commonsense. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9224, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Appendix

A Data Analysis Details

A.1 Quantitying grammatical errors

We ask two postgraduate students experts in linguistics to correct grammatical errors in a sample of 9900 captions, 900 of which are shared between the two experts. They are shown the image-caption pairs and they are asked to edit the caption whenever they identify a grammatical error. The most common errors reported by the annotators are:

- Misuse of prepositions;
- Wrong verb conjugation;
- Pronoun omissions.

In order to quantify the extent to which the corrected captions differ from the original ones, we compute the Levenshtein distance (Levenshtein, 1966) between them.

We observe that 22.5% of the sample have been edited and only 5% with a Levenshtein distance greater than 10. This suggests a reasonable level of grammatical quality overall, with no substantial grammatical issues. This can also be observed from the Levenshtein distance distribution reported in Figure 11. Moreover, the human evaluation is quite reliable as we observe a moderate inter-annotator agreement ($\alpha = 0.507$, (Krippendorff, 2018)) computed over the shared sample.

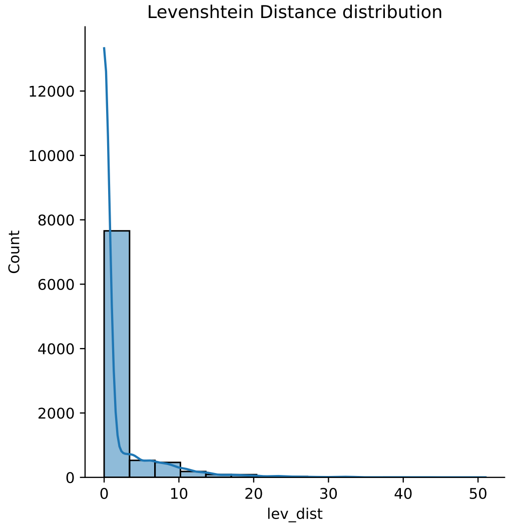


Figure 11: Distribution of the Levenshtein distance computed between the original and the corrected high-level captions in a sample of 9900 captions.

A.2 PMI analysis examples

The PMI analysis can provide interesting insight into the connection between object-level and high-level captions on all the three axes available.

On the *scene* axis, for instance, the PMI gives some clues on the extent to which an object can be considered diagnostic for a scene. For instance, two semantically similar scenes like *restaturant* (see Figure 12) and *kitchen* (see Figure 14) share several diagnostic objects, as we would expect. However, we can identify important semantic nuances: the scene *restaurant* contains objects related to the food (i.e. *pizza, cheese, wine, sandwhich*) whereas *kitchen* contains objects related to the preparation of food (i.e. *stove, oven, tray, refrigerator*). Another example is shown in Figure 13, where the most relevant objects for the action *look* encompass a wide variety of contexts, like looking at a screen or a device (e.g. *device, screen, cellphone*) or entertainment (e.g. *zoo, zebra, giraffe*). For more examples see Table 5, where are shown the top most relevant objects for the top three lemmas in the *scene, action* and *rationale* axes.

These semantic differences, while quite easy for humans to interpret, are not usually present in object-centric V&L datasets. They are made explicit and easy to identify in the HL dataset, where captions with different levels of abstraction are aligned with the same image.

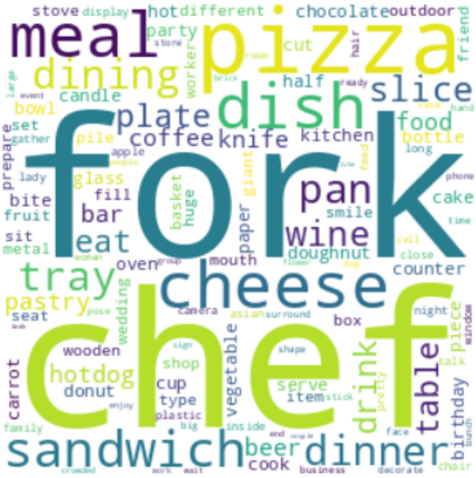


Figure 12: Most informative objects for the word *restaurant* in the *scene* axis. Font size is proportional to PMI.

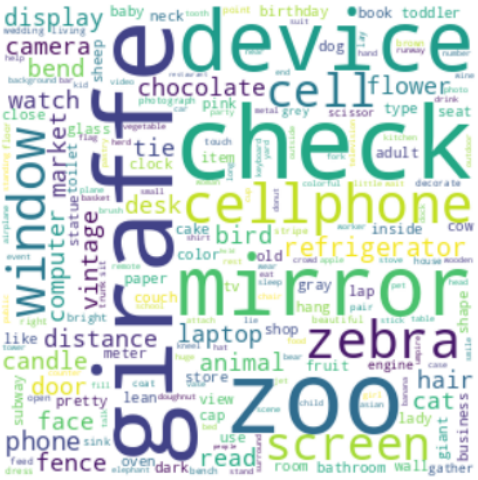


Figure 13: Most informative objects for the word *look* in the *action* axis. Font size is proportional to PMI.

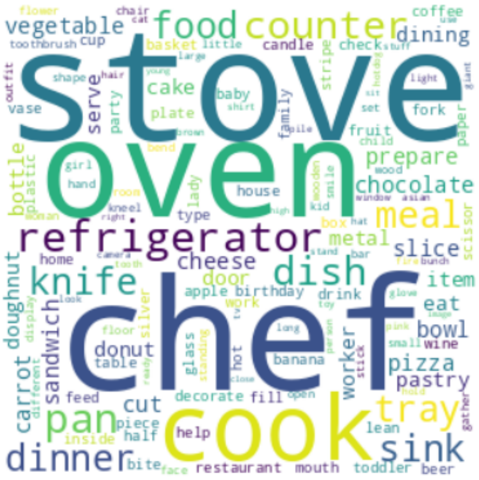


Figure 14: Most informative objects for the word *kitchen* in the *scene* axis. Font size is proportioanl to PMI.

Axis	Top Lemmas	Top Objects (PMI)
scene	street	intersection, decker, meter
	room	living, wii, nintendo
	road	traffic, decker, intersection
action	play	nintendo, wii, swing
	ride	rider, carriage, wave
	hold	controller, remote, rain
rationale	want	mirror, bathroom, sink
	enjoy	wave, kite, ocean
	fun	wii, nintendo, controller

Table 5: Top most informative objects of the top most frequent lemmas in the three axes (*scene, action, rationale*) according to PMI.

A.3 Quantifying Lexical and Semantic Diversity

In Section 4.2, we showed that in the presence of low confidence, there can be variation or disagreement among high-level captions given by different annotators for the same axis. In such cases, the captions focus on different aspects or refer to different interpretations. Although this phenomenon has been observed for captions with a low confidence score, it is conceivable that it might also happen with high-confidence captions, for example, two captions annotated by different annotators, while differing in the interpretation of an image, could nevertheless be considered highly likely. To quantify this phenomenon, in this section we further expand our analysis by studying the lexical and semantic diversity of our captions.

Purity score We leverage the BLEURT score (Sellam et al., 2020), a trainable metric used to evaluate semantic differences in Natural Language Generation, to compute a score measuring the semantic diversity among the high-level captions associated with an image. To do so, we first compute such scores across each axis, and then we combine them to obtain a final score for the item. In this way, we can unpack the semantic diversity item-wise and axis-wise.

Let C be the set of high-level captions of a given axis (e.g. scenes) for a given image. For simplicity, we do not report the index of the image and the axis in the following notation. We compute the BLEURT score of the caption as follows:

$$s_i = BLEURT(c_i, ref) \tag{2}$$

where s_i is the resulting BLEURT score, c_i is a high-level caption, and ref is the set of reference captions defined as follows:

$$ref := \{c_j \mid c_j \in C \ and \ j \neq i\} \tag{3}$$

In other words ref is the set of remaining captions along the axis and therefore, s_i is measuring the semantic diversity of the caption with respect to the other captions along the same axis.

By averaging the caption-wise scores across a single axis and across all the axes we obtain a *purity score* measuring the semantic consistency both axis-wise and item-wise.

Diversity score Along the same lines, we propose the *diversity score*, to measure the lexical diversity of the captions. The *diversity score* follows the same logic implemented to compute the *purity score* introduced in the previous paragraph, but the BLEURT score in Eq. 2 is replaced by the BLEU score (Papineni et al., 2002) and then normalized between 0 (similar) and 1 (very different). Our score is similar in spirit to self-BLEU (Zhu et al., 2018) as it measures the similarity of the captions within their own distribution. However, its computation concerns only axis-wise and item-wise captions.

A.3.1 Results and discussion

As shown in Figure 15 the purity scores obtained are mostly negative, this is due to lexical variations, which the BLEURT score is known to be sensitive to (Sellam et al., 2020). However, BLEURT is not defined in any specific interval thus, it is usually hard to interpret (Sellam et al., 2020) if not considered in relative terms. Based on that, we use it to compare the semantic purity across items and axes within our dataset. As shown in Figure 15, *action* and *scene* share similar purity score distributions whereas the *rationale* is more skewed to the left than the other axes. This shows that the rationales feature a higher semantic diversity (lower overall BLEURT) than the other axes.

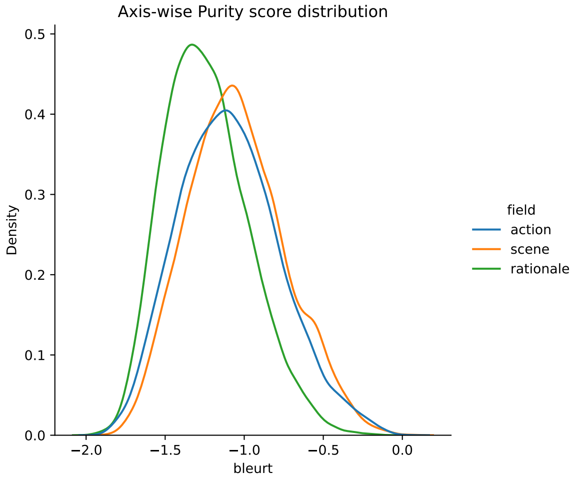


Figure 15: Axis-wise purity score distribution.

The *rationale* axis is also the one featuring the highest lexical diversity, whereas the *scene* and the *action* have similar distributions. This is shown in Figure 16 where the *rationale* density estimate (in green) has a higher peak skewed on the right-hand side than *scene* and *action*density estimate (respectively in orange and blue).

We have similar observations for both *purity* and the *diversity* scores and this confirms what was observed in the confidence score analysis in Section 4.2, namely that the task of determining the rationale of an action from a static image produces more variation and divergent interpretations leading to higher semantic and lexical diversity. Moreover, we find that both the *diversity* and the *purity* scores positively correlate with the confidence scores (See Figure 17).

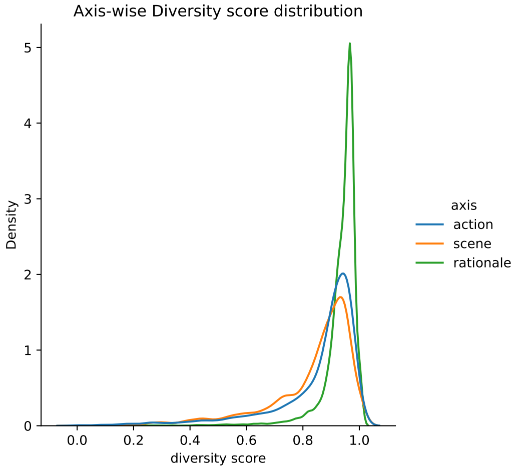


Figure 16: Axis-wise diversity score distribution. The scores have been normalized between 0 and 1.

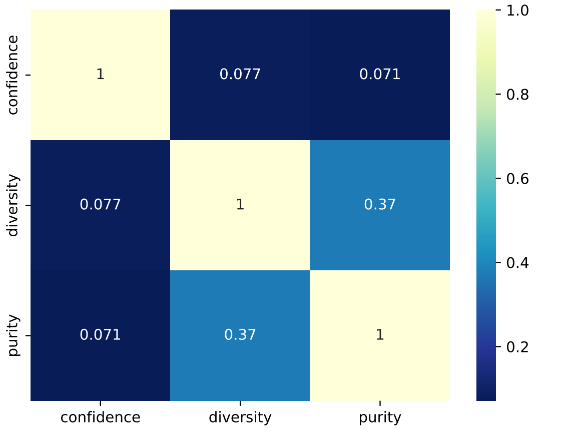


Figure 17: Pearson correlation between confidence, diversity and purity scores.

A.3.2 Item-based analysis

An item in the HL dataset is an image along with all the high-level captions of all the axes. For instance, Figures 18 and 19 show the item-wise *diversity score* and *purity score* distribution respectively, along with their average value across the whole dataset. An item on the right-hand side of the distribution is systematically more consistent across its axes with respect to the measure considered (*purity* or *diversity*). This information can

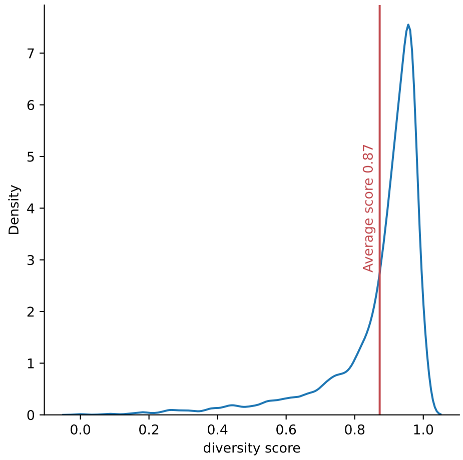


Figure 18: Item-wise diversity score distribution.

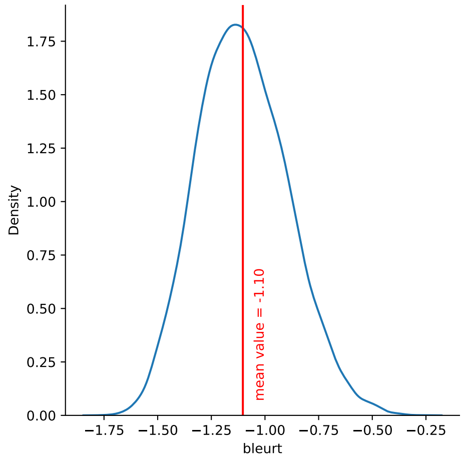


Figure 19: Item-wise purity score distribution.

be combined with confidence scores to perform a more fine-rained sample selection. For example in zero-shot testing, we might want to use a hard sample to test our model with, we can select items with similar lexicons, low-semantic purity, and low confidence scores.

B Narative Caption Generation Task Details

B.1 Few-shots Prompting Data Generation

We test an alternative data generation pipeline by leveraging the in-context learning capabilities featured by the most recent large language models (LLM) (Brown et al., 2020; Maeng et al., 2017; Touvron et al., 2023). This data generation approach has the advantage of not requiring any model fine-tuning.

We design a prompt for our task and we use it to generate data from the recently developed LLaMA model (Touvron et al., 2023). The prompt consists of the task description, followed by an example and the inputs of the task written in natural language. The full prompt is shown in Figure 20. The resulting output is then post-processed to extract the generated high-level caption.

> Given three sentences merge them into one sentence, and make sure that the sentence is grammatically correct. Here is an example:’in a beach’,’ holding an umbrella’,’ so they won’t get a sunburn’ <holding an umbrella in the beach so that they won’t get a sunburn.>\n The three sentences are: ’**scene**’,’**action**’,’**rationale**’ <

Figure 20: Prompt used for the data generation. The parts in bold are replaced with the corresponding high-level descriptions for the given sample.

Discussion As described in Section 6, we build baseline image captioning models starting from GIT-base and fine-tuning on the LLaMA- and T5-generated synthetic data. The best model is chosen on a combination of qualitative models’ output inspections and automatic metrics (SacreBLEU (Post, 2018), ROUGE-L (Lin, 2004) and Cider (Vedantam et al., 2015)) computed over the gold data.

In Table 6 we show the results of the evaluation based on the automatic metrics. First, we observe that the performance of the pre-trained model (PRE) is extremely poor, in the high-level caption generation task, highlighting the substantial difference between captions of this kind with traditional object-centric captioning the pre-trained model is trained on.

Second, focusing on the fine-tuned models, we observe that GIT fine-tuned on T5-generated data performs better than the LLaMa-based counterpart on the automatic metrics. We argue that the model trained on T5-generated synthetic data benefits from the exposure of the data generator to the gold data distribution. However, we point out that the few-shot data generation pipeline remains a valid alternative as it achieves comparable performance without requiring any further fine-tuning.

C Annotation Costs

In this section, we report the costs related to the data collection.

High-level caption collection Overall 1033 participants took part in the caption data collection, they were paid \$ 0.04 per item corresponding to the hourly minimum rate in the United Kingdom. In total, the data collection cost \$ 1938.

Model	SacreBLEU	ROUGE-L	Cider
GIT(PRE)	1.23	11.91	18.88
GIT(T5)	**11.07**	**31.37**	**74.79**
GIT(LLaMA)	10.96	24.71	65.05

Table 6: Automatic metrics computed over the gold annotated high-level captions; the scores are the average results of 5 runs using the same decoding parameters for all models. We compare the pre-trained model (PRE) with the model finetuned on T5-generated (T5) and LLaMA-generated (LLaMA) data.

Confidence Scores collection The qualification task for confidence scores led to the recruitment of 53 annotators. We found that this task was harder than the high-level caption annotation in terms of complexity but not in terms of execution time which was indeed shorter. Therefore, in order to encourage good quality annotations, we pay $ 0.04 per item. Considering the time needed to perform the task, this corresponds to 4 times the hourly rate of the minimum wage in the United Kingdom. The qualification task and the data collection cost respectively $ 93 and $ 1938.

D Annotation Details

D.1 Pilot

We run a pilot study with the double goal of collecting feedback and defining the task instructions. The pilot is run with 9 participants who were trained on the task, with high proficiency in English and a background in computer science and linguistics.

With the results from the pilot we design a beta version of the task and we run a small batch of cases on the crowd-sourcing platform. We manually inspect the results and we further refine the instructions and the formulation of the task before finally proceeding with the annotation in bulk. The final annotation form is shown in Figure 22. It is important to notice that the instructions, shown in Figure 21 are always visible to the workers.

Figure 23 shows the annotation form used for the confidence score collection. Also in this case, the instructions are always visible to the worker and each image is presented along with the original question and the answer.

E Additional Data Examples

In Table 7 we show further examples of images and their corresponding captions in the HL Dataset.

Instructions:
You are going to see some pictures. Each picture involves one or more people (’the subject’). You will be asked some questions about the picture
Don’t think too much, feel free to give your personal interpretation using your knowledge or common sense.
Try to answer using full English sentences. **If you’re not sure what the answer could be, give your best guess.** Avoid using expressions like ”I think” or ”I suppose” or ”Maybe.
Do not propose options or possibilities saying for instance: something ”or” something else. **Make your best guess** and state the one you choose.
Write a statement, **don’t write a one-word answer**, avoid acronyms or slangs and write a **full sentence**.

1. **Where is the picture taken**: give your best guess about the type of place where the action is happening (for example, ”in a ski resort”);
2. **What is the subject doing**: Try to describe what the people are doing as concisely as possible.
   If there is more than one person, try to choose a description that captures what all of them are doing (for example, ”They are skiing”)
3. **Why is the subject doing it**: here, write your best guess about why the person or persons are doing the action (for example, ”They are on a family holiday”)

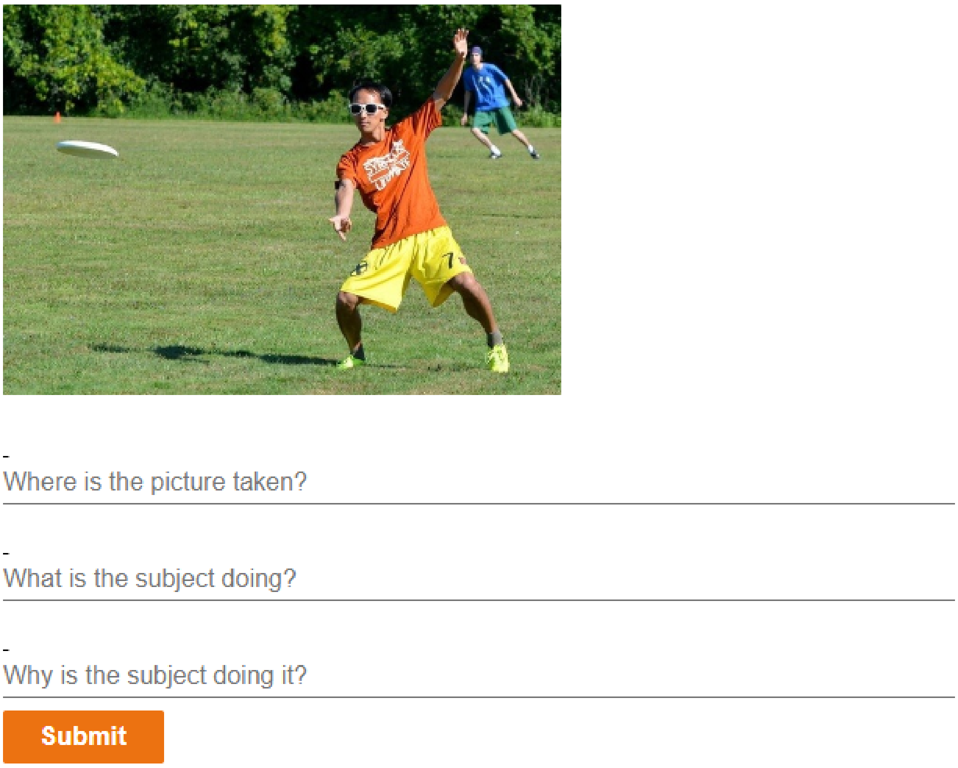
The **What** question and the **Why** question **cannot have the same** answer.

The answers must be **written correctly in English**, check the spell and most importantly **don’t forget the subject of the sentence in your answer** (he, she, it, they)

Figure 21: Final version of the instructions presented to the workers during the collection of the high-level captions. These instructions are always visible to the annotators.

F Examples of Narrative Caption generations

In Figure 24 we show examples of narrative caption generations from our fine-tuned baselines.

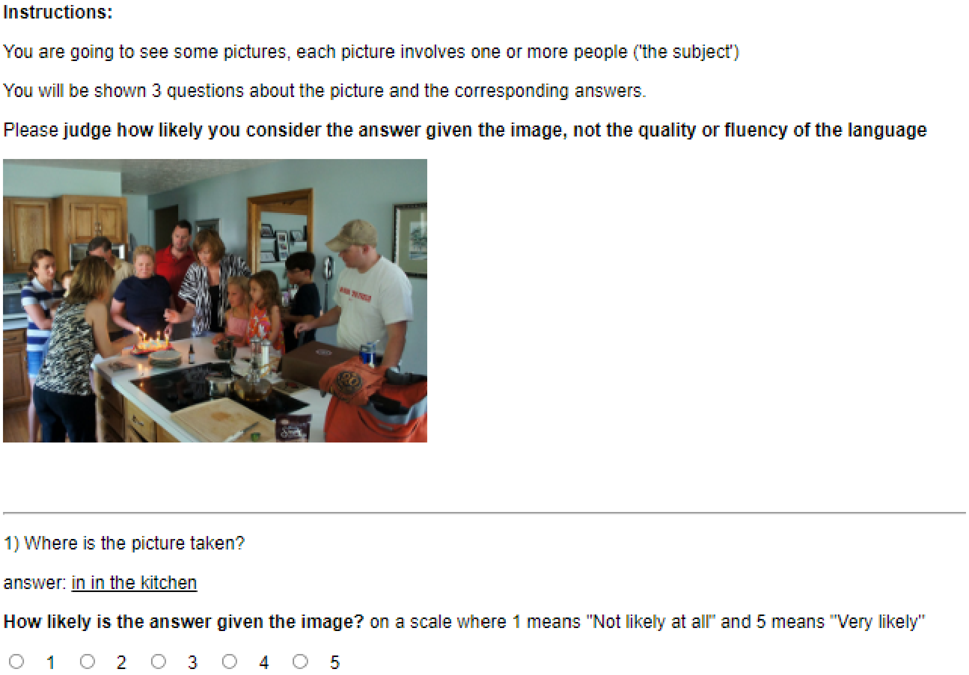


-
Where is the picture taken?

-
What is the subject doing?

-
Why is the subject doing it?

Figure 22: Annotation form presented to the worker during the high-level captions collection. The instructions (shown in Figure 21), are always visible to the annotators.



Instructions:

You are going to see some pictures, each picture involves one or more people ('the subject')

You will be shown 3 questions about the picture and the corresponding answers.

Please **judge how likely you consider the answer given the image, not the quality or fluency of the language**

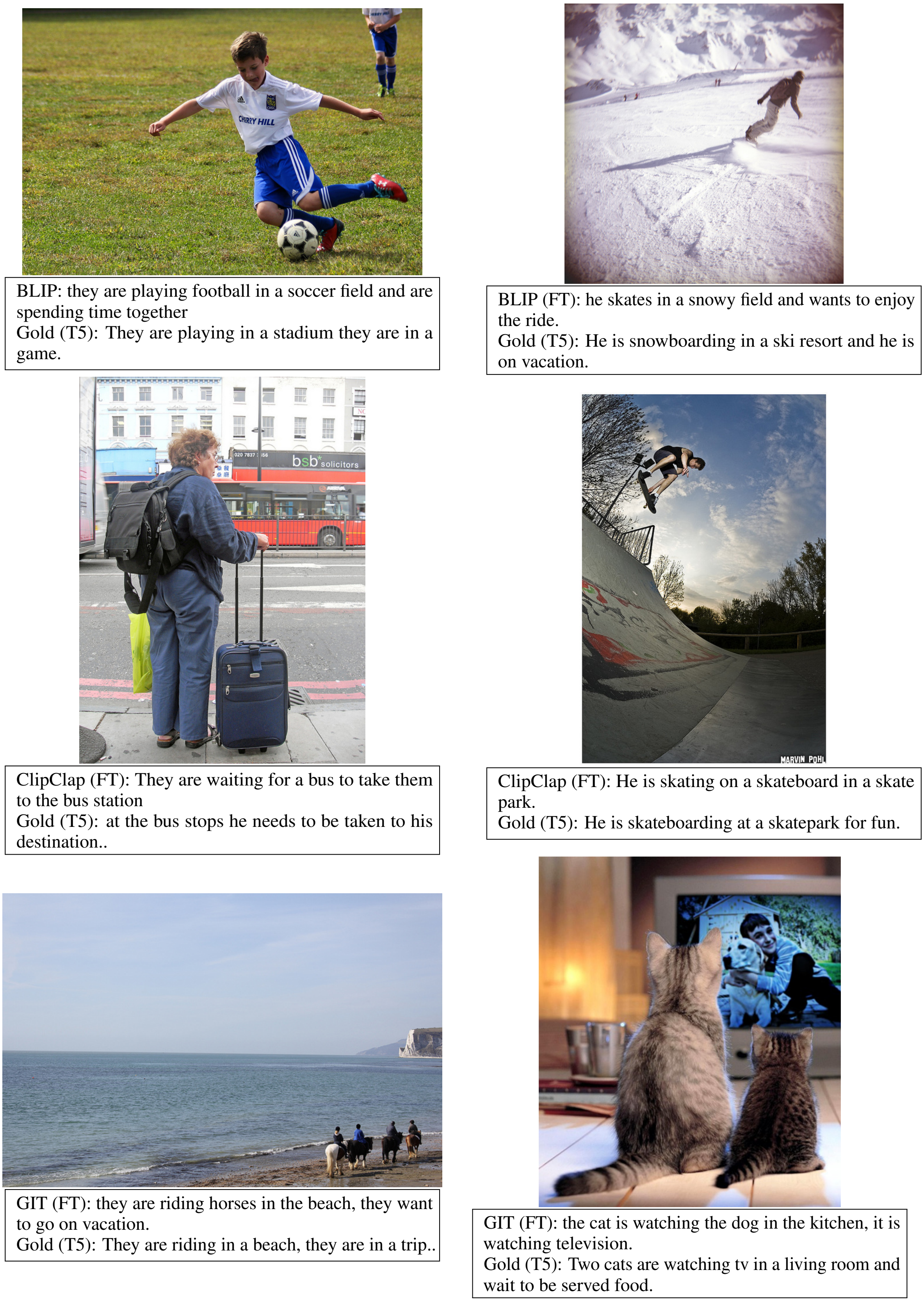
1) Where is the picture taken?

answer: in in the kitchen

How likely is the answer given the image? on a scale where 1 means "Not likely at all" and 5 means "Very likely"

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 23: The confidence scores annotation form. We show the instructions, the image, the question, and the corresponding answer.



BLIP: they are playing football in a soccer field and are spending time together
Gold (T5): They are playing in a stadium they are in a game.

BLIP (FT): he skates in a snowy field and wants to enjoy the ride.
Gold (T5): He is snowboarding in a ski resort and he is on vacation.

ClipClap (FT): They are waiting for a bus to take them to the bus station
Gold (T5): at the bus stops he needs to be taken to his destination..

ClipClap (FT): He is skating on a skateboard in a skate park.
Gold (T5): He is skateboarding at a skatepark for fun.

GIT (FT): they are riding horses in the beach, they want to go on vacation.
Gold (T5): They are riding in a beach, they are in a trip..

GIT (FT): the cat is watching the dog in the kitchen, it is watching television.
Gold (T5): Two cats are watching tv in a living room and wait to be served food.

Figure 24: Examples of captions generated by the fine-tuned (FT) models and corresponding T5-generated (T5) data on the narrative caption generation task.

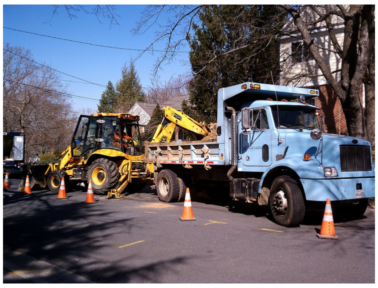
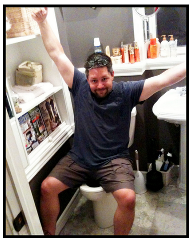
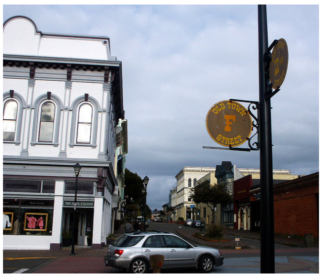
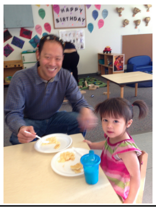
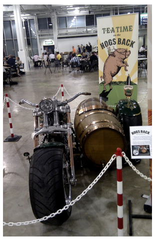
Image	Axis	Caption
	scene	the picture is taken in a construction site
	action	he is operating machinery
	rationale	he is clearing up debris with the machine.
	object-centric (COCO)	A blue flatbed truck with a yellow backhoe behind on a residential street.
	scene	The photo is taken in a toilet
	action	the subject is sitting on the toilet seat.
	rationale	doing it just for fun
	object-centric (COCO)	A man in blue shirt sitting on toilet next to sink and mirror.
	scene	the picture is taken at old town street
	action	one car is in the picture to turn to old town
	rationale	they are coming to old town
	object-centric (COCO)	A car driving on a street in the town center
	scene	in the restaurant.
	action	they are having their snacks.
	rationale	to taste it.
	object-centric (COCO)	A dad and his daughter eating a meal at a small table.
	scene	this is inside a garage
	action	the bike is just standing alone.
	rationale	no one is working on or trying to ride the bike.
	object-centric (COCO)	Custom motorcycle has a wooden barrel as a sidecar

Table 7: Examples of instances of the High-Level Dataset. It is shown one of the three captions available for each of the three axes collected: *scene, action, rationale*, aligned with the object-centric captions from COCO.