Learning Human-Human Interactions in Images from Weak Textual Supervision

Morris Alper and Hadar Averbuch-Elor
Tel Aviv University

Abstract

Interactions between humans are diverse and context-dependent, but previous works have treated them as categorical, disregarding the heavy tail of possible interactions. We propose a new paradigm of learning human-human interactions as free text from a single still image, allowing for flexibility in modeling the unlimited space of situations and relationships between people. To overcome the absence of data labelled specifically for this task, we use knowledge distillation applied to synthetic caption data produced by a large language model without explicit supervision. We show that the pseudo-labels produced by this procedure can be used to train a captioning model to effectively understand human-human interactions in images, as measured by a variety of metrics that measure textual and semantic faithfulness and factual groundedness of our predictions. We further show that our approach outperforms SOTA image captioning and situation recognition models on this task. We will release[1] *our code and pseudo-labels along with* **Waldo and Wenda**, *a manually-curated test set for still image human-human interaction understanding.*

1. Introduction

"No man is an island entire of itself." -*John Donne*

Humans are social beings. As such, interactions among people are ubiquitous and diverse, affected by various factors including social context and cultural norms. Reasoning about these interactions is crucial for gaining a holistic understanding of visual scenes depicting people. However, in spite of significant progress in analyzing isolated human actions [31, 78, 85] and relationships between entities and objects [29, 90], far less attention has been devoted towards an automatic understanding of human-human interactions (HHI). This is despite the importance of this task for applications such as interactive robotics, social behaviour understanding, and captioning systems for the visually impaired.

There are a number of factors that make the analysis of HHI difficult. The space of possible interactions between people is vast and requires understanding social context and physically non-local relationships, as illustrated in Figure 1. In addition, images depicting HHI may have multiple interpretations, some of which may be simultaneously correct. For example, the image on the left might depict "celebrating a wedding" as well as "dancing". Contextual cues such as the cake in the background of the image provide additional information that hints at the depicted HHI.

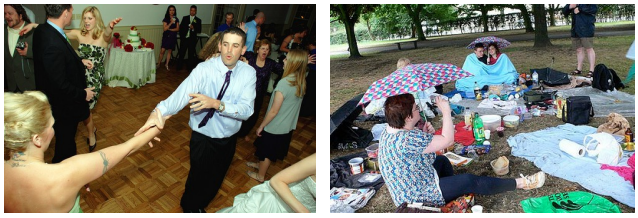


Figure 1. How would you describe the interactions depicted in these images? There are unlimited possible interactions between people which cannot be easily described by a fixed set of categories or actions. Context plays a crucial role, as in the left image where the clothing and cake in the background help to interpret the depicted interaction. Moreover, interactions may be involve participants at a physical distance as in the image on the right. To model the heavy tail of possible interactions, we propose to learn HHI as free text (see below[2] for predictions using our method).

Prior works targeting HHI understanding focus on a small fixed number of interactions; representative works include [67, 69, 46, 32], all of whose models are trained to recognize no more than ten interaction classes. In this work, we are interested in modeling the heavy tail of possible HHI to better understand the rich variety of ways in which people interact. To this aim, we propose to model HHI understanding as free text generation; since HHI are not confined to a fixed set of categories or even to a syntactic class such as verbs, HHI as free text enables the expression of an infinite variety of possible interactions. Furthermore, in contrast to previous works that frequently rely on extra context such as video data [64], we use a *single* image with no additional information (during inference), making our method more widely applicable. We focus on what Stergiou and

[1] via our project page https://learning-interactions.github.io

[2] A model fine-tuned on our pseudo-labels yields "dancing" and "having a picnic".

Poppe [64] term *dyadic* interactions—pairwise interactions between two people. Our goal is to identify the most salient dyadic interaction given an image of two or more people interacting.

One of the primary challenges to modeling HHI is a scarcity of labelled data for this particular task. There are only a handful of relatively small datasets specific to HHI, and larger video datasets for action recognition are lacking in coverage of interactions (see Table 1). To better model the heavy tail of possible HHI, we leverage the abundance of high-quality images of people and associated textual captions available on the Internet. In particular, we use the Who's Waldo dataset [14] that contains 270K image-caption pairs from Wikimedia Commons depicting people captured in a broad range of situations. Unlike many other image captioning datasets, Who's Waldo focuses on human-centric situations which are described using real-world captioned Internet data, and thus is more relevant to HHI understanding. However, it is extremely challenging to learn HHI from raw Internet captions directly, due to significant noise introduced by clutter and irrelevant details. To overcome this, we infer interactions from the original captions by applying knowledge distillation to synthetic data generated by a large language model, without explicit supervision. This approach allows for creating accurate pseudo-labels that provide textual descriptions of the HHI depicted in the images. We will release these pseudo-labels along with a manually annotated test set containing 1K image-interaction pairs from diverse Internet images which we name *Waldo and Wenda*, a new benchmark for our paradigm of HHI understanding as free text on still images, capturing the heavy tail of human-human interactions.

We demonstrate the utility of these pseudo-labels for learning HHI from images by training captioning models and using them as targets for a language modelling objective. We provide qualitative and quantitative analysis on the *Waldo and Wenda* test set; in addition, we evaluate this method on a larger scale by applying it to verb prediction on the imSitu situation recognition dataset [80], which we filter to select for images relevant to HHI.

Because we predict HHI as free text rather than categorically as in previous works, we propose a set of evaluation metrics chosen to measure important aspects of predicted HHI quality, namely textual similarity, factual groundedness, and verb similarity. Our evaluation shows that our HHI pseudo-labels allow for generating meaningful HHI free text descriptions from images, as measured by these metrics. We also show that learning on these pseudo-labels captures HHI substantially more effectively than either using existing SOTA image captioning models as-is or than training on interactions extracted with naive syntactic parsing. Explicitly stated, our key contributions are:

- A new paradigm and benchmark for HHI understanding from images—*i.e.*, predicting interactions as free text—allowing to better understand the vast variety of ways in which people interact.
- A method for isolating HHI from noisy Internet captions using knowledge distillation applied to a large language model, and a set of pseudo-labels generated by this method.
- An evaluation framework with metrics that capture HHI understanding, and results demonstrating that training image captioning models on these pseudo-labels can allow for modeling the heavy tail of possible HHI across various situations and configurations more effectively than SOTA image captioning and situation recognition models.

2. Related Work

Human action recognition. Human actions span a range from simple to complex. These include simple actions ("running"), human-object interactions ("dribbling a ball"), human-human interactions ("shaking hands"), and group actions ("gathering"). Because of the dynamic nature of actions, a large portion of work on action recognition uses video data [66, 68, 7, 89, 72]. Other approaches use other modalities such as depth or skeleton data [81, 51, 78]. Among video-based approaches, some use shallow approaches separating feature representation of action videos and classification of these features, while others use end-to-end trainable networks (see [31, 85] for detailed surveys). Works on human-object interactions (HOI) may use separate modules such as human and object detectors and relation modules [9, 21, 17], pose and gaze estimation [38, 70, 76], or graph neural networks applied to scene graphs [53, 77, 88, 39]. One line of recent work on HOI uses end-to-end models, frequently with transformer architectures [65, 90, 11, 29, 10]. In our work, we aim to predict the most salient interaction between the pictured individuals in an end-to-end manner from still image data alone.

HHIs are a subset of human actions which pose particular challenges to automatic recognition, due to non-locality, context dependency, and ambiguity. A number of works have explicitly tackled HHI recognition, as surveyed by Stergiou and Poppe [64]. As with general action recognition, these approaches most commonly use video data as input [46, 20, 71, 62, 36]. However, a few works have tackled the more challenging task of HHI recognition in still images. Some of these use classical computer vision methods to estimate human locations and poses in photos for predicting HHI [79, 8, 1]. Xiong *et al.* [75] use a CNN architecture with human, face, and object detection features for event recognition. These works all treat HHI as categorical, predicting them from a small set of predefined interaction

classes. In contrast, we use free text to describe HHIs allowing for more flexibility than categorical recognition.

HHI datasets. Most existing datasets of HHI or with subsets representing HHI classes only include a small number of interaction categories. The majority consist of video data, either curated [58, 83, 20, 23] or YouTube-based [63, 45, 87].

There are few image datasets dedicated to human actions, of which HHI are a subset. Ronchi and Perona [57] introduce the Visual Verbnet dataset consisting of images with dense verb annotations. Yatskar *et al.* [80] introduce the imSitu dataset for image situation recognition, involving recognizing the action portrayed in a still image (often with a human participant or participants) as well as predicting semantic roles for observed entities. In both cases the labels are selected from a fixed set of categories—single verbs in the case of imSitu, verbs or phrases containing verbs (*e.g.* "shake hands") for Visual Verbnet. Other image datasets such as Visual Genome [33] contain labeled entities, objects and their relationships, but focus more on general objects rather people and their interactions.

See Table 1 for a comparison of the most related datasets with our proposed HHI dataset. Unlike prior datasets, ours represents HHI as free text and not as fixed categories.

In-context learning with large language models (LLMs). The recent explosive growth in size and NLP benchmark performance of LLMs has led to their use as foundation models for use on downstream tasks [4]. Models such as GPT-3 show an emergent *in-context learning* property, whereby they may solve new tasks when prompted with only a few examples of a new task, or even just with a task description, without any parameter updates [6, 15]. The output of such models may then be used as supervised training data for conventional model fine-tuning. The idea of training on data generated using in-context learning to create a large training data set has been successfully applied to achieve state-of-the-art results on the SuperGLUE NLP benchmark by Wang *et al.* [73]. In our case, we use this data to perform *sequence-level knowledge distillation* – transferring the knowledge exhibited by such a large model into a smaller model by training on its output sequences [30, 22].

The use of LLM-generated synthetic data for multimodal learning has been explored by Brooks *et al.* [5], who use caption pairs generated by GPT-3 as auxiliary data for training a conditional diffusion model to perform image editing. Their method uses hundreds of manually labelled pairs of texts as training data; however, our pseudo-labelling method uses no explicit supervision, instead using syntactic parsing to generate automatic seeds for our synthetic data generation pipeline.

Dataset	#Seq	#HHI Classes
Curated videos		
UT-Interaction [58]	60	6
TV Human Interaction [48]	300	4
Hollywood2 [42]	3669	4
ShakeFive2 [20]	153	5
SBU Kinect [83]	300	8
AVA [23]	∼57.6k	13
NTU RGB+D (120) [60, 40]	∼114k	26
YouTube-based videos		
Kinetics [28, 63]	∼500k	11
Moments in Time [45]	∼800k	32
HACS [87]	∼50k	23
Still images		
imSitu [80]	126k	50*
Visual Verbnet [57]	10k	52*
Who's Waldo [14] (*w/ our labels*)	127k	∞* (free text)

*The number of HHI classes for Visual Verbnet includes verbs in the *communication*, *contact* and *social* categories, which sometimes mark solo actions or human-object interactions. The imSitu dataset contains a total of 504 verbs. We estimate the number of HHI interactions using an automatic methodology detailed in Section 5. Our free text pseudo-labels are limited to the types of interactions available in Who's Waldo.

Table 1. **Comparison of HHI datasets**. Prior datasets usually capture video data and target a small number of interaction classes. Several datasets focus on human actions, some of which include HHI. We denote the number of video/image samples with #Seq, and the number of HHI classes with #HHI Classes (values are taken from Stergiou and Poppe [64] where relevant). In our work, we devise a technique for generating HHI pseudo-labels for Who's Waldo [14], a dataset containing real-world image–caption pairs, allowing for modeling the heavy tail of HHI.

3. LLM-Based HHI Inference from Captions

To model the heavy tail of possible HHI using free text, leverage weak supervision in the form of image captions. We turn to Who's Waldo [14], a dataset containing image–caption pairs depicting human-centric scenes scraped from Wikimedia Commons (with names masked using their suggested `[NAME]` token). As illustrated in Figure 2, the mentions of the depicted HHI are embedded in detailed textual captions, and do not directly correspond to syntactic structures such as verbs in the text. For instance, the first depicted caption is long and the only relevant detail is the phrase "gets at [sic] high five"; the last depicted caption contains no verb (while the noun phrase "Ski Tour" hints at the relevant interaction). These captions are thus inadequate for training an HHI understanding model directly, as a captioning model fine-tuned on them mainly learns to attend to details that are irrelevant for our task (as shown in Section 5.3). We therefore present a large language model (LLM)-based abstractive text summarization technique that produces clean interaction texts from the original Internet

Hospital Corpsman 1st Class [NAME], U.S. Fleet Forces Sea Sailor of the Year, gets at high five from Master Chief Petty Officer of the Navy (MCPON) [NAME] during a Navy Exchange meet and greet at the Navy Memorial. [NAME] will be meritoriously promoted to chief petty officer.

[NAME] and [NAME] playing as doublebass-duo in Club W71, 2018.

[NAME], [NAME] and [NAME], 2017 Ski Tour Canada, Quebec City.

Parsing-based: [⋆] getting five from [⋆] | [⋆] and [⋆] playing | ∅

Final: [⋆] getting a high five from [⋆] | [⋆] playing doublebass with [⋆] | [⋆] skiing with [⋆]

Figure 2. **HHI distilled from raw Internet captions alongside their corresponding images.** On top we show several images and captions from the Who's Waldo dataset [14], with [⋆] denoting masked named person entities. Our syntactic parsing approach allows for extracting an initial partial set of interactions (first row). We refine and enlarge this initial set using our abstractive summarization model which yields our final HHI pseudo-labels (second row). While the original captions possibly contain many additional details or no verb-based interaction at all (for example, see the rightmost image), our abstractive HHI pseudo-labels succeed in describing HHI visible in the associated images.

captions, without explicit supervision.

Our unsupervised pseudo-labelling approach operates in three stages, illustrated in Figure 3: (1) We extract syntactic parsing-based interactions from captions from the Who's Waldo dataset, as well as constructing new synthetic interaction texts. (2) We prompt an LLM using the interaction–caption pairs from Who's Waldo along with the new interactions. The output synthetic captions are filtered using a pretrained natural language inference (NLI) model and various textual heuristics, to select for those that correspond to the new interactions. (3) We train an abstractive summarization model on these synthetic caption–interaction pairs; this model learns to output HHI from noisy Internet captions. As seen in Figure 2, these interaction pseudo-labels accurately describe the HHI visible in their associated images. Below, we provide more details for each stage (Sections 3.1–3.3). We then present *Waldo and Wenda*, our manually-curated HHI test set, in addition to statistics and an ethical discussion (Section 3.4).

3.1. Constructing interaction texts

We first define a rule-based approach for extracting interactions via syntactic parsing. Specifically, we extract the first verb in the caption with a [NAME] subject along with its direct objects and the heads of its prepositional arguments. This roughly corresponds to an interaction, although it may sound unnatural. This is also limited to captions containing verb phrases. We apply this procedure to captions from Who's Waldo to obtain corresponding parsing-based interactions.

We also construct new synthetic interactions by first applying this parsing procedure to scraped texts of news articles from the CC-News dataset [24] (from Common Crawl, containing text without image data), and then using the output interaction texts to prompt the large (1.3B-parameter) language model GPT-Neo [3, 18], which produces a set of diverse and more natural-sounding interactions.

3.2. Synthetic caption data generation

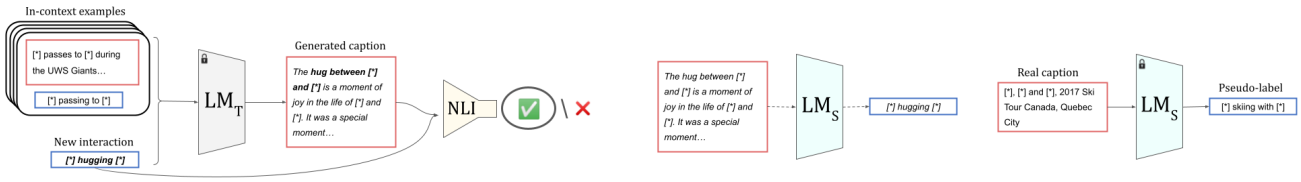
Using the caption–interaction pairs from Who's Waldo and the new synthetic interactions as seeds, we generate synthetic caption–interaction pairs using in-context learning with GPT-Neo. This allows us to create a larger and more diverse set of caption–interaction pairs than by using caption–interaction pairs directly from Who's Waldo. These pairs serve as the teacher model outputs used for knowledge distillation in the following section.

At each step, the language model is shown a prompt beginning with multiple randomly-selected examples of caption–interaction pairs from Who's Waldo. This provides context for the model to understand the task at hand–associating interactions with captions that contain them. We use ten examples in each prompt to balance between the providing sufficient context with computational considerations. The prompt ends with a new desired interaction, and the language model proceeds to generate a caption corresponding to this interaction. We filter these results using a pretrained NLI model and various textual heuristics detailed further in the supplementary material, ensuring that the output caption logically is properly formatted and logically entails the corresponding interaction.

3.3. Knowledge distillation for summarization

Using the synthetic data generated in the previous stage, we fine-tune a smaller (220M-parameter) student T5 model, a sequence-to-sequence transformer network whose pretraining tasks include text summarization [56]. We use the synthetic captions (with the task prefix "summarize:") as input and the synthetic interactions of the target text for fine-tuning. Empirically, we find that our fine-tuned student model is able to summarize captions and output valid interactions even when the caption does not contain a verb or has a syntactic structure that the syntactic parsing-based method could not process.

We apply this model to the captions in the Who's Waldo



Synthetic caption generation with teacher LM and NLI filtering | Fine-tuning student LM | Pseudo-label inference

Figure 3. **LLM-Based HHI Extraction from Captions.** We generate synthetic interaction-caption pairs via in-context learning (left), use them to fine-tune a summarization model (center), and then use this model to producing HHI pseudo-labels for captions in Who's Waldo (right), as detailed in Section 3. Captions are shown in red boxes, interaction texts in blue, and synthetic texts in italic letters. LM_T and LM_S indicate teacher and student language models respectively.

dataset to create pseudo-labels representing interactions as free text. See Figure 2 for examples of such pseudo-labels.

3.4. Our HHI dataset

Using our learned abstractive summarization model, we may generate interaction pseudo-labels from Who's Waldo captions. Out of the $\sim 270k$ samples in Who's Waldo, we use only those $\sim 130k$ containing at least two human face detections, using the detections provided by Cui *et al.* [14]. We filter out duplicate and near-duplicate images, those with high similarity to test set images, and samples with pseudo-labels that do not pass a few simple text-based filtering rules, including enforcing the format of `[NAME]`, followed by a present continuous verb ("-ing"), and including another `[NAME]` token. We are left with $\sim 126k$ images with pseudo-labels in total, which we hereby refer to as pHHI.

The *Waldo and Wenda* Benchmark. We also create *Waldo and Wenda*[3], an HHI test set containing 1K manually curated image–interaction text pairs. In order to test generalization to HHI understanding across a wide variety of natural images, we include data from three sources: (1) 300 images from Who's Waldo, (2) 300 images from COCO Captions [12], (3) 400 images from Conceptual Captions [61]. The images are selected from the validation and test splits of the relevant datasets. As the distribution of HHI in natural photographs is highly imbalanced—for instance, images in captioning datasets often display people standing side by side and posing for photographs—we curate this test set to represent a wide variety of interactions and to reflect performance on the long tail of uncommon HHI. Examples of images from *Waldo and Wenda* can be seen in Figures 1, 2, and 4.

Dataset Statistics. Overall, our pHHI training dataset contains 126,696 pairs of images and pseudo-labels. These labels contain 1,263 unique verbs and 16,136 unique interactions. The majority of the images (59.3%) only contain two detected people, with less than 5% of the images containing more than six detected people. The *Waldo and Wenda* test set contains 1,000 images along with their manually written ground truth HHI labels. These include 238 unique verbs and 575 unique interaction labels.

Ethical considerations. Our dataset inherits a diverse representation of people (ages, ethnicities, geographic etc.) from the Who's Waldo dataset [14]. Furthermore, we use their provided name masking to mitigate biases (*e.g.*, gender biases). We verify that all manually-curated test samples are neutral in nature and do not contain lurid or negative material. We perform similar verification on external test data, as described in Section 5, to avoid exposure to harmful or offensive behaviors. Furthermore, our pseudo-labels and test set will only be made available for academic purposes.

4. Learning HHI from Still Images

In the previous section, we demonstrated how we can obtain free text HHI pseudo-labels from the Internet captions of the Who's Waldo dataset [14]. We proceed to show how we use these to supervise learning HHI from still images via the paradigm of image captioning.

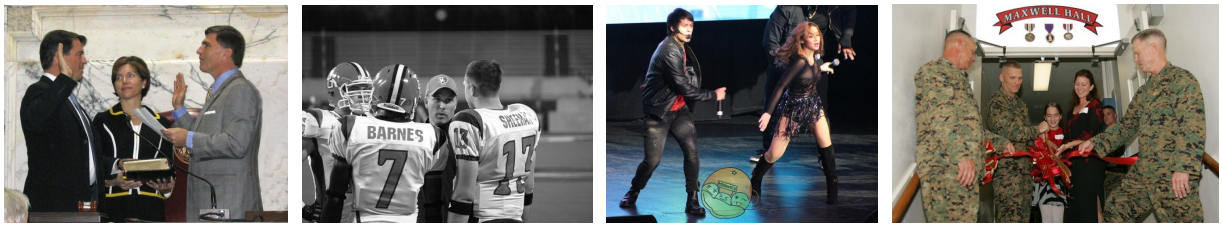
4.1. Models considered

After obtaining a set of images and pseudo-labels, we consider the task of HHI in the framework of image captioning. Given (image, pseudo-label) pair (I, L), we train an encoder-decoder network M to maximize the predicted conditional likelihood of L using a cross-entropy loss. During inference, we use autoregressive beam search decoding to generate text token by token, given I as input.

In order to evaluate the utility of transfer learning from general image captioning to our HHI understanding setting, we evaluate two choices for the model M:

(1) Vanilla encoder-decoder (EncDec). In this setting, we fine-tune a simple encoder-decoder model. We use the image encoder of pretrained CLIP-ViT [54], with its pooled embedding output followed by a single linear projection layer to match the hidden dimension of the decoder. For

[3]Wenda appears in the *Where's Waldo?* book series as Waldo's girlfriend.



CLIPCap (CC)	person, left, and person, right, receive a standing ovation for their service.	friday night rivals was for high school vs game!	wallpaper with a concert and a well dressed person entitled pop artist.	person, left, shakes hands with person, daughter of person, during a ribbon cutting ceremony.
EncDec (pHHI)	[*] administering the oath to [*]	[*] coaching [*]	[*] performing with [*]	[*] cutting the ribbon with [*]
GT	[*] swearing in [*]	[*] huddling with [*]	[*] dancing with [*]	[*] cutting a ribbon with [*]

Figure 4. **Results on the *Waldo and Wenda* test set.** We compare results obtained by a baseline, our vanilla encoder-decoder technique (trained on our pHHI data), and the ground truth labels in *Waldo and Wenda*, with [*] denoting [NAME] tokens that represent person entities. As illustrated above, our method generates text describing the HHI depicted in the image, without attending to other irrelevant details. In comparison, the SOTA captioning model CLIPCap used as-is may not output an interaction at all (middle two images). We also observe that our model predicts HHI that may require both a verb and other arguments to adequately understand (leftmost and rightmost images).

the decoder we use pretrained GPT-2 [55] with a causal language modelling head and cross-attention over the encoder output. Consistent with previous works on fine-tuning vision-and-language models [41, 44, 84], we freeze the weights of the image encoder as we fine-tune it on our pHHI data. By considering this model that was not previously trained on image captioning, we aim to evaluate the extent to which our pHHI aid in learning to understand the semantics of HHI in images (rather than simply cueing a captioning model to the correct surface form of HHI labels).

(2) Fine-tuned captioner. The second approach we consider is to apply transfer learning to a SOTA captioning model by fine-tuning it on our pHHI data. Because the Conceptual Captions (CC) dataset is more people-centric than COCO and thus closer to our use case, we pick CLIPCap [44] pretrained on CC as the base model for fine-tuning. Consistent with CLIPCap's training method, we freeze its image encoder and fine-tune the model on our pHHI data.

4.2. Training and Decoding

For all models, we use cross-entropy loss and consistent hyperparameter settings. For each model, we decode using beam search with 32 beams. We report metric values for the top 1, 5, and 8 beams.

5. Evaluation

5.1. Test Datasets

We evaluate our models on the following datasets:

***Waldo and Wenda*.** As detailed in Section 3.4, this consists of 1,000 images with manually-written ground truth labels. Examples of ground truth labels along with model predictions can be seen in Figure 4. We report metric values averaged over the three data sources (Who's Waldo, Conceptual Captions, COCO) of *Waldo and Wenda* in Table 2. We also show a breakdown of data source in Table 3.

***imSitu-HHI*.** We use an 8,021-sample subset of the imSitu [80] situation recognition benchmark, which we refer to as *imSitu-HHI*, to perform a large-scale evaluation of our models. Although imSitu does not contain free text HHI labels, it does contain categorical verb labels which can be used for comparison. Additionally, as the majority of images in imSitu do not depict HHI, we first filter for relevant samples as follows: We use person detections from YoloV5 [16] to select for images containing at least two humans. We further filter to select only samples with semantic frames containing at least two human participants. Finally, due to the noisy nature of this filtering, we only use verbs supported by at least 100 images in this filtered subset, as these verbs are most likely to describe HHI. We use these verb labels as the ground truth and evaluate predictions with the verb similarity metric as described below.

5.2. Baseline comparisons

We compare our approach to two types of SOTA model of that do not use our pHHI data as baselines:

(1) Pretrained captioner. The first baseline approach that we test is the use of a SOTA model that has already been pretrained for image captioning. We test the recent captioning models ExpansionNet v2 (ENv2) [27] and CLIPCap [44]. We use these captioners as-is and evaluate our metrics on their outputs with beam search decoding. CLIPCap is avilable with pretrained weights for both COCO [12]

and Conceptual Captions (CC) [61], and thus we test both models. ENv2 only uses COCO weights.

(2) Pretrained situation recognition model. We provide a comparison to the results of the CoFormer model [13] for grounded situation recognition with pretrained weights. Unlike weakly-supervised models trained on our pseudo-labels, which were generated from natural captions, CoFormer is supervised by training on the manually-labelled SWiG dataset [50], an extension of imSitu which includes grounding information for arguments that are visible in the accompanying images, and the model predicts the relevant verb, arguments, and grounding information given an image. We evaluate CoFormer by using its predicted verb, discarding semantic frame and grounding predictions since these semantic arguments do not directly map to the text of a human-human interaction string. See the supplementary material for details on how we insert its verb predictions into text prompts for metric calculations.

5.3. Ablations

In order to ablate the effect of our pseudo-labelling, we also report results of a captioning model fine-tuned on the entire text of the captions provided in Who's Waldo (listed in 2 and 4 under training data as "WW"). In the supplementary material we also provide a detailed comparison with results when training directly on the syntactic parsing-based seeds described in Section 3.1.

5.4. Metrics

A number of metrics have been proposed for natural language generation tasks, measuring various aspects of text quality [25, 19]. As no prior works (to the best of our knowledge) predict HHI as free text, we propose a set of metrics that evaluate various relevant aspects of generated text:

Textual similarity. We use the BLEURT [59] metric to measure similarity to the ground truth interaction. This is a learned metric for text generation which measures similarity between the text output by a model and the reference text. Because our test set is relatively small and the reference texts are short, this better reflects textual similarity than ngram-based metrics such as BLEU [47] which have high variance and must be averaged over large datasets, as is shown in detail in the supplementary material.

Factual groundedness. A key property of generated text is whether it is *consistent* or *contradictory* with respect to the ground truth (such as a source document in the case of summarization, or a reference caption in the case of image captioning) [34]. This may be quantified by using the scores output by a natural language inference (NLI) model, in order to measure the degree of factual groundedness or hallucination in generated text [43, 35]. For example, given an image with ground truth label *[NAME] sitting next to [NAME]*, the prediction *[NAME] standing with [NAME]* logically contradicts the reference label and thus is a factual hallucination. To measure this, we use scores (p_e, p_c) from a pretrained NLI model to estimate the factual groundedness of the predicted text, where p_e is the probability of entailment and p_c is the probability of contradiction. We treat the image caption from *Waldo and Wenda* as the premise and the model's prediction as the hypothesis for NLI inference. For test items sourced from COCO Captions, in which images correspond to multiple reference captions, we use the first reference as the premise for this calculation.

Verb similarity. We calculate the average cosine similarity of the predicted and ground truth verbs in GloVe [49] embedding space. The motivation for this metric is that a prediction may be valid or nearly valid even if it is not identical to the ground truth label as long as the semantic distance between the verbs is small (e.g. "hugging" vs. "embracing"). To evaluate this on free text predictions, we either select the first non-[NAME] word in the output (for models trained

Method	Training Data	BL↑	p_e ↑	p_c ↓	sim↑
Results@1					
CoFormer	SWiG	0.34	0.18	0.57	0.35
ENv2	COCO	0.27	0.25	<u>0.33</u>	0.41
CLIPCap	COCO	0.28	<u>0.34</u>	0.37	<u>0.42</u>
CLIPCap	CC	0.27	0.18	0.38	0.35
CLIPCap	CC+WW	0.26	0.16	0.40	0.17
EncDec	pHHI	<u>0.38</u>	0.30	0.37	0.41
CLIPCap	CC+pHHI	**0.42**	**0.41**	**0.32**	**0.46**
Results@5					
ENv2	COCO	0.31	0.39	0.19	0.46
CLIPCap	COCO	0.31	0.47	0.24	0.46
CLIPCap	CC	0.33	0.32	0.20	0.47
CLIPCap	CC+WW	0.33	0.29	0.24	0.27
EncDec	pHHI	<u>0.51</u>	<u>0.61</u>	<u>0.09</u>	<u>0.59</u>
CLIPCap	CC+pHHI	**0.57**	**0.71**	**0.07**	**0.65**
Results@8					
ENv2	COCO	0.32	0.43	0.17	0.48
CLIPCap	COCO	0.32	0.50	0.21	0.46
CLIPCap	CC	0.35	0.36	0.16	0.49
CLIPCap	CC+WW	0.35	0.33	0.21	0.31
EncDec	pHHI	<u>0.54</u>	<u>0.65</u>	<u>0.19</u>	<u>0.65</u>
CLIPCap	CC+pHHI	**0.59**	**0.76**	**0.04**	**0.69**

Table 2. Results on ***Waldo and Wenda***. The listed metrics are BLEURT (BL) and NLI scores (p_e, p_c) and verb embedding similarity (sim). CC+WW/pHHI indicates models that were initialized with pretrained CC weights and subsequently fine-tuned on Who's Waldo captions or on pHHI respectively. Best results are in bold, and second best are underlined. Results are aggregated across the three data sources of *Waldo and Wenda*. For models using beam search, we report results for top 1, 5, and 8 beams.

Method	Training Data	WW				CC				COCO			
		BL↑	p_e ↑	p_c ↓	sim↑	BL↑	p_e ↑	p_c ↓	sim↑	BL↑	p_e ↑	p_c ↓	sim↑
Results@1													
CoFormer	SWiG	0.40	0.29	0.29	0.34	0.33	0.16	0.68	0.38	0.29	0.10	0.74	0.33
Env2	COCO	0.24	0.20	**0.28**	0.38	0.26	0.20	0.45	0.41	0.31	0.36	**0.26**	0.45
CLIPCap	COCO	0.27	0.37	0.35	0.39	0.26	0.26	0.43	0.43	0.31	0.38	0.33	0.44
CLIPCap	CC	0.30	0.24	0.42	0.36	0.25	0.18	0.38	0.40	0.26	0.11	0.34	0.30
CLIPCap	CC+WW	0.28	0.12	0.59	0.27	0.26	0.21	0.34	0.15	0.23	0.15	0.28	0.10
EncDec	pHHI	0.41	**0.38**	0.30	0.42	0.38	0.30	0.44	0.42	0.34	0.22	0.36	0.38
CLIPCap	CC+pHHI	**0.42**	**0.38**	0.33	**0.45**	**0.44**	**0.44**	**0.33**	**0.47**	**0.40**	**0.40**	0.30	**0.47**
Results@5													
Env2	COCO	0.28	0.30	0.17	0.42	0.30	0.32	0.28	0.46	0.35	0.55	0.12	0.51
CLIPCap	COCO	0.31	0.50	0.21	0.42	0.29	0.38	0.28	0.46	0.34	0.52	0.23	0.49
CLIPCap	CC	0.36	0.41	0.23	0.46	0.30	0.31	0.23	0.50	0.32	0.23	0.14	0.43
CLIPCap	CC+WW	0.33	0.20	0.50	0.33	0.34	0.37	0.13	0.28	0.33	0.31	0.09	0.20
EncDec	pHHI	0.55	0.61	0.11	0.63	0.51	0.65	0.10	0.59	0.46	0.56	0.07	0.56
CLIPCap	CC+pHHI	**0.57**	**0.64**	**0.10**	**0.64**	**0.60**	**0.75**	**0.06**	**0.68**	**0.53**	**0.74**	**0.05**	**0.63**
Results@8													
Env2	COCO	0.29	0.34	0.15	0.42	0.31	0.35	0.25	0.48	0.36	0.59	0.10	0.53
Env2	COCO	0.32	0.53	0.18	0.43	0.30	0.41	0.25	0.47	0.35	0.55	0.21	0.50
CLIPCap	CC	0.38	0.47	0.17	0.48	0.32	0.35	0.19	0.53	0.34	0.26	0.11	0.45
CLIPCap	CC+WW	0.34	0.22	0.48	0.35	0.36	0.42	0.10	0.33	0.35	0.34	0.06	0.25
EncDec	pHHI	**0.60**	0.69	**0.06**	**0.69**	0.55	0.72	0.05	0.64	0.50	0.66	0.04	0.61
CLIPCap	CC+pHHI	**0.60**	**0.70**	0.07	0.68	**0.63**	**0.81**	**0.03**	**0.72**	**0.55**	**0.78**	**0.03**	**0.67**

Table 3. Results on ***Waldo and Wenda*** **split by data source** – Who's Waldo (WW), Conceptual Captions (CC), and COCO Captions. For models using beam search, we report results for top 1, 5, and 8 beams.

on pHHI) or extract its first verb using a syntactic parsing model. If syntactic parsing does not yield a verb, the zero vector is used as the given embedding.

5.5. Results and Discussion

For *Waldo and Wenda*, we report all of the metrics described above. For *imSitu-HHI*, we only use the verb similarity metric since the ground truth label is a single verb. We report average similarity over all samples in *imSitu-HHI* as well as displaying averages for the most-supported verbs. See Tables 2–4 for quantitative results, and see Figure 4 for a visual comparison on *Waldo and Wenda*. Note that we do not include CoFormer in the table of *imSitu-HHI* results since it was trained directly on some of these items; see the supplementary material for analysis of CoFormer on in-distribution and out-of-distribution images in *imSitu-HHI*.

Overall we see that training on our pseudo-labels improves performance on our benchmarks. In Tables 2 and 3, showing results on *Waldo and Wenda*, the best-performing model by all metrics is CLIPCap fine-tuned with our pseudo-labels. This holds across data sources, as seen in Table 3, showing that this improvement generalizes to images beyond those originating in the Who's Waldo dataset. This model is also the best-performing on average and across a majority of verb categories on *imSitu-HHI* as seen in Table 4. Qualitative comparison shows that the captioning models used as-is output text that is far from the ground truth HHI labels, containing many irrelevant details and not necessarily describing an interaction. This can be seen in Figure 4, where the CLIPCap (CC) captions contain many hallucinated, non-factual details.

While transfer learning with pretrained CLIPCap yields the best results, we also observe that the vanilla Encoder-Decoder fine-tuned on our pseudo-labels also performs well, achieving the second-best BLEURT score on *Waldo and Wenda* and second-best verbal similarity metrics overall and across many verb categories on *imSitu-HHI*. We infer that our pseudo-labels do impart semantic knowledge of HHI beyond simply cueing existing captioning models to the surface form of HHI labels. Nevertheless, CLIPCap fine-tuned on pHHI does generalize better across the data from all sources in *Waldo and Wenda* and to *imSitu-HHI* which is entirely out-of-distribution for this model.

We also note that the metrics improve dramatically for both datasets when considering 5 or 8 beams. This is consistent with the fact that beam search using models fine-tuned on pHHI outputs a list of diverse candidate interactions, allowing a more directed search in the space of HHI descrip-

		Average sim.	socializing	distributing	teaching	communicating	interviewing	lecturing	training	providing	instructing	giving	pushing	helping	asking	coaching	talking
Results@1																	
ENv2	COCO	0.22	0.19	0.07	0.22	0.21	0.28	0.19	0.25	0.26	0.11	0.45	**0.38**	0.29	0.44	0.28	0.28
CLIPCap	COCO	0.23	0.18	0.07	0.26	0.24	0.29	0.17	0.26	0.24	0.10	0.44	0.36	0.25	**0.46**	0.30	0.63
CLIPCap	CC	0.27	0.16	**0.25**	0.37	0.26	0.21	0.23	0.28	0.35	0.16	0.46	0.31	0.37	0.38	0.26	0.53
CLIPCap	CC+WW	0.09	0.02	0.08	0.12	0.05	**0.30**	0.08	0.10	0.09	0.06	0.25	0.07	0.09	0.11	0.05	0.08
EncDec	pHHI	0.28	0.19	0.21	0.34	0.27	0.23	0.24	0.35	0.38	0.17	0.60	0.36	0.34	**0.46**	**0.76**	0.64
CLIPCap	CC+pHHI	**0.32**	**0.21**	**0.25**	**0.56**	**0.33**	0.27	**0.30**	**0.43**	**0.44**	**0.19**	**0.66**	**0.38**	**0.44**	**0.46**	0.65	**0.70**
Results@5																	
ENv2	COCO	0.26	0.21	0.10	0.26	0.23	0.30	0.21	0.27	0.31	0.13	0.49	0.45	0.35	0.49	0.33	0.31
CLIPCap	COCO	0.25	0.19	0.09	0.29	0.26	0.31	0.20	0.27	0.26	0.12	0.47	0.41	0.29	0.48	0.32	0.66
CLIPCap	CC	0.35	0.21	0.30	0.48	0.33	0.28	0.31	0.37	0.43	0.21	0.56	0.42	0.48	0.47	0.31	0.64
CLIPCap	CC+WW	0.18	0.05	0.15	0.22	0.14	**0.60**	0.16	0.17	0.22	0.11	0.46	0.15	0.21	0.22	0.15	0.24
EncDec	pHHI	0.40	**0.30**	**0.35**	0.49	0.41	0.39	0.33	0.51	0.51	0.23	0.79	0.47	0.49	0.57	**0.92**	0.86
CLIPCap	CC+pHHI	**0.44**	0.29	**0.35**	**0.85**	**0.44**	0.41	**0.40**	**0.56**	**0.56**	**0.27**	**0.88**	**0.48**	**0.56**	**0.58**	0.86	**0.92**
Results@8																	
ENv2	COCO	0.28	0.22	0.12	0.27	0.24	0.30	0.21	0.28	0.32	0.13	0.51	0.47	0.37	0.50	0.35	0.34
CLIPCap	COCO	0.26	0.20	0.10	0.31	0.27	0.31	0.21	0.28	0.26	0.12	0.48	0.43	0.31	0.49	0.32	0.67
CLIPCap	CC	0.37	0.23	0.31	0.51	0.34	0.31	0.32	0.40	0.45	0.22	0.59	0.45	0.51	0.49	0.32	0.70
CLIPCap	CC+WW	0.21	0.06	0.16	0.25	0.16	**0.62**	0.16	0.20	0.24	0.12	0.51	0.20	0.23	0.26	0.18	0.32
EncDec	pHHI	0.44	**0.33**	**0.38**	0.55	0.44	0.42	0.34	0.54	0.56	0.25	0.85	0.49	0.56	0.59	**0.94**	0.91
CLIPCap	CC+pHHI	**0.47**	**0.33**	0.37	**0.90**	**0.46**	0.41	**0.43**	**0.60**	**0.59**	**0.29**	**0.92**	**0.50**	**0.59**	**0.61**	0.91	**0.96**

Table 4. Results on ***imSitu-HHI***. In addition to the average verb embedding similarity between predicted verbs and the ground truth verb, we also present mean similarities for the most common 15 verbs in *imSitu-HHI*. Best results are in bold, and second best are underlined. For models using beam search, we report results for top 1, 5, and 8 beams.

tors, while beam search applied to captioning models as-is tends to produce many slight variations of the same long caption.

6. Conclusion

We present a new framework for learning to understand human-human interactions in still images using weak supervision from textual captions. We demonstrate the use of knowledge distillation applied to a large language model without explicit supervision to produce pseudo-labels that can serve as targets for predicting interactions as free text. We show that training on these pseudo-labels enables HHI understanding beyond that of SOTA captioning and situation recognition models, and we provide the *Waldo and Wenda* as a new benchmark for this task.

There are various avenues for future research to extend our work. One possible direction is the incorporation of visual grounding into HHI understanding. We predict the most salient interaction in an image, which we assume to be the interaction the one that is described or suggested in its accompanying caption. It remains to localize the participants, including generalizing to group interactions where more than two participants are visible. Another important aspect that remains to be explored is the hierarchical nature of interactions. For example, the generic HHI label "meeting" is valid for almost every image, while "shaking hands" is more specific and valid for a subset of those images. Further research could extend our results to hierarchical prediction of multiple HHI labels for a single image.

Finally, we note the importance of style-content disentanglement in HHI prediction, which our work does not explicitly consider. Scene cues in images can be important for correctly identifying HHI, as illustrated in Figure 1, but also may be misleading. For instance, an image of soldiers in uniform is more likely to depict "saluting", but HHI is only valid if the image actually contains a salute. Future work on disentangling style and content shows promise for improving the robustness of HHI understanding models.

Acknowledgements. We thank Ron Mokady for providing helpful feedback. This work was supported by a research gift from Meta and the Alon fellowship.

References

[1] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 401–416. Springer, 2014.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, Mar. 2021. If you use this software, please cite it using these metadata.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

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

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[8] Ishani Chakraborty, Hui Cheng, and Omar Javed. 3d visual proxemics: Recognizing human interactions in 3d from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3406–3413, 2013.

[9] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

[10] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.

[11] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021.

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

[13] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19659–19668, 2022.

[14] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021.

[15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[16] Glenn Jocher et. al. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Oct. 2021.

[17] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.

[18] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[19] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

[20] Coert van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *International Workshop on Human Behavior Understanding*, pages 116–133. Springer, 2016.

[21] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018.

[22] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[23] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[24] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223, March 2017.

[25] Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*, 2022.

[26] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[27] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast

end to end training for image captioning. *arXiv preprint arXiv:2208.06551*, 2022.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[29] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021.

[30] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.

[31] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[32] Yu Kong, Yunde Jia, and Yun Fu. Interactive phrases: Semantic descriptions for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1775–1788, 2014.

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[34] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019.

[35] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

[36] Dong-Gyu Lee and Seong-Whan Lee. Human interaction recognition framework based on interacting body part attention. *Pattern Recognition*, 128:108645, 2022.

[37] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[38] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019.

[39] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. Visual-semantic graph attention networks for human-object interaction detection. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1441–1447. IEEE, 2021.

[40] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[41] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning, 2022.

[42] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.

[43] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

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

[45] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[46] Khadidja Nour el houda Slimani, Yannick Benezeth, and Feriel Souami. Human interaction recognition based on the co-occurence of visual words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 455–460, 2014.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[48] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, volume 1, page 33, 2010.

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[50] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020.

[51] Liliana Lo Presti and Marco La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016.

[52] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for mt. In *Proceedings of EMNLP*, 2021.

[53] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

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

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[57] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. *arXiv preprint arXiv:1506.02203*, 2015.

[58] Michael S Ryoo and JK Aggarwal. Ut-interaction dataset, icpr contest on semantic description of human activities (sdha). In *IEEE International Conference on Pattern Recognition Workshops*, volume 2, page 4, 2010.

[59] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

[60] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[61] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018.

[62] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1110–1118, 2019.

[63] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

[64] Alexandros Stergiou and Ronald Poppe. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding*, 188:102799, 2019.

[65] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[66] Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1250–1257. IEEE, 2012.

[67] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Multi-stream pose convolutional neural networks for human interaction recognition in images. *Signal Processing: Image Communication*, 95:116265, 2021.

[68] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[69] Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. Hands-on: deformable pose and motion models for spatiotemporal localization of fine-grained dyadic interactions. *EURASIP Journal on Image and Video Processing*, 2018(1):1–16, 2018.

[70] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019.

[71] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3048–3056, 2017.

[72] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[73] Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

[74] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[75] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1609, 2015.

[76] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2019.

[77] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[78] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[79] Yi Yang, Simon Baker, Anitha Kannan, and Deva Ramanan. Recognizing proxemics in personal photos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3522–3529. IEEE, 2012.

[80] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[81] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. A survey on human motion analysis from depth data. In *Time-of-flight and depth imaging. sensors, algorithms, and applications*, pages 149–187. Springer, 2013.

[82] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

[83] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012.

[84] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[85] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005, 2019.

[86] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[87] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019.

[88] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 843–851, 2019.

[89] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020.

[90] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021.

Appendix

We refer readers to the interactive visualizations at our project page that show results for all presented models on the two test sets we examine (*Waldo and Wenda* and *imSitu-HHI*). In this document, we describe additional experiments and results (Section A) and provide additional details (Section B).

A. Additional Results and Comparisons

A.1. CoFormer on *imSitu-HHI*

The CoFormer grounded situation recognition model, whose results on *Waldo and Wenda* are reported in the main paper, was trained on the SWiG dataset, which extends the imSitu dataset with grounding information. [13, 50, 80] Since *imSitu-HHI* also includes some of this data, CoFormer's performance on *imSitu-HHI* is not comparable to the out-of-distribution performance of the other models we consider. Nevertheless, we can use its performance on *imSitu-HHI* as a rough upper bound for this task. We report its performance on all of *imSitu-HHI*, which includes some of its training data, as well as on the intersection of *imSitu-HHI* with imSitu's train, dev, and test sets alone. See Table 5 for these metrics and a comparison to the Enc-Dec model trained on our pseudo-labels. As expected, CoFormer's performance is much higher on its own training data, and generally outperforms our model by this metric on imSitu. However, CoFormer was trained using the verb labels from imSitu, while our model, trained without supervision from manually-labelled data, is being evaluated out-of-distribution and without regard to the additional text in its predictions besides the predicted verb.

Model	Data	Eval split	sim.
CoFormer	SWiG	all (8021)	0.63
EncDec	pHHI	all (8021)	0.28
CoFormer	SWiG	train (4906)	0.73
EncDec	pHHI	train (4906)	0.34
CoFormer	SWiG	dev (1549)	0.50
EncDec	pHHI	dev (1549)	0.27
CoFormer	SWiG	test (1566)	0.48
EncDec	pHHI	test (1566)	0.28

Table 5. CoFormer results on ***imSitu-HHI*** as described in Section A.1, with Enc-Dec model for comparison. CoFormer was trained with supervision from the imSitu train set, while our models did not see any of these samples during training; therefore, we treat the CoFormer model performance as an upper bound for achievable verb similarity on this dataset in the out-of-distribution setting. The "Data" column shows the model's training data. The "Eval split" column gives the evaluation data split used and its size - either the entire 8,021-sample *imSitu-HHI* subset of imSitu, or else its intersection with imSitu's train, dev, or test sets. The average verb embedding similarity is shown as "sim.". Note that SWiG here refers to the train set of imSitu along with grounding data. Enc-Dec model results refer to top-1 predictions.

A.2. Extended-*imSitu-HHI* results

In Section B.6, we described the construction of the 8,021-sample *imSitu-HHI* dataset, a subset of the full imSitu dataset. One of its design choices was the final filtering of verbs by number of supported images, to use only those verbs with at least 100 images after filtering for human detections and semantic arguments. We now present results on an extended version of this dataset where we lower the threshold for the required number of images supporting a verb and thus keep a larger subset of imSitu.

See Table 6 for quantitative results. We observe that decreasing the minimum required support of verbs increases the number of unique verbs dramatically, but has a minimal impact on the verb embedding similarity metric when lowered from 100 to 50. However, lower thresholds more significantly impact the verb similarity scores. This comports with the observation that verbs with higher support values are more likely to represent HHI.

We include examples of verbs with support values at different levels to illustrate this intuition:

Verbs with support ≥ 180: *socializing, distributing, teaching, communicating, interviewing, lecturing, training, providing, instructing, giving, pushing, helping, asking, coaching, selling, talking, educating*

Verbs with support $\in [100, 120]$: *imitating, offering, plunging, pitching, reassuring, autographing, clapping, ignoring, dousing, speaking, operating, wheeling, loading*

Verbs with support $\in [50, 55]$: *repairing, chasing, drumming, applauding, breaking, eating, climbing, officiating, carting, deflecting, building, measuring*

Verbs with support $\in [20, 25]$: *colliding, guarding, submerging, twirling, rocking, miming, clearing, calming, sowing, massaging, nuzzling, butting, tasting, waxing, clenching, knocking, scooping, stacking, vaulting, shopping*

Verbs with support $\in [1, 2]$: *curtsying, coughing, reading, crawling, surfing, dialing, erasing, slipping, marching, frying, dripping, phoning, mopping, bulldozing, sharpening, walking, landing, boating, circling, boarding, skipping, shivering, signing, flapping, crouching, sneezing, raking, launching, protesting, piloting, unplugging, ejecting, praying, typing, stitching, watering, queuing*

A.3. Training on syntactic parsing-based seeds

To ablate the effect of our pseudo-labelling, we compare to results when training directly on syntactic parsing-based seeds. As described in the main paper, these can sometimes be extracted from Who's Waldo captions when they fit a particular syntactic pattern, specifically containing an interaction verb with arguments representing the relevant partic-

Support	Verbs	Samples	sim@1	sim@5	sim@8
≥ 100	50	$\sim 8k$	0.28	0.40	0.44
≥ 50	98	$\sim 11k$	0.28	0.40	0.43
≥ 20	178	$\sim 14k$	0.26	0.38	0.41
≥ 0	359	$\sim 15k$	0.25	0.37	0.40

Table 6. Results of the EncDec model on extended-*imSitu-HHI*, as described in Section A.2.

		Waldo and Wenda				*imSitu-HHI*		
Method	Data	BL	sim	n_i	n_v	sim	n_i	n_v
EncDec	pHHI	0.38	0.41	298	100	0.28	1468	245
CLIPCap	CC+pHHI	0.42	0.46	158	86	0.32	325	133
EncDec	SP	0.33	0.36	126	66	0.24	216	82
CLIPCap	CC+SP	0.41	0.44	123	78	0.29	268	129

Table 7. Comparison of results when training on syntactic parsing-based seeds ("SP") versus our pseudo-labels ("pHHI"), as described in Section A.3. "BL" refers to BLEURT and "sim" refers to verb embedding similarity. On *Waldo and Wenda*, results are aggregated across data sources.

ipants.

Out of the $\sim 126k$ images from Who's Waldo that we used, only $\sim 23k$ have captions that yield a syntactic parsing-based seed (while pseudo-labels could be assigned to all of them). Therefore in this ablation the models train on $< 20\%$ the number of images used to train the models with pseudo-labelling.

We compare results on *Waldo and Wenda* and *imSitu-HHI* when training only on these seeds versus training on our pseudo-labels in Table 7. In addition to the textual similarity metrics, we include two simple measures of diversity: the number of unique interaction texts in the predictions (n_i) and the number of unique predicted verbs (n_v) across all test items. Although diversity metrics are less meaningful for comparisons to the output of captioning models used as-is, since their outputs are highly detailed, they can be used in this case since the models under comparison all output predictions of roughly the same length and level of detail. Models trained on pHHI show higher similarity to the ground truth labels as seen in the reported textual similarity metrics. In addition, we see a significant increase in diversity relative to training on syntactic parsing seeds. This suggests that the large increase in training data provided by pseudo-labelling allows models to represent a larger space of interactions, consistent with our goal in modelling the heavy tail of possible HHI. This is also illustrated in Figure 5, which compares outputs of two models (both pretrained on CC captions)—one trained with our pseudo-labels and the other with the set of syntactic parsing-based seeds.

Method	Data	BL	BE	BA	SC
ENv2	COCO	0.27	0.87	-6.25	**0.24**
CLIPCap	COCO	0.28	0.87	-7.24	**0.24**
CLIPCap	CC	0.27	0.86	-6.66	0.23
CLIPCap	CC+WW	0.26	0.85	-5.90	0.22
EncDec	pHHI	0.38	0.92	-3.53	0.22
CLIPCap	CC+pHHI	**0.42**	**0.93**	**-3.34**	0.22

Table 8. Comparison of BLEURT and additional neural metrics on captioning models, aggregated across data sources in *Waldo and Wenda*, as described in Section A.4. Metrics shown are BLEURT (BL), BERTScore (BE), BARTScore (BA), and SummaC (SC).

A.4. Additional neural metrics

In addition to BLEURT, we report metrics for additional neural metrics for natural language generation. Fmor easuring textual similarity between predictions and ground truth HHI labels, we provide results for BERTScore [86] and BARTScore [82]. We also measure factuality of predictions relative to ground truth captions (similar to the NLI scores reported in the main paper) using the model SummaC [35]. Table 8 for results on captioning models, aggregated over data sources in *Waldo and Wenda*.

BERTScore uses the default pretrained checkpoint for English provided by the Hugging Face `evaluate` library[4], and we report the output F1 score. BARTScore uses the model trained on ParaBank2 provied in the official BARTScore repository[5]. SummaC scores use the default checkpoint and settings for SummaC-Conv provided in its official repository[6]. For all of these models, we replace `[NAME]` with the text "person" as needed, just as we do for calculating BLEURT scores (see Section B.9)

We see the textual similarity metrics (BERTScore, BARTScore) pattern similarly to BLEURT in supporting the use of our pHHI as training data. SummaC scores are slightly higher for captioning models trained on COCO and used as-is, possibly reflecting generic text that is closer to ground truth captions though not necessarily effective at capturing HHI.

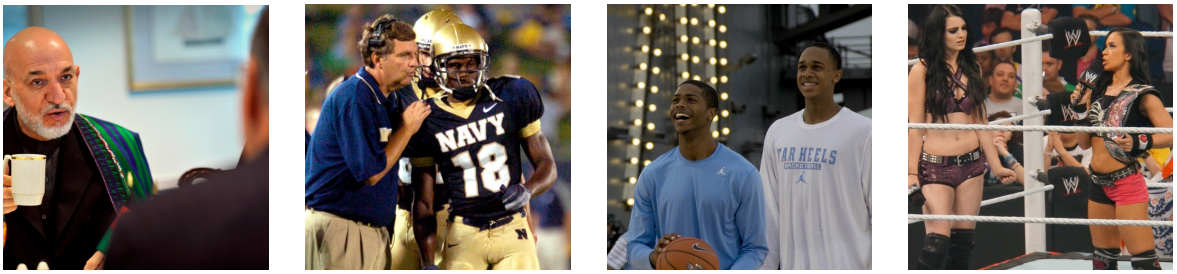
A.5. Ngram-based metrics

In this section we discuss the use of BLEURT [59] as our main textual metric rather than ngram-based metrics such as BLEU [47]. Ngram-based metrics are common in text generation tasks such as machine translation, comparing predicted texts to a ground truth reference (or multiple references). They have the advantage of being simple and fast to calculate, but focus on surface forms of text rather than underlying semantics.

[4] https://huggingface.co/spaces/evaluate-metric/bertscore

[5] https://github.com/neulab/BARTScore

[6] https://github.com/tingofurro/summac



CLIPCap (CC+pHHI)	[⋆] being interviewed by [⋆]	[⋆] coaching [⋆]	[⋆] playing basketball with [⋆]	[⋆] wrestling with [⋆]
CLIPCap (CC+SP)	[⋆] talking with [⋆]	[⋆] talking with [⋆]	[⋆] playing with [⋆]	[⋆] playing with [⋆]

Figure 5. Examples of diverse predictions on *Waldo and Wenda* from a model trained with our pseudo-labels, compared to predictions when trained on syntactic-parsing based seeds ("SP"). See Section A.3 for details.

We provide a comparison of BLEU and BLEURT scores in Table 9, aggregated across data sources in *Waldo and Wenda*. We provide scores for a captioning model (CLIPCap trained on Conceptual Captions) and a model fine-tuned on our pseudo-labels. Although the latter has a higher BLEU score, its extremely low value (0.06) is due to the fact that only 72 out of 1,000 predictions achieve a nonzero BLEU score relative to the ground truth labels. Because BLEU measures ngram precision and the ground truth labels are short, it returns zero unless the prediction is a near-perfect textual match. This effectively ignores the vast majority of predictions, unlike BLEURT which [59] show to have a robust correlation with human judgements of semantic similarity at the sentence level.

We additionally provide scores for the METEOR metric [2], which uses unigram alignment statistics and incorporates both precision and recall. It also uses stemming and synonym matching to provide some robustness relative to changes in the surface forms of semantically similar texts. Although METEOR does not ignore most predictions as does BLEU, we find that it underperforms BLEURT in capturing semantic similarity in our setting. The baselines in Table 9 are calculated by (1) using the constant text "[NAME] meeting with [NAME]", and (2) randomizing the order of the predictions of the model fine-tuned on our pseudo-labels. Both baselines achieve a relatively high METEOR score, while BLEURT decreases significantly and approaches the BLEURT score of the plain captioning model. This suggests that METEOR is biased towards measuring surface similarity rather than underlying semantics, consistent with the findings of [59] who explicitly compare METEOR and BLEURT. This can also be seen in the qualitative examples in Table 10 of prediction (CLIPCap CC+pHHI) and ground truth pairs from *Waldo and Wenda* where METEOR and BLEURT differ strongly in magnitude.

Method	Data	BLEU	METEOR	BLEURT
CLIPCap	CC	0.00	0.13	0.27
CLIPCap	CC+pHHI	0.06	0.46	0.42
baseline (constant)		0.00	0.36	0.29
baseline (jumbled)		0.00	0.43	0.33

Table 9. Comparison of ngram-based metrics and BLEURT, aggregated across data sources in *Waldo and Wenda*, as described in Section A.5.

We replace [NAME] with the text "person" as needed to calculate these scores, just as we do for calculating BLEURT scores (see Section B.9).

A.6. Ablation of few-shot learning for synthetic caption generation

In order to ablate few-shot examples used when generating synthetic captions (see Section B.4), we split our synthetic caption-interaction pairs into two non-overlapping folds, train summarization models on each of these folds and then generate pseudo-labels with each model. We fine-tune CLIPCap+CC on these pseudo-labels and evaluate the resulting models on *Waldo and Wenda*, as shown in Table 11. The negligible differences across all metrics suggest that our method is robust to the particular (randomly selected) few-shot examples used in training the summarizer.

A.7. Qualitative results

See our project page for an interactive visualization of the results of all of the considered models on the *Waldo and Wenda* 1,000-item test set and on the 8,021-item *imSitu-HHI* dataset.

Ground truth	Prediction	METEOR	BLEURT
[⋆] wrestling with [⋆]	[⋆] competing against [⋆]	0.25	0.65
[⋆] giving signatures to [⋆]	[⋆] signing autographs with [⋆]	0.20	0.57
[⋆] arguing with [⋆]	[⋆] driving with [⋆]	0.64	0.28
[⋆] making sandcastles with [⋆]	[⋆] working with [⋆]	0.77	0.24

Table 10. Comparison of METEOR and BLEURT scores on selected examples from *Waldo and Wenda*, as described in Section A.5. Predictions are from CLIPCap trained on Conceptual Captions and fine-tuned on our pseudo-labels.

Method	Training Data	BL↑	p_e ↑	p_c ↓	sim↑
CLIPCap	CC+pHHI$_1$	0.39	0.35	0.35	0.43
CLIPCap	CC+pHHI$_2$	0.39	0.35	0.37	0.42

Table 11. Few-shot learning ablation. pHHI$_1$ and pHHI$_2$ refer to pseudo-labels produced from summarizers trained on two non-overlapping splits of our synthetic caption-interaction data.

B. Additional Details

B.1. Scraping additional captions from CC-News

In order to find additional caption texts for use in our knowledge distillation process, we use the CC-News dataset as available via Hugging Face datasets[7], containing the text of $\sim 708k$ scraped English language news articles from 2017 through 2019 [24]. These frequently include the text of captions accompanying images in news articles. To roughly filter for these captions, we select lines of $\leq 1,000$ characters that contain any of the following textual patterns: "(left)", "(right)", "(center)", ", left,", ", right,", ", center,", ", centre,", ", pictured,", "PHOTO: ", "Photo by", "Image copyright", "Getty ", "AP Photo", "AP Image".

In captions that we extract, we remove those patterns along with the following, so that the extracted captions will not all contain common substrings: "(Image ...)", "(Photo ...)", "(AP Photo ...)", "(Credit ...)", "[Image ...]", "[Featured Image ...]", "Getty Images", "Image copyright ... Image caption", "Photo:", "FILE PHOTO:", "Image (number) of (number)".

Finally, we discard captions that did not contain an interaction as extracted in Section B.2. This left us with 6,212 captions. Examples of such captions from CC-News include the following (patterns detected and removed are shown in red strike-through text):

- Northern Ireland's Corry Evans~~, left,~~ and Germany's Toni Kroos battle for the ball during their 2018 World Cup Group C qualifying soccer match at Windsor Park, Belfast, Thursday, Oct. 5, 2017. (Brian Lawless/PA via AP)
- Arizona Coyotes defenseman Luke Schenn (2) and Los Angeles Kings left winger Kyle Clifford (13) reach for the puck during the second period of an NHL hockey game in Los Angeles on Saturday, Feb. 3, 2018.~~ (AP Photo/Reed Saxon)~~
- ~~Image copyright~~ Kalpana Vaughan Wilson ~~Image caption~~ Kalpana Wilson pictured with daughter Clara shortly after giving birth

B.2. Syntactic parsing-based interactions

We use syntactic parsing with spaCy's `en_core_web_trf` model to extract interactions from CC-News and Who's Waldo captions using the procedure described below. As described in Section B.3, the parsing-based interactions from CC-News are used as seeds to generate more novel interaction texts. Then, as further described in B.5, the interactions from Who's Waldo and novel interactions are used to generate synthetic interaction-caption pairs for use in training a summarization model.

For each caption, we search for verbs that it contains. For each verb lemma V, we consider all of its children in the syntactic parse tree. For child node X, we extract the text of X's syntactic head. If X is a preposition, we also extract the head of its complement, and for any determined noun we also extract the text of its determiner. We filter out any such X containing named entities of types `DATE`, `GPE`, `FAC`, `ORG`, `LOC`, or `TIME`, and if X contains coordinated human named entities ("`NAME` and `NAME`") we include both of them. We mask all human name entities using the special token `[NAME]`. We concatenate all of these together, including V in present continuous form, to form an extractive interaction text. Finally we filter for such texts containing at least two `NAME` entities, with at least one of them being a syntactic subject[8].

Also note that since captions from Who's Waldo already have human names masked as `[NAME]`, we first replaced these tokens with generic names ("Adam, Bob, ...") before applying syntactic parsing, so that the input text would be valid English.

Among CC-News captions, 6,212 captions include such interactions. Interactions extracted from CC-News captions include the following:

[7] https://huggingface.co/datasets/cc_news

[8] The default entity labels in this parsing model use the label `PERSON` for human entities, but we use `NAME` for consistency with later sections.

- **CC-News caption:** Chinese President Xi Jinping (L) and First Lady Peng Liyuan bid farewell as they board their plane to depart from the Julius Nyerere International Airport in Dar es Salaam, Tanzania, March 25, 2013. REUTERS/Thomas Mukoya/File Photo
  - **Extracted interaction:** [NAME] and [NAME] bidding farewell
- **CC-News caption:** Colombia's Radamel Falcao jumps for the ball with England's Harry Maguire during the round of 16 match between Colombia and England at the 2018 soccer World Cup in the Spartak Stadium, in Moscow, Russia, Tuesday, July 3, 2018.
  - **Extracted interaction:** Colombia [NAME] jumping for the ball with England [NAME] during the match
- **CC-News caption:** Chuck Munro and Brian Alexander of Spraying Systems welcome Eric Vetters of ProCorr to their booth at NACE 2018 in Phoenix.
  - **Extracted interaction:** [NAME] and [NAME] welcoming [NAME] to their booth

In addition, 22,637 captions from Who's Waldo include such interactions. Interactions extracted from Who's Waldo captions include the following:

- **Who's Waldo caption:** Chief of Naval Operations Adm. [NAME] speaks at the Navy and Marine Corps Relief Society ball with Vice Commandant of the Marine Corps Gen. [NAME] at the Washington Hilton.
  - **Extracted interaction:** [NAME] speaking at the ball with [NAME] at the Hilton
- **Who's Waldo caption:** [NAME] and [NAME] discuss Ancestry at the Maltz Performing Arts Center
  - **Extracted interaction:** [NAME] and [NAME] discussing Ancestry at the Center
- **Who's Waldo caption:** NASA astronaut [NAME] (left) and Japan Aerospace Exploration Agency (JAXA) astronaut [NAME], both Expedition 20 flight engineers, perform a check of the Synchronized Position Hold, Engage, Reorient, Experimental Satellites (SPHERES) Beacon / Beacon Tester in the Destiny laboratory of the International Space Station.
  - **Extracted interaction:** [NAME] and [NAME] performing a check in the laboratory

Note that these extracted interactions may contain prepositional phrases. We remove prepositional phrases from results when generating synthetic interaction-caption pairs, as described in Section B.4.

B.3. Generating novel interaction texts

Among the 6,212 CC-News captions with interactions, we have only 3,146 unique interaction texts as extracted by the parsing-based model described above. In order to have access to a richer set of interactions for training the subsequent summarization model, we use text generation with a large language model to generate more interactions similar to those extracted from CC-News captions with the above method, using the parsing-based interactions as seeds. We use few-shot prompting by providing 10 random newline-separated parsing-based interactions from CC-News captions as a prompt to the large language model GPT-Neo-1.3B [3, 18] and generating until the next newline. We use nucleus sampling [26] with $p = 0.95$, as well as a constraint to prevent repeated trigrams. We also replace [NAME] mask tokens with generic names ("Alex, Bailey, ...") so that the input text is more natural English and thus more in distribution for the language model. We discard texts that do not pass the following filters:

- Text contains "Alex" and "Bailey" in order, exactly once, and no other names.
- Text does not contain uppercase letters, besides in names.
- Text must contain a word ending in "-ing".
- Text does not end with " the" or " a".

Finally, we re-mask names with the token [NAME]. In this way we generate $\sim 116k$ novel interaction texts used for synthetic interaction-caption pairs as described in Section B.4.

Examples of such randomly generated interaction texts include the following:

- [NAME] kissing [NAME] after a win
- [NAME] handing [NAME] an autograph sheet
- [NAME] congratulating [NAME] in victory
- [NAME] calling [NAME] in a business suit
- [NAME] hugging [NAME]
- [NAME] telling [NAME] he'll have
- [NAME] catching a short pass from [NAME] during a play

- [NAME] receiving a high five from [NAME] in the post
- [NAME] giving [NAME] congratulations for a goal during a period
- [NAME] telling [NAME] that he's glad he came out to see him
- [NAME] as [NAME] is being picked
- [NAME] shooting over [NAME] during practice
- [NAME] saying to [NAME] what he is going to do
- [NAME] watching [NAME] celebrate with teammates as the ceremony began
- [NAME] walking with [NAME] around the deep area

As mentioned above, these may contain prepositional phrases, which are removed later as discussed in Section B.4.

B.4. Synthetic interaction-caption pair generation

Using syntactic parsing-based caption-interaction pairs from Who's Waldo data, described in Section B.2, and novel interaction texts from CC-News, described in Section B.3, we use few-shot learning to generate training data for an abstractive summarization model as follows:

For each inference iteration, we construct a few-shot prompt by selecting 10 interaction-caption pairs $(I_1, C_1), \cdots, (I_k, C_k)$ using captions from Who's Waldo and syntactic parsing-based interaction texts, and a single novel CC-News based interaction I^*. For each pair (I_i, C_i), as well as in I^*, we replace [NAME] tokens with random names using the random-name library[9] library. We then construct a prompt containing the following texts, in order and newline-separated:

- For $i = 1, \cdots, k$:
  - "Caption of image showing I_i"
  - C_i
- "Caption of image showing I^*:"

We input this prompt to GPT-Neo-1.3B [3, 18] and generate text until a newline is output. We generate using nucleus sampling [26] with $p = 0.95$, temperature 0.7, a constraint to prevent repeated trigrams, and a maximum output length of 200 tokens.

Denote the output of generation by C^*. The pairs (I^*, C^*) generated by this method are noisy, so we select for valid synthetic interaction-caption using the following filters:

- C^* must contain the same random names that were used for I^* in the prompt
- C^* must entail I^* ($p_e > 0.5$), as measured by the entailment probability p_e calculated by a pre-trained NLI model. We use BART-large [37] fine-tuned on the MNLI dataset [74] (using the facebook/bart-large-mnli checkpoint from Hugging Face model hub[10]).
- I^* must contain a verb, checked using spaCy's en_core_web_trf syntactic parsing model.
- I^* may not contain any of the following banned substrings, which are common artifacts that do not reflect interactions: "photo", "image", "picture", "in this", "In this"

Finally, we postprocess each I^* with the following steps:

- Remove prepositional phrases that do not contain [NAME]. For example: "[NAME] meeting with [NAME] at a hotel" → "[NAME] meeting with [NAME]".
- Normalize subjects of verbs containing two or more people joined by "and", "with", "&" and/or commas, by replacing them with "with [NAME]" at the end of an interaction. For example: "[NAME] and [NAME] meeting" → "[NAME] meeting with [NAME]".

It total, we generate 62,176 synthetic interaction-caption pairs with this method. Examples of such pairs include the following:

1. **Caption:** Estella, a member of the Women's Auxiliary Fire Corps, hugs Lorne, the President of the United States, at a ceremony honoring firefighters at the White House in Washington, D.C. on Sept. 30, 2012.
   Interaction: [NAME] hugging [NAME]
2. **Caption:** Angelia shoots the puck in the face of Gladi during a game on April 27, 2012, at the St. Louis Blues home rink in St. Paul, Minn.
   Interaction: [NAME] shooting the puck against [NAME]
3. **Caption:** Emmie receives a letter in her mailbox from Jacinthe.
   Interaction: [NAME] receiving a letter from [NAME]
4. **Caption:**The hug between Bettye and Hester is a moment of joy in the life of Hester and Bettye. It was a special moment for all of them. It is a special memory for Bettye, and it is a great moment for Hester, and

[9] https://github.com/dominictarr/random-name

[10] https://huggingface.co/facebook/bart-large-mnli

that's how it should be.
Interaction: [NAME] hugging [NAME]

5. **Caption:** Kippie, who attended the conference, asked Paulie to make an official statement on the issue of the military's role in the US Embassy in Timor-Leste. Paulie stated that he would not comment on the matter.
Interaction: [NAME] pressuring [NAME]

Note that although the interaction often contains the same verb as the accompanying caption, it may also contain a verb based on non-verbal cues ("hugging" in example 4 above, with the noun "hug" in the caption) or even based on the general meaning of the synthetic caption ("pressuring" in example 5 above).

B.5. Pseudo-label generation

Using the synthetic interaction-caption pairs (I, C) described and illustrated in Section B.4, we fine-tune a pre-trained T5 model [56] using the "summarize:" task prefix on these pairs, using each I as the target. We use `T5-base` and fine-tune for 3 epochs with batch size 8, initial learning rate $5e-5$ with linear schedule, AdamW optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$, and maximum gradient norm of 1.0, and otherwise default hyperparameter settings as defined in the Hugging Face summarization model training script.[11]

After fine-tuning, we apply this model to each caption in the Who's Waldo dataset corresponding to samples with ≥ 2 facial detections, as provided in the dataset, to create pseudo-labels. We filter these to only keep those pseudo-labels beginning with [NAME], followed by a present progressive verb ("-ing"), followed by more text containing exactly one additional [NAME]. We filter out examples containing any of the banned substrings "photo", "image", or "picture" since these often are artifacts that do not reflect interactions.

Finally, in order to avoid data leakage with the test set, we remove any samples with captions identical to those in the test set, or with identical date-time metadata fields (since these often are images taken from the same event).

In total, this procedure yielded 126,696 pseudo-labels for Who's Waldo, including 1,263 unique verbs, and 16,136 unique interactions.

Examples of such pseudo-labels created from Who's Waldo captions include the following:

Caption: The Assistant Commandant of the Marine Corps, Gen. [NAME], [NAME], left, poses for a photo with Master Sgt. [NAME] during the U.S. Marine Corps Command, Control, Communications and Computers (C4) annual awards dinner in Arlington, Va., April 17, 2014. The awards presented included the Gen. [NAME] for outstanding communications leadership, the James Hamilton Information Technology Management Civilian Marine of the Year Award, the Pfc. Herbert A Littleton Non-Commissioned Officer Trophy for operational communications excellence, and the Lt. Col. [NAME] Memorial Unit Award.
Pseudo-label: [NAME] posing with [NAME]

Caption: [NAME] and [NAME] at Governor [NAME] annual address in February 2016
Pseudo-label: [NAME] standing next to [NAME]

Caption: With Italian Prime Minister [NAME].
Pseudo-label: [NAME] talking with [NAME]

Caption: [NAME] at the Gothenburg Book Fair 2014.
Pseudo-label: [NAME] standing with [NAME]

Caption: Commemoration of 150th birth anniversary of [NAME], organized by the Ministry of Culture, Government of India.
Pseudo-label: [NAME] congratulating [NAME]

Caption: General [NAME], Air Force Chief of Staff, addresses the 347th Wing personnel. Senator [NAME] is standing next to the general.
Pseudo-label: [NAME] standing next to [NAME]

Caption: Luge World Cup Men 2017/18 in Altenberg: Flower Ceremony – [NAME], [NAME], [NAME]
Pseudo-label: [NAME] congratulating [NAME]

Caption: US Reality TV Star And Fashion Expert [NAME] in Sydney, by [NAME] 'How Do I Look' was the topic of conversation at King's Cross Barrio Chino tonight. US reality television star [NAME] and host of the 'How Do I Look' show was the main attraction. The red carpet came out as [NAME] and a few familiar Sydney faces did their walks and poses.
Pseudo-label: [NAME] talking to [NAME]

Caption: Crown [NAME] and [NAME] of Sweden during the inauguration of the Northern Link in Stockholm November 30, 2014.
Pseudo-label: [NAME] standing next to [NAME]

Caption: [NAME], french politician, Brive la Gaillarde book fair, France, 2010 11 06
Pseudo-label: [NAME] attending [NAME]'s book fair

[11] As of v4.18.0, script available at https://github.com/huggingface/transformers/blob/31ec2cb2badfbdd4c1ac9c6c9b8a74e974984206/examples/pytorch/summarization/run_summarization.py

Caption: [NAME] during 2013 World Championships in Athletics in Moscow.
Pseudo-label: [NAME] standing with [NAME]

Caption: [NAME] shakes hands with Vice President [NAME] shortly after becoming a U.S. citizen during a naturalization ceremony on Camp Victory in Baghdad, July 4, 2010. [NAME], assigned to the 82nd Airborne Division's 307th Brigade Support Battalion, 1st Advise and Assist Brigade, is originally from Colombia.
Pseudo-label: [NAME] shaking hands with [NAME]

Caption: A bit of 'Underbelly' blurb that we got hold of (thanks [NAME] - author of Razor) reads...Back in the day the East Village was called 'The Tradesman's Arms', a bloodhouse with sawdust on the floor to soak up the spit and vomit, hard stools at the bar and a dozen cheap wooden tables with chairs scattered around". The cast of Underbelly Razor and special guests partied into the night celebrating the Underbelly Razor Uncut DVD release at the very same place that crime queens [NAME], [NAME], along with [NAME] frequented back in their heyday. Strutting the blood red carpet was all of the Razor cast, including [NAME], better known now as our vice queen [NAME], [NAME] who played [NAME], [NAME] ([NAME]), [NAME], better recognised as the [NAME], [NAME], aka the suave [NAME]' [NAME] and [NAME], who we know as [NAME]. [NAME] tells us of the former glory days of 'The Arms', recounted from the many interviews he conducted, compiling the book, [NAME]. The red carpet event brought out the inner gangster in a few of us with [NAME] stating she would consider more 'Underbelly Razor' type roles under the right circumstances, [NAME] telling us to watch out for his uncut and fight scenes, and [NAME] saying he was a "fashionable gangster".
Pseudo-label: [NAME] hitting the red carpet with [NAME]

Caption: [NAME] and wife [NAME]
Pseudo-label: [NAME] sitting with [NAME]

Caption: [NAME] at 2017 European Athletics U23 Championships
Pseudo-label: [NAME] standing with [NAME]

Caption: [NAME], coach of the french feminine ski-jumping team 2010
Pseudo-label: [NAME] coaching [NAME]

Caption: [NAME] on the red carpet for 'Gods of Egypt' in New York City on February 24, 2016.
Pseudo-label: [NAME] standing with [NAME]

Caption: SEOUL (July 6, 2009) Chief of Naval Operations (CNO) Adm. [NAME] receives the National Security Merit Tongil Medal for his outstanding and meritorious service rendered to the Republic of Korea. [NAME] is on an official visit to the U.S. 7th Fleet area of responsibility to strengthen global maritime partnerships.
Pseudo-label: [NAME] receiving [NAME]'s award

Caption: [NAME], a retired United States Marine Lieutenant Colonel, and administrator at the State University of New York's Maritime College, being promoted to two-star general in New York's Military Forces.
Pseudo-label: [NAME] being promoted by [NAME]

Caption: Pabradė, Lithuania – Maj. Gen. [NAME], Pennsylvania's adjutant general, shakes hands with Maj. Gen. [NAME] in an APC 113 used by the Lithuanian Army while preparing to tour the training grounds. [NAME] visited the exercise Amber Hope 2011 June 22 while conducting his first trip to Lithuania as Pennsylvania's adjutant general.
Pseudo-label: [NAME] shaking hands with [NAME]

B.6. *imSitu-HHI* details

We form *imSitu-HHI*, an 8,021-sample subset of the imSitu dataset [80], as described here.

Because we only use this data to evaluate our models, and in order to have a sufficiently large sample size in the final subset, we use all data from imSitu dataset (train, validation and test set combined together). In total this includes 126,102 samples. Using person detections from a pretrained YoloV5 model (`ultralytics/yolov5` checkpoint[12], pretrained on MS COCO)[16], we discard samples whose images have less than two person detections. We also filter using the semantic frame data from imSitu, to select for samples with at least two human participants. Since arguments are not directly labelled as human or non-human, we use NLI-based filtering to select for human arguments. There are 146,347 unique argument texts in imSitu. For each such arugment A, we apply a pretrained NLI model (BART-large finetuned on MNLI, as described in Section B.4) to the following pair of texts:

- **Premise:** This is a *A*.
- **Hypothesis:** This is a human.

The model returns an entailment probability p_e for each such text pair, and we classify *A* as a human participant if $p_e > 0.5$. We remove all samples containing less than two arguments that are classified as human.

[12] https://hub.docker.com/r/ultralytics/yolov5

13,560 of the unique argument texts are classified as human, including the following examples:

- alpha
- desk sergeant
- Alfred the Great
- chief justice
- Gregory Pincus
- Pablo Neruda
- Spanish people
- abidance
- friend
- Cline

Examples of the remaining argument texts not classified as human include the following:

- sugar beet
- barouche
- water development
- St. John's
- stopper
- horsehair
- stripe
- advocator
- readjustment
- flamingo plant

It can be seen that the arguments have a very heavy-tailed distribution, with many rare or highly specific texts, and the NLI filtering contains noise. However we find this filtering to be a useful heuristic in addition to other forms of filtering.

We filter out samples containing the following verbs with negative or inappropriate connotations: *ailing, apprehending, arresting, attacking, bandaging, begging, biting, bothering, brawling, burning, clawing, complaining, confronting, crying, destroying, detaining, disciplining, dissecting, exterminating, frisking, frowning, gambling, grieving, grimacing, handcuffing, hanging, hitting, hunting, interrogating, misbehaving, mourning, panhandling, peeing, pinching, poking, pooing, pouting, punching, restraining, scolding, shooting, slapping, spanking, spearing, spying, stinging, striking, stripping, subduing, urinating, weeping, whipping*

After these filtering criteria, we are left with 15,207 samples. These samples include 359 out of the 504 unique verbs found in imSitu. The number of images supporting each verb gives an estimate of the likelihood of the given verb to describe a scenario with multiple human participants and thus gives us an estimate of its affinity to human-human interactions (HHI).

The verbs with the highest support are "socializing" (270 images), "distributing" (261 images), "teaching" (252 images), "communicating" (251 images), and "interviewing" (244 images). Among the least-supported verbs, which have only a single image as support, are "slipping", "skipping", "boarding", "reading", and "erasing".

Finally, to select for verbs that represent HHI, only use samples with verbs that are supported by at least 100 images. This leaves us with the 8,021 *imSitu-HHI* dataset. This contains the following 50 verbs:

- socializing (270 images)
- distributing (261 images)
- teaching (252 images)
- communicating (251 images)
- interviewing (244 images)
- lecturing (241 images)
- training (228 images)
- providing (223 images)
- instructing (217 images)
- giving (213 images)
- pushing (201 images)
- helping (200 images)
- asking (195 images)
- coaching (192 images)
- selling (185 images)
- talking (185 images)
- educating (183 images)
- buying (170 images)
- filming (161 images)
- assembling (157 images)
- encouraging (157 images)

- serving (156 images)
- dragging (155 images)
- baptizing (153 images)
- carrying (150 images)
- flinging (149 images)
- unloading (149 images)
- crowning (145 images)
- patting (138 images)
- examining (132 images)
- nagging (131 images)
- tickling (131 images)
- admiring (129 images)
- shaking (123 images)
- pinning (122 images)
- videotaping (122 images)
- arranging (121 images)
- imitating (119 images)
- offering (116 images)
- plunging (116 images)
- pitching (115 images)
- reassuring (114 images)
- autographing (112 images)
- ignoring (109 images)
- clapping (109 images)
- dousing (107 images)
- speaking (104 images)
- operating (103 images)
- wheeling (103 images)
- loading (102 images)

B.7. Training details

For training CLIPCap [44], we use checkpoints for the MLP mapping CLIPCap variant with fine-tuned GPT2 decoder, trained on Conceptual Captions.[13]

For the Enc-Dec model, we initialize the CLIP encoder with checkpoint `vit-base-patch32` and the GPT2 decoder with checkpoint `gpt2` (base), as available in the Hugging Face transformers library.

We trained all models with batch size 16, AdamW optimizer with learning rate $1e-5$ and $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay 0.1. For pretrained CLIPCap fine-tuned on our pseudo-labels, we trained for two epochs, CLIPCap trained on entire Who's Waldo captions was trained for three epochs, and the simple Enc-Dec model was trained for 17 epochs.

For models fine-tuned on our pseudo-labels, we use sample weights during training. In particular, we multiply the loss for samples with label L by $c(L)^{-1/4}$, where $c(L)$ is the count of occurrences of label L in our training data. In order to prevent overfitting to repeated captions in training data, we also use a multiplier of $c(C)^{-1}$ applied to training samples with caption C, where $c(C)$ gives the number of times caption C occurs verbatim in the training data. (See Section B.5 for details on how we filter out samples with captions that are repeated in the test set.)

B.8. Baseline model details

As in B.7, pretrained CLIPCap baselines use the MLP mapping variant with fine-tuned GPT2 decoder; in this case, using both the COCO and Conceptual Captions checkpoints. For ExpansionNetV2 [27], we initialize with the weights of the ensemble model pretrained on COCO (`rf_model.pth`)[14]. For CoFormer, we use the publically-available pretrained checkpoint for inference[15].

B.9. Metric calculation details

All reported BLEURT metrics use the BLEURT-20 checkpoint which more accurately predicts semantic similarity than the original BLEURT model [52]. For all BLEURT calculations involving texts containing `[NAME]` slots in either the predicted or ground truth text, we replace `[NAME]` with the text "person" so that the texts are in distribution for BLEURT.

NLI metrics (p_e, p_c) use BART-large [37] fine-tuned on the MNLI dataset [74] (using the `facebook/bart-large-mnli` checkpoint from Hugging Face model hub).

[13] Available at https://github.com/rmokady/CLIP_prefix_caption.

[14] Available at https://github.com/jchenghu/expansionnet_v2.

[15] Available at https://github.com/jhcho99/CoFormer.

Verb similarity scores use GloVe [49] word embeddings, specifically the `glove-wiki-gigaword-200` model available via Gensim. For models trained on our pseudo-labels, the model typically outputs the verb as the first word token, so we could use it for this metric directly. For captioning models not trained on our pseudo-labels, we virst extract a verb from their outputs for this metric using spaCy's `en_core_web_trf` model. We find the first verb lemma in the given text and convert it to present continuous form ("-ing"). For texts not containing a verb, we use the zero vector as their verb embedding.

B.10. CoFormer evaluation details

Since the CoFormer baseline model does not output free text, we elaborate here on the evaluation method used to compare it to the other methods under consideration.

For all tasks, we evaluate CoFormer by using its predicted verb, discarding semantic frame and grounding predictions. This is because these semantic arguments do not directly map to the text of a human-human interaction string, so we cannot directly compare them using text-based metrics.

The results for CoFormer on *Waldo and Wenda* reported in the main paper are calculated by inserting its predicted verbs into a text prompt and treat this as the predicted interaction. We use two different prompt templates for evaluation:

- P_1: "_ Ving _", where V denotes the given verb. This is most appropriate for transitive verbs ("_ greeting _").
- P_2: "_ Ving with _", where V denotes the given verb. This is most appropriate for intransitive verbs ("_ dancing with _").

Because P_1 or P_2 may be more appropriate depending on the verb, the reported metrics are aggregated by using the best (maximum or minimum, depending on the metric) score among both prompt templates for each sample.

We also note that we are discarding predicted semantic frame arguments from CoFormer's predictions that could be important to understanding the depicted interaction. However, they do not map directly to a single interaction string. Our approach has the advantage of directly inserting additional context into the predicted string using valid English syntax.

C. Image Attribution

- COCO val2014, ID 503278 / CC BY-NC-ND 2.0
- COCO val2014, ID 369122 / CC BY-NC-ND 2.0
- Photo by Jennifer A. Villalovos / Public domain
- Leandre Gramss double double bass 14 by Schorle / CC BY-SA 4.0
- 2017 Ski Tour Canada Quebec city 17 by Cephas / CC BY-SA 4.0
- UWS Giants vs. Eastlake NEAFL round 17, 2015 159 by Amy Mergard / CC BY 2.0
- Gansler swearing in by Doug Gansler / CC BY 2.0
- 20091112 Freddie Barnes huddling by PhotoBen27 / CC BY 2.0
- Enrique and Maja in Toronto 2014 02 by 001Jrm / CC BY-SA 3.0
- USMC-051115-M-9876R-032 by Slick-o-bot / Public domain
- Photo by Glenn Fawcett / Public domain
- Photo by Damon J. Moritz / Public domain
- Photo by Karolina A. Martinez / Public domain
- AJ Challenges Paige by Miguel Discart / CC BY-SA 2.0