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 [29, 73, 79] and relationships between entities and objects [27, 83], 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.

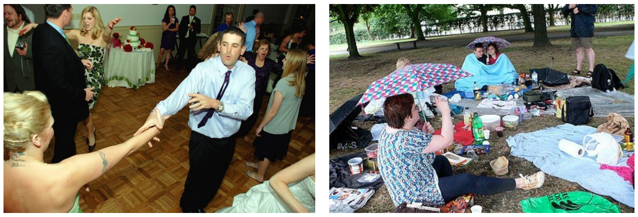


Figure 1. How would you describe the interactions depicted in these images? There are unlimited possible interactions between people which cannot be easily described by a fixed set of categories or actions. Context plays a crucial role, as in the left image where the clothing and cake in the background 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 [63, 65, 43, 30], 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 [60], 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 [60] 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 [13] 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 [75], 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 [62, 64, 6, 82, 68]. Other approaches use other modalities such as depth or skeleton data [76, 48, 73]. 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 [29, 79] for detailed surveys). Works on human-object interactions (HOI) may use separate modules such as human and object detectors and relation modules [8, 20, 16], pose and gaze estimation [35, 66, 71], or graph neural networks applied to scene graphs [49, 72, 81, 36]. One line of recent work on HOI uses end-to-end models, frequently with transformer architectures [61, 83, 10, 27, 9]. 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 [60]. As with general action recognition, these approaches most commonly use video data as input [43, 19, 67, 58, 34]. 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 [74, 7, 1]. Xiong *et al.* [70] 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 [54, 77, 19, 22] or YouTube-based [59, 42, 80].

There are few image datasets dedicated to human actions, of which HHI are a subset. Ronchi and Perona [53] introduce the Visual Verbnet dataset consisting of images with dense verb annotations. Yatskar *et al.* [75] 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 [31] 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 [3]. 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 [5, 14]. 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.* [69]. 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 [28, 21].

The use of LLM-generated synthetic data for multimodal learning has been explored by Brooks *et al.* [4], 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.

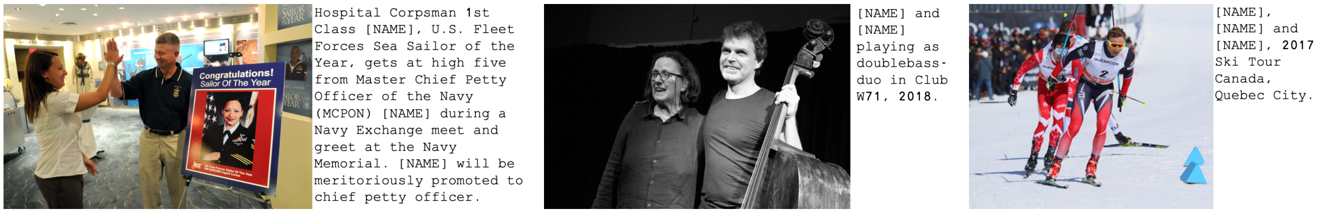
Dataset	#Seq	#HHI Classes
Curated videos		
UT-Interaction [54]	60	6
TV Human Interaction [45]	300	4
Hollywood2 [39]	3669	4
ShakeFive2 [19]	153	5
SBU Kinect [77]	300	8
AVA [22]	∼57.6k	13
NTU RGB+D (120) [56, 37]	∼114k	26
YouTube-based videos		
Kinetics [26, 59]	∼500k	11
Moments in Time [42]	∼800k	32
HACS [80]	∼50k	23
Still images		
imSitu [75]	126k	50*
Visual Verbnet [53]	10k	52*
Who's Waldo [13] (*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 [60] where relevant). In our work, we devise a technique for generating HHI pseudo-labels for Who's Waldo [13], 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 [13], 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 [13], 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 [23] (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 [2, 17], which produces a set of diverse and more natural-sounding interactions.

3.2. Synthetic caption data generation

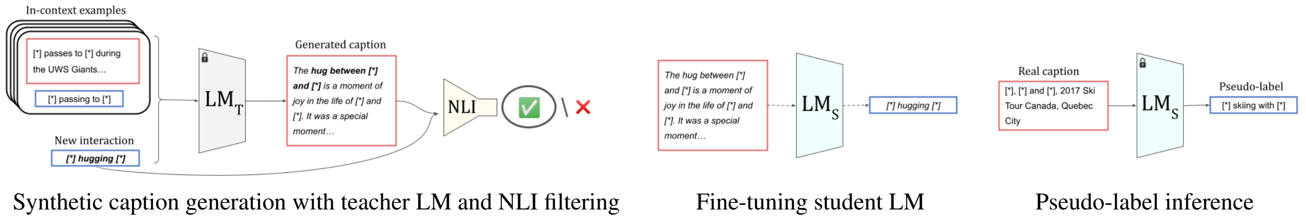
Using the caption–interaction pairs from Who's Waldo and the new synthetic interactions as seeds, we generate synthetic caption–interaction pairs using in-context learning with GPT-Neo. This allows us to create a larger and more diverse set of caption–interaction pairs than by using caption–interaction 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 [52]. 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


In-context examples
[*] passes to [*] during the UWS Giants…
[*] passing to [*]
New interaction
[*] hugging [*]
LM_T
Generated caption
The hug between [*] and [*] is a moment of joy in the life of [*] and [*]. It was a special moment…
NLI
✅ \ ✕
The hug between [*] and [*] is a moment of joy in the life of [*] and [*]. It was a special moment…
LM_S
[*] hugging [*]
Real caption
[*], [*] and [*], 2017 Ski Tour Canada, Quebec City
LM_S
Pseudo-label
[*] skiing with [*]

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.* [13]. 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 [11], (3) 400 images from Conceptual Captions [57]. 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 [13]. 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 [13]. We proceed to show how we use these to supervise learning HHI from still images via the paradigm of image captioning.

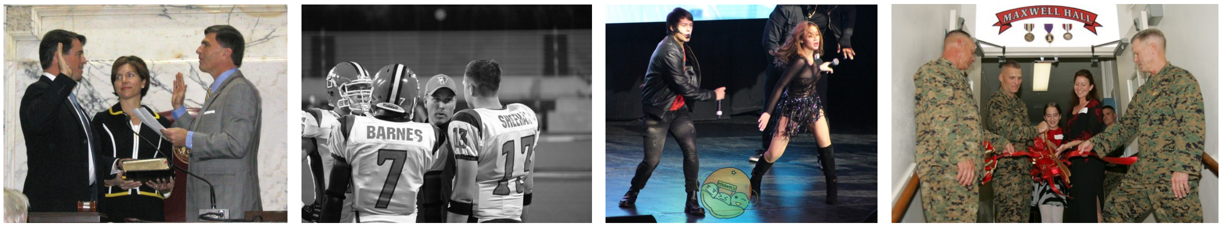
4.1. Models considered

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 [50], 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 [51] with a causal language modelling head and cross-attention over the encoder output. Consistent with previous works on fine-tuning vision-and-language models [38, 41, 78], 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 [41] 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 [75] 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 [15] 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) [25] and CLIPCap [41]. 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 [11] and Conceptual Captions (CC) [57], 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 [12] 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 [47], 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 [24, 18]. 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 [55] 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 [44] 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) [32]. 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 [40, 33]. 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.

Method	Train Data	BL ↑	p_e ↑	p_c ↓	sim ↑
Results@1					
CoFormer	SWiG	0.41*	0.33*	0.28*	0.35
ENv2	COCO	0.27	0.25	0.33	0.41
CLIPCap	COCO	0.28	0.34	0.37	0.42
CLIPCap	CC	0.27	0.18	0.38	0.35
CLIPCap	CC+WW	0.26	0.16	0.40	0.17
EncDec	pHHI	0.38	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	0.51	0.61	0.09	0.59
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	0.54	0.65	0.19	0.65
CLIPCap	CC+pHHI	**0.59**	**0.76**	**0.04**	**0.69**

*Evaluated by using the best of two prompt templates for each item, as described in the supplementary material.

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.

Verb similarity. We calculate the average cosine similarity of the predicted and ground truth verbs in GloVe [46] 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").

		WW				CC				COCO			
Method	Train Data	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*	0.35*	0.34	0.45*	0.40*	0.43*	0.38	0.37*	0.30*	0.33*	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**

*Evaluated by using the best of two prompt templates for each item, as described in the supplementary material.

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.

To evaluate this on free text predictions, we either select the first non-`[NAME]` word in the output (for models trained 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

		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.

on pHHI outputs a list of diverse candidate interactions, allowing a more directed search in the space of HHI descriptors, 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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.