Learning Human-Human Interactions in Images from Weak Textual Supervision —Supplementary Material—

Morris Alper and Hadar Averbuch-Elor
Tel Aviv University

Contents

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. [3, 14, 19] 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 1 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 1. 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.

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

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 2 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*

		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

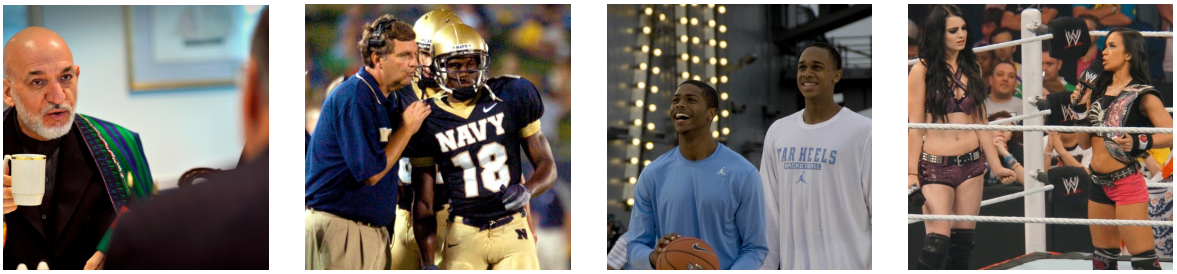
Table 3. Comparison of results when training on syntactic parsing-based seeds ("SP") versus our pseudo-labels ("pHHI"), as described in Section A.3. "BL" refers to BLEURT and "sim" refers to verb embedding similarity. On *Waldo and Wenda*, results are aggregated across data sources.

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 participants.

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 3. 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 1, 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.



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

Figure 1. 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.

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 [21] and BARTScore [20]. 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 [9]. Table 4 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[1], and we report the output F1 score. BARTScore uses the model trained on ParaBank2 provied in the official BARTScore repository[2]. SummaC scores use the default checkpoint and settings for SummaC-Conv provided in its official repository[3]. 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 [17] as our main textual metric rather than ngram-based metrics such as BLEU [12]. 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.

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 4. 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).

We provide a comparison of BLEU and BLEURT scores in Table 5, 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 [17] show to have a robust correlation with human judgements of semantic similarity at the sentence level.

We additionally provide scores for the METEOR metric [1], which uses unigram alignment statistics and incorporates both precision and recall. It also uses stemming and synonym matching to provide some robustness rela-

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

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

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

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

tive 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 5 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 [17] who explicitly compare METEOR and BLEURT. This can also be seen in the qualitative examples in Table 6 of prediction (CLIPCap CC+pHHI) and ground truth pairs from *Waldo and Wenda* where METEOR and BLEURT differ strongly in magnitude.

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 7. 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.

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[4], containing the text of $\sim 708k$ scraped English language news articles from 2017 through 2019 [6]. 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

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

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 6. 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 7. 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.

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[5].

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:

- – **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.

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

  - **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 [2, 5] and generating until the next newline. We use nucleus sampling [7] 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[6] 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 [2, 5] and generate text until a newline is output. We generate using nucleus sampling [7] with $p = 0.95$, temperature 0.7, a constraint to prevent repeated trigrams, and a maximum output length of 200 tokens.

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

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 [10] fine-tuned on the MNLI dataset [18] (using the `facebook/bart-large-mnli` checkpoint from Hugging Face model hub[7]).
- 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 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 [16] 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.[8]

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.

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

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

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

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 [19], 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[9], pretrained on MS COCO)[4], 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.

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

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.

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 [11], we use checkpoints for the MLP mapping CLIPCap variant with fine-tuned GPT2 decoder, trained on Conceptual Captions.[10]

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 [8], we initialize with the weights of the ensemble model pretrained on COCO (`rf_model.pth`)[11]. For CoFormer, we use the publicly available pretrained checkpoint for inference[12].

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 [15]. 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 [10] fine-tuned on the MNLI dataset [18] (using the `facebook/bart-large-mnli` checkpoint from Hugging Face model hub). To calculate these metrics, `[NAME]` slots in texts are filled with an underscore character ("_").

Verb similarity scores use GloVe [13] 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

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

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

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

- 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

References

[1] 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.

[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] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

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

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