Who are you referring to?
Weakly supervised coreference resolution with multimodal grounding

Arushi Goel\textsuperscript{1}, Basura Fernando\textsuperscript{2}, Frank Keller\textsuperscript{1}, and Hakan Bilen\textsuperscript{1}
\textsuperscript{1}School of Informatics, University of Edinburgh, UK
\textsuperscript{2}CFAR, IHPC, A*STAR, Singapore.

Abstract

Coreference resolution aims at identifying words and phrases which refer to same entity in a text, a core tool in natural language processing. In this paper, we propose a novel task, resolving coreferences in multimodal data, long-form textual descriptions of visual scenes. Most existing image-text datasets only contain short sentences without coreferent expressions, or coreferences are not annotated. To this end, we first introduce a new dataset, Flickr30k-Coref in which coreference chains and bounding box localization of these chains are annotated. We propose a new technique that learns to identify coreference chains through weakly supervised grounding from image-text pairs and a regularization using prior linguistic knowledge. Our model yields large performance gains over prior work in coreference resolution and weakly supervised grounding of long-form text descriptions.

1. Introduction

Coreference resolution (CR) is the task of identifying all referring expressions (mentions) in a document and determining which of them refer to the same entity. It is a core element in many natural language processing (NLP) tasks, including question answering, document summarization, and machine translation. CR is a challenging problem due to the various forms that referring expressions can take in unconstrained documents. The most common form of coreference is pronominal anaphora, where a pronoun (e.g., she) refers back to noun phrases earlier in the text (e.g., the back side lady in Figure 1). Authors in [46] provide a recent and detailed review of CR in text data.

Despite the large body of existing work, CR research has almost exclusively focused on text data. However, reference to objects in a visual scene through pronouns or noun phrases commonly occurs in day-to-day speech, and often visual information is crucial for determining which pronouns, proper names, and noun phrases corefer. In this paper, we propose a new task: multimodal coreference resolution, i.e., resolving the coreference of referring expressions in text that is paired with an image. This form of CR is crucial for applications in multimodal reasoning, human-robot interaction, and virtual assistance. Aside from pronominal anaphora, phrase mentions can also corefer with other phrase mentions (e.g., a lady, this person, and Mary can all refer to the same entity). Here, we are particularly interested in identifying coreference chains with multiple mentions ((a lady, her), (the back side lady, she, her)). In a visual context, we need to ground coreference chains, i.e., link them to the correct visual referent (an object bounding box), as indicated by color-coding in Fig. 1. This even applies to singleton coreference chains (underlined in Figure 1 e.g., black hair cut sheet).

Prior CR methods [3, 15, 23, 24] without image understanding are trained on annotated text and structured training data. For example the fact that the first lady can refer to Michelle Obama can be inferred from knowledge bases [41, 48, 55]. Also, the reference of anaphora follows clear linguistic rules [22], e.g., her can refers to any female entity that appeared earlier in the text. However, state-of-the-art textual coreference resolvers [15, 23] fail to disambiguate references correctly in our multimodal setting (as shown in Figure 1). Standard image/text datasets such as COCO or
Flickr30k [5, 20, 27, 37] only contain short descriptions (typically single sentences), which include very few or no cases of coreference. For instance, the Flickr30k caption for the image in Figure 1 is *A woman holding scissors and giving a haircut*. These datasets therefore neither allow learning nor evaluating CR. To address this gap, we introduce the Flickr30k-Coref dataset, which is based on the Localized Narratives dataset [38], where we augment the existing long-form descriptions with coreference chain annotation. Furthermore, our new dataset also includes bounding box annotations, making it possible to evaluate the grounding of coreference chains.

Manual annotation of large datasets is expensive. Hence we focus on learning CR from only paired image-text data under weak supervision, and use coreference chain and bounding box annotations in Flickr30k-Coref only for evaluation. Concretely, we learn grounding mentions from paired image-text descriptions by finding the most correlated mention and image region pairs and contrasting them to the rest. We leverage the inferred grounding to resolve coreferring mentions by matching similar mention-region pairs. Our method also exploits linguistic knowledge through multiple priors to favor more linguistically plausible solutions. We report extensive experiments on the Flickr30k-Coref and demonstrate that our method not only brings significant improvements in CR but also large gains in weakly supervised visual grounding.

At a high-level, our work can be seen as a step towards expanding image-text understanding from single to multiple sentences. Our key contributions are:

- a new task that aims at resolving coreferences in multimodal long form textual descriptions,
- Flickr30k-Coref dataset enabling the evaluation of coreference chains in text and the localization of bounding boxes in images, which is provided with multiple baselines and exhaustive analysis in multimodal CR for fostering future work,
- a new method that learns to resolve coreferences while jointly grounding them from weak supervision and exploiting prior linguistic knowledge,
- rigorous experimental evaluation showing significant improvement over the prior work not only in CR but also in weakly supervised grounding of complex phrases in long-form text, a form of disambiguation that has been underexplored in visual grounding.

2. Related Work

CR in NLP has a long history of rule-based and machine learning-based methods. Early methods [14, 40] used hand-engineered rules to parse dependency trees, which outperformed all the learning-based methods. But recently neural network methods [9, 15, 23, 50, 51] have achieved significant performance gains. The key idea in their work is to identify all mentions in a document either by parsing them using parsers or by learning them end-to-end with the coreference resolution system and then learn a distribution over all the possible antecedents for each mention. It is worth noting that their training requires gold standard (ground-truth) coreference chains obtained from large datasets such as OntoNotes [39] and PreCo [6]. More recently, large transformer based models such as SpanBERT [15] are used by incorporating a span-based masked prediction objective for pre-training and show improvements on the downstream task of coreference resolution.

There are few prior work that focuses on CR for solving specific visual tasks. For instance, [42] and [45] link character mentions in TV shows or movie descriptions to their occurrence in the video. Another line of work [18] exploits CR to relate texts to 3D scenes. An orthogonal direction of research is to resolve coreference in visual dialog [19] for developing better question-answering systems but limited to specific scenarios such as linking only pronouns. Unlike them, we focus on learning coreferences from long-form text using weak supervision.

A related line of work is [11, 26, 52, 54] that aims to ground nouns and more complex phrases in image parts. In visual phrase grounding [17, 53], the main objective is to localize a single image region given a textual query. State of the art approaches either use a two-stage [7, 28, 52] method by first encoding object regions and then maximizing the similarity with the textual query or a single stage method [11, 16, 25] where the entire image is encoded and then a localization is learned using a standard regression function to find the bounding box coordinates. Single stage approaches are very similar to object detection [13, 43] with a conditioned input. The models are trained on visual grounding datasets such as ReferItGame [17], Flickr30K Entities [37], or RefCOCO [53]. Though these methods can localize a single phrase in the image (achieving up to 80% accuracy [11]), they fail to disambiguate coreferences (*a lady and the back side lady* that refer to different regions/entities in Figure 1), which is the key contribution of our work.

Weak supervision has recently been used in [29–32] for referring expression grounding with the main focus on reconstruction-based losses, a loss for language query reconstruction given the image proposal information or a consistency network where both the query and the image proposal features are reconstructed. Other methods [10, 12, 49] use contrastive learning by creating many negative queries (based on word replacement) or by mining negative image regions for a given query. Our method also employs contrastive learning but for learning CR from weak supervision.

---

1Our code and dataset will be made publicly available.
3. Flickr30k-Coref

The Flickr30k-Coref dataset contains 1880 images from the Localized Narratives dataset [38] that come with long-form text descriptions and mouse traces. These images are a subset of the test and validation set of the Flickr30k dataset [37]. We added coreference chains to the narratives and bounding box localizations to the images, which we grounded in the coreference chains. This includes singletons (i.e., coreference chains of length one). Fig. 1 shows an example image from the Flickr30k-Coref dataset.

**Annotation procedure.** The annotation involved three steps: (1) marking the mentions (sequences of words) that refer to a localized region in the image, (2) identifying coreference chains for the marked mentions including (a) pronominal words such as him, who that are used to refer to other mentions, and (b) phrases that refer to the same entity (e.g., a lady and that person), (3) drawing bounding boxes in the image for the coreference chains identified in steps (1) and (2). We created an annotation interface based on LabelStudio [1], an HTML-based tool that allows us to combine text, image, and bounding box annotation. More details are provided in the supplementary material.

| Dataset          | Noun phrases | Pronouns | Coreference chains | Bounding boxes |
|------------------|--------------|----------|--------------------|----------------|
| Flickr30k Entities [37] | 15252        | -        | -                  | 17234          |
| RefCOCO [53]      | 10668        | -        | -                  | 10668          |
| Flickr30k-Coref (Ours) | 19587       | 1659     | 3310               | 21246          |

Table 1. Statistics of relevant noun phrases, pronouns, coreference chains and bounding boxes on Flickr30k Entities [37], RefCOCO [53] and Flickr30k-Coref.

**Dataset statistics.** We split the 1880 images in the Flickr30k-Coref dataset into a test and validation set using the split of [37]. Specifically, we have 1000 images in the test set and 880 images in the validation set. It is important to note that the Localized Narratives dataset also has noisy text descriptions due to a lot of abstract references such as I can see . . . that are not a part of the coreference chain and cannot be grounded on the image. We elaborate on the filtering process for these phrases in the supplementary material. Overall, the Flickr30k-Coref dataset has 19,587 noun phrase mentions, 1659 pronominal pronouns, 3310 coreference chains and 21,246 bounding boxes. In Table 1, we compare the statistics of Flickr30k-Coref with other related datasets.

4. Method

4.1. CR for text data

Given a sentence containing a set of mentions (i.e., referential words or phrases), the task of CR is to identify which mentions refer to the same entity. This is fundamentally a clustering problem. In this work, we use the off-theshelf NLP parser [2] to obtain the mentions. Formally, let $S = \{m_1, m_2, \ldots, m_{|S|}\}$ denote a sentence with $|S|$ mentions, where each mention $m$ contains a sequence of words, $\{w_1, w_2, \ldots, w_{|m|}\}$. We assign a label $y_{ij}$ to each mention pair $(m_i, m_j)$, which is set to 1 when the pair refers to the same entity, and to −1 otherwise. We want to learn a compatibility function, a deep network $f$ that scores high if a pair refers to the same entity, and low otherwise.

Given a training set $D$ that contains $|D|$ sentences with their corresponding labels, one can learn $f$ by optimizing a binary cross-entropy loss:

$$
\min_f \sum_{S \in D} \sum_{i=0}^{|S|-1} \sum_{j=i+1}^{|S|} \log(y_{ij}(\sigma(f(m_i, m_j))) - \frac{1}{2}) + \frac{1}{2})
$$

(1)

where $\sigma$ is the sigmoid function. Note that prior CR methods [15, 22, 23] require large labeled datasets for training and are limited to only a single modality, text. These methods typically also combine the learning with fixed rules based on recency and grammatical principles [22].

4.2. CR for image and text data

**Problem definition.** Next, we extend the standard CR learning problem to image-text data in the absence of coreference labels. Let $(I, S)$ denote an image-text pair where $S$ describes an image $I$ as illustrated in Figure 1, and assume that coreference labels for mention pairs are not present. As in Sec. 4.1, our goal is to identify the mentions that refer to the same entity in an image-text pair. Each image is defined by $|I|$ regions $I = \{r_1, r_2, \ldots, r_{|I|}\}$ which are obtained by running a pretrained object detector [44] on the image. Each region $r$ is described by its bounding box coordinates $b$, the text embedding for the detected object category $o$, and the visual features $v$. We provide further details in Section 4.3.

**Weak supervision.** In this paper, we use “weak supervision” to refer to a setting where we have no coreference label for mention pairs and no grounding of mentions (i.e., bounding boxes are not linked to phrases in the text). Moreover, in contrast to the output space of the object detector (a restricted set of object categories), the sentences describing our images come from unconstrained vocabulary. Hence, an object instance in a sentence can be referred to with a synonym or may not even be present in the object detector vocabulary [21, 27]. Finally, the object detector can only output category-level labels and hence cannot localize object instances based on the more specific instance-level descriptions provided by the sentences. For instance in Figure 1, back side lady and a lady both are labeled as women by the object detector.

Furthermore, along with image and text data, we explore the use of an auxiliary modality, the mouse trace segments provided in the Localized Narratives dataset [38].
The mouse traces are a sequence of 2D points over time that relate to a region in the image when describing the scene. As the text in Localized Narratives is transcribed the speech of the annotators, the mouse traces are synced with spoken words, which we denote as $T = \{t_1, t_2, \ldots, t_{|T|}\}$ where $|T| = |S|$. These features are stacked with textual features. Further details are given in Section 4.3.

**Weakly supervised grounding.** In the weakly supervised setting, the key challenge is to replace the coreference label supervision with an alternative one. We hypothesize that each mention in a corefering pair corresponds to the same image region, and it is possible to learn a joint image-text space which is sufficiently rich to capture such correlations. Concretely, let $g(m, r)$ denote an auxiliary function that is instantiated as a deep network and outputs a score for the mention $m$ being located at region $r$ in image $I$. The grounding score for each mention can be converted into probability values by normalizing them over all regions:

$$\tilde{g}(m, r) = \frac{\exp(g(m, r))}{\sum_{r' \in I} \exp(g(m, r'))}. \quad (2)$$

The compatibility function $f$ can be defined as a sum product of a pair’s grounding probabilities over all regions:

$$f(m, m') = \sum_{r \in I} \tilde{g}(m, r) \tilde{g}(m', r). \quad (3)$$

In words, mention pairs with similar region correlations yield bigger compatibility scores and are hence more likely to corefer to each other. For inference, we compute $f(m, m')$ for all the pairs and threshold them to predict their pairwise coreference labels. We also use fixed rules from NLP literature to eliminate erroneous labels, which are explained below. The key intuition is that we leverage the grounding information of mentions as anchors to relate coreferring mentions (e.g., *back side lady* and *she*).

As no ground-truth bounding box for each mention is available for learning the grounding $g$, we pose grounding as a weakly supervised localization task as in [12, 49]. To this end, we impute the missing bounding boxes by taking the highest scoring region for a given mention $m$ at each training iteration:

$$r_m = \arg \max_{r \in I} g(m, r). \quad (4)$$

Then we use $r_m$ as the pseudo-ground truth to learn $g$ as following:

$$\min_g \sum_{(I, S) \in D} \sum_{m \in S} -\log \left( \frac{\exp(g(m, r_m))}{\sum_{r' \in I} \exp(g(m, r'_m))} \right) \quad (5)$$

where $r'_m = \arg \max_{r \in I} g(m, r)$ is the highest scoring region in image $I'$ for mention $m$. For each mention, we treat the highest scoring region in the original image as positive and other highest scoring regions across different images as negatives, and optimize $g$ for discriminating between the two. However, as the denominator requires computing $g$ over all training samples at each iteration, which is not computationally feasible, we instead sample the negatives only from the randomly sampled minibatch.

**Linguistic constraints.** Learning the associations between textual and visual features helps with disambiguating coreferring mentions, especially when mentions contain visually salient and discriminative features. However, resolving coreferences when it comes to pronouns (e.g., *her, their*) or ambiguous phrases (e.g., *one man or another man*)
remains challenging. To address such cases, we propose to incorporate a regularizer into the compatibility function \( f(m, m') \) based on various linguistic priors. Concretely, we construct a look-up table for each mention pair \( q(m, m') \) based on the following set of rules [22]:

(a) **Exact String Match.** Two mentions corefer if they exactly match and are noun phrases (not pronouns).

(b) **Pronoun Resolution.** Based on the part-of-speech tags for the mentions, we set \( q(m, m') \) to 1 if \( m \) is a pronoun and \( m' \) is the antecedent noun that occurs before the pronoun.

(c) **Distance between mentions.** Smaller distance is more likely to indicate coreference since mentions to occur close together if they refer to the same entity.

(d) **Last word match.** In certain cases, the entire phrases might not match but only the last word of the phrases.

(e) **Overlap between mentions.** If two mentions have one or more overlapping words, then they are likely to corefer.

Finally, we include \( q(m, m') \) as a regularizer in Eq. (5):

\[
\min_g \sum_{(I,S) \in D} \sum_{m \in S} \left( -\log \left( \frac{\exp(g(m, r_m))}{\sum_{r \in I} \exp(g(m, r_m))} \right) \right) + \lambda \sum_{m \in S} \|f(m, m') - q(m, m')\|_F^2
\]

where \( \lambda \) is a scalar weight for the Frobenius norm term. Note that \( f \) is a function of \( g \) (see Eq. (3)). We show in Section 6 that incorporating this term results in steady and significant improvements in CR performance.

### 4.3. Network modules

Our model (shown in Figure 2) consists of an image encoder \( e_i \) and text encoder \( e_t \) to extract linguistic and visual information, and a cross-attention module \( a \) to integrate them.

**Image encoder** \( e_i \) takes in a \( d_r \)-dimensional vector for each region \( r \) that consists of a vector consisting of bounding box coordinates \( b \in \mathbb{R}^4 \), text embedding for the detected object category \( o \in \mathbb{R}^{d_o} \) and textual features \( v \in \mathbb{R}^{d_v} \). The regions are extracted from a pretrained object detector [44] for the given image \( I \). The image encoder applies a nonlinear transformation to this vector to obtain a \( d \)-dimensional embedding for each region \( r \).

**Text encoder** \( e_t \) takes in the multiple mentions from a parsed multi-sentence image description \( S \) produced by an NLP parser [2] and outputs a \( d \)-dimensional embedding for each word in the parsed mentions. Note that the parser does not only extract nouns but also pronouns as mentions.

**Mouse trace encoder** \( e_m \) takes in the mouse traces for each mention parsed above after it is preprocessed into a 5D vector of coordinates and area, \((x_{\text{min}}, x_{\text{max}}, y_{\text{min}}, y_{\text{max}}, \text{area})\) [35] and outputs a \( d_m \)-dimensional embedding. In [4, 38], mouse trace embeddings have been exploited for image retrieval and captioning, however, we use them to resolve coreferences. In this case, we augment each mention embedding extracted from \( e_t \) with the mouse trace encoding and apply additional nonlinear transformations (Text-Trace Encoder in Fig. 2) before they are fed into the cross-attention module.

**Cross-attention module** \( a \) takes in the text embeddings for all the words in each mention \( \{e_t(w) \ \forall w \in m\} \) and returns a \( d \)-dimensional vector for each \( m \) by taking a weighted average of them based on their correlations with the image regions. Concretely, in this module, we first compute the correlation between each word \( w \) (or joint word-mouse trace) and all regions, take the highest correlation over the regions through an auxiliary function \( \bar{a} \):

\[
\bar{a}(w) = \max_{r \in I} \left( \frac{\exp(e_t(w) \cdot e_i(r))}{\sum_{r' \in I} \exp(e_t(w) \cdot e_i(r'))} \right)
\]

where \( \cdot \) is dot product. The transformation can be interpreted as probability of word \( w \) being present in image \( I \). Then we compute a weighted average of the word embeddings for each mention \( m \):

\[
a(m) = \sum_{w \in m} \bar{a}(w)e_t(w).
\]

Similarly, \( a(m) \) can be seen as probability of mention \( m \) being present in image \( I \).

**Scoring function** \( g(m, r) \) can be written as a dot product between the output of the attention module and region embeddings:

\[
g(m, r) = a(m) \cdot e_i(r). \tag{9}
\]

While taking a dot product between the two embeddings seemingly ignores the correlation between text and image data, the region embedding \( e_i(r) \) encodes the semantic information about the detected object category in addition to other visual features and hence results in a high score only when the mention and region are semantically close. Further implementation details about the modules can be found in Section 5.2 and the supplementary.

### 5. Experiments

**Datasets.** We train our models on the Flickr30k subset with Localized Narratives annotations [38] and evaluate them on the proposed Flickr30k-Coref dataset. For training, we have 30k images with the long form text description for each image. The evaluation dataset, Flickr30k-Coref has 1000 images for test and 880 images for validation.

**5.1. Evaluation metrics.**

**Coreference evaluation.** There are a number of metrics that are proposed in the NLP literature to measure the task
of coreference resolution [23, 34]. In this paper, we use the link-based metrics MUC [47] and BLANC [43] to evaluate the coreference links predicted by our model. We denote the predicted chains from the model as \( R \) and the ground truth chains from human labeling as \( K \). Below, we explain how the two metrics compute precision and recall:

**MUC F-measure.** The MUC F-measure is based on the number of coreference links (pairs of mentions) common to the output and ground-truth chains. The main step is to compute the partitions with respect to the two chains:

\[
\text{MUC-R} = \frac{\sum_{i=1}^{N_r} (|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_r} (|K_i| - 1)}
\]

\[
\text{MUC-P} = \frac{\sum_{i=1}^{N_r} (|R_i| - |p^*(R_i)|)}{\sum_{i=1}^{N_r} (|R_i| - 1)}
\]

where \( K_i \) is the \( i \)th ground-truth chain and \( p(K_i) \) is the set of partitions created by intersecting \( K_i \) with the output chains; \( R_i \) is the \( i \)th output chain and \( p^*(R_i) \) is the set of partitions created by intersecting \( R_i \) with the ground-truth chains; and \( N_k \) and \( N_r \) are the total number of ground-truth and output chains, respectively.

**BLANC metric.** Let \( C_k \) and \( C_r \) be the pairs of coreference links respectively, and \( N_k \) and \( N_r \) be the set of non-coreference links in the ground-truth and output respectively. The BLANC Precision and Recall for coreference links is calculated as follows:

\[
\text{BLANC-P} = \frac{\sum_{i=1}^{N_r} (|R_i| - |p^*(R_i)|)}{\sum_{i=1}^{N_r} (|R_i| - 1)}
\]

\[
\text{BLANC-R} = \frac{\sum_{i=1}^{N_r} (|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_r} (|K_i| - 1)}
\]

where \( K_i \) is the \( i \)th ground-truth chain and \( p(K_i) \) is the set of partitions created by intersecting \( K_i \) with the output chains; \( R_i \) is the \( i \)th output chain and \( p^*(R_i) \) is the set of partitions created by intersecting \( R_i \) with the ground-truth chains; and \( N_k \) and \( N_r \) are the total number of ground-truth and output chains, respectively.

**Grounding evaluation.** For evaluating the grounding of phrases in images, we consider a prediction to be correct if the IoU (Intersection over Union) score between the predicted bounding box and the ground truth box is larger than 0.5 [12, 49]. Following [16], if there are phrases with multiple ground truth boxes (e.g., several people), we use the any-box protocol i.e., if any ground truth bounding box overlaps the predicted bounding box, it is a correct prediction. We report percentage accuracy for phrase grounding evaluation.

5.2. Implementation details.

**Inputs and modules.** For the image branch of our model, we extract bounding box regions, visual features and object class labels using the Faster-RCNN object detector [44]. We use Glove embeddings [36] to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow [38] and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. There are two variations of our model, one with a two-layer MLP and another with a transformer encoder. For the two-layer MLP, all the modules i.e., image encoder, text encoder, trace encoder and joint text-trace encoder are feed-forward MLP blocks with a non-linear ReLU layer in between. For the transformer one, we replace the MLP image encoder with a stack of two transformer encoder layers. Each transformer encoder layer includes a multi-head self attention layer and an FFN. There are two heads in the multi-head attention layer, and two FC layers followed by ReLU activation layers in the FFN. The output channel dimensions of these two FC layers are 2048 and 1024, respectively. Similarly, we replace the joint text-trace encoder with the same settings for the transformer. The input to the joint text-trace encoder comes from the separate text and trace encoder branches. We add a special embedding to the learned embeddings following [8] to distinguish between the two modalities (text and trace) in the transformer encoder.

**Training details.** The whole architecture is trained end-to-end with the AdamW [33] optimizer. For training the MLP, the learning rate is set to 1e-4, batch size is set to 64, weight decay is 1e-4 and the loss coefficient \( \lambda \) is set to 0.01. We train the transformer encoders with the learning rate of 3e-5, batch size of eight, weight decay of 0.01 and the loss coefficient \( \lambda \) of 0.001. In both the settings we train the model for 60 epochs and choose the best performing model based on the validation set.

6. Results

6.1. Quantitative results

**Coreference resolution.** In Table 8, we compare the performance of our method to text-only baselines and a weakly supervised multimodal baseline (MAF) [49] in CR. For the text-only baselines, we consider a rule-based method [22] and an end-to-end trained neural CR method [23] trained on large-scale supervised language coreference dataset. We train the MAF baseline on the Flickr30k Localized Narratives training data. For our method, we report results with the transformer encoder (Tr) trained with the language prior regularization (Reg) in Section 4. Note that, unless described otherwise, our method corresponds to the variant that use the transformer variant, trained with image, text and mouse traces with the regularizer.

The text-only baselines fail to resolve conferences from long-form textual descriptions. Specifically, these models are trained to follow strict rules such as the pronominal pronouns mentioned later in the sentence are likely to be linked to the pronoun that is mentioned before (as also shown in Fig. 1). This kind of prior is harmful in our setting. Compared to the multimodal baseline (MAF), our transformer encoder based method trained with the regularizer, with or without mouse traces, surpasses the performance of the baselines in all metrics. The consistent gains in perfor-
mance shows the effectiveness of our proposed method in resolving coreferences. It is important to note that the extremely low BLANC scores obtained by the text baselines is due to the fact that they only produce coreference chains (multi-mention entities) and therefore achieve lower scores, whereas our method also produces singletons (chains of length one). In contrast, the MUC scores only measure coreference linking between pairs, so they are unaffected by singleton mentions.

**Grounding.** We also evaluate our models on the task of visual grounding as shown in Table 3. MAF [49] is a weakly supervised model for phrase grounding and they evaluate their method on the Flickr30k Entities [37] dataset. Compared to the performance on Flickr30k Entities dataset (reported as 61%) where the textual descriptions are significantly shorter, the performance of even the strongest baseline method is significantly lower on our Flickr30k-Coref dataset. This clearly shows that grounding is much more difficult when we are dealing with long-form descriptions rather than single sentences.

| Method     | Arch | Reg | Noun Phrases | Pronouns | Overall |
|------------|------|-----|--------------|----------|---------|
| MAF [49]   | MLP  | X   | 25.58        | 22.36    | 24.91   |
|            | Tr   |     | 27.62        | 23.46    | 26.75   |
|            | MLP  | ✓   | 27.73        | 24.66    | 27.09   |
|            | Tr   | ✓   | 30.27        | 25.96    | 29.36   |

Table 3. Grounding accuracy (%) for noun phrases and pronouns and the overall accuracy on the Flickr30k-Coref dataset.

Our method significantly outperforms the baseline on grounding accuracy for both the noun phrase mentions and pronouns. Hence, we are not only able to resolve coreferences but also ground the mentions to the image with a higher level of contextual reasoning.

### 6.2. Ablation Study

**Incorporating mouse traces.** In Table 9 we show the importance of incorporating mouse traces in addition to the knowledge from the textual encoder. Adding the mouse traces improves performance significantly on both coreference resolution and grounding. They provide a discriminative location prior to the textual mentions, which helps to learn a better compatibility score. As an example consider Figure 2, where the same textual mention *this person* points to two different visual regions. Here, mouse traces can provide a strong signal for disambiguation. With the self-attention in the joint text-trace encoder transformer, we can leverage the power of mouse traces even more compared to MLP.

Choice of constraint loss function. In Table 5, we compare the performance of our model when trained with different variants of the language constraint. In particular, we experiment with the L1 loss and MSE loss and compare it to our proposed Frobenius norm minimization. We achieve large performance gains with the Frobenius norm, as it imposes a stronger constraint on the learned coreference matrix.

| Loss Fn. | CR MUC-F1 | BLANC-F1 | Grounding Acc(%) |
|----------|-----------|----------|------------------|
| L1 loss  | 21.24     | 66.67    | 27.65            |
| MSE loss | 25.30     | 68.75    | 28.50            |
| Frobenius Norm | 29.13 | 70.43 | 29.36 |

Table 5. Performance of our method with variants of the language prior regularizer function.

Importance of cross attention module. Table 6 shows the model performance for two settings: (1) directly averaging the word features or (2) attending over the words by using the image as the query as discussed in Sec. 4. Both the grounding accuracy and the coreference evaluation get a boost in performance for visually aware word features. More often than not, the word phrases are relatively short (e.g., *the machine*) and hence the model does not always learn to disambiguate better with attention for the grounding. On the other hand, this technique is especially useful

---

| Method          | Arch | Reg | Noun Phrases | Pronouns | Overall |
|-----------------|------|-----|--------------|----------|---------|
| Rule-Based [22] | ✓    | ✓   | ✗            | ✗        | 5.6     |
| Neural-Coref [23]| ✓  | ✓  | ✗            | ✗        | 21.7    |

Table 2. Coreference resolution performance on the Flickr30k-Coref dataset. *MT denotes mouse trace*.
for coreference resolution because the flow of visual information to the word features acts as a prior to cluster mentions that refer to the same region but with different phrase entities in text (e.g., the machine and an equipment).

| Attention Type | CR MUC-F1 | BLANC-F1 | Grounding Acc(%) |
|----------------|-----------|----------|------------------|
| Average        | 25.80     | 69.28    | 28.83            |
| Cross attention| 29.13     | 70.43    | 29.36            |

Table 6. Results on our image text trace transformer with and without a cross attention module.

6.3. Qualitative Results

Figure 3 qualitatively analyzes coreference resolution and visual grounding from our image text trace transformer architecture trained with the combined loss function or only with contrastive loss. The first three columns shows the results from our final method. The model correctly resolves and localizes phrases such as a person, who, the people, them and a baby trolley, the baby trolley. For some sentences, the model correctly localizes an object which refers to the water sprinkler but fails to correctly localize them to the same location. For these cases, it is hard for language rules to provide a strong prior as they are abstract concepts that frequently occur in natural language and are difficult for the model to disambiguate. This also shows the challenging problem of ambiguous references we are dealing with, indicating the great potential for developing models with strong contextual reasoning.

In the third and fourth columns, we compare the models trained with and without the language prior loss, respectively. The model trained with the language prior correctly disambiguates the relationship between their and their bicycles, clustering them in the same region, whereas the other model fails to correctly localize their bicycles and randomly assigns a region for their (last column, first image). Moreover, in the follow-up sentence of this image description, if the model has no prior imposed, it fails to correctly resolve behind that, where that refers to the banner in red color.

7. Conclusion

In this paper, we introduced a new task along with a dataset and customized solution. Existing multi-modal datasets has largely ignored the task of CR, clustering mention pairs referring to the same entity. Hence we introduced Flickr30k-Coref dataset that contains images with long-form descriptions annotated with coreference chains and their grounding in the images. We formulated the problem of learning CR by using weak supervision from image-text pairs, using visual grounding as an anchor to disambiguate coreference chains. Moreover, we imposed linguistic rules while learning these chains to avoid clusters that are grammatically incorrect. We demonstrated strong experimental results in multiple settings. We hope that our proposed task definition, dataset and the weakly supervised method will advance the research in multi-modal understanding.
References

[1] Labelstudio. https://labelstud.io/. 3, 11
[2] Spacy. https://spacy.io/. 3, 5, 11, 12
[3] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 294–303, 2008. 1
[4] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12136–12146, 2021. 5
[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568, 2021. 2
[6] Hong Chen, Zhenhua Fan, Hao Lu, Alan L Yuille, and Shu Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Eric Bengtson and Dan Roth. Understanding the value of Spacy. https://spacy.io/ https://labelstud.io/
[7] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In Proceedings of the IEEE International Conference on Computer Vision, pages 824–832, 2017. 2
[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Kevin Clark and Christopher D Manning. Improving coreferences of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 104–120. Springer, 2020. 6
[9] Kevin Clark and Christopher D Manning. Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:1606.01323, 2016. 2
[10] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2601–2610, 2019. 2
[11] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1769–1779, 2021. 2
[12] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In European Conference on Computer Vision, pages 752–768. Springer, 2020. 2, 4, 6
[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 2
[14] Jerry R Hobbs. Resolving pronoun references. Lingua, 44(4):311–338, 1978. 2
[15] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77, 2020. 1, 2, 3
[16] Ashwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdeetr-modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1780–1790, 2021. 2, 6
[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787–798, 2014. 2
[18] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3558–3565, 2014. 2
[19] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 153–169, 2018. 2
[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kananidis, 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. 2
[21] Alina Kuznetsova, Hassan Rom, Neil Aldrinn, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. International Journal of Computer Vision, 128(7):1956–1981, 2020. 3, 11
[22] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task. In Proceedings of the fifteenth conference on computational natural language learning: Shared task, pages 28–34, 2011. 1, 3, 5, 6, 7, 13
[23] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045, 2017. 1, 2, 3, 6, 7, 13
[24] Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 687–692, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 1
[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jeng-Neng Hwang, et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10965–10975, 2022. 2
[26] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. Advances in
Appendix

8. Annotation Details

Localized Narratives dataset. Tuset et al. [38] proposed the Localized Narratives dataset, a new form of multimodal image annotations connecting vision and language. In particular, the annotators describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Hence, each image is described with a natural language description attending to different regions of the image. In addition to textual descriptions (obtained using speech to text conversion), we additionally have mouse traces for the words which also correspond to different regions of the image.

The Localized Narratives dataset is built on top of COCO [27], Flickr30k [37], ADE20k [56] and Open Images [21]. The statistics of the individual datasets are shown in Table 7.

| Localized Narratives Subsets [38] | #images | #Captions | #words/capt. |
|----------------------------------|---------|-----------|--------------|
| COCO                             | 123,287 | 142,845   | 41.8         |
| Flickr30k                        | 31,783  | 32,578    | 57.1         |
| ADE20k                           | 22,210  | 22,529    | 43.0         |
| Open Images                      | 671,469 | 675,155   | 34.2         |

Table 7. Statistics of Localized Narratives for COCO, Flickr30k, ADE20k, and Open Images.

Annotation tool and analysis. We develop an HTML based interface on the Label Studio annotation tool [1]. Figure 4 shows the annotation interface from Label Studio. We hired 6 high quality annotators (all from computer science background) for an average of 54 hours of annotation time. The annotators were trained with the exact description of the task and given a pilot study before proceeding with the complete annotations. The pilot study was useful to correct and retrain annotators if needed. As shown in Figure 4, the annotators had to select a mention in the caption with a given label (C1, C2, etc.) in Step 1 and draw a bounding box in the image for the selected mention in Step 2 (with the same label).

For Step 1, if the mention is coreferring then it is selected with the same label to define coreference chains. It is important to note that the captions are pre-marked with noun phrases parsed from [2]. The annotators are instructed to correct the phrases if they are wrong (e.g. for a mention glass windows, the parser parses glass and windows as two different mentions rather than belonging to the same label/cluster) and remove the phrases that do not correspond to region in the image.

In Step 2, if there are plural mentions such as two men, we ask the annotators to draw two separate bounding boxes for this. In the case of mentions such as several people if the people are less than five, they are instructed to draw separate bounding boxes otherwise a group bounding box (covering all the people).
Given the challenging nature of the task, we doubly annotate 30 images with coreference chains and bounding boxes to compute the inter-annotator agreement. More specifically, for the coreference chain we compute **Exact Match** which denotes whether the coreference chains annotated by the two annotators are the same. We get an exact match of 79.9% in the coreference chains, which is a high agreement given the complexity of the task. For the bounding box localization, we compute the **Intersection over Union (IoU)** to compute the overlap between the two annotations. It is considered to be correct/matching if the IoU is above 0.6. We achieve bounding box accuracy of 81% on this subset of images. This analysis shows good agreement between the annotators given the subjective nature and complexity of the task.

**Flickr30k-Coref dataset.** In total, we annotate all the 1000 test images and 880 validation images (out of 1000) in the Flickr30k dataset. The text descriptions from the Localized Narratives dataset are very noisy with a lot of words/sequence of words. We manually filter phrases such as - *in this image, in the front, in the background, we can see, i can see, in this picture.* If there are some other mentions that are pre-marked and not filtered, we ask the annotators explicitly to filter them out. By doing this, we make sure that the dataset is clear of any unnecessary and noisy mentions.

All the words that are marked as mentions and are not noun phrases (as detected by the part of speech tagger [2]) are considered as pronouns *e.g. them, they, their, this, that, which, those, it, who, he, she, her, him, its.*

**Statistics for the Flickr30k-Coref.** In Figure 5, we show the statistics for the frequency of pronouns in the dataset. Few pronouns (*e.g. he, it, them*) are more frequent than the others. Overall, the occurrence of pronouns is frequent to conduct a fair evaluation of the coreference based models. Similarly in Figure 6, we evaluate how many mentions occur in the coreference chains. Coreference chains with 2 and 3 mentions have a very high frequency in the dataset. There are few chains that have longer mentions (*e.g. 6 and 7*). Hence, we can safely conclude that the dataset is a powerful tool to evaluate coreference chains and learn complex coreferencing and grounding models. Moreover, the average length of the mentions (excluding pronouns) is 1.93.

9. **Additional Quantitative Results**

In Table 8, and Table 9 we present an extension of results from Table 2 and Table 4 in the main paper respectively.

More specifically, in Tab. 8 we add results for the MLP
Table 8. Coreference resolution performance on the Flickr30k-Coref dataset.

| Method         | Text | Image | MT | Arch | Reg | MUC-R | MUC-P | MUC-F1 | BLANC-R | BLANC-P | BLANC-F1 |
|----------------|------|-------|----|------|-----|-------|-------|--------|---------|---------|----------|
| Rule-Based [22] | ✔    | ✗     | ✗  | -    | -   | 5.6   | 10.13 | 6.4    | 3.3     | 4.1     | 4.9      |
| Neural-Coref [23] | ✔    | ✗     | ✗  | -    | -   | 21.7  | 39.4  | 24.4   | 12.0    | 58.6    | 15.5     |
| MAF [49]       | ✔    | ✗     | ✗  | -    | -   | 29.4  | 14.8  | 17.7   | 68.1    | 66.4    | 63.6     |
| Ours           | ✔    | ✔     | ✔  | MLP  | ✗   | 24.36 | 17.55 | 18.60  | 67.46   | 68.08   | 66.00    |
|                | ✔    | ✔     | ✔  | Tr   | ✗   | 33.78 | 20.42 | 23.30  | 71.22   | 69.55   | 67.73    |
| Ours           | ✔    | ✔     | ✔  | MLP  | ✔   | 35.20 | 20.08 | 23.24  | 71.72   | 69.19   | 67.25    |
|                | ✔    | ✔     | ✔  | Tr   | ✔   | 38.98 | 22.06 | 26.29  | 73.33   | 70.13   | 68.55    |
|                | ✔    | ✔     | ✔  | Tr   | ✔   | 44.11 | 24.44 | 29.13  | 76.34   | 71.78   | 70.43    |

In Tab. 9, we evaluate the performance for both coreference resolution and grounding for different model architectures, training with/without regularization and with/without mouse traces (MT). Adding mouse traces during training improves the grounding performance significantly as it provides a very strong signal to disambiguate complex mentions better. Combination of both the mouse trace and regularizer (for both MLP and Transformer) improves on coreferencing and overall grounding accuracy.

Table 9. Exhaustive evaluation for our method under different settings.

| Arch | Reg | MT | MUC-F1 | BLANC-F1 | Acc(%) |
|------|-----|----|--------|----------|--------|
| ✗    | ✗   |    | 19.9   | 66.3     | 22.74  |
| ✗    | ✔   |    | 18.6   | 66.0     | 24.91  |
| ✗    | ✗   |    | 21.8   | 66.8     | 26.01  |
| ✔    | ✗   |    | 23.2   | 67.2     | 27.09  |
| ✗    | ✗   |    | 22.2   | 66.5     | 25.97  |
| ✗    | ✔   |    | 23.3   | 67.7     | 26.75  |
| ✔    | ✗   |    | 26.3   | 68.6     | 28.31  |
| ✔    | ✔   |    | 29.1   | 70.4     | 29.36  |

10. Additional Qualitative Results

In Fig. 7, we show additional qualitative results from our proposed method. The model correctly chains mentions and grounds them to the correct entities in the image even for complex and ambiguous cases. Our model finds coreferences for people (e.g. [a man, his]) or for objects (e.g. [a barbecue grill, it]). Moreover, it also finds links for plurals such as [two men, them]. There is a huge potential in learning to disambiguate the mentions in the descriptions and this work paves the way for future research.

variant of our model architecture and the impact of regularizer on coreference resolution performance. Adding the regularization using language rules consistently improves performance in predicting correct coreference chains.
Figure 7. Additional qualitative results for coreference chains. For each image, we show the predicted coreference chain (mentions more than 2) and the grounding results for the corresponding mentions in the chain. The colored mentions in the descriptions are the ground-truth coreference chains.