

Edited Media Understanding Frames: Reasoning About the Intents and Implications of Visual Disinformation

Jeff Da[♦] Maxwell Forbes^{♦♥} Rowan Zellers^{♦♥} Anthony Zheng[♦]
Jena D. Hwang[♦] Antoine Bosselut^{♦◇} Yejin Choi^{♦♥}

[♦]Allen Institute for Artificial Intelligence [♥]University of Michigan

[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington

[◇]Stanford University

jeffda.com/edited-media-understanding

Abstract

Understanding manipulated media, from automatically generated ‘deepfakes’ to manually edited ones, raises novel research challenges. Because the vast majority of edited or manipulated images are benign, such as photoshopped images for visual enhancements, the key challenge is to understand the complex layers of underlying intents of media edits and their implications with respect to disinformation.

In this paper, we study **Edited Media Understanding Frames**, a new conceptual formalism to understand visual media manipulation as structured annotations with respect to the intents, emotional reactions, effects on individuals, and the overall implications of disinformation. We introduce a dataset for our task, EMU, with 56k question-answer pairs written in rich natural language. We evaluate a wide variety of vision-and-language models for our task, and introduce a new model PELICAN, which builds upon recent progress in pretrained multimodal representations. Our model obtains promising results on our dataset, with humans rating its answers as accurate 48.2% of the time. At the same time, there is still much work to be done – and we provide analysis that highlights areas for further progress.

1 Introduction

The modern ubiquity of powerful image-editing software has led to a variety of new disinformation threats. From AI-enabled “deepfakes” to low-skilled “cheapfakes,” attackers edit media to engage in a variety of harmful behaviors, such as spreading disinformation, creating revenge porn, and committing fraud (Paris and Donovan, 2019; Chesney and Citron, 2019; Kietzmann et al., 2020, c.f.). Accordingly, we argue that it is important to develop systems to help spot harmful manipulated media. The rapid growth and virality of social

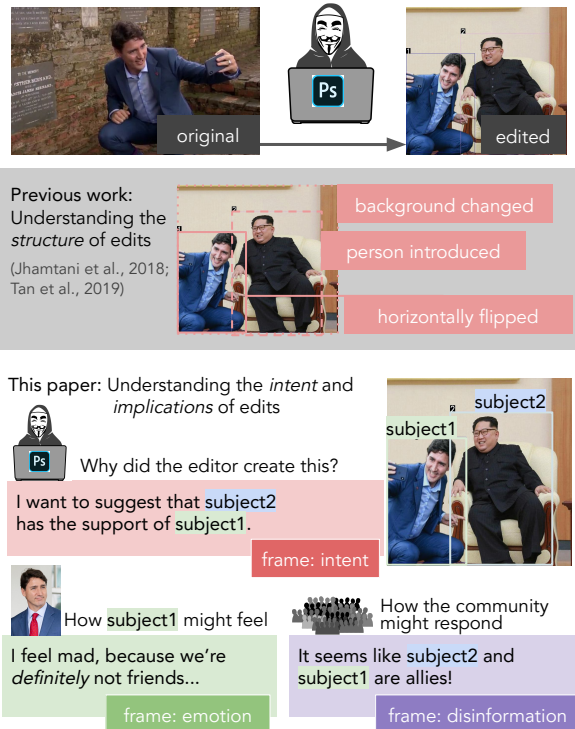


Figure 1: **Edited Media Understanding Frames.** Given a manipulated image and its source, a model must generate natural language answers to a set of open-ended questions. Our questions test the understanding of the *what* and *why* behind important changes in the image – like that subject1 appears to be on good terms with subject2.

media requires as such, especially as social media trends towards visual content (Gretzel, 2017).

Identifying *whether* an image or video has been digitally altered (i.e., “digital forgery detection”) has been a long-standing problem in the computer vision and media forensics communities. This has enabled the development of a suite of detection approaches, such as analyzing pixel-level statistics and compression artifacts (Farid, 2009; Bianchi and Piva, 2012; Bappy et al., 2017) or identifying

“what” the edit was (Tan et al., 2019).

However, little work has been done on “why” an edit is made, which is necessary for identifying harm. Darkening someone’s skin in a family photo because background light made them seem quite pale is generally harmless. While such color rebalancing is common, darkening Barack Obama’s (or Rafael Warnock’s) skin in campaign ads was clearly meant as a harmful edit by the editor that did it.¹ We choose to focus on the “why” – we define a schema for approaching the problem of intent and provide a rich set of natural language responses. We also make a significant contribution towards the “what:” we include a physical-change question, provide rationales based in physical changes, and give structured annotations (bounding boxes) on what was changed in the edit.

We introduce Edited Media Understanding Frames (EMU), a new conceptual formalism that captures the notions of “why” and “what” in image editing for language and vision systems (Figure 1). Following literature on pragmatic frames (Sap et al., 2017, 2020; Forbes et al., 2020)—derived from frame semantics (Baker et al., 1998)— we formalize EMU frames along six dimensions that cover a diverse range of inferences necessary to fully capture the scope of visual disinformation. We delve into the concept of *intention* as discussed by the fake news literature (Rashkin et al., 2017; Shu et al., 2017; Zhou and Zafarani, 2020) to capture editor’s *intent* such as motivation for edit and intent to deceive, as well as the resulting *implications* of the edited content. For every dimension we collect both a classification label and a free-form text explanation. For example, for frame *intent*, a model must classify the intent of the edit, and describe why this classification is selected.

We then introduce a new dataset for our task, EMU, with 56k annotations over 8k image pairs. To kickstart progress on our task, we introduce a new language and vision model, PELICAN, that leverages recent progress in pretrained multimodal representations of images and text (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019). We compare our model to a suite of strong baselines, including a standard VLP model (Zhou et al., 2019), and show key improvement in terms of ability to reason about co-referent subjects in the edit. Nevertheless, our task is far from solved: a significant gap remains

¹How Georgia’s Senate race pits the Old South against the New South. <https://www.politico.com/news/2020/12/05/georgia-senate-old-new-south-442423>

between the best machine and human accuracy.

Our contributions are thus as follows. First, we introduce a new task of Edited Media Understanding Frames, which requires a deep understanding of *why* an image was edited, and a corresponding dataset, EMU, with 56k captions that cover diverse inferences. In addition, we introduce a new model, PELICAN, improving over competitive language-and-vision transformer baselines. Our empirical study demonstrates promising results, but significant headroom remains. We release our dataset at jeffda.com/edited-media-understanding to encourage further study in discovering pragmatic markers of disinformation.

2 Defining Edited Media Understanding Frames

Through an edit e on source image i (e.g. “ $e = x$ is edited into a room full of drugs”), an editor can cause harm to the subject x ’s *mental state* (*mental state*: “ x is angry about e ”) and *effect* x ’s image (*effect*: “ e makes x seem dishonest”) (Rashkin et al., 2016). The editor does this through the *intention* of the edit (*intent*: “ e intends to harm x ’s image”) and changing the *implications* of the image (*implication*: “ e frames x as a drug cartel member”) (Forbes et al., 2020; Sap et al., 2020; Paris and Donovan, 2019).

To this end, we collect edits e and source images i from Reddit’s `r/photoshoppbattles` community. There is no readily available (large) central database of harmful image edits, but `r/photoshoppbattles` is replete with suitable complex and culturally implicative edits (e.g., reference to politics or pop culture). This provides us with relevant image edits at a reasonable cost without advocacy for dangerous training on *real* harmful image edits. Keeping the source image i in the task allows us to sustain the tractability of the image edit problem (Tan et al., 2019; Jhamtani and Berg-Kirkpatrick, 2018).

2.1 Edited Media Understanding Frames: Task Summary

Given an edit $e: I_S \rightarrow I_E$, we define an edited media understanding frame $\mathcal{F}(\ast)$ as a collection of typed dimensions and their polarity assignments: (i) **physical** $P(I_S \rightarrow I_E)$: the changes from $I_S \rightarrow I_E$, (ii) **intent** $N(E \rightarrow I_E)$: whether the Editor E implied malicious intent in $I_S \rightarrow I_E$, (iii) **implication** $M(E \rightarrow I_E)$: how E might use I_E to

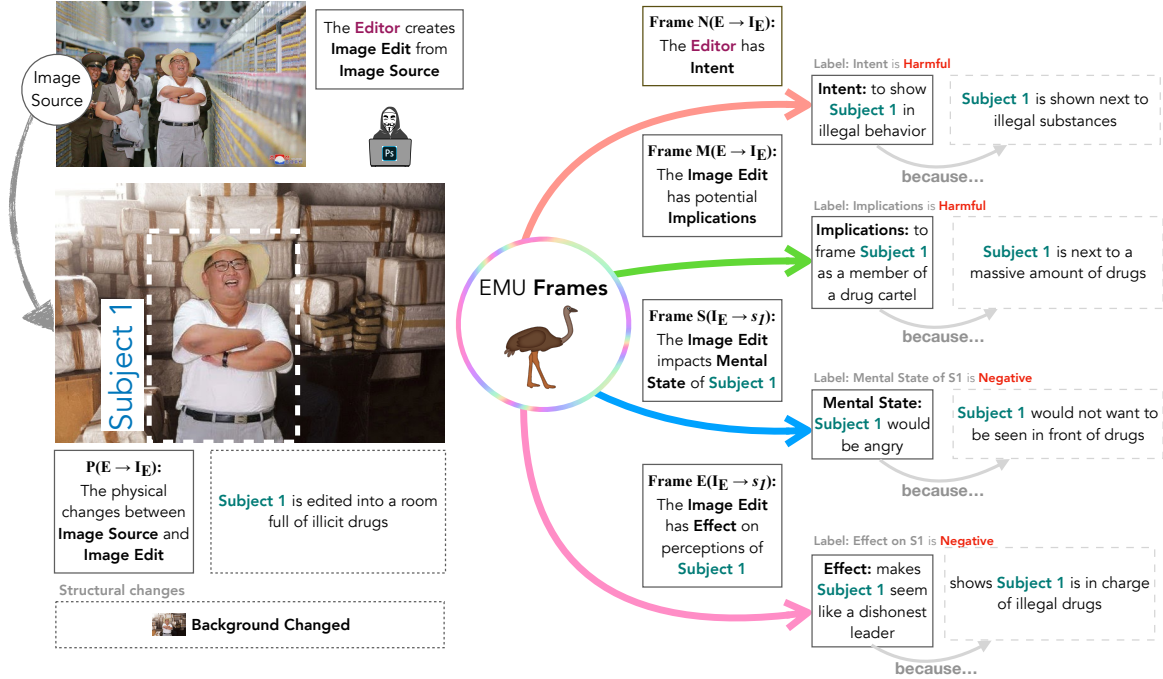


Figure 2: An example from EMU. Given a source image and its edit, and a list of main subjects in the image, we collect a label \mathbf{I} and natural language responses (reponse to frame \mathbf{y} and rationale \mathbf{r} to applicable open-ended questions \mathbf{q} covering each of five frames $f \in \mathcal{F}$). We also collect structural annotations \mathbf{a}_i highlighting the edited sections of the image.

mislead, (iv) **mental state** $S(I_E \rightarrow s_i)$: whether the predicate I_E impacts the emotion of a role s_i , (v) **effect** $E(I_E \rightarrow s_i)$: the effect of I_E on s_i . We assume frames can be categorized as harmful or not harmful with polarity $\mathbf{I} \in \{+, -\}$. Each polarity \mathbf{I} can be interpreted with *reason* \mathbf{y} , and that each reason can be supported with *rationale* \mathbf{r} .

Technically, a model is given the following as input:

- A source image I_S , and an edited image I_E .
- A list of important subjects: expressed as bounding boxes \mathbf{b}_i for each subject.
- An open-ended question \mathbf{q} associated with $\mathcal{F}(\ast)$; e.g., “How might subject3 feel upon seeing this edit?”
- A list of annotated boxes $\mathbf{a}_i \in I_E$ marking the objects in the image that were *introduced* and *modified*, and a true/false label denoting if the background was changed.

A model must produce the polarity classification $\mathbf{I}' \in \{+, -\}$, interpretation of the polarity (response \mathbf{y}') and rationale for interpolation \mathbf{r}' . (For the physical frame, only \mathbf{y} needs to be generated). Figure 2 shows an example of our task configuration. The lexicon of the label is fixed for each $F(\ast)$ (e.g. for $N(\ast)$, $- \rightarrow$ harmful, $+ \rightarrow$ harmless).

3 EMU: A Corpus of Edited Media Understanding Frames

Sourcing Image Edits We source our image edits from the [r/photoshopbattles](#) community on Reddit which hosts regular Photoshop competitions, where given a *source photo*, members submit a comment with their own *edited photo*.

We collect 8K image edit pairs (source and edited photo pairs) from this community by, first, manually curating a list of more than 100 terms describing people frequently appearing in Photoshop battles posts. Then, we screen over 100k posts for titles that contain one or more of these search terms resulting in 20k collected image pairs. Additionally, we run an object detector (He et al., 2017) to ensure that is at least one person present in each image as a means for ensuring that annotators do not see image pairs without any subjects.

Annotating Image Edits We ask a group of vetted crowd workers to identify the main subjects in an image edit and answer open-ended questions in natural language. Each image is annotated by 3 independent crowd workers.

Crowd workers are first presented with a numbered set of people bounding boxes (produced by Mask R-CNN (He et al., 2017)) over the edited

FRAME	Notation	Related Question
PHYSICAL	$P(I_S \rightarrow I_E)$	What changed in this image edit?
INTENT	$N(E \rightarrow I_E)$	Why would someone create this edit?
IMPLICATION	$M(E \rightarrow I_E)$	How might this edit be used to mislead?
MENTAL STATE [of subjectX]	$S(I_E \rightarrow s_i)$	How might this image edit make subjectX feel?
EFFECT [on subjectX]	[on $E(I_E \rightarrow s_i)$]	How could this edit mislead public perception of subjectX?

Table 1: Questions for each of the frames in Edited Media Understanding Frames. Each frame is associated with a question that allows human annotators to address the frame, and models to generate l, y, r for the given frame.

image and are asked to **select subjects** that are significant to the edits (as opposed, say, a crowd in the background). Once subjects are selected, the annotators are asked to assign **classification labels** for each of the five possible question types and provide **free-form text** answers for each question (when applicable). For the classification label, we retain the majority vote (Fleiss $\kappa = 0.67$). In a separate and final pass, we explicitly identify which portions of the modified image is *introduced* or *altered* by asking the workers to label the most important sections of the modified image and selecting one of the two labels. The statistics of the dataset are shown in Figure 3.

4 Modeling Edited Media Understanding Frames

In this section, we present a new model for Edited Media Understanding Frames, with a goal of kick-starting research on this challenging problem. As described in Section 2, our task differs from many standard vision-and-language tasks both in terms of format and required reasoning: a model must take as input two images (a source image and its edit), with a significant change of implication added by the editor. A model must be able to answer questions, grounded in the main subjects of the image, describing these changes. The answers are either boolean labels, or open-ended natural language – including explainable rationales.

4.1 Our model: PELICAN

For Edited Media Understanding Frames, *not all image regions are created equal*. Not only is the subject referred to in the question (e.g. `subject1`) likely important, so too are all of the regions in

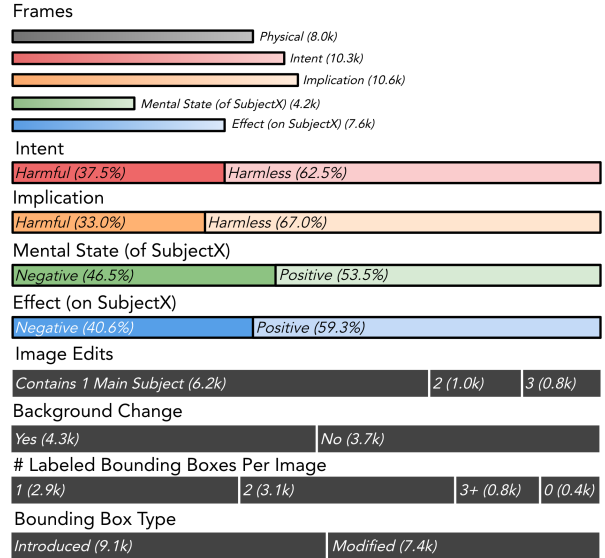


Figure 3: Statistics for EMU. We consider five question types, which in aggregate require a strong understanding of the image edit. The first three types are subject agnostic, though annotations refer explicitly to subjects through subject tags; two (with `subjectX`) are subject-specific.

the image edit that are *introduced* or *altered*. We propose to use the annotations that collected for these regions as additional signal for the model to highlight where to attend.² Not only should a model likely attend to these important regions, it should prioritize attending to regions *nearby* (such as objects that an edited person is interacting with).

We propose to model the (likely) importance of an image region through graph propagation. We will build a directed graph with all regions of the image, rooted at a subject mentioned by the question (e.g. `subject1`). We will then *topologically sort* this graph; each region is then given an embedding corresponding to its sorted position – similar to the position embedding in a Transformer. This will allow the model to selectively attend to important image regions in the image edit. We use a different position embedding for the image source, and do not perform the graph propagation here (as we do not have *introduced* or *altered* annotations); this separate embedding captures the inductive bias that the edited is more important than the source.

²These annotations are collected from workers, but in theory, it would be possible to train a model to annotate regions as such. To make our task as accessible and easy-to-study as possible, however, we use the provided labels in place of a separate model however.

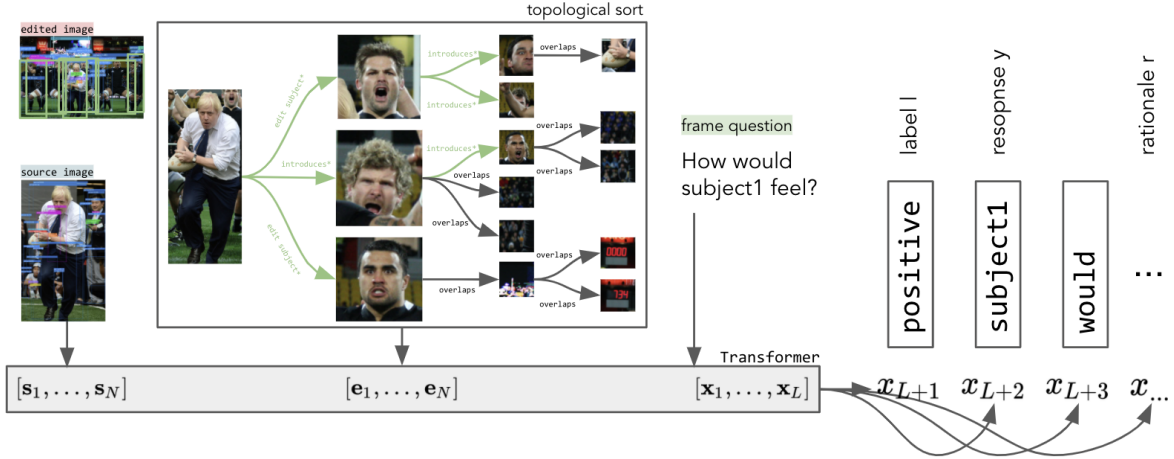


Figure 4: Overview of PELICAN. Our model takes as input all regions s from the source image and e from the edited image. We order the regions in e using a topological sort of overlapping boxes, rooted at subject1. The green regions marked with an asterisk are additional regions that were introduced, and were labeled through annotators. This ordering allows the model to selectively attend to important image regions in generating an answer to the visual question about subject1.

4.2 Model details and Transformer integration

In this section, we describe integrating our *importance embeddings* with a multimodal transformer.

Let the source image be I_S and I_E . We use the backbone feature extractor ϕ (Faster-RCNN feature extractor (Ren et al., 2015; Anderson et al., 2018)) to extract N regions of interest for each region:

$$[s_1, \dots, s_N] = \phi(I_S) \quad [e_1, \dots, e_N] = \phi(I_E). \quad (1)$$

We note that some of these regions in e_1, \dots, e_N are provided to the model (as annotated regions in the image); the rest are detected by ϕ . These, plus the language representation of the question, are passed to the Transformer backbone T :

$$[z_1 \dots z_{N+L}] = T([s_1 \dots s_N], [e_1, \dots, e_N], [x_1 \dots x_L]) \quad (2)$$

Important for EMU, $z_{2N+1}, \dots, z_{2N+L}$ serve as language representations. Training under a left-to-right language modeling objective, we can predict the next *next token* x_{L+1} using the representation z_{N+L} .

4.2.1 Prioritization Embeddings from Topological Sort

Transformers require *position embeddings* to be added to each image region and word – enabling it to distinguish which region is which. We supplement the position embeddings of the regions

$\{e_1 \dots e_N\}$ in the edited image I_E with the result of a topological sort.

Graph definition. We define the graph over image regions in the edited image as follows. We begin by sourcing a seed region $s \in \{e_1 \dots e_N\}$. Let $G = (V, E)$, where each $v \in V$ represents metadata of some $r_i \in \phi(I_E)$, defined as $v_i \in m(I_E)$ for simplicity, s.t.:

$$v_i = \{x_1, y_1, x_2, y_2, s_i, l_i\} \quad (3)$$

where x_1, y_1, x_2, y_2 represents the bounding box of r_i , $s_i \in \{1, 0\}$ denoting if r_i is a subject of I_E , and $l_i \in \{\text{introduced}, \text{altered}\}$ denoting the label of r_i .

We build the graph iteratively: for each iteration, we define an edge $e = \{v, u\}; u \in V$ s.t.:

$$\forall v \in m(I_E), \forall u \in V, E = E \cup (u, v) \in E' \quad (4)$$

We define E' as the set of edges (u, v) in which u and v are *notationally similar*. We define three cases in which this is true: if $s_i \in u_i \wedge s_j \in v_j$, if $l_i \in u_i = l_j \in v_j$, and if $x_1, y_1, x_2, y_2 \in u_i$ and $x_3, y_3, x_4, y_4 \in u_i$ overlaps, in which the percentage overlap is defined by standard intersection-over-union:

$$\frac{\min\{x_4, x_2\} - \max\{x_3, x_1\}}{\min\{y_4, y_2\} - \max\{y_3, y_1\}} \quad (5)$$

We cap the number of outgoing edges at 3, and prevent cycles by allowing edges only to unseen image regions. In cases where there are more than

three possible edges, we add edges in the order defined in the previous paragraph, and break overlap ties via maximum overlap.

To produce embeddings, we run topological sort over the directed graph to assign each image region an embedding, then assign an embedding to each image region based on the ordered index. The embedding is zeroed out for image regions that are missing from the DAG, and from the source image (which are unlabeled). We include bounding box and class labels. To generate text and classification labels, we attach the embeddings onto the input for an encoder-decoder structure.

5 Experimental Results on EMU

In this section, we evaluate a variety of strong vision-and-language generators on EMU. Similar to past work on VQA, we rebalance our test set split ensuring a 50/50 split per question type of maliciously labeled captions. We provide two human evaluation metrics – head-to-head, in which generated responses are compared to human responses, and accuracy, in which humans are asked to label if generated responses are accurate in regards to the given edit.

5.1 Baselines

In addition to evaluating PELICAN, we compare and evaluate the performance of various potentially high-performing baselines on our task.

a. Retrieval. For a retrieval baseline, which generally performs well for generation-based tasks, we use features from ResNet-158 (He et al., 2016), defined as ϕ , to generate vectors for each I_E in the test set. We then find the most similar edited image I_T in the training set \mathbf{T} via cosine similarity:

$$\operatorname{argmax}_{I_T \in \mathbf{T}} \frac{\phi(I_E) \cdot \phi(I_T)}{\|\phi(I_E)\| \times \|\phi(I_T)\|} \quad (6)$$

We use the captions associated with the most similar image in the training set.

b. GPT-2 (Radford et al., 2019). As a text-only baseline, we use the 117M parameter model from GPT-2, fine-tuned on the captions from our dataset. Since the images are not taken into consideration, we generate from the seeds associated with each question type and use the same captions for all images in the test set.

c. Cross-Modality GPT-2. We test a unified language-and-vision model on our dataset. Similar to (Alberti et al., 2019), we append the visual

features $\phi(I_S)$ and $\phi(I_E)$ to the beginning of the token embeddings from GPT-2 (117M). For the questions involving a subject, we append an additional vector $\phi(r)$, where r is the region defined by the bounding box for that subject.

d. Dynamic Relational Attention (Tan et al., 2019). We test the best model from previous work on image edits on our task, Dynamic Relational Attention. We train the model from scratch on our dataset, using the same procedure as (Tan et al., 2019). We seed each caption with the relevant question.

e. VLP (Zhou et al., 2019). We test VLP, a pre-trained vision-and-language transformer model. For image captioning, VLP takes a single image as input and uses an off-the-shelf object detector to extract regions, generation a caption using sequence-to-sequence decoding and treating the regions as a sequence of input tokens.

To generate a caption for a particular question type, we fix the first few generated tokens to match the prefix for that question type. We fine-tune VLP starting from weights pre-trained on Conceptual Captions (3.3m image-caption pairs) (Sharma et al., 2018) and then further trained on COCO Captions (413k image-caption pairs) (Lin et al., 2014).

5.2 Quantitative Results and Ablation Study

We present our results in Table 2. We calculate generative metrics (e.g. METEOR) by appending the rationale to the response. Generations from PELICAN are preferred over human generations 14.0% of the time, with a 0.86 drop in perplexity compared to the next best model. To investigate the performance of the model, we run an ablation study on various modeling attributes, detailed in Table 3. First, we investigate the effect of pretraining (on Conceptual Captions (Sharma et al., 2018; Zhou et al., 2019)). We find that performance drops without pretraining (53.47%), but surprisingly still beats other baselines. This suggests that the task requires more pragmatic inferences than the semantic learning typically gained from pre-training tasks. Second, we ablate the importance of including annotated (**a**) features from the dataset when creating the directed graph, relying on a seed from a random R-CNN region (54.44%). We also ablate our use of topological sort and a directed graph by suggesting a simple (but consistent) order for image regions (54.91%). Finally, we ablate including the visual regions from the source image. The performance is

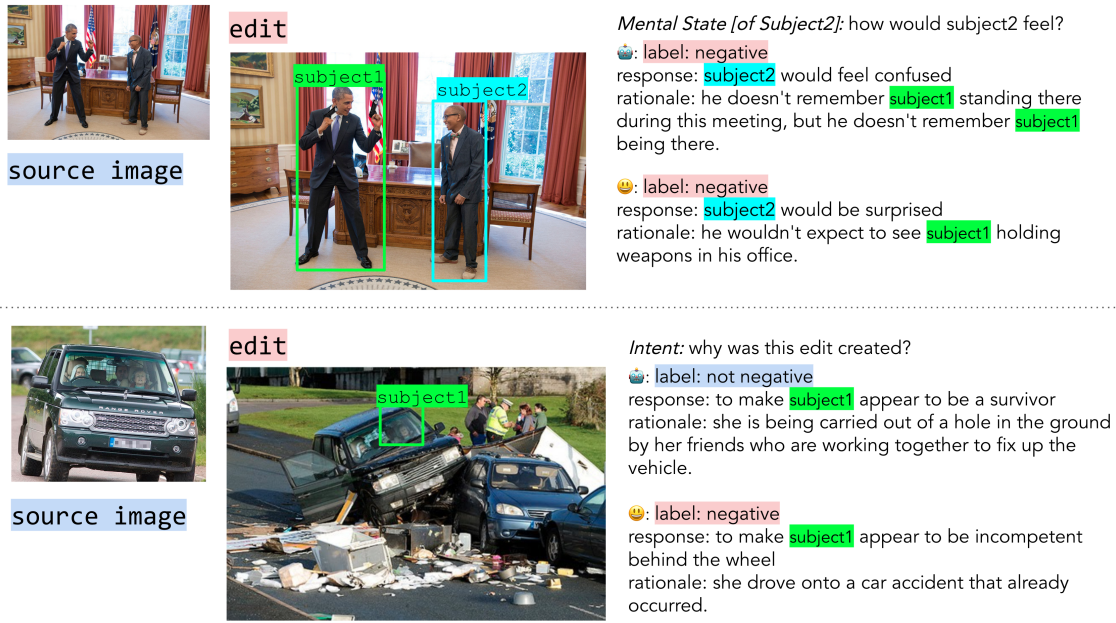


Figure 5: Generation examples from PELICAN, marked with results from human evaluation. PELICAN is able to correctly reference marked figures and is able to infer intent accordingly across each question type.

Model	Automated metrics				Human evaluation	
	Perplexity ↓	ROUGE-L ↑	METEOR ↑	Accuracy ↑	Head-to-Head ↑	Accurate % ↑
Humans	n/a	n/a	n/a	89.8	50.0	95.2
Retrieval Baseline	n/a	11.5	7.2	51.9	4.4	20.6
GPT-2 (Radford et al., 2019)	26.6	10.3	6.2	50.0	0.0	3.0
Cross-Modality GPT-2	22.1	12.0	7.9	51.0	4.1	10.4
Dynamic RA (Tan et al., 2019)	23.1	13.2	8.9	51.8	5.3	12.4
VLP (Zhou et al., 2019)	12.3	18.5	10.5	53.2	9.3	20.3
PELICAN REAL (ours)	11.6	19.5	10.8	54.1	11.3	25.5
PELICAN (ours)	11.0	22.1	11.6	55.4	14.6	48.2

Table 2: Experimental results on EMU. We compare our model, PELICAN, with several strong baseline approaches. We calculate generative metrics (e.g. METEOR) by appending the rationale to the response. PELICAN REAL describes a version of PELICAN trained on EMU without additional human annotation (6.1).

Model	Auto Eval	Human Eval
	Accuracy ↓	Accuracy ↑
PELICAN	55.40	48.2
physical	n/a	60.5
intent	55.2	43.0
implication	60.1	49.9
mental state [of subjectx]	54.6	42.5
effect [on subjectx]	53.7	41.1
– pretraining	54.6	44.0
– annotated features	54.4	40.1
– directed graph	54.9	45.2
– source image	55.3	47.5

Table 3: Ablation study for PELICAN. We also explore the performance of PELICAN across each frame type.

similar (55.35%), suggesting that PELICAN would be able to perform in real-world settings in which only the edited image is present (e.g. social media posts).

5.3 Qualitative Results

Last, we present qualitative examples in Figure 5. PELICAN is able to correctly understand image pairs which require mostly surface level understanding - for example, in the top example, it is able to identify that the gun and action implies negative context, but misunderstands the response with regards to the situation. In the bottom example, we show that PELICAN is able to refer to subject1 correctly, but misinterprets the situation to be non-negative.

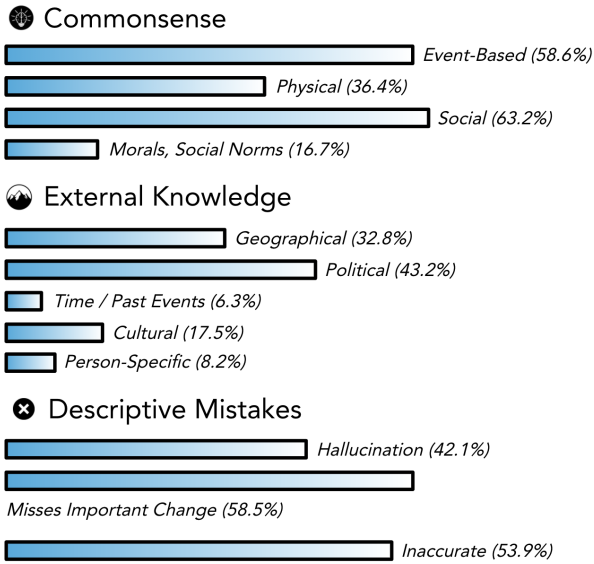


Figure 6: Failure cases from PELICAN, trained on EMU. Commonsense is the largest differentiator between human understanding and model-based analysis of disinformation.

6 Future Implications

6.1 EMU in the Real World

To study if EMU is helpful in real-world settings, we train a model of PELICAN on EMU with only the edited image. In this setting, the model must hypothesize which parts of the image were edited and discern the main subjects in the image. At test time, we generate captions for each of the 5 intention-based question types. Results of this version of PELICAN is in Table 2.

While this evaluation scheme is crude, we find that this version of PELICAN is still able to outperform previous models without usage of the source image. This suggests potential for generations from EMU-trained models in *human-assisted* settings. In an initial human study (given PELICAN REAL captions, classify the edit as disinformation – were the captions helpful in your decision?) we find that annotators label as helpful 71.5% of the time. Additionally, annotators tended more often to pick the gold label (89.1% → 95.2%).

6.2 Failure Cases in Current Models and Avenues for Future Research

EMU also helps us understand what current vision-and-language models are missing for use on disinformation, by analyzing the reasons and rationales generated. We ask annotators to compare PELICAN-generated captions marked as “worse”

and human captions. Category details are included in the appendix. Figure 6 shows our results. Overall, current models primarily lack the commonsense (event-based and social) to accurately describe disinformation. Geographical (location-based) and political (e.g. knowledge about the job of a president) external knowledge is also a missing component.

PELICAN also still makes mistakes in description-related attributes: describing something other than the important change and an inaccuracy (e.g. wrong color) are the most common. Specific information – such as information relating to a specific person in the image (i.e. requiring a model to identify the person in the image), and information about a past event – are the least critical, suggesting that efforts should be focused first on general intelligence rather than named-entity lookup.

7 Related Work

Language-and-Vision Datasets Datasets involving images and languages cover a variety of tasks, including visual question answering (Agrawal et al., 2015; Goyal et al., 2017), image caption generation (Lin et al., 2014; Young et al., 2014; Krishna et al., 2016), visual storytelling (Park and Kim, 2015; Bosselut et al., 2016), machine translation (Elliott et al., 2016), visual reasoning (Johnson et al., 2017; Hudson and Manning, 2019; Suhr et al., 2019), and visual common sense (Zellers et al., 2019).

Two-image tasks Though most computer vision tasks involve single images, some work has been done on exploring image pairs. The NLVR2 dataset (Suhr et al., 2019) involves yes-no question answering over image pairs. Neural Naturalist (Forbes et al., 2019) tests fine-grained captioning of bird pairs; (Jhamtani and Berg-Kirkpatrick, 2018) identifies the difference between two similar images.

Image Edits There has been some computer vision research studying image edits. Unlike our EMU dataset, however, much of this work has focused on modeling lower-level image edits wherein the *cultural implications* do not change significantly between images. For example, (Tan et al., 2019) predicts image editing requests (generate ‘change the background to blue’ from a pair of images). Past work has also studied learning to perform image adjustments (like colorization and enhancement) from a language query (Chen et al., 2017; Wang et al., 2018). Hateful Meme Challenge

(Kiela et al., 2020) is a recent work challenging models to classify a meme as hateful or not.

8 Conclusion

We present Edited Media Understanding Frames—a language-and-vision task requiring models to answer open-ended questions that capture the intent and implications of an image edit. Our model, PEL-ICAN, kickstarts progress on our dataset – beating all previous models and with humans rating its answers as accurate 48.2% of the time. At the same time, there is still much work to be done – and we provide analysis that highlights areas for further progress.

Acknowledgements

The authors would like to thank Ryan Qiu for help with analysis, and the Amazon Mechanical Turk community for help with annotation. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1256082, and in part by NSF (IIS-1714566), DARPA CwC through ARO (W911NF15-1-0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), NSF (IIS-1714566), and the Allen Institute for AI.

9 Ethical Considerations

In constructing the EMU dataset, great care was taken to ensure that crowd-workers are compensated fairly for their efforts. To this end, we monitored median HIT completion times for each published batch, adjusting the monetary reward such that at least 80% of workers always received >\$15/hour, which is roughly double the minimum wage in the United States (the country of residence for most Amazon Mechanical Turk workers). This included the qualification and evaluation rounds. The following data sheet summarized relevant aspects of the data collection process (Bender and Friedman, 2018):

A. CURATION RATIONALE: Selection criteria for the edits included in the presented dataset are discussed Section 3. We selected the highest-rated posts on Reddit, and collected metadata data from annotators marking if the edit is NSFW or offensive.

B. LANGUAGE VARIETY: The dataset is available in English, with mainstream US Englishes being the dominant variety, as per the demographic of Amazon Mechanical Turk workers.

C. SPEAKER DEMOGRAPHIC: N/A

D. ANNOTATOR DEMOGRAPHIC: N/A

E. SPEECH SITUATION: All frames were collected and validated over a period of about 12 weeks, between November and January 2020, through the Amazon AMT platform. Workers were given regular, detailed feedback regarding the quality of their submissions and were able to address any questions or comments to the study’s main author via Email or Slack.

F. TEXT CHARACTERISTICS: In line with the intended purpose of the dataset, the included edits describe social interactions related (but not limited to) platonic and romantic relationships, political situations, as well as cultural and social contexts.

G. RECORDING QUALITY: N/A

H. OTHER: N/A

Lastly, we want to emphasize that our work is strictly scientific in nature, and serves the exploration of machine reasoning alone. It was not developed to offer guidance on misinformation or to train models to classify social posts as misinformation. Consequently, the inclusion of malicious image edits could allow adversaries to train malicious agents to produce visual misinformation. We are aware of this risk, but also want to emphasize that the utility of these agents allow useful negative training signal for minimizing harm that may be caused by agents operating in visual information. It is, therefore, necessary for future work that uses our dataset to specify how the collected examples of both negative and positive misinformation are used, and for what purpose.

References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.
- Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. *ArXiv*, abs/1908.05054.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Collin F. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.
- Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. 2017. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979.
- Emily M. Bender and B. Friedman. 2018. Data statements for nlp: Toward mitigating system bias and enabling better science.
- Tiziano Bianchi and Alessandro Piva. 2012. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017.
- Antoine Bosselut, Jianfu Chen, David Warren, Hananeh Hajishirzi, and Yejin Choi. 2016. Learning prototypical event structure from photo albums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1769–1779.
- Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2017. Language-based image editing with recurrent attentive models. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8721–8729.
- Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753.
- Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *ArXiv*, abs/1605.00459.
- Hany Farid. 2009. A survey of image forgery detection. *IEEE Signal Processing Magazine*, 26(2):16–25.
- M. Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.
- Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge J. Belongie. 2019. Neural naturalist: Generating fine-grained image comparisons. In *EMNLP/IJCNLP*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.
- Ulrike Gretzel. 2017. The visual turn in social media marketing.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.
- Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *EMNLP*.
- Johanna E. Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790.
- Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
- Britt S. Paris and Joan M. Donovan. 2019. Deepfakes and cheap fakes. Technical report, Data and Society.
- Cesc C. Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NIPS*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. [Truth of varying shades: Analyzing language in fake news and political fact-checking](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.
- Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. *arXiv: Computation and Language*.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.
- Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Alane Suhr, Stephanie Zhou, Iris D. Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *ACL*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP*.
- Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *ACL*.
- Hai Wang, Jason D. Williams, and SingBing Kang. 2018. Learning to globally edit images with textual description. *ArXiv*, abs/1810.05786.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.
- Luowei Zhou, Hamid Palangi, Lefei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059.
- Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Appendices

A.1 Reproducibility of Experiments

We provide downloadable source code of all scripts, and experiments, at `to-be-provided`. We use two Titan X GPUs to train and evaluate all models, except Dynamic Relational Attention (Tan et al., 2019), which was trained on a single Titan Xp GPU. For GPT-2 (Radford et al., 2019), we use the 117M parameter model, taking 5 hours to train. Our configuration of VLP (Zhou et al., 2019) has 138,208,324 parameters, taking 6 hours to train. Our model, PELICAN, has 138,208,324 parameters, taking 6 hours to train. Our Dynamic Relational Attention model has 55,165,687 parameters, taking 10 hours to train.

A.2 Reproducibility of Hyperparameters

For models using GPT-2 as their underlying infrastructure, we use a maximum sequence length of 1024, 12 hidden layers, 12 heads for each attention layer, and 0.1 dropout in all fully connected layers. For Dynamic Relational Attention (Tan et al., 2019), we use a batch size of 95, hidden dimension size of 512, embedding dimension size of 256, 0.5 dropout, Adam optimizer, and a $1e-4$ learning rate. We used early stopping based on the BLEU score on the validation set at the end of every epoch; the test scores reported are for a model trained for 63 epochs. For all models relying on VLP as their underlying infrastructure, we use 30 training epochs, 0.1 warmup proportion, 0.01 weight decay, 64 batch size.

A.3 Reproducibility of Datasets

Our dataset has 39338 examples in the training set and 4268 and 3992 examples in the development and test sets respectively. All training on additional datasets (e.g. (Zhou et al., 2019)) matches their implementation exactly. Our train/val/test splits were chosen at random, during the annotation period. No data was excluded, and no additional pre-processing was done. A downloadable link is available at `to-be-provided` after publication.

A.4 Data Collection

For reference and reproducibility, we show the full template used to collect data in Figure 9.

We also show our human evaluation process in Figure 10.

Distribution across subject counts

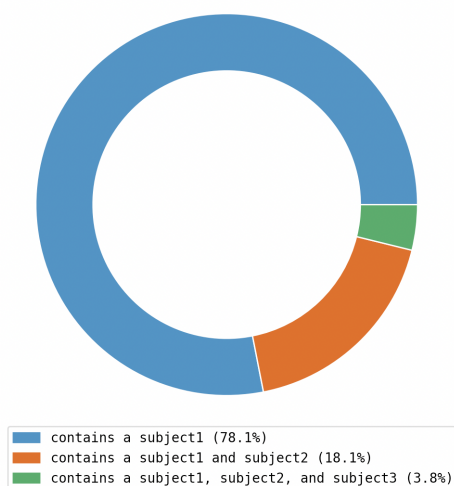


Figure 7: Subject distribution. To highlight our decision for a 3 subject limit, we show that the majority of images contains 1-2 subjects.

A.5 Additional Annotation Details

For an image pair (consisting of an image edit and a source image), we 1) ask the annotator to identify and index the main subjects in the image, 2) prime the annotator by asking them to describe the physical change in the image, 3) ask a series of questions for each main person they identified, and 4) ask a series of questions about the image as a whole. For each question we require annotators to provide both an answer to the question and a rationale (e.g. the physical change in the image edit that alludes to their answer). This is critical, as the rationales prevent models from guessing a response such as “would be harmful” without providing the proper reasoning for their response. We ask annotators to explicitly separate the rationale from the response by using the word “because” or “since” (however, we find that the vast majority of annotators naturally do this, without being explicitly prompted). For the main subjects, we limit the number of subjects to 3. This also mitigates a large variation in workload between image pairs, which was gathered as potentially problematic from annotator feedback. We limit the number of captions per type to 3. We find that a worker chooses to provide more than one label for a type in only a small proportion of cases, suggesting that usually, one caption is needed to convey all the information about the image edit relating to that type .

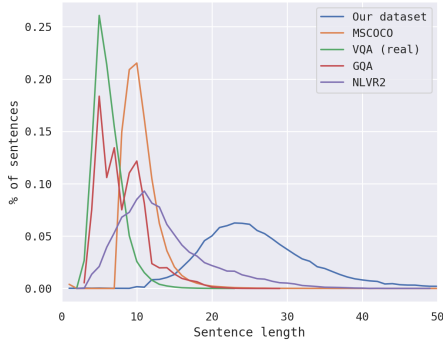


Figure 8: Language sentence length distribution, measured in words, across other language-and-vision datasets. The natural language answers in EMU show a high degree of complexity, with an average sentence length of 26.45 words.

A.6 Lexical Analysis

Word-Level Statistics We analyze the lexical statistics of this dataset. We remove stop words as words such as “him”. We show that different types require different language in their response. In addition, we highlight that many of the rationales involve people, suggesting that understanding social implications is critical to solving this task.

A.7 Motivation for EMU Task Definition

We begin by motivating and contextualising our problem. A key insight is that we need to think into the future – since the task is important but difficult, we aim to structure EMU such that it can help models learn how to understand misinformation (by providing the source image, grounding captions, and additional annotations) without oversimplifying the task.

Frames. We ask models a series of questions about the *what and why* of the image edit. We arrived on these questions by first asking annotators to explain the image edits without prompting. Then, we bucketed the responses into similar categories, motivating us to create questions based on the parts of edits humans naturally focused on. In our task, we consider six open-ended question types – *physical*, *intent*, *implication*, *emotion* [of SubjectX], *attack* [on SubjectX], and *disinformation*. Descriptions of each are in Figure 2. Each type focuses on a different aspect of the image edit, and is related one-to-one with an open-ended question q . Each question type may also reference a specific entity b . In these cases, the answer to the question would differ based on the main subject

Figure 9: Example of our annotation process.


Figure 10: Example of our evaluation process.

referred.

Labels. For each q , we ask models to provide both a classification label \mathbf{l} and a generated answer (response \mathbf{y} and rationale \mathbf{r}) for a given image edit. Visual misinformation is not a closed form problem – the potential label-space and responses for an malicious edit are ever-changing with recent events. Thus, we suggest that models need to produce a generated answer. However, we also want models to go beyond simple answering – we want them to answer *for the right reasons*, in an explainable way. Thus, we require models to generate a *rationale* explaining why its answer is true. For example, a good rationale explains that the perception of subject1 could be injured because a gun was added to subject1’s hand. Our evaluation recruits human raters to compare generated answers and rationales \mathbf{y}/\mathbf{r} to those written by annotators. To account for the current difficulty of evaluating generation, we

Rationales		Responses									
		intent		implication		disinformation		emotion [of SubjectX]		attack [on SubjectX]	
holding	4.21%	fun	4.83%	public	3.07%	movie	2.93%	confused	7.62%	likes	3.00%
face	4.09%	powerful	1.13%	think	2.12%	woman	2.12%	amused	4.38%	hates	2.21%
wearing	3.17%	funny	1.09%	man	1.75%	new	1.92%	embarrassed	3.88%	loves	1.36%
man	2.64%	hero	1.02%	fun	1.68%	game	1.23%	upset	3.50%	wants	1.35%
appears	2.41%	movie	1.01%	disgrace	1.25%	real	1.23%	proud	2.61%	doesn't	1.31%

Table 4: Lexical statistics. Statistics for each dimension represent omit the rationale, and statistics for the rationale are reported separately.



of these labels in our modeling section.

A: subject1 would ...

B: subject1 might ...

[1] Which caption gives a better analysis of the edit?

Definitely A

Slightly A

Slightly B

Definitely B

[2] How accurate is the worse caption?

Slightly Accurate

Not Accurate

Figure 11: Our template for human evaluations. Each annotator is shown an edited image, the source image, and is asked to compare a human annotated captions and a machine annotated caption.

include a binary classification label **I** for each of the “why” answers to allow for a simple checkpoint evaluation metric of model progress.

Grounding. Each explanation is grounded to bounding boxes \mathbf{a}_i of the people in the edited image. Similar to past work in vision-and-language (Zellers et al., 2019), annotators write captions that refer to the bounding box (for example, subject1would be angry). This allows precise reference in visually complex edits.

Additional annotations. Finally, we provide annotators for bounding boxes of *introduced* and *modified* regions in edited images. These bounding boxes provide the *syntax* of the change in a machine digestible format (bounding boxes + labels). We conduct initial exploration of the empirical benefit