Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs

Ana Marasović[†◇] **Chandra Bhagavatula**[†] **Jae Sung Park**[◇†]
Ronan Le Bras[†] **Noah A. Smith**[◇†] **Yejin Choi**[◇†]

[†]Allen Institute for Artificial Intelligence
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
Seattle, WA, USA
{anam,chandrab,jamesp,ronanlb}@allenai.org
{jspark96,nasmith,yejin}@cs.washington.edu

Abstract

Natural language rationales could provide intuitive, higher-level explanations that are easily understandable by humans, complementing the more broadly studied lower-level explanations based on gradients or attention weights. We present the first study focused on generating natural language rationales across several complex visual reasoning tasks: visual commonsense reasoning, visual-textual entailment, and visual question answering. The key challenge of accurate rationalization is comprehensive image understanding at all levels: not just their explicit content at the pixel level, but their contextual contents at the semantic and pragmatic levels. We present RATIONALE^{VT} TRANSFORMER, an integrated model that learns to generate free-text rationales by combining pretrained language models with object recognition, grounded visual semantic frames, and visual commonsense graphs. Our experiments show that free-text rationalization is a promising research direction to complement model interpretability for complex visual-textual reasoning tasks. In addition, we find that integration of richer semantic and pragmatic visual features improves visual fidelity of rationales.

1 Introduction

Explanatory models based on natural language rationales could provide intuitive, higher-level explanations that are easily understandable by humans (Miller, 2019). In Figure 1, for example, the natural language rationale given in free-text provides a much more informative and conceptually relevant explanation to the given QA problem compared to the non-linguistic explanations that are often provided as localized visual highlights on the image. The latter, while pertinent to what the vision component of the model was attending to, cannot provide the full scope of rationales for such complex reasoning tasks as illustrated in Figure 1. Indeed, explanations for higher-level conceptual reasoning can be best conveyed through natural language, as has been studied in recent literature on (visual) NLI (Do et al., 2020; Camburu et al., 2018), (visual) QA (Wu and Mooney, 2019; Rajani et al., 2019), playing arcade games (Ehsan et al., 2019), fact checking (Atanasova et al., 2020), image classification (Hendricks et al., 2018), motivation prediction (Vondrick et al., 2016), and self-driving cars (Kim et al., 2018).

In this paper, we present the first focused study on generating natural language rationales across several complex visual reasoning tasks: visual commonsense reasoning, visual-textual entailment, and visual question answering. Our study aims to *complement* the more broadly studied lower-level explanations such as attention weights and gradients in deep neural networks (Simonyan et al., 2014; Zhang et al., 2017; Montavon et al., 2018, among others). Because free-text rationalization is a challenging research question, we assume the gold answer for a given instance is given and scope our investigation to justifying the gold answer.

The key challenge in our study is that accurate rationalization requires comprehensive image understanding at all levels: not just their basic content at the pixel level (recognizing "waitress", "pancakes", "people" at the table in Figure 1), but their contextual content at the semantic level (understanding the structural relations among objects and entities through action predicates such as "delivering" and "pointing to") as well as at the pragmatic level (understanding the "intent" of the pointing action is to tell the waitress who ordered the pancakes).

We present RATIONALE^{VT} TRANSFORMER, an integrated model that learns to generate free-text rationales by combining pretrained language models based on GPT-2 (Radford et al., 2019) with visual features. Besides commonly used features derived from object detection (Fig. 2a), we explore two

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2810–2829
November 16 - 20, 2020.

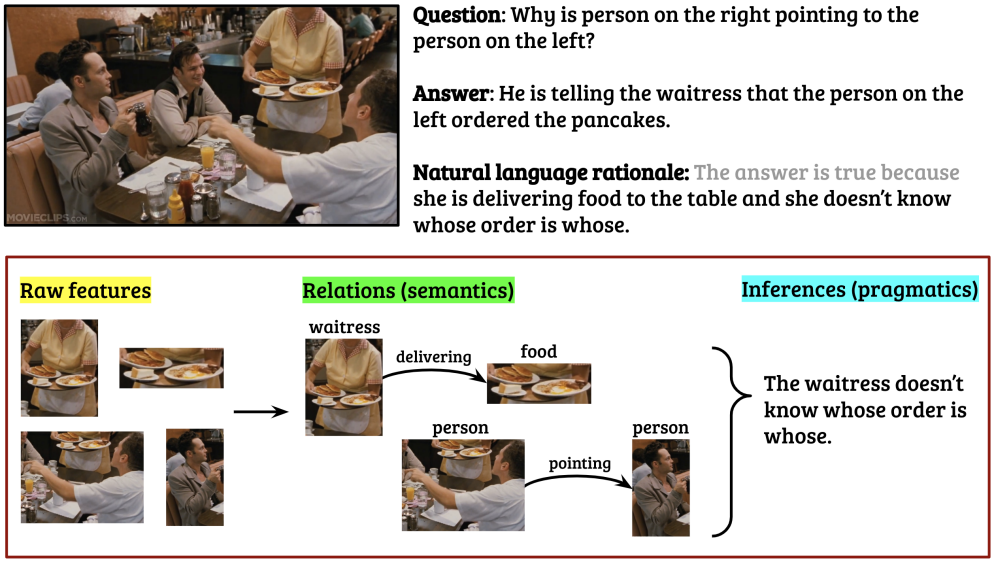


Figure 1: An illustrative example showing that explaining higher-level conceptual reasoning cannot be well conveyed only through the attribution of raw input features (individual pixels or words); we need natural language.

new types of visual features to enrich base models with semantic and pragmatic knowledge: (i) visual semantic frames, i.e., the primary activity and entities engaged in it detected by a grounded situation recognizer (Fig. 2b; Pratt et al., 2020), and (ii) commonsense inferences inferred from an image and an optional event predicted from a visual commonsense graph (Fig. 2c; Park et al., 2020).[1]

We report comprehensive experiments with careful analysis using three datasets with human rationales: (i) visual question answering in VQA-E (Li et al., 2018), (ii) visual-textual entailment in E-SNLI-VE (Do et al., 2020), and (iii) an answer justification subtask of visual commonsense reasoning in VCR (Zellers et al., 2019a). Our empirical findings demonstrate that while free-text rationalization remains a challenging task, newly emerging state-of-the-art models support rationale generation as a promising research direction to complement model interpretability for complex visual-textual reasoning tasks. In particular, we find that integration of richer semantic and pragmatic visual knowledge is important for generating rationales with higher visual fidelity, especially for tasks that require higher-level concepts and richer background knowledge.

Our code, model weights, and the templates used for human evaluation are publicly available.[2]

[1]Figures 2a–2c are taken and modified from Zellers et al. (2019a), Pratt et al. (2020), and Park et al. (2020), respectively.

[2]https://github.com/allenai/visual-reasoning-rationalization

2 Rationale Generation with RATIONALE^{VT} TRANSFORMER

Our approach to visual-textual rationalization is based on augmenting GPT-2's input with output of external vision models that enable different levels of visual understanding.

2.1 Background: Conditional Text Generation

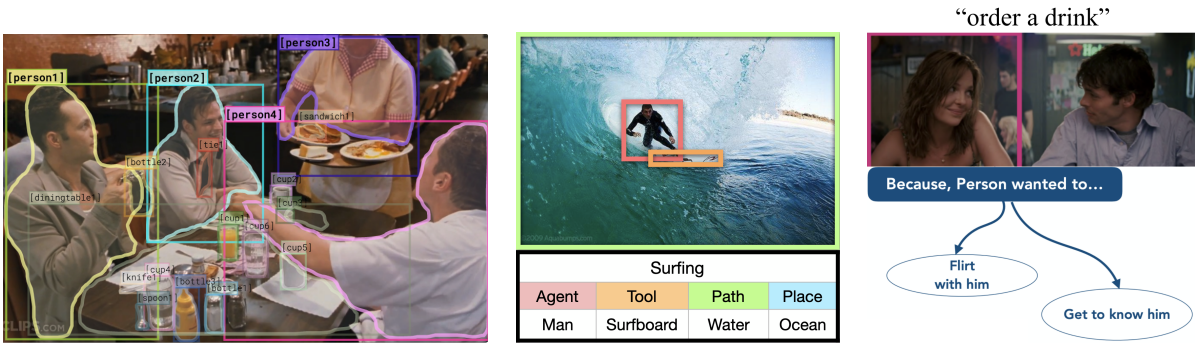
The GPT-2's backbone architecture can be described as the decoder-only Transformer (Vaswani et al., 2017) which is pretrained with the conventional language modeling (LM) likelihood objective.[3] This makes it more suitable for generation tasks compared to models trained with the masked LM objective (BERT; Devlin et al., 2019).[4]

We build on pretrained LMs because their capabilities make free-text rationalization of complex reasoning tasks conceivable. They strongly condition on the preceding tokens, produce coherent and contentful text (See et al., 2019), and importantly, capture some commonsense and world knowledge (Davison et al., 2019; Petroni et al., 2019).

To induce conditional text generation behavior, Radford et al. (2019) propose to add the context tokens (e.g., question and answer) before a special token for the generation start. But for visual-textual tasks, the rationale generation has to be conditioned not only on textual context, but also on an image.

[3]Sometimes referred to as density estimation, or left-to-right or autoregressive LM (Yang et al., 2019).

[4]See Appendix §A.1 for other details of GPT-2.



(a) Object Detection (b) Grounded Situation Recognition (c) Visual Commonsense Graph

Figure 2: An illustration of outputs of external vision models that we use to visually adapt GPT-2.

	Object Detector	**Grounded Situation Recognizer**	**Visual Commonsense Graphs**
Understanding	Basic	Semantics	Pragmatics
Model name	Faster R-CNN (Ren et al., 2015)	JSL (Pratt et al., 2020)	VISUALCOMET (Park et al., 2020)
Backbone	ResNet-50 (He et al., 2016)	RetinaNet (Lin et al., 2017), ResNet-50, LSTM	GPT-2 (Radford et al., 2019)
Pretraining data	ImageNet (Deng et al., 2009)	ImageNet, COCO	OpenWebText (Gokaslan and Cohen, 2019)
Finetuning data	COCO (Lin et al., 2014)	SWiG (Pratt et al., 2020)	VCG (Park et al., 2020)
UNIFORM	"non-person" object labels	top activity and its roles	top-5 *before, after, intent* inferences
HYBRID	Faster R-CNN's object boxes' representations and coordinates	JSL's role boxes' representations and coordinates	VISUALCOMET's embedding for special tokens that signal the start of *before, after, intent* inference

Table 1: Specifications of external vision models and their outputs that we use as features for visual adaptation.

2.2 Outline of Full-Stack Visual Understanding

We first outline types of visual information and associated external models that lead to the full-stack visual understanding. Specifications of these models and features that we use appear in Table 1.

Recognition-level understanding of an image begins with identifying the objects present within it. To this end, we use an object detector that predicts objects present in an image, their labels (e.g., "cup or "chair"), bounding boxes, and the boxes' hidden representations (Fig. 2a).

The next step of recognition-level understanding is capturing relations between objects. A computer vision task that aims to describe such relations is situation recognition (Yatskar et al., 2016). We use a model for *grounded* situation recognition (Fig. 2b; Pratt et al., 2020) that predicts the most prominent activity in an image (e.g., "surfing"), roles of entities engaged in the activity (e.g., "agent" or "tool"), the roles' bounding boxes, and the boxes' hidden representations.

The object detector and situation recognizer focus on recognition-level understanding. But visual understanding also requires attributing mental states such as beliefs, intents, and desires to people participating in an image. In order to achieve this, we use the output of VISUALCOMET (Fig. 2c; Park et al., 2020), another GPT-2-based model that generates commonsense inferences, i.e. events before and after as well as people's intents, given an image and a description of an event present in the image.

2.3 Fusion of Visual and Textual Input

We now describe how we format outputs of the external models (§2.2; Table 1) to augment GPT-2's input with visual information.

We explore two ways of extending the input. The first approach adds a vision model's *textual* output (e.g., object labels such as "food" and "table") before the textual context (e.g., question and answer). Since everything is textual, we can directly embed *each* token using the GPT-2's embedding layer, i.e., by summing the corresponding token, segmentation, and position embeddings.[5] We call this kind

[5]The segment embeddings are (to the best of our knowl-

Dataset	Task	Train	Dev	Expected Visual Understanding
VCR (Zellers et al., 2019a)	visual commonsense reasoning (question answering)	212,923	26,534	higher-order cognitive, commonsense, recognition
E-SNLI-VE (Do et al., 2020) ¬ NEUTRAL	visual-textual entailment	511,112† 341,095	17,133† 13,670	higher-order cognitive, recognition
VQA-E (Li et al., 2018)	visual question answering	181,298	88,488	recognition

Table 2: Specifications of the datasets we use for rationale generation. Do et al. (2020) re-annotate the SNLI-VE dev and test splits due to the high labelling error of the *neutral* class (Vu et al., 2018). Given the remaining errors in the training split, we generate rationales only for entailment and contradiction examples. †Do et al. (2020) report 529,527 and 17,547 training and validation examples, but the available data with explanation is smaller.

of input fusion UNIFORM.

This is the simplest way to extend the input, but it is prone to propagation of errors from external vision models. Therefore, we explore using vision models' *embeddings* for regions-of-interest (RoI) in the image that show relevant entities.[6] For each RoI, we sum its visual embedding (described later) with the three GPT-2's embeddings (token, segment, position) for a special "unk" token and pass the result to the following GPT-2 blocks.[7] After all RoI embeddings, each following token (question, answer, rationale, separator tokens) is embedded similarly, by summing the three GPT-2's embeddings and a visual embedding of the entire image.

We train and evaluate our models with different fusion types and visual features separately to analyze where the improvements come from. We provide details of feature extraction in App. §A.4.

Visual Embeddings We build visual embeddings from bounding boxes' hidden representations (the feature vector prior to the output layer) and boxes' coordinates (the top-left, bottom-right coordinates, and the fraction of image area covered). We project bounding boxes' feature vectors as well as their coordinate vectors to the size of GPT-2 embeddings. We sum projected vectors and apply the layer normalization. We take a different approach for VISUALCOMET embeddings, since they are not related to regions-of-interest of the input image (see §2.2). In this case, as visual embeddings, we use VISUALCOMET embeddings that signal to start generating *before*, *after*, and *intent* inferences, and since there is no representation of the entire image, we do not add it to the question, answer, rationale, separator tokens.

Hypothesis: A dog plays with a tennis ball.
Label: Entailment.
Rationale: A dog *jumping* is how he plays.

Textual premise: A brown dog is *jumping* after a tennis ball.

Figure 3: An illustrative example of the entailment artifact in E-SNLI-VE.

3 Experiments

For all experiments, we visually adapt and fine-tune the original GPT-2 with 117M parameters. We train our models using the language modeling loss computed on rationale tokens.[8]

Tasks and Datasets We consider three tasks and datasets shown in Table 2. Models for VCR and VQA are given a question about an image, and they predict the answer from a candidate list. Models for visual-textual entailment are given an image (that serves as a premise) and a textual hypothesis, and they predict an entailment label between them. The key difference among the three tasks is the level of required visual understanding.

We report here the main observations about how the datasets were collected, while details are in the Appendix §A.2. Foremost, only VCR rationales are

edge) first introduced in Devlin et al. (2019) to separate input elements from different sources in addition to the special separator tokens.

[6]The entire image is also a region-of-interest.

[7]Visual embeddings and object labels do not have a natural sequential order among each other, so we assign position zero to them.

[8]See Table 7 (§A.5) for hyperparameter specifications.

human-written for a given problem instance. Rationales in VQA-E are extracted from image captions relevant for question-answer pairs (Goyal et al., 2017) using a constituency parse tree. To create a dataset for explaining visual-textual entailment, E-SNLI-VE, Do et al. (2020) combined the SNLI-VE dataset (Xie et al., 2019) for *visual-textual* entailment and the E-SNLI dataset (Camburu et al., 2018) for explaining *textual* entailment.

We notice that this methodology introduced a data collection artifact for entailment cases. To illustrate this, consider the example in Figure 3. In visual-textual entailment, the premise is the image. Therefore, there is no reason to expect that a model will build a rationale around a word that occurs in the textual premise it has never seen ("jumping"). We will test whether models struggle with entailment cases.

Human Evaluation For evaluating our models, we follow Camburu et al. (2018) who show that BLEU (Papineni et al., 2002) is not reliable for evaluation of rationale generation, and hence use human evaluation.[9] We believe that other automatic sentence similarity measures are also likely not suitable due to a similar reason; multiple rationales could be plausible, although not necessarily paraphrases of each other (e.g., in Figure 4 both generated and human rationales are plausible, but they are not strict paraphrases).[10] Future work might consider newly emerging *learned* evaluation measures, such as BLEURT (Sellam et al., 2020), that could learn to capture non-trivial semantic similarities between sentences beyond surface overlap.

We use Amazon Mechanical Turk to crowdsource human judgments of generated rationales according to different criteria. Our instructions are provided in the Appendix §A.6. For VCR, we randomly sample one QA pair for each movie in the development split of the dataset, resulting in 244 examples for human evaluation. For VQA and E-SNLI-VE, we randomly sample 250 examples from their development splits.[11] We did not use any of these samples to tune any of our hyperparameters. Each generation was evaluated by 3 crowdworkers. The workers were paid ∼$13 per hour.

Baselines The main objectives of our evaluation are to assess whether (i) proposed visual features help GPT-2 generate rationales that support a given answer or entailment label better (**visual plausibility**), and whether (ii) models that generate more plausible rationales are less likely to mention content that is irrelevant to a given image (**visual fidelity**). As a result, a text-only GPT-2 approach represents a meaningful baseline to compare to.

In light of work exposing predictive data artifacts (e.g., Gururangan et al., 2018), we estimate the effect of artifacts by reporting the difference between visual plausibility of the text-only baseline and plausibility of its rationales assessed without looking at the image (**textual plausibility**). If both are high, then there are problematic lexical cues in the datasets. Finally, we report estimated **plausibility of human rationales** to gauge what has been solved and what is next.[12]

3.1 Visual Plausibility

We ask workers to judge whether a rationale supports a given answer or entailment label in the context of the image (*visual plausibility*). They could select a label from {*yes, weak yes, weak no, no*}. We later merge *weak yes* and *weak no* to *yes* and *no*, respectively. We then calculate the ratio of *yes* labels for each rationale and report the average ratio in a sample.[13]

We compare the text-only GPT-2 with visual adaptations in Table 3. We observe that GPT-2's visual plausibility benefits from some form of visual adaptation for all tasks. The improvement is most visible for VQA-E, followed by VCR, and then E-SNLI-VE (all). We suspect that the minor improvement for E-SNLI-VE is caused by the entailment-data artifact. Thus, we also report the visual plausibility for entailment and contradiction cases separately. The results for contradiction hypotheses follow the trend that is observed for VCR and VQA-E. In contrast, visual adaption does not help rationalization of entailed hypotheses. These

[9] This is based on a low inter-annotator BLEU-score between three human rationales for the same NLI example.

[10] In Table 8 (§A.5), we report automatic captioning measures for the best RATIONALE^{VT} TRANSFORMER for each dataset. These results should be used only for reproducibility and not as measures of rationale plausibility.

[11] The size of evaluation sample is a general problem of generation evaluation, since human evaluation is crucial but expensive. Still, we evaluate ∼2.5 more instances per each of 24 dataset-model combinations than related work (Camburu et al., 2018; Do et al., 2020; Narang et al., 2020); and each instance is judged by 3 workers.

[12] Plausibility of human-written rationales is estimated from our evaluation samples.

[13] We follow the related work (Camburu et al., 2018; Do et al., 2020; Narang et al., 2020) in using yes/no judgments. We introduced weak labels because they help evaluating cases with a slight deviation from a clear-cut judgment.

			VCR	E-SNLI-VE			VQA-E
				Contradiction	Entailment	All	
		Baseline	53.14	46.85	**46.76**	46.80	47.20
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	54.92	58.56	36.45	46.27	54.40
		Situation frames	56.97	59.16	38.13	**47.47**	50.93
		VISCOMET text inferences	**60.93**	53.75	29.26	40.13	53.47
	HYBRID	Object regions	47.40	**60.96**	34.05	46.00	59.07
		Situation roles regions	47.95	51.95	37.65	44.00	**63.33**
		VISCOMET embeddings	59.84	48.95	32.13	39.60	54.93
		Human (estimate)	87.16	80.78	76.98	78.67	66.53

Table 3: Visual plausibility of random samples of generated and human (gold) rationales. Our baseline is text-only GPT-2. The best model is boldfaced.

findings, together with the fact that we have already discarded neutral hypotheses due to the high error rate, raise concern about the E-SNLI-VE dataset. Henceforth, we report entailment and contradiction separately, and focus on contradiction when discussing results. We illustrate rationales produced by RATIONALE^{VT} TRANSFORMER in Figure 4, and provide additional analyses in the Appendix §B.

3.2 Effect of Visual Features

We motivate different visual features with varying levels of visual understanding (§2.2). We reflect on our assumptions about them in light of the visual plausibility results in Table 3. We observe that VISUALCOMET, designed to help attribute mental states, indeed results in the most plausible rationales for reasoning in VCR, which requires a high-order cognitive and commonsense understanding. We propose situation frames to understand relations between objects which in turn can result in better recognition-level understanding. Our results show that situation frames are the second best option for VCR and the best for VQA, which supports our hypothesis. The best option for E-SNLI-VE (contradiction) is HYBRID fusion of objects, although UNIFORM situation fusion is comparable. Moreover, VISUALCOMET is less helpful for E-SNLI-VE compared to objects and situation frames. This suggests that visual-textual entailment in E-SNLI-VE is perhaps focused on recognition-level understanding more than it is anticipated.

One fusion type does not dominate across datasets (see an overview in Table 9 in the Appendix §B). We hypothesize that the source domain of the pretraining dataset of vision models as well as their precision can influence which type of fusion works better. A similar point was recently raised by Singh et al. (2020). Future work might consider carefully combining both fusion types and multiple visual features.

		VCR	E-SNLI-VE (contrad.)	E-SNLI-VE (entail.)	VQA-E
	Baseline	70.63	74.47	**46.28**	68.27
UNIFORM	Object labels	75.14	**77.48**	37.17	64.80
	Situation frames	74.18	72.37	38.61	61.73
	VISCOMET text	73.91	71.17	32.37	69.73
HYBRID	Object regions	69.26	69.97	33.81	68.53
	Situation roles regions	69.81	62.46	38.61	69.47
	VISCOMET embd.	**81.15**	74.77	32.37	**76.53**
	Human (estimate)	90.71	79.58	74.34	64.27

Table 4: Plausibility of random samples of human (gold) and generated rationales assessed **without** looking at the image (*textual plausibility*). The best model is boldfaced.

3.3 Textual Plausibility

It has been shown that powerful pretrained LMs can reason about textual input well in the current benchmarks (e.g., Zellers et al., 2019b; Khashabi et al., 2020). In our case, that would be illustrated with a high plausibility of generated rationales in an evaluation setting where workers are instructed to ignore images (*textual plausibility*).

We report textual plausibility in Table 4. Text-only GPT-2 achieves high textual plausibility (relative to the human estimate) for all tasks (except the entailment part of E-SNLI-VE), demonstrating good reasoning capabilities of GPT-2, when the context image is ignored for plausibility assessment. This result also verifies our hypothesis that generating a textually plausible rationale is easier for models than producing a visually plausible ratio-

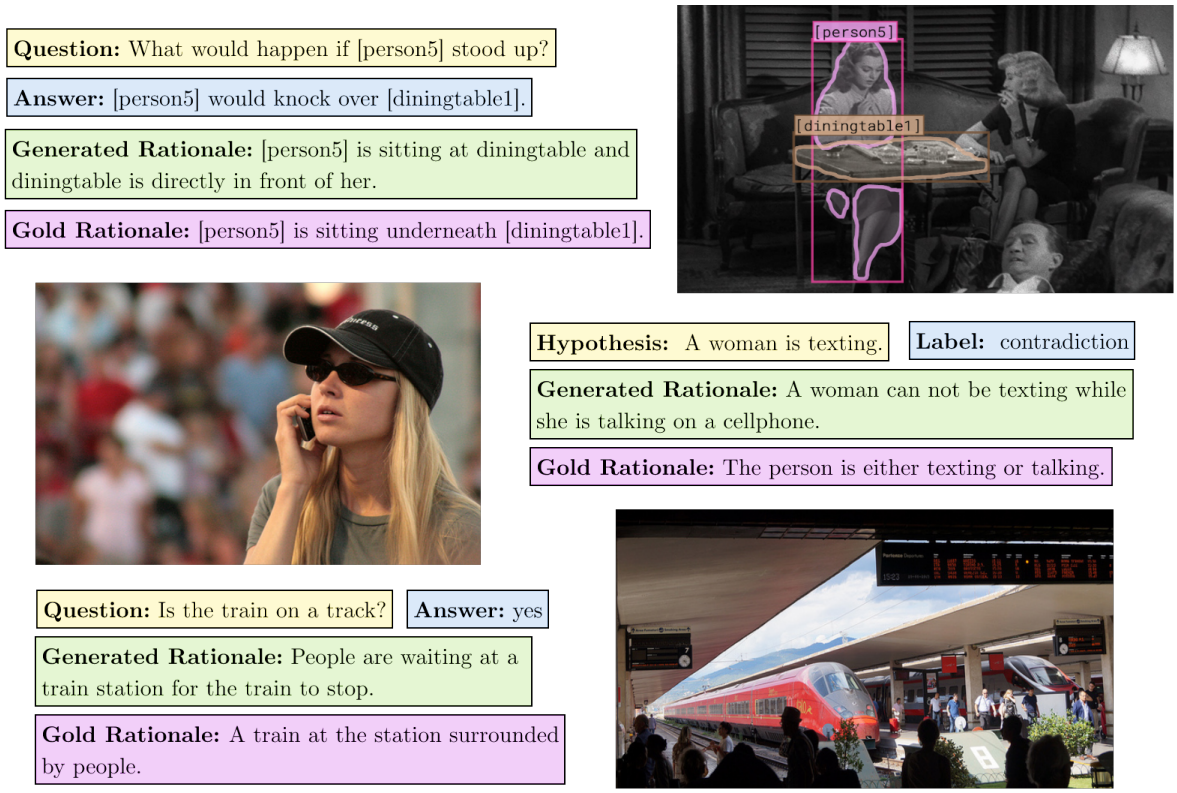


Figure 4: RATIONALE^{VT} TRANSFORMER generations for VCR (top), E-SNLI-VE (contradiction; middle), and VQA-E (bottom). We use the best model variant for each dataset (according to results in Table 3).

nale. For example, GPT-2 can likely produce many statements that contradict "the woman is texting" (see Figure 4), but producing a visually plausible rationale requires conquering another challenge: capturing what is present in the image.

If both textual and visual plausibility of the text-only GPT-2 were high, that would indicate there are some lexical cues in the datasets that allow models to ignore the context image. The decrease in plausibility performance once the image is shown (cf. Tables 3 and 4) confirms that the text-only baseline is not able to generate visually plausible rationales by fitting lexical cues.

We notice another interesting result: textual plausibility of visually adapted models is higher than textual plausibility of the text-only GPT-2. The following three insights together suggest why this could be the case: (i) the gap between textual plausibility of generated and human rationales shows that generating textual plausible rationales is not solved, (ii) visual models produce rationales that are more visually plausible than the text-only baseline, and (iii) visually plausible rationales are usually textually plausible (see examples in Figure 4).

3.4 Plausibility of Human Rationales

The best performing models for VCR and E-SNLI-VE (contradiction) are still notably behind the estimated visual plausibility of human-written rationales (see Table 3). Moreover, plausibility of human rationales is similar when evaluated in the context of the image (visual plausibility) and without the image (text plausibility) because (i) data annotators produce visually plausible rationales since they have accurate visual understanding, and (ii) visually plausible rationales are usually textually plausible. These results show that generating visually plausible rationales for VCR and E-SNLI-VE is still challenging even for our best models.

In contrast, we seem to be closing the gap for VQA-E. In addition, due in part to the automatic extraction of rationales, the human rationales in VQA-E suffer from a notably lower estimate of plausibility.

3.5 Visual Fidelity

We investigate further whether visual plausibility improvements come from better visual understanding. We ask workers to judge if the rationale mentions content unrelated to the image, i.e., anything

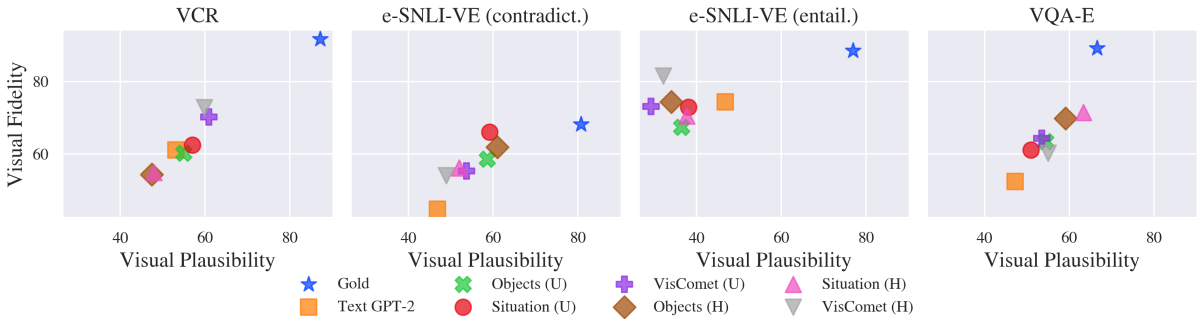


Figure 5: The relation between visual plausibility (§3.1) and visual fidelity (§3.5) . We denote UNIFORM fusion with (U) and HYBRID fusion with (H).

			VCR			E-SNLI-VE (contradict.)			E-SNLI-VE (entail.)			VQA-E		
			Plaus.	Fidelity	r	Plaus.	Fidelity	r	Plaus.	Fidelity	r	Plaus.	Fidelity	r
		Baseline	53.14	61.07	0.68	46.85	44.74	0.53	**46.76**	74.34	0.50	47.20	52.40	0.61
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	54.92	60.25	0.73	58.56	58.56	0.55	36.45	67.39	0.58	54.40	63.47	0.54
		Situation frames	56.97	62.43	0.78	59.16	**66.07**	0.37	38.13	72.90	0.51	50.93	61.07	0.53
		VISCOMET text	**60.93**	70.22	0.62	53.75	55.26	0.45	29.26	73.14	0.49	53.47	64.27	0.66
	HYBRID	Object regions	47.40	54.37	0.67	**60.96**	61.86	0.40	34.05	74.34	0.31	59.07	69.87	0.53
		Situ. roles regions	47.95	54.92	0.66	51.95	56.16	0.45	37.65	70.50	0.59	**63.33**	**71.47**	0.62
		VISCOMET embd.	59.84	**72.81**	0.72	48.95	54.05	0.48	32.13	**81.53**	0.41	54.93	60.27	0.59
		Human (estimate)	87.16	91.67	0.58	80.78	68.17	0.28	76.98	88.49	0.43	66.53	89.20	0.35

Table 5: Visual plausibility (Table 3; §3.1), visual fidelity (§3.5), and Pearson's r that measures linear correlation between the visual plausibility and fidelity. Our baseline is text-only GPT-2. The best model is boldfaced and the second best underlined.

that is not directly visible and is unlikely to be present in the scene in the image. They could select a label from {*yes, weak yes, weak no, no*}. We later merge *weak yes* and *weak no* to *yes* and *no*, respectively. We then calculate the ratio of *no* labels for each rationale. The final **fidelity** score is the average ratio in a sample.[14]

Figure 5 illustrates the relation between visual fidelity and plausibility. For each dataset (except the entailment part of E-SNLI-VE), we observe that visual plausibility is larger as visual fidelity increases. We verify this with Pearson's r and show moderate linear correlation in Table 5. This shows that models that generate more visually plausible rationales are less likely to mention content that is irrelevant to a given image.

4 Related Work

Rationale Generation Applications of rationale generation (see §1) can be categorized as text-only, vision-only, or visual-textual. Our work belongs to the final category, where we are the first to try to generate rationales for VCR (Zellers et al., 2019a). The bottom-up top-down attention (BUTD) model (Anderson et al., 2018) has been proposed to incorporate rationales with visual features for VQA-E and E-SNLI-VE (Li et al., 2018; Do et al., 2020). Compared to BUTD, we use a pretrained decoder and propose a wider range of visual features to tackle comprehensive image understanding.

Conditional Text Generation Pretrained LMs have played a pivotal role in open-text generation and conditional text generation. For the latter, some studies trained a LM from scratch conditioned on metadata (Zellers et al., 2019c) or desired attributes of text (Keskar et al., 2019), while some fine-tuned an already pretrained LM on commonsense knowledge (Bhagavatula et al., 2020) or text attributes (Ziegler et al., 2019a). Our work belongs to the latter group with focus on conditioning on comprehensive image understanding.

Visual-Textual Language Models There is a surge of work that proposes visual-textual pretraining of LMs by predicting masked image regions and tokens (Tan and Bansal, 2019; Lu et al., 2019;

[14]We also study assessing fidelity from phrases that are extracted from a rationale (see Appendix B).

Chen et al., 2019, to name a few). We construct input elements of our models following the VL-BERT architecture (Su et al., 2020). Despite their success, these models are not suitable for generation due to pretraining with the masked LM objective. Zhou et al. (2020) aim to address that, but they pretrain their decoder from scratch using 3M images with weakly-associated captions (Sharma et al., 2018). This makes their decoder arguably less powerful compared to LMs that are pretrained with remarkably more (diverse) data such as GPT-2. Ziegler et al. (2019b) augment GPT-2 with a feature vector for the entire image and evaluate this model on image paragraph captioning. Some work extend pretrained LM to learn video representations from sequences of visual features and words, and show improvements in video captioning (Sun et al., 2019a,b). Our work is based on fine-tuning GPT-2 with features that come from visual object recognition, grounded semantic frames, and visual commonsense graphs. The latter two features have not been explored yet in this line of work.

5 Discussion and Future Directions

Rationale Definition The term *interpretability* is used to refer to multiple concepts. Due to this, criteria for explanation evaluation depend on one's definition of interpretability (Lipton, 2016; Doshi-Velez and Kim, 2017; Jacovi and Goldberg, 2020). In order to avoid problems arising from ambiguity, we reflect on our definition. We follow Ehsan et al. (2018) who define AI *rationalization* as a process of generating rationales of a model's behavior as if a human had performed the behavior.

Jointly Predicting and Rationalizing We narrow our focus on improving generation models and assume gold labels for the end-task. Future work can extend our model to an end-to-end (Narang et al., 2020) or a pipeline model (Camburu et al., 2018; Rajani et al., 2019; Jain et al., 2020) for producing both predictions and natural language rationales. We expect that the *explain-then-predict* setting (Camburu et al., 2018) is especially relevant for rationalization of commonsense reasoning. In this case, relevant information is not in the input, but inferred from it, which makes extractive explanatory methods based on highlighting parts of the input unsuitable. A rationale generation model brings relevant information to the surface, which can be passed to a prediction model. This makes rationales intrinsic to the model, and tells the user what the prediction should be based on. Kumar and Talukdar (2020) highlight that this approach resembles post-hoc methods with the label and rationale being produced jointly (the *end-to-end predict-then-explain* setting). Thus, all but the pipeline predict-then-explain approach are suitable extensions of our models. A promising line of work trains end-to-end models for joint rationalization and prediction from weak supervision (Latcinnik and Berant, 2020; Shwartz et al., 2020), i.e., without human-written rationales.

Limitations Natural language rationales are easily understood by lay users who consequently feel more convinced and willing to use the model (Miller, 2019; Ribera and Lapedriza, 2019). Their limitation is that they can be used to persuade users that the model is reliable when it is not (Bansal et al., 2020)—an ethical issue raised by Herman (2017). This relates to the pipeline predict-then-explain setting, where a predictor model and a post-hoc explainer model are completely independent. However, there are other settings where generated rationales are intrinsic to the model by design (end-to-end predict-then-explain, both end-to-end and pipeline explain-then-predict). As such, generated rationales are more associated with the reasoning process of the model. We recommend that future work develops rationale generation in these settings, and aims for *sufficiently faithful* models as recommended by Jacovi and Goldberg (2020), Wiegreffe and Pinter (2019).

6 Conclusions

We present RATIONALE^{VT} TRANSFORMER, an integration of a pretrained text generator with semantic and pragmatic visual features. These features improve visual plausibility and fidelity of generated rationales for visual commonsense reasoning, visual-textual entailment, and visual question answering. This represents progress in tackling important, but still relatively unexplored research direction; rationalization of complex reasoning for which explanatory approaches based solely on highlighting parts of the input are not suitable.

Acknowledgments

The authors thank Sarah Pratt for her assistance with the grounded situation recognizer, Amandalynne Paullada, members of the AllenNLP team, and anonymous reviewers for helpful feedback.

This research was supported in part by NSF (IIS1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), and gifts from Allen Institute for Artificial Intelligence.

References

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 *CVPR*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *ACL*.

Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv:2006.14779.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *ICLR*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *EMNLP*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *NeurIPS*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning UNiversal Image-TExt Representations.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense Knowledge Mining from Pretrained Models. In *EMNLP*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-SNLI-VE-2.0: Corrected Visual-Textual Entailment with Natural Language Explanations. arXiv:2004.03744.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *AIES*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *IUI*.

Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *EMNLP (Findings)*.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus.

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. In *CVPR*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding Visual Explanations. In *ECCV*.

Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. In *Symposium on Interpretable Machine Learning @ NeurIPS*.

Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? In *ACL*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *ACL*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv:1909.05858.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, P. Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. In *EMNLP (Findings)*.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *ECCV*.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *ACL*.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining Question Answering Models through Text Generation. arXiv:2004.05569.

Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *ECCV*.

Tsung-Yi Lin, Priyal Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

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. In *ECCV*.

Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. In *WHI*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process.*, 73:1–15.

Sharan Narang, Colin Raffel, Katherine J. Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. arXiv:2004.14546.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visual Commonsense Graphs: Reasoning about the Dynamic Context of a Still Image. In *ECCV*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *EMNLP*.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded Situation Recognition. In *ECCV*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Mireia Ribera and Àgata Lapedriza. 2019. Can We Do Better Explanations? A Proposal of User-Centered Explainable AI. In *ACM IUI Workshop*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. 2016. Movie Description. *International Journal of Computer Vision*, 123:94–120.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do Massively Pretrained Language Models Make Better Storytellers? In *CoNLL*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

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*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *EMNLP-IJCNLP*, Hong Kong, China.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *EMNLP*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR Workshop Track*.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are We Pretraining It Right? Digging Deeper into Visio-Linguistic Pretraining. arXiv:2004.08744.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. arXiv:1906.05743.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*.

Hao Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NeurIPS*.

Carl Vondrick, Deniz Oktay, H. Pirsiavash, and A. Torralba. 2016. Predicting Motivations of Actions by Leveraging Text. In *CVPR*.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. 2018. Grounded Textual Entailment. In *COLING*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *EMNLP-IJCNLP*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv:1910.03771.

Jialin Wu and Raymond Mooney. 2019. Faithful Multimodal Explanation for Visual Question Answering. In *BlackboxNLP @ ACL*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv:1901.06706.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *CVPR*.

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. *TACL*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019c. Defending Against Neural Fake News. In *NeurIPS*.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2017. Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, 126:1084–1102.

Luowei Zhou, Hamid Palangi, Lefei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019a. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.

Zachary M. Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M. Rush. 2019b. Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv:1908.06938.

A Experimental Setup

A.1 Deatils of GPT-2

Input to GPT-2 is text that is split into subtokens[15] (Sennrich et al., 2016). Each subtoken embedding is added to a so-called positional embedding that signals the order of the subtokens in the sequence to the transformer blocks. The GPT-2's pretraining corpus is OpenWebText corpus (Gokaslan and Cohen, 2019) which consists of 8 million Web documents extracted from URLs shared on Reddit. Pretraining on this corpus has caused degenerate and biased behaviour of GPT-2 (Sheng et al., 2019; Wallace et al., 2019; Gehman et al., 2020, among others). Our models likely have the same issues since they are built on GPT-2.

A.2 Details of Datasets with Human Rationales

We obtain the data from the following links:

- https://visualcommonsense.com/download/
- https://github.com/virginie-do/e-SNLI-VE
- https://github.com/liqing-ustc/VQA-E

Answers in VCR are full sentences, and in VQA single words or short phrases. All annotations in VCR are authored by crowdworkers in a single data collection phase. Rationales in VQA-E are extracted from relevant image captions for question-answer pairs in VQA v2 (Goyal et al., 2017) using a constituency parse tree. The overall quality of VQA-E rationales is 4.23/5.0 from human perspective.

The E-SNLI-VE dataset is constructed from a series of additions and changes of the SNLI dataset for *textual* entailment (Bowman et al., 2015). The SNLI dataset is collected by using captions in Flickr30k (Young et al., 2014) as textual premises and crowdsourcing hypotheses.[16] The E-SNLI dataset (Camburu et al., 2018) adds crowdsourced explanations to SNLI. The SNLI-VE dataset (Xie et al., 2019) for *visual-textual* entailment is constructed from SNLI by replacing textual premises with corresponding Flickr30k images. Finally, Do et al. (2020) combine SNLI-VE and E-SNLI to produce a dataset for explaining *visual-textual* entailment. They reannotate the dev and test splits due to the high labelling error of the *neutral* class in SNLI-VE that is reported by Vu et al. (2018).

[15] Also known as wordpieces or subwords.

[16] Captions tend to be literal scene descriptions.

A.3 Details of External Vision Models

In Table 6, we report sources of images that were used to train external vision models and images in the end-task datasets.

A.4 Details of Input Elements

Object Detector For UNIFORM fusion, we use labels for objects other that people because *person* label occurs in every example for VCR. We use only a single instance of a certain object label, because repeating the same label does not give new information to the model. The maximum number of subtokens for merged object labels is determined from merging all object labels, tokenizing them to subtokens, and set the maximum to the length at the ninety-ninth percentile calculated from the VCR training set. For HYBRID fusion, we use hidden representation of all objects because they differ for different detections of objects with the same label. These representations come from the feature vector prior to the output layer of the detection model. The maximum number of objects is set to the object number at the 99th percentile calculated from the VCR training set.

Situation Recognizer For UNIFORM fusion, we consider only the best verb because the top verbs are often semantically similar (e.g. *eating* and *dining*; see Figure 13 in Pratt et al. (2020) for more examples). We define a structured format for the output of a situation recognizer. For example, the situation predicted from the first image in Figure 4, is assigned the following structure "`<|b_situ|> <|b_verb|> dining <|e_verb|> <|b_agent|> people <|e_agent|> <|b_place|> restaurant <|e_place|> <|e_situ|>`". We set the maximum situation length to the length at the ninety-ninth percentile calculated from the VCR training set.

VISUALCOMET The input to VISUALCOMET is an image, question, and answer for VCR and VQA-E; only image for E-SNLI-VE. Unlike situation frames, top-k VISUALCOMET inferences are diverse. We merge top-5 before, after, and intent inferences. We calculate the length of merged inferences in number of subtokens and set the maximum VISUALCOMET length to the length at the ninety-ninth percentile calculated from the VCR training set.

Dataset	Image Source
COCO	Flickr
E-SNLI-VE	Flickr (SNLI; Bowman et al., 2015)
ImageNet	different search engines
SWiG	Google Search (imSitu; Yatskar et al., 2016)
VCG, VCR	movie clips (Rohrbach et al., 2016), Fandango†
VQA-E	Flickr (COCO)

Table 6: Image sources. † https://www.youtube.com/user/movieclips

A.5 Training Details

We use the original GPT-2 version with 117M parameters. It consists of 12 layers, 12 heads for each layer, and the size of a model dimension set to 768. We report other hyperaparametes in Table 7. All of them are manually chosen due to the reliance on human evaluation. In Table 8, for reproducibility, we report captioning measures of the best RATIONALE^{VT} TRANSFORMER variants. Our implementation uses the HuggingFace `transformers` library (Wolf et al., 2019).[17]

A.6 Crowdsourcing Human Evaluation

We perform human evaluation of the generated rationales through crowdsourcing on the Amazon Mechanical Turk platform. Here, we provide the full set of **Guidelines** provided to workers:

- First, you will be shown a (i) Question, (ii) an Answer (presumed-correct), and (iii) a Rationale. You'll have to judge if the rationale supports the answer.
- Next, you will be shown the same question, answer, rationale, and an associated image. You'll have to judge if the rationale supports the answer, in the context of the given image.
- You'll judge the grammaticality of the rationale. Please ignore the absence of periods, punctuation and case.
- Next, you'll have to judge if the rationale mentions persons, objects, locations or actions unrelated to the image—i.e. things that are not directly visible and are unlikely to be present to the scene in the image.
- Finally, you'll pick the NOUNS, NOUN PHRASES and VERBS from the rationale that are unrelated to the image.

We also provide the following additional **tips**:

- Please ignore minor grammatical errors—e.g. case sensitivity, missing periods etc.
- Please ignore gender mismatch—e.g. if the image shows a male, but the rationale mentions female.
- Please ignore inconsistencies between person and object detections in the QUESTION / ANSWER and those in the image—e.g. if a pile of papers is labeled as a laptop in the image. Do not ignore such inconsistencies for the rationale.
- When judging the rationale, think about whether it is plausible.
- If the rationale just repeats an answer, it is not considered as a valid justification for the answer.

B Additional Results

We provide the following additional results that complement the discussion in Section 3:

- a comparison between UNIFORM and HYBRID fusion in Table 9,
- an investigation of fine-grained visual fidelity in Table 11,
- additional analysis of RATIONALE^{VT} TRANSFORMER to support future developments.

Fine-Grained Visual Fidelity At the time of running human evaluation, we did not know whether judging visual fidelity is a hard task for workers. To help them focus on relevant parts of a given rationale and to make their judgments more comparable, we give workers a list of nouns, noun phrases, as well as verb phrases with negation, without adjuncts. We ask them to pick phrases that are unrelated to the image. For each rationale, we calculate the ratio of nouns that are relevant over the number of all nouns. We call this **"entity fidelity"** because extracted nouns are mostly concrete (opposed to abstract). Similarly, from noun phrases

[17] https://github.com/huggingface/transformers

Computing Infrastructure	Quadro RTX 8000 GPU
Model implementation	https://github.com/allenai/visual-reasoning-rationalization

Hyperparameter	**Assignment**
number of epochs	5
batch size	32
learning rate	5e-5
max question length	19
max answer length	23
max rationale length	50
max merged object labels length	30
max situation's structured description length	17
max VISUALCOMET merged text inferences length	148
max input length	93, 98, 123, 102, 112, 241
max objects embeddings number	28
max situation role embeddings number	7
dimension of object and situation role embeddings	2048
decoding	greedy

Table 7: Hyperparameters for RATIONALE^{VT} TRANSFORMER. The length is calculated in number of subtokens including special separator tokens for a given input type (e.g., begin and end separator tokens for a question). We calculate the maximum input length by summing the maximum lengths of input elements for each model separately. A training epoch for models with shorter maximum input length ∼30 minutes and for the model with the longest input ∼2H.

judgments, we calculate **"entity detail fidelity"**, and from verb phrases **"action fidelity"**. Results in Table 11 show close relation between the overall fidelity judgment and entity fidelity. Furthermore, for the case where the top two models have close fidelity (VISUALCOMET models for VCR), the fine-grained analysis shows where the difference comes from (in this case from action fidelity). Despite possible advantages of fine-grained fidelity, we observe that is less correlated with plausibility compared to the overall fidelity.

Additional Analysis We ask workers to judge grammatically of rationales. We instruct them to ignore some mistakes such as absence of periods and mismatched gender (see §A.6). Table 10 shows that the ratio of grammatical rationales is high for all model variants.

We measure similarity of generated and gold rationales to question (hypothesis) and answer. Results in Tables 12–13 show that generated rationales repeat the question (hypothesis) more than human rationales. We also observe that gold rationales in E-SNLI-VE are notably more repetitive than human rationales in other datasets.

In Figure 6, we show that the length of generated rationales is similar for plausible and implausible rationales, with the exception of E-SNLI-VE for which implausible rationales tend to be longer than plausible. We show that plausible rationales tend to rationalize slightly shorter textual context in VCR (question and answer) and E-SNLI-VE (hypothesis).

Finally, in Figure 7, we show that there is more variation across {*yes, weak yes, weak no, no*} labels for our models than for human rationales.

In summary, future developments should improve generations such that they repeat textual context less, handle long textual contexts, and produce generations that humans will find more plausible with high certainty.

	VCR VisualComet UNIFORM	E-SNLI-VE (contradict.) Situation Frame UNIFORM	E-SNLI-VE (entail.) Text-Only GPT-2	VQA-E Situation Frame HYBRID
BLEU-1	20.98	32.18	33.09	36.64
BLEU-2	12.15	20.35	22.55	22.48
BLEU-3	7.52	13.90	15.78	14.33
BLEU-4	4.98	9.50	11.37	9.47
METEOR	12.21	19.29	20.09	19.33
ROUGE-L	23.08	27.25	27.74	35.31
CIDEr	37.22	71.37	73.35	94.89

Table 8: We report standard automatic captioning measure for the best RATIONALE^{VT} TRANSFORMER for each dataset (according to results in Table 3; §3.1), except for E-SNLI-VE for which we use UNIFORM fusion of situation frames instead of object labels, because they have comparable plausibility, but situation frames result in better fidelity. We use the **entire development sets** for this evaluation.

		UNIFORM	HYBRID
VCR	Objects	7.51	-
	Situation frame	9.02	-
	VisualComet	1.09	-
E-SNLI-VE (contradiction)	Objects	-	2.40
	Situation frame	7.21	-
	VisualComet	4.80	-
E-SNLI-VE (entailment)	Objects	2.40	-
	Situation frame	0.48	-
	VisualComet	-	2.88
VQA-E	Objects	-	4.67
	Situation frame	-	12.40
	VisualComet	-	1.47

Table 9: Comparison of HYBRID and UNIFORM fusion visual plausibility results that are reported in Table 3 (§3.1). The number shows the difference in visual plausibility between the fusion type in a given column and the other column. The number is placed in the column with better fusion type for a given task and feature.

			VCR	E-SNLI-VE (contradict.)	E-SNLI-VE (entail.)	VQA-E
		Baseline	92.49	94.29	<u>86.81</u>	96.53
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	92.62	**96.10**	**87.05**	97.20
		Situation frames	92.62	94.89	86.33	95.07
		VisComet text inferences	<u>94.54</u>	94.89	82.97	97.73
	HYBRID	Object regions	93.03	<u>95.50</u>	84.65	96.67
		Situation roles regions	90.03	94.59	86.33	<u>96.67</u>
		VisComet embeddings	**96.31**	95.20	84.65	**98.13**
		Human (estimate)	95.22	87.69	86.33	94.67

Table 10: The ratio of grammatically correct rationales (according to human evaluation) in random samples of gold and generated rationales. The most grammatical model is **boldfaced** and the model that produces the most plausible rationales (according to the evaluation in Table 3; §3.1) is <u>underlined</u>.

		VCR	Fidelity	Entity Fidelity	Entity Detail Fidelity	Action Fidelity
		Baseline	61.07	75.32	65.88	61.36
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	60.25	77.45	69.29	66.67
		Situation frames	62.43	77.70	66.49	61.54
		VISUALCOMET text inferences	70.22	**79.91**	**75.74**	69.63
	HYBRID	Object regions	54.37	73.86	58.50	59.36
		Situation frames	54.92	73.88	62.22	60.80
		VISUALCOMET embeddings	**72.81**	79.89	75.25	**74.41**
		Human (estimate)	91.67	94.79	93.60	91.58

		E-SNLI-VE (contradiction)	Fidelity	Entity Fidelity	Entity Detail Fidelity	Action Fidelity
		Baseline	44.74	73.21	65.05	52.19
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	58.56	78.23	68.27	70.03
		Situation frames	**66.07**	**82.52**	71.72	71.11
		VISUALCOMET text inferences	55.26	79.24	72.00	**73.65**
	HYBRID	Object regions	61.86	82.08	73.33	65.56
		Situation frames	56.16	79.87	68.78	64.29
		VISUALCOMET embeddings	54.05	77.37	**79.00**	62.91
		Human (estimate)	68.17	83.07	80.85	72.71

		E-SNLI-VE (entailment)	Fidelity	Entity Fidelity	Entity Detail Fidelity	Action Fidelity
		Baseline	74.34	82.99	93.08	94.59
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	67.39	84.31	93.46	95.59
		Situation frames	72.90	84.69	92.77	95.05
		VISUALCOMET text inferences	73.14	82.66	94.77	99.55
	HYBRID	Object regions	74.34	**86.28**	**95.00**	96.75
		Situation frames	70.50	84.77	92.78	95.83
		VISUALCOMET embeddings	**81.53**	85.60	94.65	**99.10**
		Human (estimate)	88.49	94.81	90.11	93.50

		VQA-E	Fidelity	Entity Fidelity	Entity Detail Fidelity	Action Fidelity
		Baseline	52.40	74.44	74.24	67.20
RATIONALE^{VT} TRANSFORMERS	UNIFORM	Object labels	63.47	83.84	**84.34**	78.14
		Situation frames	61.07	81.82	78.52	73.85
		VISUALCOMET text inferences	64.27	77.71	71.49	66.18
	HYBRID	Object regions	69.87	86.98	79.08	**84.75**
		Situation frames	**71.47**	**89.04**	78.75	80.87
		VISUALCOMET embeddings	60.27	77.40	76.72	64.58
		Human (estimate)	89.20	94.92	94.21	92.67

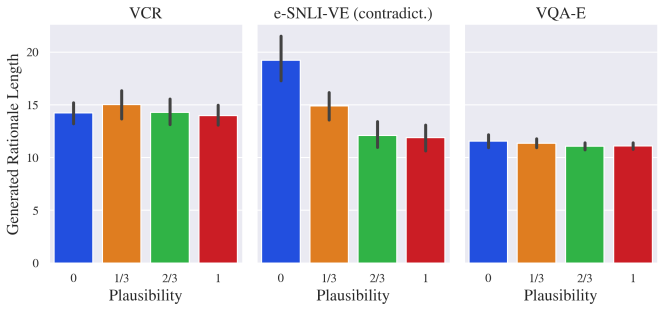
Table 11: RATIONALE^{VT} TRANSFORMER visual fidelity with respect to extracted nouns (entity fidelity), noun phrases (entity detail fidelity), and verbs phrases (action fidelity).

		VCR	E-SNLI-VE (contradict.)	E-SNLI-VE (entail.)	VQA-E
Question or Hypothesis	BLEU-1	20.25	32.57	37.71	13.49
	BLEU-2	9.78	23.29	32.93	5.69
	BLEU-3	6.48	15.92	29.59	2.46
	BLEU-4	4.58	10.94	26.83	0.97
	METEOR	14.05	30.25	38.47	13.13
	ROUGE-L	19.64	37.45	42.93	15.44
	Content Word Overlap	23.22	53.81	48.11	18.96
Answer	BLEU-1	27.67			4.96
	BLEU-2	19.07			1.50
	BLEU-3	12.97			0.49
	BLEU-4	9.83			0.00
	METEOR	20.22			13.38
	ROUGE-L	31.62			10.07
	Content Word Overlap	30.09			11.66

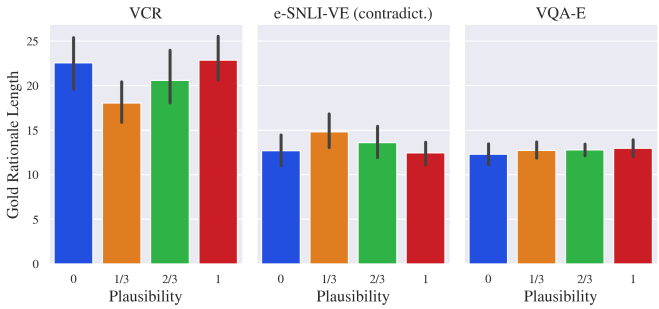
Table 12: Similarity between question and **generated** rationale (upper part) and similarity between answer and **generated** rationale (lower part). For each dataset, we use rationales from the best RATIONALE^{VT} TRANSFORMER (according to results in Table 3; §3.1), except for E-SNLI-VE for which we use UNIFORM fusion of situation frames instead of object labels, because they have comparable plausibility, but situation frames result in better fidelity. We use this model for both E-SNLI-VE parts. We use the same samples of data as in the main evaluation.

		VCR	E-SNLI-VE (contradict.)	E-SNLI-VE (entail.)	VQA-E
Question or Hypothesis	BLEU-1	11.66	31.01	33.14	10.10
	BLEU-2	5.20	19.76	24.09	3.45
	BLEU-3	3.37	12.91	18.39	1.27
	BLEU-4	2.36	7.99	14.15	0.56
	METEOR	11.49	24.69	27.19	11.44
	ROUGE-L	13.88	37.33	41.02	12.07
	Content Word Overlap	13.68	47.70	43.95	14.38
Answer	BLEU-1	15.29			4.00
	BLEU-2	8.13			0.69
	BLEU-3	4.16			0.00
	BLEU-4	2.29			0.00
	METEOR	16.35			11.16
	ROUGE-L	19.87			8.47
	Content Word Overlap	18.01			9.26

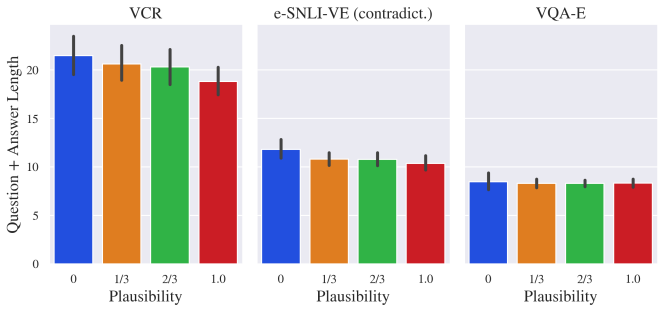
Table 13: Similarity between question and **gold** rationale (upper part) and similarity between answer and **gold** rationale (lower part). We use the same samples of data as in the main evaluation.



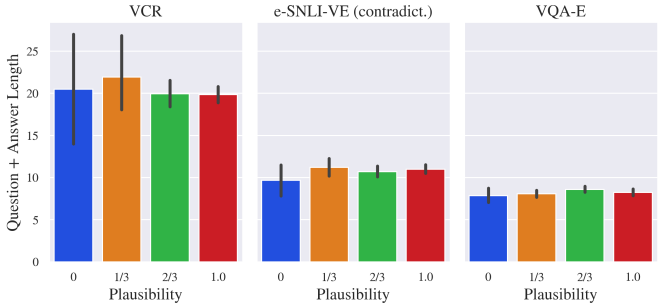
(a) The mean and variance of the **length of generated rationale** with respect to visual plausibility of **generated rationales**. The length of generated rationales is similar for plausible and implausible rationales, with exception of E-SNLI-VE for which implausible rationales tend to be longer.



(b) The mean and variance of the **length of gold rationale** with respect to visual plausibility of **generated rationales**. Rationale generation is not affected by gold rationale length.

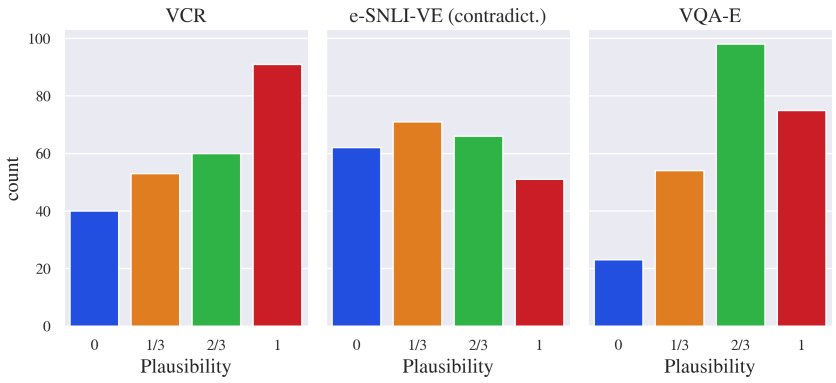


(c) The mean and variance of the **merged question and answer** or just **hypothesis** with respect to visual plausibility of **generated rationales**. Plausible rationale tend to rationalize slightly shorter textual context in VCR and E-SNLI-VE.

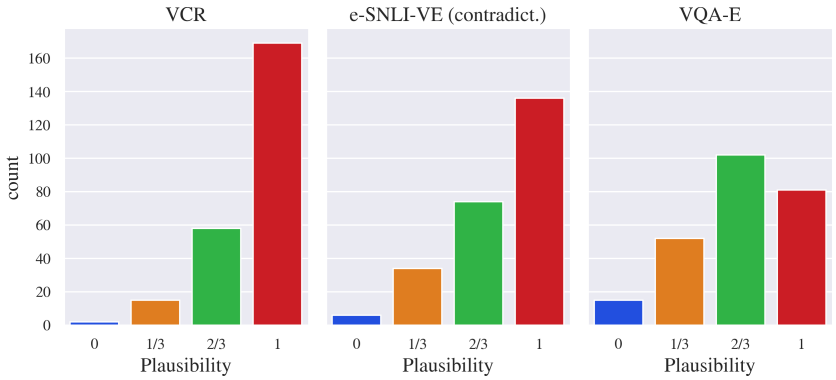


(d) The mean and variance of the **merged question and answer** or just **hypothesis** with respect to visual plausibility of **gold rationales**. The small number of implausible VCR examples also tend to rationalize slightly longer textual contexts, in contrast to E-SNLI-VE.

Figure 6: Analysis of plausibility of rationales with respect to input length. Plausibility value is 0 for unanimously implausible, 1 for unanimously plausible, 1/3 for majority vote for implausible, and 2/3 for majority vote for plausible. For each dataset in 6a–6c, we use rationales from the best RATIONALE^{VT} TRANSFORMER (according to results in Table 3; §3.1), except for E-SNLI-VE for which we use UNIFORM fusion of situation frames instead of object labels, because they have comparable plausibility, but situation frames result in better fidelity. We use this model for both E-SNLI-VE parts. We use the same samples of data as in the main evaluation.



(a) Plausibility variation for **generated** rationales. For each dataset, we use rationales from the best RATIONALE^{VT} TRANSFORMER (according to results in Tables 3; §3.1), except for E-SNLI-VE for which we use UNIFORM fusion of situation frames instead of object labels, because they have comparable plausibility, but situation frames result in better fidelity.



(b) There is less variation for **gold** rationales.

Figure 7: Analysis of variation of plausibility judgments. Plausibility value is 0 for unanimously implausible, 1 for unanimously plausible, 1/3 for majority vote for implausible, and 2/3 for majority vote for plausible. We use the same samples of data as in the main evaluation.