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).

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.

[15] Also known as wordpieces or subwords.

[16] Captions tend to be literal scene descriptions.

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$^{\text{VT}}$ 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$^{\text{VT}}$ 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$^{\text{VT}}$ 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	E-SNLI-VE (contradict.)	E-SNLI-VE (entail.)	VQA-E
	VISUALCOMET UNIFORM	Situation Frame UNIFORM	Text-Only GPT-2	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	86.81	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	94.54	94.89	82.97	97.73
	HYBRID	Object regions	93.03	95.50	84.65	96.67
		Situation roles regions	90.03	94.59	86.33	96.67
		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 underlined.

		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

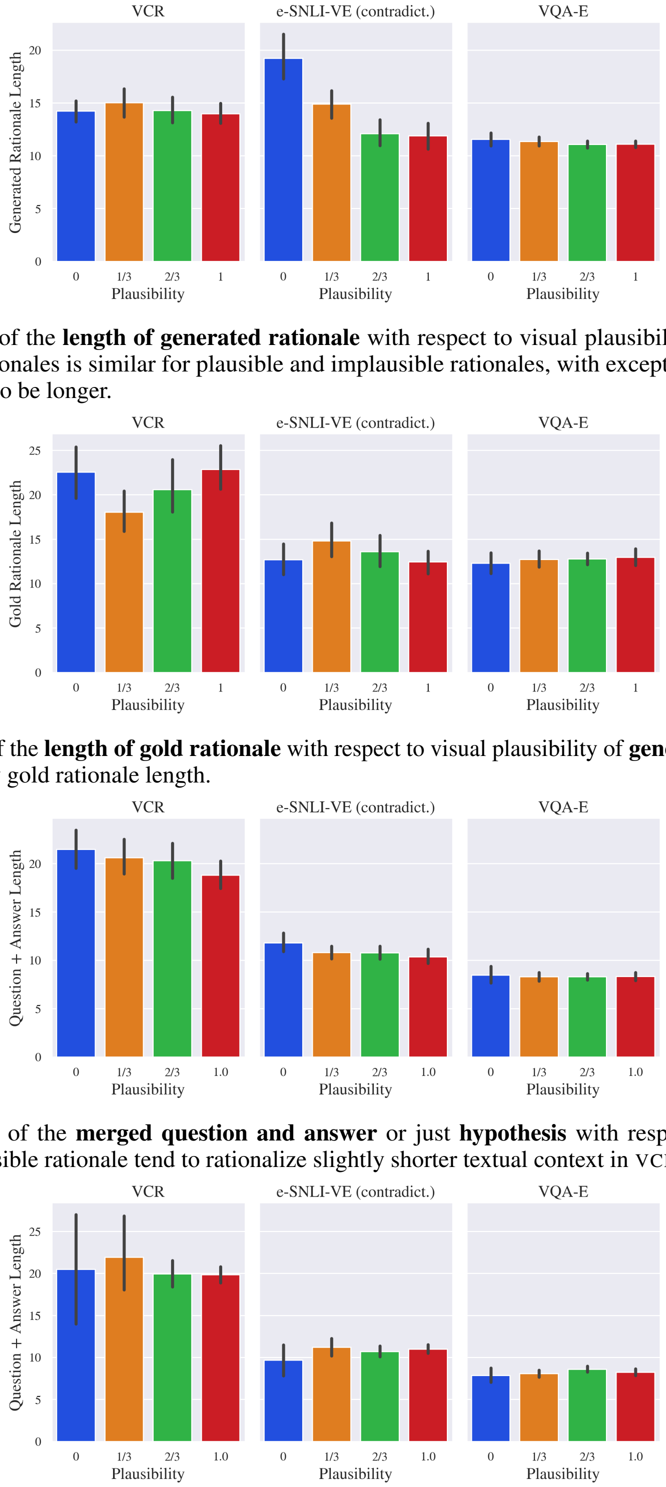
Table 11: Rationale^{VT} Transformer visual fidelity with respect to extracted nouns (entity fidelity), noun phrases (entity detail fidelity), and verbs phrases (action fidelity).

		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.



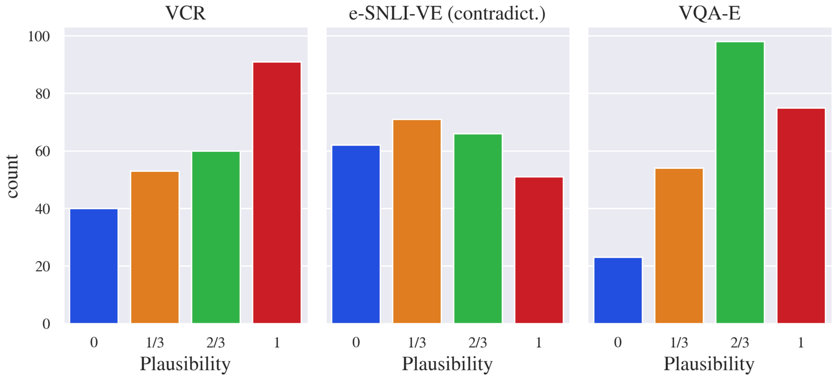
(a) The mean and variance of the **length of generated rationale** with respect to visual plausibility of **generated rationales**. The length of generated rationales is similar for plausible and implausible rationales, with exception of E-SNLI-VE for which implausible rationales tend to be longer.

(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.

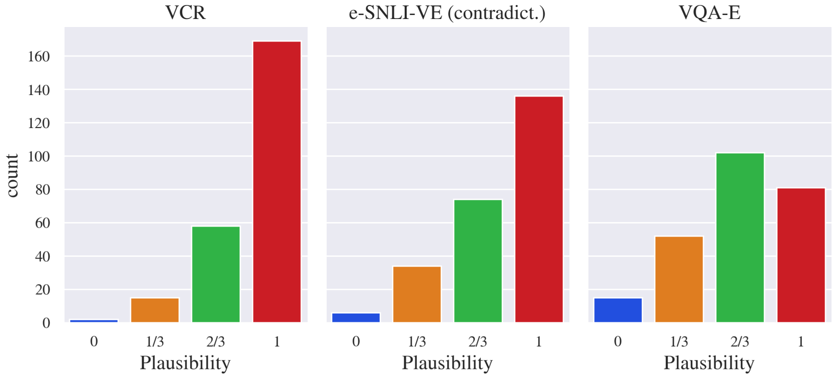
(c) The mean and variance of the **merged question and answer** or just **hypothesis** with respect to visual plausibility of **generated rationales**. Plausible rationale tend to rationalize slightly shorter textual context in VCR and E-SNLI-VE.

(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$^{\text{VT}}$ 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.