On Advances in Text Generation from Images Beyond Captioning: A Case Study in Self-Rationalization

Shruti Palaskar[†]   **Akshita Bhagia**[‡]   **Yonatan Bisk**[†]

Florian Metze[†]   **Alan W Black**[†]   **Ana Marasović**[‡]

[†]Carnegie Mellon University, Pittsburgh, PA, USA
[‡]Allen Institute for AI, Seattle, WA, USA
{spalaska,ybisk,fmetze,awb}@cs.cmu.edu, {akshitab,anam}@allenai.org

Abstract

Integrating vision and language has gained notable attention following the success of pretrained language models. Despite that, a fraction of emerging multimodal models is suitable for *text generation* conditioned on images. This minority is typically developed and evaluated for image captioning, a text generation task conditioned solely on images with the goal to describe what is explicitly visible in an image. In this paper, we take a step back and ask: How do these models work for more complex generative tasks, conditioned on both text and images? Are models based on joint multimodal pretraining, visually adapted pretrained language models, or models that combine these two approaches, more promising for such tasks? We address these questions in the context of *self-rationalization* (jointly generating task labels/answers and free-text explanations) of three tasks: (i) visual question answering in VQA-X, (ii) visual commonsense reasoning in VCR, and (iii) visual-textual entailment in E-SNLI-VE. We show that recent advances in each modality, CLIP image representations and scaling of language models, do not consistently improve multimodal self-rationalization of tasks with multimodal inputs. We also observe that no model type works universally the best across tasks/datasets and finetuning data sizes. Our findings call for a backbone modelling approach that can be built on to advance text generation from images and text beyond image captioning.

1 Introduction

The pretrain-finetune paradigm has proved to be a seminal moment in NLP. Inspired by this success, there has been an explosion of interest to similarly pretrain models to learn *joint* representations of text and other modalities, often images (Su et al., 2020; Lu et al., 2019; Chen et al., 2019; Li et al., 2020a; Tan and Bansal, 2019; Li et al., 2020b). To make joint models better fit for text generation conditioned on images, captioning has been included as one of the pretraining tasks (Zhou et al., 2020; Gupta et al., 2022; Wang et al., 2022). Captioning is also the only generative task used to evaluate and compare joint models, for which minor improvements are reported relative to classification tasks (Cho et al., 2021). Moreover, captioning is conditioned only on an image. This leads us to ask: Do recent advances transfer to more complex generative tasks? How about generation conditioned on both images and text? Another line of work skips joint pretraining and directly modifies and finetunes a pretrained language model apt for generation (e.g., GPT-2; Radford et al., 2019) on multimodal datasets (Park et al., 2020; Sollami and Jain, 2021; Eichenberg et al., 2021). This approach has distinct benefits and downsides compared to models based on joint pretraining (see Table 1), but to the best of our knowledge, these two families of models are rarely compared. This leads us to other questions: Given a new generative task, is one of the two approaches more promising? Do models that combine both work the best?

We study these questions through the lens of a newly emerging and challenging task of *self-rationalization* (Wiegreffe et al., 2021): jointly generating both the task label/answer and a free-text explanation for the prediction. This task is a good testbed for our analysis. Foremost, there is a growing interest in self-rationalizing tasks with multimodal inputs such as answering questions about images in the VQA dataset (Park et al., 2018) and the VCR dataset (Zellers et al., 2019), and visual-textual entailment in the E-SNLI-VE dataset (Kayser et al., 2021). Generating explanations for these tasks presents different levels of difficulty. Explaining VQA instances is similar to captioning since corresponding explanations describe *visible* information in the image that is relevant to the context (Fig. 1, left). On the other hand, VCR instances require higher-order reasoning about *unstated* information such as commonsense (Fig. 1,

Figure 1: Self-rationalization tasks.

right). Moreover, since VCR answers are full sentences, VCR self-rationalization consists of two generative sub-tasks.

We evaluate the following models: (i) a joint vision-language model, VLP (Zhou et al., 2020), (ii) a pretrained language model, T5 (Raffel et al., 2020), that we visually adapt only through finetuning, and (iii) VL-T5/VL-BART (Cho et al., 2021), a combination of the previous two approaches. Namely, VL-T5/VL-BART are developed from T5 and BART (Lewis et al., 2020) by doing a second round of pretraining, using multimodal datasets and objectives. We finetune all three types of models for self-rationalization in two data settings: (i) using entire finetuning datasets (high-resource setting), and (ii) using 30% of the training data (low-resource setting).

We first present an analysis of the factors influencing performance of the visually adapted T5: the choice of image representation (§4.1) and T5's model size (§4.2). We demonstrate that recent advances in representing images, namely CLIP features, can be easily leveraged to get more accurate visually adapted T5 models. However, these improvements are not realized for a more complex sub-task of explanation generation. Moreover, unlike for text generation conditioned only on text (including self-rationalization of tasks with textual inputs; Marasović et al. 2022) we do not see clear performance improvements from scaling visually adapted PLMs. Finally, the main comparison of the three model types described above (§4.3) shows that no model type works universally the best across tasks and data sizes. These re-

sults demonstrate that researchers and practitioners working on text generation conditioned on images beyond image captioning still need to deal with basic modeling decisions such as the choice of backbone models and their size. We hope they spur research on more extensive multimodal model comparisons on variety of generative tasks and experimental setups, as well as on finding ways to realize benefits of different model families (Table 1) with a single approach.

2 Text Generation from Images: Models

Vision-and-language (VL) learning currently comprises two families of models: (i) *joint VL models* that are pretrained from scratch using data with both modalities (§2.1), and (ii) *vision-adapted language models*—pretrained language models adapted to the visual modality through finetuning using end-task multimodal datasets (§2.2). Some models combine these two approaches to some extent, so they could be the best of both worlds (§2.3).

In Table 1, we overview reasons for why one model family might be preferred over the other. For generative tasks conditioned on images, including self-rationalization of VL tasks, the choice of the best base model family is not obvious. The aim of this paper is to find whether such a choice exists.

2.1 VLP: Joint Vision-and-Language Model

We use VLP (Zhou et al., 2020) to analyze importance of benefits of joint VL pretraining from scratch relative to other approaches.

VLP is a shared encoder-decoder transformer (Vaswani et al., 2017) of size similar to BERT-base

Visually Adapted PLMs: Finetuned language models (pretrained only with text) with image features		
Pros	1.	Available PLMs are much larger than joint VL models. This is beneficial since text generation improves with model size (Brown et al., 2020), incl. self-rationalization (Marasović et al., 2022).
	2.	Due to huge pretraining corpora, PLMs have been shown to capture some world (Petroni et al., 2019) and commonsense knowledge (Davison et al., 2019) which is beneficial for self-rationalization as it often requires *inferring* relevant information from the inputs (see Fig. 1).
	3.	Easy to plug and play latest PLMs and image representations as there is no training from scratch.
Cons	1.	Visual information may get suppressed, especially when finetuning multimodal data is limited.
	2.	The number of images in finetuning datasets is much smaller than in pretraining multimodal datasets (e.g., Conceptual Captions with 3.3M image-caption pairs vs. ∼80K images in VCR).
Joint VL Models: Models based on joint pretraining of visual and textual features		
Pros	1.	Joint pretraining from scratch allows the model to learn equally from each modality without starting with any priors. Thus, this joint pretraining should give tighter coupling between modalities.
	2.	Expected to be more robust to limited multimodal datasets used for end-task finetuning.
Cons	1.	Text inputs in multimodal pretraining dataset are often simple language captions.
	2.	Joint models are smaller than PLMs.
	3.	If captioning is only *generative* pretrain task, they are *not* designed for generation conditioned on *both* images and text.
	4.	Requires training from scratch.

Table 1: Summary of pros and cons of training vision and language (VL) jointly from scratch (e.g., VLP) versus adapting pretrained language models (PLM) to visual features (e.g., VA-T5). Some models are combinations of these two approaches (VL-T5, VL-BART). See §2 for more information about these models.

(110M parameters; Devlin et al., 2019). It is pretrained with objectives similar to both masked and standard language modeling. Thus, it is suitable for discriminative as well as generative tasks. A given input image is represented with vector representations of a fixed number of regions obtained with an off-the-shelf object detector (Wu et al., 2019). During finetuning, the same object detector representations should be used. The Conceptual Captions dataset (Sharma et al., 2018), containing about 3M web-accessible images and associated captions, is used for pretraining.

2.2 VA-T5: Vision-Adapted Pretrained LM

Vision-adapted training involves starting with a pretrained language model (PLM) and adapting it to VL tasks as a finetuning step. We start with T5 (Raffel et al., 2020)—a PLM that is commonly used for self-rationalization (Wiegreffe et al., 2021; Marasović et al., 2022)—and finetune it for self-rationalization of VL tasks. Specifically, we concatenate image representations with representations of textual inputs, feed the resulting input in the subsequent PLM's layers, and train the model using language modeling loss on generated answer and explanation tokens.

A question that emerges is: what kind of image

representations should be used? While vector representations of image regions extracted with an object detector are the most frequent choice, most recently advantages of representations from the CLIP model (Radford et al., 2021) have been demonstrated for various applications (Shen et al., 2022). Moreover, we wonder whether using automatic image captions is the way to visually adapt a PLM given that the input will then be completely textual, i.e, in the modality a PLM has seen before. Thus, in §4.1, we compare three different features: (i) auto-generated captions from the off-the-shelf image captioning model (VL-T5 model; Cho et al., 2021), (ii) object features from a pre-trained R-CNN model (Ren et al., 2015a), and (iii) CLIP features (obtained from the last layer of ViT-B/32).

Besides exploring different image representations, visually adapting a PLM allows us to study model scaling. In §4.2, we study the following sizes of T5: Base (220M parameters), Large (770M), and 3B. We refer to visually adapted T5 as VA-T5.

2.3 VL-T5 / VL-BART: Combined Models

Another approach is to start with a PLM, do a second round of pretraining to learn joint VL representations, and finally finetune the model to learn the

Dataset	Task	# Samples train/val/test	Avg. Answer Len train/val/test	Avg. Explanation Len train/val/test
VCR	Visual Commonsense Reasoning	212.9K / 26.5K / 25.2K	7.54 / 7.65 / 7.55	16.16 / 16.19 / 16.07
E-SNLI-VE	Visual Entailment	402K / 14K / 15K	1/1/1	12.3/13.3/13.2
VQA-X	Visual Question Answering	29.5K / 1.5K / 2K	1.03/1.05/1.03	8.6/9.0/9.2

Table 2: Specifications of the self-rationalization datasets.

end-task. This approach can be seen both as a joint model and a visually adapted PLM, and thus may offer benefits from the both model families.

To compare this approach with the others, we use VL-T5 and VL-BART (Cho et al., 2021). VL-BART, a multimodal extension of BART-Base (139M parameters; Lewis et al., 2020), also follows an encoder-decoder transformer but does not share the parameters between the encoder and decoder as is done in VLP. To represent images, it uses the Faster R-CNN model for object detection (Ren et al., 2015b). Specifically, vector representations of a fixed number of image regions are are concatenated with text embeddings and fed into VL-BART, which is then pretrained using masked language modeling objectives in addition to new objectives such as visual question answering, image-text matching, visual grounding and grounded captioning. VL-T5 is similar in spirit to VL-BART, where it is initialized with a T5-Base model (220M parametes; Raffel et al., 2020) correspondingly. T5 is trained for various downstream tasks jointly, whereas BART exploits a task-specific encoder-decoder set up for sequence generation tasks. Both VL-BART and VL-T5 are pre-trained with MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), VQA v2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and Visual7W (Zhu et al., 2016), leading to a total of 9.18M image-text pairs on 180K unique images.

3 Experimental Setup

In this section, we describe tasks, datasets, and evaluation setup for self-rationalization of vision-and-language tasks introduced in prior work (Marasović et al., 2020; Kayser et al., 2021).

3.1 Tasks and Datasets

The dataset statistics are given in Table 2, where the average answer and explanation lengths hint on differences in complexities of each task. The three datasets represent different levels of required reasoning (see examples in Figure 1), e.g., are the

images representing simpler scenarios (Flickr30K) or complex movie scenes (VCR).

VQA-X (Park et al., 2018) is the extension of the widely-used Visual Question Answering v1 (Antol et al., 2015) and v2 (Goyal et al., 2017) datasets, with corresponding free-text explanations. The images here are originally sourced from the MSCOCO dataset (Lin et al., 2014), and the answers are collected for open-ended questions about these images that require vision, language, and commonsense knowledge to answer.

E-SNLI-VE (Kayser et al., 2021) is a dataset for visual-textual entailment, the task of predicting whether a statement is entailed, contradicted, or neutral given an image that serves as a premise. E-SNLI-VE combines annotations from two datasets: (i) SNLI-VE (Xie et al., 2019), collected by replacing the textual premises of SNLI (Bowman et al., 2015) with Flickr30K images (Young et al., 2014), and (ii) E-SNLI (Camburu et al., 2018), a dataset of crowdsourced free-text explanations for SNLI.

VCR (Zellers et al., 2019) is a carefully crowd-sourced dataset of answers and explanations for visual scenes extracted from Hollywood movies. Thus, the visual context in this data is more complex than MSCOCO or Flickr30K images, leading to more complex answers and explanations. Zellers et al. instructed crowdworkers to first annotate answers for a given question-image pair, and then showed the annotated answer along with the question-image pair to a different set of annotators to get the corresponding explanation. This dataset was orginally introduced in a classification setting: given a question about an image, pick the correct answer from four choices, and then pick the correct explanation again from four choices. Dua et al. (2021) propose instead to generate both the answer and explanation. This is more realistic than the multiple-choice setting which is restricted to a user giving answer choices. In this paper, we also generate VCR answers and explanations.

3.2 Evaluation Metrics

Self-rationalization requires evaluating two sub-tasks: correctness of predicted answers/labels, and quality of generated explanations. For the former, we use **accuracy** for E-SNLI-VE and VQA-X. Evaluating VCR answers is more complicated as they are full sentences. Following Dua et al. (2021), given a generated answer, we normalize text (remove articles, punctuation, lowercase), and count the number of overlapping words with the four available answer choices in the VCR dataset. We select an answer candidate with the highest overlap as the predicted answer. **Proxy accuracy** is the accuracy that is computed between the correct answer candidate and predicted answer candidate. Dua et al. (2021) do not report the correlation between proxy accuracy and human judgments of answer plausibility. To this end, for 600 VCR instances, we ask 5 crowdworkers to respond to "Given the image and the question, is this answer likely?" with yes, weak yes, weak no, or no, and map answers to 1, 2/3, 1/3, and 0, respectively. **Answer plausibility** is the average of scores of 5 annotators. In §4, we report the average answer plausibility across 600 instances, as well as proxy accuracy. Spearman's correlation coefficient between the proxy accuracy and answer plausibility is 0.56 ($p < 0.028$) indicating a moderate correlation between them.

Due to unreliability of automatic evaluation measures, human evaluation has been used to evaluate free-text explanation generation (Camburu et al., 2018; Kayser et al., 2021; Clinciu et al., 2021). We ask 3 annotators whether an explanation justifies an answer/label given an image, questions/hypothesis, and a generated answer/label. Annotators can again pick one of the four options (yes, weak yes, weak no, no), and the four options are assigned numerical values (1, 2/3, 1/3, 0). We average scores of 3 annotators to get the plausibility of an individual explanation, and report the average **explanation plausibility** in a sample of 300 instances. Following Kayser et al. (2021), we select the first 300 instances for which the answer/label is correctly generated. For E-SNLI-VE, we select an equal number of examples for each label to produce a balanced evaluation set.

We perform all human evaluation using Amazon Mechanical Turk. Each batch of evaluation contains 10 samples. We pay $1.5 per batch to each annotator for VCR evaluations and $1 per batch for E-SNLI-VE and VQA-X each as VCR tasks

require longer time to complete.

Kayser et al. (2021) report that all automated metrics are weakly correlated with explanation plausibility (per humans), but that BERTscore is most correlated. Therefore, we report **BERTscores** for completeness and reproducibility. Following Kayser et al. (2021), we set the BERTscore of incorrectly predicted instances to 0.

4 Results

In this section, we study whether there is a base model family that is more suitable for text generation conditioned on images and text. Specifically, we compare: (i) a joint vision-and-language model, VLP (§2.1), (ii) a visually adapted PLM, VA-T5 (§2.2), and (iii) VL-T5 / VL-BART, a combination of the previous two model families (§2.3), for self-rationalization of three tasks overviewed in Figure 1 and Table 2. The benefits and downsides of (i) and (ii) are outlined in Table 1.

Before presenting the outcomes of this comparison in §4.3, in §4.1–§4.2 we study the impact of the choice of image features and model size on VA-T5's performance.

4.1 VA-T5: Analysis of Image Features

Visually adapting PLMs allows us to combine different image features with PLM's text representations. In this section, we analyze VA-T5-Base (finetuned with the full training data) with different image features in the input: auto-generated captions, object, and CLIP features (see §2.2 for more information). We also report a control setting where no image features are used (*None*). In Table 3a, we report *answer* (proxy) accuracy and plausibility (for VCR), and in Tables 3b and 3c *explanation* BERTscore and plausibility. Evaluations metrics are described in §3.2. Hyperparameters are reported in the Appendix.

Results We observe that CLIP features give the best accuracy scores for all three datasets (see Table 3a). This result demonstrates the benefit of visual adaptation: new advances in image representations, such as CLIP in this case, can be effortlessly used, unlike with joint models for which we need to retrain the model jointly from scratch with these new representations.

In §2.2, we hypothesize that captioning could be a straightforward way to bridge two modalities to get the most out of a PLM that is already well-positioned to solve the end-task in one modality

Task \ Feat.	None	Captions	Objects	CLIP
VCR	54.5	56.2	57.8	**58.1**
	29.4	34.4	37.4	**41.4**
E-SNLI-VE	67.6	71.7	72.5	**74.7**
VQA-X	43.4	73.8	72.3	**74.7**

(a) Answer (Proxy) Accuracy / Plausibility[†]

Task \ Feat.	None	Captions	Objects	CLIP
VCR	85.3	85.7	85.6	**86.0**
E-SNLI-VE	89.1	**89.3**	**89.3**	**89.3**
VQA-X	90.7	**91.2**	91.1	90.9

(b) Explanation BERTscore

Task \ Feat.	None	Captions	Objects	CLIP
VCR	15.6	19.6	**21.0**	**21.0**
E-SNLI-VE	65.4	**66.6**	65.7	65.2
VQA-X	70.8	75.5	**76.4**	75.8

(c) Explanation Plausibility

Table 3: A comparison of image features (Captions, Objects, CLIP) for VA-T5-Base (§2.2). † For VCR answers, we report both proxy accuracy (1st row) and plausibility (2nd row; §3.2).

(text). However, with the exception of VQA-X, captions give the worst accuracy scores relative to object and CLIP features.

We turn to evaluation of plausibility of generated explanations that paints a different picture (see Table 3c). We observe that CLIP and object features perform similarly for VCR and VQA-X—object features are even slightly better for VQA-X. In other words, advances from CLIP features diminish for a more complex task of explanation generation. Captions work the best for generating E-SNLI-VE explanations with VA-T5-Base, but not for the other two datasets. However, E-SNLI-VE is an outlier in a different way too. It is the only task for which having no image features is better for explanation generation than having CLIP features, and just slightly worse (0.3 points) than having object features. Notably, CLIP/object features require combining vectors from different models while captions are represented with the same pretrained word embeddings as the rest of the input. We thus explore whether the way that layer normalization is applied to the concatenated vectors is crucial, but we find that it is not (see Appendix). We leave further analysis for why visual adaptation gives only minor improvements for generation of E-SNLI-VE

explanations with VA-T5 relative to other datasets to future work.

BERTscore results (Table 3b) are mixed. According to them, CLIP features are the best for VCR and the worst for VQA-X. Moreover, the differences between BERTscore values obtained with different features are very small (0.0–0.4) which makes these results hard to interpret.

4.2 VA-T5: Analysis of Model Size

Another advantage of visually adapting PLMs is that we can use larger model sizes since PLMs are typically more frequently scaled relative to joint models. The benefits of scaling the model and pretraining data size are outlined in Table 1. We explore three model sizes for VA-T5 (§2.2): Base (220M), Large (770M), and 3B. We use CLIP features to visually adapt T5 for these experiments since they give more accurate VA-T5 models, while generating explanations that are similarly plausible to those generated by T5 adapted with object features (see §4.1). We use the full training sets to finetune VA-T5 models in this section.

Results Scaling the model size from 220M (Base) to 770M (Large) parameters gives more accurate models for VCR and VQA-X, but further scaling to 3B parameters degrades performance. This is in contrast to self-rationalization of text-only inputs where performance monotonically increases with the T5's model size (Marasović et al., 2022). E-SNLI-VE is an exception with no clear pattern between the model size and accuracy. Moreover, explanation plausibility decreases with the model size (Base > Large > 3B) for E-SNLI-VE and VQA-X. This is also in contrast to observations in text-only self-rationalization. The exact opposite is true for the plausibility of generated VCR explanations which increases with the model size (3B > Large > Base). Notably, in Table 1, we report that the larger model and data size are correlated with capturing more world and commonsense knowledge, and generating VCR explanations requires more inferring about information that is unstated in the input relative to generating E-SNLI-VE or VQA-X explanations. This might explain the difference between why scaling is beneficial for VCR and not for other datasets.

Unlike accuracy and plausibility for which model scaling is helpful at least to some extent, BERTscore values decrease monotonically with scaling the model size. While we have reserva-

Task \ Model Size	Answer Accuracy[†]			Explanation BERTscore			Explanation Plausibility		
	Base	Large	3B	Base	Large	3B	Base	Large	3B
VCR	58.1	**59.1**	**59.1**	**86.0**	**86.0**	85.8	21.0	24.4	**25.8**
	41.4	**45.8**	42.2	-	-	-	-	-	-
E-SNLI-VE	**74.7**	74.4	68.0	**89.3**	**89.3**	89.2	**65.2**	65.0	64.1
VQA-X	74.7	**75.6**	75.1	**90.9**	90.6	90.0	**75.8**	72.3	69.3

Table 4: A comparison of VA-T5-CLIP model sizes (§2.2). † For VCR answers, we report both proxy accuracy (1st row) and plausibility (2nd row; §3.2).

tions about this evaluation metric to begin with given its weak correlation with human judgements of explanation plausibility, BERTscore values are increasing monotonically for self-rationalization with textual inputs as expected (Marasović et al., 2022). Thus, we see this result as another evidence of the difficulty of visually adapting larger models rather than the limitations of BERTscore as an evaluation metric.

A better understanding of what is the bottleneck for visually adapting larger PLMs is needed. We might need other ways to visually adapt besides the simple input changes that have been done so far.

4.3 Joint Models vs. Visually Adapted PLMs

We turn to our main comparison between a joint model (VLP), a visually adapted PLM (VA-T5-CLIP), and combined models (VL-BART, VL-T5). Given that joint models might be advantageous when finetuning data is limited, we compare the models when finetuned with: (i) the entire training sets (high-resource data setting), and (ii) 30% of the training data (low-resource data setting).

Results In a high-resource setting (see Table 5a), the best *answer* accuracy/plausibility is achieved by a different model for each dataset. To illustrate, VA-T5 (a visually adapted PLM) obtains the best VCR answer proxy accuracy, VLP (a joint model) VCR answer plausibility, VL-T5 (a combined model) E-SNLI-VE accuracy, and VL-BART (another combined model) VQA-X accuracy.

Explanation plausibility results are slightly more consistent (see Table 5a). Namely, VL-BART generates most plausible explanations for E-SNLI-VE and VQA-X, and is only behind VLP for VCR. The best explanation BERTscore is also achieved with VL-BART for all tasks. However, the relative order of model types (joint, visually adapted, combined) across tasks is still mixed. Specifically, for VCR: joint > combined (both) > adapted (all);

for E-SNLI-VE: combined (both) > adapted (all) > joint; for VQA-X: combined > adapted > joint > adapted > combined > adapted. Moreover, future work should consider investigating how its performance changes when other image features are used during finetuning since one advantage of visually adapting PLMs is the ability to plug and play with recent advances in representation learning of both modalities.

It is not necessarily concerning that the results are mixed given the unique benefits and downsides of these models (see Table 1) that could be relevant for one task and not another. However, observed cross-task differences in results are not intuitive. For example, visually adapting T5 and combined models give worse VCR answer and explanation plausibility compared to plausibility obtained with the joint model, VLP. Since generating VCR answers and explanations require the most common-sense and word knowledge relative to the other two tasks, it is reasonable to expect that this is a scenario where long pretraining on a more complex text will be beneficial, but it turns out it is not.

We now turn to results in low-resource data setting (see Table 5b). We hypothesized that joint models might work better when finetuning data is limited since there might not be enough images to appropriately visually adapt PLMs and they might still behave like (unimodal) language models. However, VA-T5 is not always underperforming relative to VLP, VL-BART, and VL-T5 in the low-resource setting. Specifically, VA-T5 is slightly better for explanation generation compared to VL-BART and VL-T5 for VCR, comparable to VLP for E-SNLI-VE, and better than all of the three models for VQA-X. It is also better than VLP for generating VCR answers. These results show that the size of finetuning data is not as detrimental for visual adaptation relative to joint models as we speculated.

As in the high-resource setting, we observe that

Model	VCR				E-SNLI-VE			VQA-X		
	Answer		*Explanation*		*Answer*	*Explanation*		*Answer*	*Explanation*	
	Acc.	Plaus.	BERTsc.	Plaus.	Acc.	BERTsc..	Plaus.	Acc.	BERTsc.	Plaus.
VLP	55.5	**50.1**	85.7	**34.0**	75.4	87.7	63.8	79.4	89.6	73.5
VA-T5-Base	58.1	41.4	86.0	21.0	74.7	**89.3**	65.2	74.7	90.9	75.8
VA-T5-Large	**59.1**	45.8	86.0	24.4	74.4	89.3	65.0	75.6	90.6	72.3
VA-T5-3B	**59.1**	42.2	85.8	25.8	68.0	89.2	64.1	75.1	90.0	69.3
VL-BART	57.8	47.7	**86.6**	29.5	75.6	**89.3**	**71.5**	**86.3**	**91.2**	**75.9**
VL-T5	58.4	44.6	85.5	28.7	**76.3**	89.1	69.0	84.9	91.0	72.2

(a) High-resource data setting.

Model	VCR				E-SNLI-VE			VQA-X		
	Answer		*Explanation*		*Answer*	*Explanation*		*Answer*	*Explanation*	
	Acc.	Plaus.	BERTsc.	Plaus.	Acc.	BERTsc..	Plaus.	Acc.	BERTsc.	Plaus.
VLP	54.7	21.1	85.9	**25.1**	73.5	87.6	63.6	71.8	89.4	72.9
VA-T5-Base	57.2	38.6	85.7	23.1	66.5	89.1	64.5	66.5	90.8	**75.8**
VA-T5-Large	57.3	37.3	85.8	21.1	68.3	89.2	62.0	67.3	90.4	73.6
VA-T5-3B	57.0	32.5	85.3	19.0	68.7	89.2	63.6	53.7	90.6	69.8
VL-BART	57.4	**42.8**	**86.4**	22.7	**74.4**	**89.3**	66.1	**85.6**	**90.9**	72.9
VL-T5	**58.5**	41.5	85.2	22.5	74.2	**89.3**	**66.6**	83.5	90.5	71.3

(b) Low-resource data setting.

Table 5: Comparison of a joint VL mode (VLP), visually adapted pretrained LM (VA-T5), and combined models (VL-BART, VL-T5) on three datasets: VCR, E-SNLI-VE, and VQA-X. We report (proxy) *answer* accuracy and plausibility (for VCR), and *explanation* BERTscore and plausibility. See §2 for more information on models and §3 for tasks, datasets, and evaluation metrics.

no model (type) works universally the best. Model ordering according to their performance sometimes stay consistent compare to the high-resource setting, e.g., for E-SNLI-VE accuracy, and sometimes changes notably, e.g., VLP gives the best VCR answer and explanation plausibility in high-resource setting, but the worst when data is limited. These results highlight the necessity to compare models, not only using a variety of tasks/datasets, but also models finetuned with different amounts of data.

We see that the differences between model performance in high-resource are smaller than in low-resource, where in some cases the gap is huge. For example, VL-BART achieves VCR answer plausibility of 42.8 in low-resource, while VLP results in only 21.1. Another example is VQA-X answer accuracy for which VL-BART achieves 85.6 and VLP 71.8. Unlike the previous example (VCR), this can be explained by the fact that VQA is one of the tasks used to pretrain VL-BART (and VL-T5). Such huge differences between models are not observed for explanation generation, so even though a model is much more accurate, the plausi-

bility of explanations for its correct answers are not that much more plausible compared to explanations of other models.

5 Conclusions

We analyze model instances from distinct model families that we expect to have unique benefits and downsides for text generation conditioned on images and text beyond image captioning. We evaluate them for self-rationalization (jointly generating labels/answers and free-text explanations) of three multimodal tasks. We observe that model scaling and latest image representations do not clearly improve visually adapted pretrained language models. Our main finding is that a single model (family) does not work universally the best across tasks and available finetuning data. Future advances in text generation conditioned on images and text should be demonstrated using a variety of tasks/datasets and dataset sizes, and we should find a method that realizes the benefits of different model families.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (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. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. 2021. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Constantin Eichenberg, Sid Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. Magma - multimodal augmentation of generative models through adapter-based finetuning. arXiv:abs/2112.05253.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2022. Towards general purpose vision systems. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, pages 32–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020a. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. In *AAAI*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2810–2829, Online. Association for Computational Linguistics.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015a. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can clip benefit vision-and-language tasks? In *ICLR*.

Michael Sollami and Aashish Jain. 2021. Multimodal conditionality for natural language generation. arXiv:2109.01229.

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*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

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*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A Appendix

A.1 VA-T5: LayerNorm Analysis

Since VA-T5 is not pretrained jointly with visual and textual representation, when we combine them for finetuning we might need to re-consider how to apply the layer normalization to concatenated representations. Table 6 contains results of no layer normalization upon fusion (concatenation of image and text representations), layer normalization only for image features (as the T5 backbone automatically normalizes text during pretraining), and layer normalization of both text and vision upon fusion. The differences between different ways of applying layer normalization are minor. Therefore, we

A.2 Hyperparameters

Adafactor was chosen as the optimizer as it converged faster than Adam, while taking up lesser memory; this was necessary for fitting the VA-T5-3B model on the GPU. We trained VQA-X for longer as its convergence was slower. The batch size for every model was picked based on the validation BLEU-4 score for the generated answer and rationale.

	Accuracy	BERTscore
No LayerNorm	72.4	**89.3**
LayerNorm(vision)	**73.0**	**89.3**
LayerNorm([vision;text])	72.6	**89.3**

Table 6: A comparison of different way of applying layer normalization in the VA-T5-CLIP-Base model finetuned and evaluated in the E-SNLI-VE dataset.

Computing Infrastructure	NVIDIA Tesla A100
Hyperparameter	**Assignment**
Number of epochs	20 or 50 (VQA-X)
Patience	5
Optimizer	Adafactor
Learning rate	5e-5
Learning rate scheduler	None
Dropout	0.1

(a) Common hyperparameters.

Model	VCR			E-SNLI-VE			VQA-X		
	CLIP	Object	Captions	CLIP	Object	Captions	CLIP	Object	Captions
VA-T5-Base	256	64	placeholder	512	128	placeholder	8	8	placeholder
VA-T5-Large	64	-	-	128	-	-	32	-	-
VA-T5-3B	4	-	-	8	-	-	8	-	-

(b) Effective batch size.

Table 7: Hyperparameters for VA-T5