FigCaps-HF: A Figure-to-Caption Generative Framework and Benchmark with Human Feedback

Ashish Singh[1], **Ashutosh Singh**[2], **Prateek Agarwal**[1], **Zixuan Huang**[1], **Arpita Singh**[1], **Tong Yu**[3], **Sungchul Kim**[3], **Victor Bursztyn**[3], **Nesreen K. Ahmed**[4], **Puneet Mathur**[3], **Erik Learned-Miller**[1], **Franck Dernoncourt**[3], **Ryan A. Rossi**[3]

[1]University of Massachusetts Amherst, [2]Northeastern University, [3]Adobe Research, [4]Cisco Research

ashishsinghw@cs.umass.edu, {dernonco,ryrossi}@adobe.com.

Abstract

Captions are crucial for understanding scientific visualizations and documents. Existing captioning methods for scientific figures rely on figure-caption pairs extracted from documents for training, many of which fall short with respect to metrics like helpfulness, explainability, and visual-descriptiveness, leading to generated captions being misaligned with reader preferences. To address this issue, we introduce **FigCaps-HF**, a new framework for figure-caption generation that can incorporate domain expert feedback in generating captions optimized for reader preferences. Our framework comprises of 1) an automatic method for evaluating the quality of figure-caption pairs, and 2) a novel reinforcement learning with human feedback (RLHF) method to optimize a generative figure-to-caption model for reader preferences. We demonstrate the effectiveness of our simple learning framework by improving performance over standard fine-tuning across different types of models. In particular, when using BLIP as the base model, our RLHF framework achieves a mean gain of 35.7%, 16.9%, 9%, and 11.4% in ROUGE, BLEU, Meteor, and CIDEr scores, respectively. Finally, we release a large-scale benchmark dataset with human feedback on figure-caption pairs to enable further evaluation and development of RLHF techniques for this problem.

Benchmark: Benchmark **Code:** Codebase
Documentation: Documentation

1 Introduction

For scientific articles, figures (graphs, plots, charts) are integral for conveying key research findings. To understand a given figure and, by extension, the scientific work itself, it becomes crucial that the corresponding captions are informative, i.e., a given caption can represent and complement the figure, situating it in the context of the article. While the importance of figure captions is universally acknowledged, writing a good caption is not trivial. More often than not, many scholarly works contain generic figure captions and lack descriptiveness, thus rendering the figure unhelpful. This has motivated extensive research into developing methods that can automatically generate captions for figures to assist researchers in writing better captions.

Existing methods treat figure-captioning as an image-to-text task, where training data is mostly extracted from publicly available scientific articles (Hsu et al., 2021). Many existing datasets, particularly those sourced from platforms like arXiv, contain low-quality captions, which are either uninformative or lack descriptiveness. Such captions can thus result in models with poor generalization and lacking alignment with reader preferences.

To address this, we introduce **FigCaps-HF**, a benchmark and learning framework for improving figure-caption generation by model alignment with reader preferences. Figure 1 describes our proposed framework, designed around two key questions: **(1)** How can we integrate expert feedback into model training without additional compute overhead? **(2)** How can we scale feedback generation while minimizing human annotation efforts?

For **(1)**, we employ offline Upside-Down Reinforcement Learning (UDRL), an offline reward-conditioned behavioral cloning method, to align model-generated captions with expert feedback. Once the reward model is trained and generates reward scores, it is no longer needed during figure-caption model training, reducing computational costs while maintaining performance.

For **(2)**, we develop a caption-rating mechanism guided by reader preference feedback to assess the quality of figure-caption pairs. Using a small, human-annotated dataset with ratings on key factors (e.g., helpfulness, OCR content, takeaway), we

Proceedings of Recent Advances in Natural Language Processing,pages 1172–1182
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-135

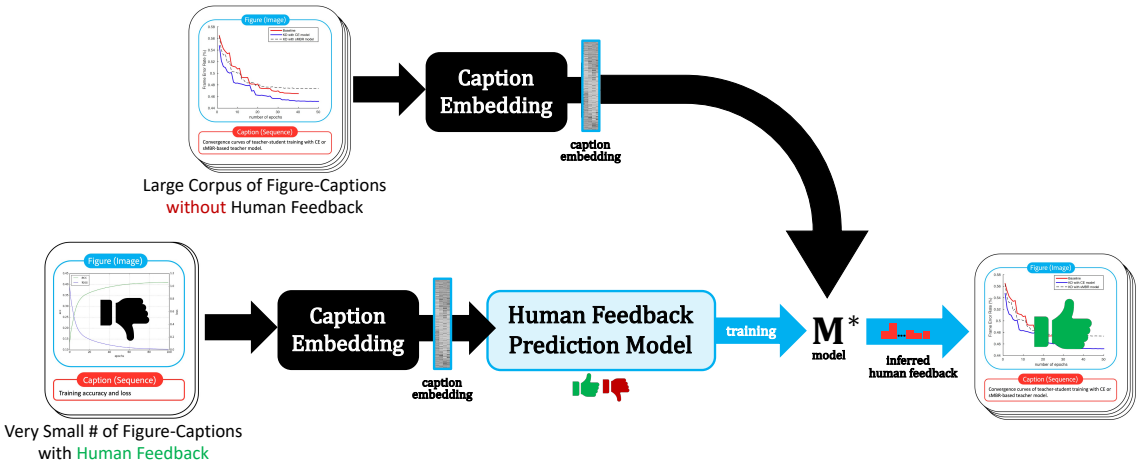

Figure 1: Our proposed framework for improved figure-captioning using Upside-Down RLHF. The framework utilizes a very small set of reader-feedback annotated figure-caption pairs to learn a calibrated figure-caption scoring model. This model is then used to fine-tune the figure-caption model conditioned on inferred feedback scores.

train an auxiliary model to predict caption quality scores. This allows us to infer scores for a larger training set, improving scalability.

Our experimental results demonstrate the effectiveness of our approach. Our trained reward model generalizes well to unseen samples. Evaluations across multiple baseline models show that our reader preference alignment framework outperforms standard supervised fine-tuning, with our best-performing model achieving a 35.7% increase in BLEU, 16.9% in ROUGE-L, 9% in METEOR, and 11.4% in CIDEr scores. Ablation studies further highlight the impact of the type and nature of preference feedback on performance.

Summary of main contributions.

- We introduce an RLHF-based framework for figure-caption generation that uses a small amount of human feedback to train an oracle model, enabling large-scale inference of feedback scores for unseen figure-caption pairs.
- We develop a method for leveraging limited human feedback to predict feedback scores for new figure-caption pairs, improving model alignment with reader preferences.
- We release a benchmark dataset to facilitate further research in figure-caption generation with RLHF, fostering advancements in this domain.

2 Background

Figure Caption Generation. Initial works in scientific figure captioning focused primarily on model design and feature engineering for caption generation. Works like (Siegel et al., 2016; Qian et al., 2021, 2020; Chen et al., 2019, 2020a,b; Hsu et al., 2021) followed a standard pipeline of utilizing a CNN-based vision-encoder to encode figure-features, followed by an LSTM/RNN based text-decoder to generate captions. For model training, (Chen et al., 2019, 2020a,b) created and used synthetic figure-caption pairs, while in (Siegel et al., 2016; Hsu et al., 2021), figure-caption pairs were extracted from publicly available scientific works. With recent advancements in multimodal learning, the standard pipeline has shifted to utilizing pre-trained transformer-based vision-language models for either zero-shot inference or supervised fine-tuning on specific domains for image-to-text generation. Recent works like (Roberts et al., 2024) have focused on benchmarking large multimodal models (LMMs) for figure-caption generation under zero-shot and fine-tuning settings. In contrast, our work is focused on improving model alignment with respect to reader preference in a simple and scalable manner. Our proposed framework is thus model agnostic and applicable to any LMM.

Figure Question Answering. A closely related task is Figure Question Answering, which formulates the more general problem of figure understanding as a visual-question answering task. There has been a variety of works in this space towards modeling (Siegel et al., 2016; Kahou et al., 2017; Li et al., 2022b; Singh and Shekhar, 2020; Zou et al., 2020; Kafle et al., 2018, 2020) as well as creating curated datasets including DVQA (Kafle et al., 2018), FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020), Leaf-QA (Chaudhry

et al., 2020), and ChartQA (Masry et al., 2022). In contrast, our work addresses caption generation and does not focus on question answering.

Learning with Human Feedback Aligning model predictions with human preference has been shown to improve task performance in various areas, including tasks like language model pretraining (Korbak et al., 2023), machine translation (Bahdanau et al., 2016; Kreutzer et al., 2018), text summarization (Stiennon et al., 2020), unlearning undesirable behaviors from language models (Lu et al., 2022), text-to-image generation (Lee et al., 2023; Zhang et al., 2023) and training RL agents (MacGlashan et al., 2017; Ibarz et al., 2018; Lee et al., 2021). In contrast to prior works, we aim at improving figure caption generation by optimizing model learning to align with domain expert feedback. However, unlike previous work that leverages on-policy RL (Schulman et al., 2017) algorithms to maximize the reward-weighted likelihood, our framework utilizes reward-conditioned behavioral cloning (Emmons et al., 2021), an offline variant of the upside-down RL method (Srivastava et al., 2019) to optimize model learning for reader preference. This provides a simpler and more controllable framework for human preference alignment. Furthermore, our feedback scheme allows for incorporating multiple feedback at different granularities as a reward signal during the model optimization step, thus improving model learning.

3 Framework

In this section, we present our framework for learning with expert feedback (Figure 1). First, we describe a standard figure-captioning pipeline (Sec. 3.1), then outline the design and training of a generalizable human-feedback prediction model (Sec. 3.2), and conclude with our feedback-aligned model training strategy using a simple RLHF framework (Sec. 3.3).

3.1 Preliminaries

Given the dataset D_w, we can then define a model f_θ, that takes in information corresponding to the figure and outputs a sequence of text as output.

Model f_θ consists of a vision encoder module to get image-based encoding and a language encoder-decoder module to encode and generate corresponding text. The weights θ can either be randomly initialized or initialized by large-scale pretrained model weights. Furthermore, the model weights corresponding to the vision encoder and text encoder-decoder models can either be initialized with separate weights or jointly trained model weights. After initialization, the model f_θ can then be trained for the task of caption generation.

Generally, for training such a model, Language Modeling (LM) loss is used as a standard training objective. Let $\{I_i, T_i\} \in D$ be the input to the model f_θ, where $I_i \in \mathbb{R}^n$ is the input figure, and T_i is the corresponding text sequence. Additionally, T_i is represented as sequence of K_j tokens from a fixed vocabulary $\mathcal{V}$: $T_i = (T_{i,1}, ...T_{i,K_j})$, where $K_j = |T_i|$. Then the training objective is defined as:

$$\mathcal{L}_{\text{LM}} = \frac{1}{K_j + 1} \sum_{j=0}^{K_j+1} H(T_{i,j} | I_i, (T_{i,0}, ..., T_{i,j-1})), \tag{1}$$

where H denotes the cross-entropy loss and $(T_{i,0}, ..., T_{i,j-1})$ represents all the tokens in the caption prior to $T_{i,j}$.

3.2 Human Feedback Prediction Model

To improve figure-caption generation, we propose to incorporate domain expert feedback into our optimization step. To generate feedback for figure-caption pairs, we thus propose to learn a feedback prediction model to score individual datasamples based on different metrics representing reader preferences. Our objective is to learn a model that can predict human feedback scores for unseen captions accurately, given a small set of training samples.

To this end, we first label a small control set D_h consisting of M figure caption pairs $\{I_w, T_w\}$ with domain experts ratings. Here we assume that $M \ll N$, i.e., the size of the control set is significantly less than the original noisy dataset. We can now train a model on D_h to predict the human expert ratings for the original dataset D_w. Specifically, given human feedback dataset D_h containing figure-caption pairs $\{I_h, T_h\} \in D_h$ and k human expert evaluation metrics for each datasample $y_i \in \{y_0, y_1, ...y_k\}$, we want to train k models $R(x_i, \theta)_k$ to predict the k scores, respectively. Here, the output of a model $R(x_i, \theta)_k(T_h)$ is a scalar quantity denoting a specific metric score for the given input caption. Thus, we formulate the scoring problem as a regression task. Specifically, we can define our human-feedback prediction model as follows:

$$R(x_i, \theta)_k(T_h) = g(l(\theta_l, x_i), \theta_g), \tag{2}$$

where, $R(x_i, \theta) : \mathbb{R}^N \rightarrow \mathbb{R}$, $l(x_i, \theta_l) : \mathbb{R}^N \rightarrow \mathbb{R}^D$ and $g(u_i, \theta_g) : \mathbb{R}^D \rightarrow \mathbb{R}$. In the above, $l(., \theta_l)$ is an embedding function that takes in input data $x_i \in \mathbb{R}^N$ and generates corresponding representation $u_i \in \mathbb{R}^D$, and $g(., \theta_l)$ is a regression function to generate the scores respectively. We only train the regression function while keeping the weights of the embedding function fixed. For training the regression function, we use mean-squared error loss, written as: $\mathcal{L}_{\text{R}} = \frac{1}{D_h} \sum_{i=1}^{D_h} (\hat{y}_i - y_i)^2$, where $\hat{y}_i$ is the predicted score while y_i is the ground-truth evaluation score. After training the human-feedback prediction models, we compute scores for all the samples in the training dataset D_w to construct our new set, which will be used for training the figure-caption model.

3.3 Reinforcement Learning with Human Feedback

We use the human-feedback prediction model as a reward model to train an image-to-text model for generating higher-quality captions, framing the problem as a reinforcement learning task. Given a dataset D_w with figure-caption pairs $\{I_w, T_w\}$, we treat figures I_w as states, captions T_w as actions, and predicted metric scores $R(T_w)$ as rewards. Our goal is to train an image-to-text model $f(\theta)$ that maps states to actions while maximizing rewards, ensuring that captions align with human judgment.

We adopt offline UDRL for its computational efficiency and robustness (Emmons et al., 2021). Here, the policy π_θ maps states (S_t) to actions (a_t) given rewards (r_t), formulating learning as a supervised problem. We sample triplets $\{S_t, a_t, r_t\}$ to construct a dataset and train π_θ using:

$$\max_{\theta} \sum_{t \in D} \mathbb{E}[\log \pi_\theta(a_t | S_t, r_t)] \tag{3}$$

Following this UDRL framework, we define our figure-to-caption model $f(\theta)$ as the policy π_θ. For each caption T_i, we compute a reward score and binarize it into control tokens: <|good|> if $R(I_i, T_i) \geq t$, otherwise <|bad|>, where t is a hyperparameter. Given this feedback, we fine-tune f_θ using:

$$\mathcal{L}_{\text{HF}} = \frac{1}{K_j + 1} \sum_{j=0}^{K_j+1} H(T_{i,j} | I_i, (c_i, T_{i,0}, ..., T_{i,j-1})) \tag{4}$$

where c_i is the control token derived from R.

4 FigCaps-HF: Figure-Captioning with Human Feedback Benchmark

We propose a new benchmark for figure-captioning with feedback. Our benchmark consists of 106,834 figure-caption pairs (Hsu et al., 2021) with feedback scores. Our dataset contains feedback based on different measures to evaluate the quality of the author-written captions for the corresponding figure. For each figure-caption pair, we evaluate the data sample based on four quality measures: **(1) Helpfulness**, **(2) Takeaway**, **(3) Visual-descriptiveness (visual)**, and **(4) Image-text (OCR)** (Huang et al., 2023). Each quality metric is selected to measure the ability of the readers to comprehend and draw inferences based on the provided figure and the corresponding caption.

We compute the feedback scores for each data sample by first annotating a small subset with domain-expert feedback and then predicting the scores for the entire dataset using the human-feedback model described in Sec. 3.2. Using this labeled subset, we train a human-feedback prediction model to generate scores for the remainder of the dataset. Unlike the subset, we retain the scores for the entire dataset as continuous values. This allows the users of the benchmark to accordingly decide their scheme for labeling each figure-caption pair based on different thresholding criteria, thus providing flexibility for fine-grained feedback.

Table 1 presents an overview of the statistics related to the actual and predicted human feedback for the captioning of the scientific figures. We see that the predicted human feedback values in our study show a diverse range, as indicated by the small standard deviation of 1 ± 0.2 and a consistent mean value across all ratings. Additionally, the alignment of the median predicted scores with the actual human feedback values indicates that the model's performance is not skewed towards any particular rating but provides an accurate assessment across the range of ratings. This suggests that the human-feedback prediction model used to infer the scores is generalizable and can accurately assess the quality of captions across various ratings. Furthermore, the proposed model provides reliable scores for captions that fall outside the typical range of scores.

We provide more details regarding the benchmark and the corresponding datasheet at Documentation.

	# Fig-Caption Pairs	Human Feedback	Median	Mean	Std	Q1	Q3
ACTUAL HUMAN FEEDBACK	**438**	**Helpfulness**	3	3.01	1.19	2	3
		Takeaway	2	2.16	1.22	1	2
		Visual	2	2.11	1.08	1	2
		OCR	4	3.83	0.80	4	4
PREDICTED HUMAN FEEDBACK	**106,834**	**Helpfulness**	2.89	2.89	1.07	2.17	3.61
		Takeaway	1.95	2.06	1.03	1.33	2.66
		Visual	1.91	2.02	1.01	1.31	2.63
		OCR	3.88	3.84	0.83	3.32	4.41

Table 1: Summary of our benchmark dataset for figure-caption generative models with RLHF.

	MODEL	#Params	ROUGE-L	BLEU-4	CIDEr	METEOR
OCR-ONLY	**Pegasus**	0.27B	0.026	4.78e-4	0.134	0.042
FIGURE-ONLY	**TrOCR**	0.23B	0.025	<0.001	0.016	0.018
	BEiT+GPT2	0.24B	0.142	0.005	0.372	0.124
	ViT + RoBERTA	0.23B	0.140	0.012	0.380	0.121
	ViT + GPT2	0.24B	0.142	0.018	0.427	0.126
FIGURE-CAPTION	**PromptCap**	0.47B	0.130	0.009	0.269	0.082
	Flamingo	1.14B	0.087	0.001	0.243	0.046
	GIT	0.17B	0.119	0.002	0.219	0.091
	BLIP	0.25B	0.130	0.014	0.438	0.132
	CLIPCap	0.15B	0.103	0.012	0.284	0.131
RLHF	**Ours-BLIP-RLHF**	0.25B	**0.152**	0.019	**0.552**	**0.145**
	Ours-ViT+GPT2-RLHF	0.24B	0.138	**0.020**	0.489	0.126

Table 2: Comparison with state-of-the-art methods. For all the metrics, higher values are better (↑).

FEEDBACK	ROUGE-L	BLEU-4	METEOR
Binary	0.152	0.019	0.145
Multi-label	0.153	**0.022**	**0.151**
Binary + Multi-label	**0.156**	0.019	0.148

Table 3: Results with different forms of feedback.

	ROUGE-L	BLEU-4	METEOR
Helpfulness	0.1520	0.0186	0.1450
Takeaway	0.1676	0.0230	0.1598
Visual	0.1678	0.0230	0.1595
OCR	0.1654	0.0223	0.1565

Table 4: Results with different human feedback metrics.

	ROUGE-L	BLEU-4	METEOR
BERT	0.1565	0.01927	0.1473
SciBERT	0.1577	0.0201	0.1509
BLIP	0.1573	0.01977	0.1494

Table 5: Results with different embedding models for the human-feedback model.

	ROUGE-L	BLEU-4	METEOR
BLIP-RLHF (append)	0.136	0.018	0.132
VIT-GPT2-RLHF (append)	0.138	0.016	0.119
BLIP-RLHF (prepend)	0.152	**0.019**	**0.145**
ViT+GPT2-RLHF (prepend)	0.138	**0.020**	0.126

Table 6: Comparing RLHF prepend to append.

	MSE
Helpfulness	0.082 ± 0.12
Visual	0.076 ± 0.20
Takeaway	0.087 ± 0.17
OCR	0.095 ± 0.13

Table 7: Evaluation of out-of-sample generalization with respect to different human feedback metrics

5 Experiments

Dataset. Our benchmark dataset follows the splits from (Hsu et al., 2021), which contains 106,834, 13,354, and 13,355 samples in train, val, and test sets, respectively. Each training sample is further augmented with feedback predictions generated using our human-feedback prediction model.

Annotation details of Human-Feedback set. We selected the annotators based on their expertise in the areas of computer vision/natural language processing and machine learning. Our annotator pool consisted of 10 Ph.D. graduates and active gradu-

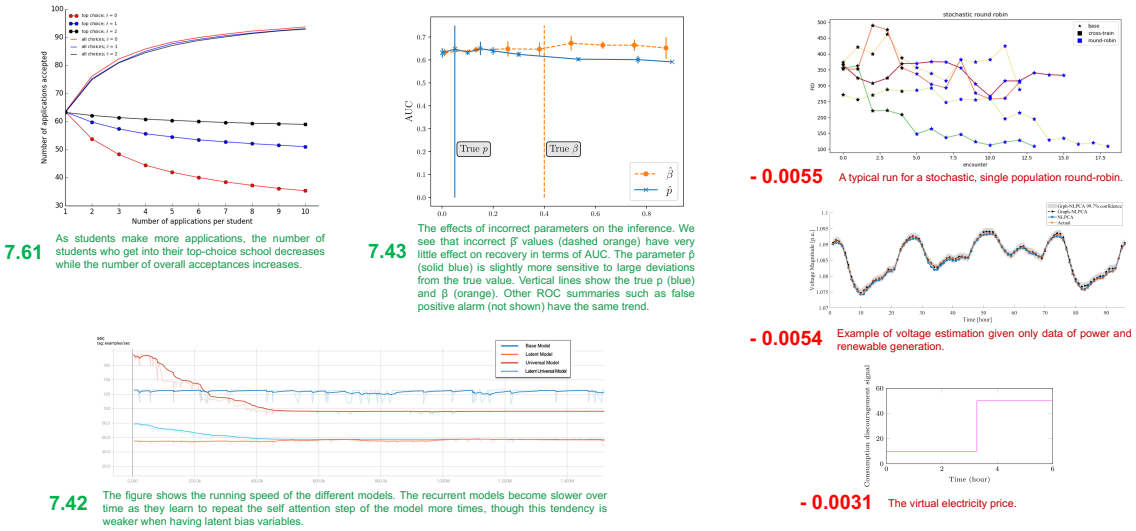

Figure 2: Results of our Human Feedback Prediction Model. Here we show the three figure-caption pairs with the highest (left; green) and smallest (right; red) "helpfulness" human feedback score from our trained HF model. Notably, the figure-caption pairs rated highly by our human-feedback predictive model are better as they mention specific takeaways, figure text, and visual details. In contrast, the figure-caption pairs with the lowest scores by our predictive model are those that are extremely vague and uninformative.

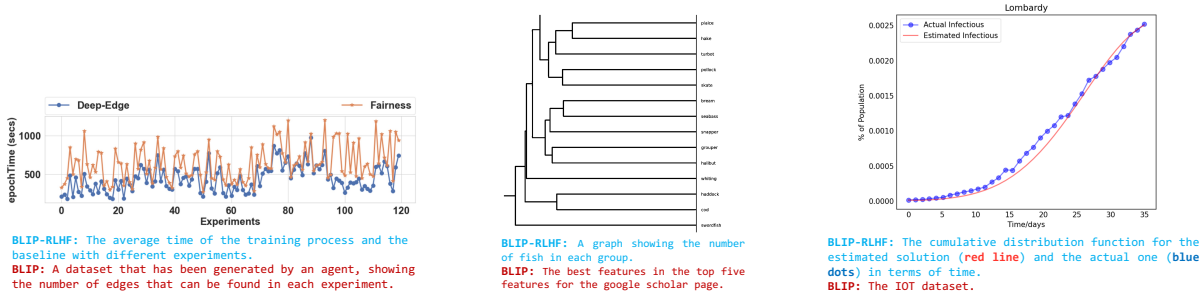

Figure 3: Generated captions from our RLHF framework using BLIP as the base model (in Blue) compared to BLIP without RLHF (in Red). Fine-tuning BLIP with human-feedback predictions significantly improves the caption quality with respect to descriptiveness while maintaining conciseness.

Training Size	MSE	Gain
25% (109)	0.579	91.72%
50% (219)	0.323	6.95%
100% (438)	0.311	2.98%
125% (657)	0.309	2.32%
200% (876)	0.302	0%

Table 8: Results varying the training size used for learning the human feedback prediction model (for inferring "Helpfulness"). Note that the gain is computed with respect to the best (lowest) MSE obtained (0.302).

ate students (no authors) with published work in the CV, NLP, and ML conferences. We randomly selected 438 figure-caption pairs from the dataset to be annotated. Each annotator was provided 2 weeks to annotate the data subset. For each sample, annotators were asked to provide ratings on a five-point Likert scale for the following attributes **[OCR, Visual, Takeaway, Helpfulness]**. For each sample, the following descriptions were provided:

- **OCR:** The caption includes named entities or important words/numbers in the figure(e.g., title, legends, labels, etc.).
- **Visual-Descriptiveness:** The caption includes some visual characteristics of the figure (e.g., color, shape, trend, etc.).
- **Takeaway:** The given caption explicitly states the high-level takeaway message or the conclusion that the figure attempted to convey.
- **Helpfulness:** The caption helped understand the message that the figure is attempting to convey.

Human-feedback prediction setup For our human-feedback prediction model, we use MCSE

(Zhang et al., 2022) as the embedding function and a 2-layer MLP as the regression function. We train the MLP layers until convergence on the human feedback set. Once the model is trained, we infer feedback scores for the train set. We select the median score from the train set as the threshold and label each sample as "good" or "bad". After prepending our captions with these annotations, we effectively train our models in a UDRL framework.

Baselines. For comparative evaluation, we select the following models as our baselines based on input: (1) OCR-only: Pegasus(Zhang et al., 2020), (2) Figure-only: TrOCR (Li et al., 2021), BeiT+GPT2, ViT+GPT2 (Dosovitskiy et al., 2021), ViT+RoBERTA (Dosovitskiy et al., 2021; Liu et al., 2019) and (3) Figure-Caption: PromptCap (Hu et al., 2022), Flamingo (Alayrac et al., 2022), GIT (Wang et al., 2022), BLIP (Li et al., 2022a) and CLIPCap (Mokady et al., 2021). We use ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BLEU-4 (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) metrics for model evaluation.

5.1 Results

We show our experimental results in Table 2. We compare our framework with the standard fine-tuning method and benchmark the performance on the Test set of our proposed benchmark. We use BLIP and ViT+GPT2 to evaluate our RLHF framework. From Table 2, we see that models trained using our proposed RLHF formulation perform better than simple fine-tuning. Specifically, for BLIP, RLHF provides a 35.7% increase in BLEU, 16.9% increase in ROUGE-L, 9% increase in METEOR, and 11.4% in CIDEr score. For ViT+GPT2, RLHF provides a 11.1% increase in BLEU and a 5.1% increase in CIDEr score.

Overall, since the performance increase is generalized among models with different pre-training strategies and overall model structure, the results show the benefits of using this simple UDRL framework for fine-tuning. Utilizing different scoring mechanisms and prompts can be further developed to take advantage of this limited supervision and further increase performance.

5.2 Qualitative Results

Figure 2 and Figure 3 show some of the qualitative results of the feedback prediction model and the figure-captioning models trained with RLHF. We provide our analysis below:

Human Feedback Prediction Model: To evaluate the generalizability of our model, we first computed the score predictions on all the figure-caption pairs. Then we ordered the figure-caption pairs by the predicted scores and selected the top-3 figure-caption pairs with the largest score, along with the bottom-3 figure-caption pairs with the lowest score. Results are provided in Figure 2. We observe that the figure-caption pairs with the largest scores are highly helpful to the reader (shown in green on the left in Figure 2), as they mention specific takeaways from the figure (*e.g.*, "as students make more applications, the number of students who get into their top-choice school decreases, while the number of overall acceptances increases."), as well as mentioning specific visual aspects that are important to the understanding of the underlying context (*e.g.*, "... Vertical lines show the true p (blue) and β (orange)"). In contrast, the figure-caption pairs scoring the lowest (bottom-3), which are shown in red on the right in Figure 2, are vague, without any takeaways, nor reference to visual elements in the figure.

Figure-Caption Generative Model: From Figure 3 we see that, qualitatively, BLIP-RLHF produces better captions compared to fine-tuned BLIP. In most cases, captions produced by BLIP (Fine-tuned) are either explaining the given figure incorrectly (Figure 3, leftmost sub-figure), not relevant (Figure 3, middle sub-figure) or are completely uninformative (Figure 3, rightmost sub-figure). On the other hand, captions produced by the BLIP-RLHF method are more faithful to the figure, capture the semantic relation between texts to summarize the phenomenon, and utilize visual attributes in explaining the figure.

5.3 Ablation Study

We provide our findings from ablation studies for different components of our framework below:

Effect of granularity of feedback labels: To evaluate how quantization levels of reward signals (Binary vs. Multi-level) impact model learning, we conducted a comparative study by modifying feedback while training the BLIP-RLHF model. First, we trained the model for 10 epochs using multi-level human feedback (Row 2), with five feedback levels (very bad, bad, neutral, good, very good) determined at the 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles to balance sample distribution. We also experimented with varying label granularity (Row

3), training with binary-label feedback for 5 epochs, followed by multi-label feedback for another 5 epochs. Results in Table 3 indicate that both approaches using finer feedback outperform simple binary feedback. Our framework demonstrates the model's ability to effectively leverage fine-grained feedback. Additionally, the experiment validates the quality of our human prediction model, which provides useful labels at different levels of granularity, enhancing performance for figure-captioning.

Comparison of different feedback types: To understand the effect of different types of feedback, we compare the results of training the BLIP-RLHF model using Helpfulness, Takeaway, Visual-descriptiveness (Visual), and Image-text (OCR) feedback scores. The results are provided in Table 4. We observe that training BLIP-RLHF with Takeaway, Visual, and OCR feedback outperforms training with Helpfulness feedback. This is expected, as the Helpfulness rating is subjective, whereas Visual and Takeaway are objective evaluation metrics. This finding highlights the importance of feedback type and suggests that further improvements can be achieved by modeling different aspects of the annotated human dataset.

Feedback prediction model architecture: We compare different embedding models (BERT, SciBERT, and BLIP) in constructing the human feedback prediction model. The results are provided in Table 5. We observe that different representations outperform our default MCSE implementation, indicating that our human feedback prediction model and downstream figure-captioning performance are sensitive to the quality of representations used. This highlights that further performance gains can be made by using different representations, for example, by encoding different modalities (text only vs joint encoding of text and vision).

Generalizability of the human feedback prediction model: To evaluate the out-of-sample generalization of our human-feedback prediction model, we conduct a 5-fold cross-validation experiment on the original 438 annotated. We repeated the above experiment 5 times. We report our results in Table 7, including mean squared error (MSE) and standard deviation. As can be seen from Table 7, our model is able to achieve good results on the validation set. This highlights that our human-feedback prediction model demonstrates out-of-sample generalization and proves the statistical significance of our model.

Varying training size: To evaluate the effectiveness of our approach when varying the number of samples used during training, we train the human feedback prediction model using 25%, 50%, 100%, 125%, and 200% of the human-annotated data. We used a held-out set of 300 samples for model evaluation of each of these models. We then trained separate models for each training set for the task of predicting the 'Helpfulness' measure. The results showing mean-squared error (MSE; lower is better) are provided in Table 8. Notably, we see the test performance of the model saturates as the number of training samples is increased. Even with 50% of the original human-annotated data, the model achieves good test results.

Effect of human feedback position: To understand the sensitivity of the model to the position of human feedback, we compare the performance of appending and pre-pending the human feedback labels in Table 6. Since our models generate text, during test time, without any human feedback label prompt, they can only rely on feedback during training. Additionally, due to the auto-regressive generation of our models, they only observe the label before generation, and for append, only observe the label after generation. Intuitively, pre-pending should work best since the generation is conditioned on the label. The results support this and show that ViT+GPT2 and BLIP perform better when trained with pre-pended human feedback.

6 Conclusion

In this work, we developed a new benchmark and methodology to improve caption generation for scientific figures. We showed that incorporating domain expert feedback in learning a model for figure-to-caption generation improves both model performance and caption quality. Our proposed framework is scalable (requires limited manual human effort in labeling) and flexible (allows for incorporating multiple reward signals at different granularities). We hope that this new benchmark dataset will allow researchers to benchmark their own methods for incorporating human feedback in figure-to-caption generation tasks and various other image-to-text generation tasks. Future work will explore techniques to incorporate multiple complementary feedback as well as different ways to quantize the reward score to leverage it as valid feedback when training the model.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Charles Chen, Ruiyi Zhang, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019. Neural caption generation over figures. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 482–485.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020a. Figure captioning with relation maps for reasoning. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1526–1534.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020b. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1545.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. 2021. Rvs: What is essential for offline rl via supervised learning? *arXiv preprint arXiv:2112.10751*.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv:2211.09699*.

Chieh-Yang Huang et al. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. *Open Review*. https://openreview.net/pdf?id=80R7RVLcsf.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. 2020. Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1498–1507.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *Sixth International Conference on Learning Representations Workshop*.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. *arXiv preprint arXiv:2302.08582*.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Kimin Lee, Laura Smith, and Pieter Abbeel. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Ying Li, Qingfeng Wu, and Bin Chen. 2022b. Multi-attention relation network for figure question answering. In *Knowledge Science, Engineering and Management: 15th International Conference, KSEM 2022, Singapore, August 6–8, 2022, Proceedings, Part II*, pages 667–680. Springer.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pages 2285–2294. PMLR.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, and Joel Chan. 2020. A formative study on designing accurate and natural figure captioning systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A. Rossi, Sana Malik, and Tak Yeon Lee. 2021. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, WWW '21, page 2792–2804, New York, NY, USA. Association for Computing Machinery.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision*, pages 664–680. Springer.

Hrituraj Singh and Sumit Shekhar. 2020. Stl-cqa: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284.

Rupesh Kumar Srivastava, Pranav Shyam, Filipe Mutz, Wojciech Jaśkowski, and Jürgen Schmidhuber. 2019. Training agents using upside-down reinforcement learning. *arXiv preprint arXiv:1912.02877*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. 2022. Mcse: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931*.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. 2023. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*.

Jialong Zou, Guoli Wu, Taofeng Xue, and Qingfeng Wu. 2020. An affinity-driven relation network for figure question answering. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.