DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback

WARNING: This paper contains qualitative examples which are offensive in nature.

https://dresslvlm.github.io/

Yangyi Chen[1,2]*, Karan Sikka[1], Michael Cogswell[1], Heng Ji[2], Ajay Divakaran[1]
[1] SRI International [2] University of Illinois Urbana-Champaign
yangyic3@illinois.edu

Abstract

*We present **DRESS**, a large vision language model (LVLM) that innovatively exploits Natural Language feedback (NLF) from Large Language Models to enhance its alignment and interactions by addressing two key limitations in the state-of-the-art LVLMs. First, prior LVLMs generally rely only on the instruction finetuning stage to enhance alignment with human preferences. Without incorporating extra feedback, they are still prone to generate unhelpful, hallucinated, or harmful responses. Second, while the visual instruction tuning data is generally structured in a multi-turn dialogue format, the connections and dependencies among consecutive conversational turns are weak. This reduces the capacity for effective multi-turn interactions. To tackle these, we propose a novel categorization of the NLF into two key types: critique and refinement. The critique NLF identifies the strengths and weaknesses of the responses and is used to align the LVLMs with human preferences. The refinement NLF offers concrete suggestions for improvement and is adopted to improve the interaction ability of the LVLMs– which focuses on LVLMs' ability to refine responses by incorporating feedback in multi-turn interactions. To address the non-differentiable nature of NLF, we generalize conditional reinforcement learning for training. Our experimental results demonstrate that DRESS can generate more helpful (9.76%), honest (11.52%), and harmless (21.03%) responses, and more effectively learn from feedback during multi-turn interactions compared to SOTA LVLMs.*

1. Introduction

Large vision-language models (LVLMs) can perceive the visual world and follow the instructions to generate user-friendly responses [6, 43, 90]. This is achieved by effectively combining large-scale visual instruction finetuning [78] with large language models (LLMs) [5, 53].

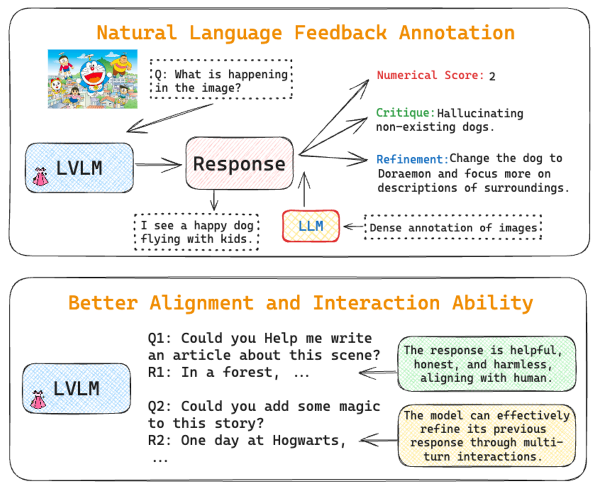

Figure 1. We instruct DRESS to improve both alignment with human preferences and interaction ability via natural language feedback, which is categorized into critique and refinement.

However, existing LVLMs solely leverage the LLMs-generated or hand-crafted datasets to achieve alignment via supervised fine-tuning (SFT) [6, 43, 78]. While it's effective at transforming LVLMs from caption generators to instruction-following models, LVLMs can still generate responses that are unhelpful, hallucinated, or even harmful (see Figure 4). This indicates that their present level of alignment with human preference is still relatively low [81]. In addition, although existing work motivates to structure visual instruction tuning samples in multi-turn formats, the connection and dependencies among various turns are weak, which restricts the interaction ability of the LVLMs. Here the interaction ability measures whether LVLMs can effectively leverage the previous context in multi-turn interactions and refine their responses [72]. These two limitations restrict the potential of LVLMs to serve as visual assistants in practice.

In this work, we introduce **DRESS**, an LVLM distinctively trained through the application of Natural

*Work done during internship at SRI International.

Language Feedback (NLF) generated by LLMs (see Figure 1). We provide LLMs with dense annotation for images and detailed guidelines, instructing them to give fine-grained feedback on the LVLM's responses. This feedback annotation considers 3H criteria– helpfulness, honesty, and harmlessness, consistent with the practice in developing human-aligned LLMs [51]. The generated feedback includes the numerical score and NLF that measure the overall quality of the responses along the 3H criteria.

In our approach, we introduce a novel categorization of NLF into two distinct types: *critique* and *refinement*. The critique NLF provides an assessment of the strengths and weaknesses of the responses, whereas the refinement NLF provides specific suggestions to LVLMs on improving their responses to align with the ground truth reference. This categorization offers a natural utilization of two types of NLF to align the LVLMs with human preferences and improve their interaction capabilities. To train the LVLMs with such feedback, we generalize the conditional reinforcement learning algorithm to address the non-differentiable nature of NLF. In particular, we train DRESS👗 to produce corresponding responses conditioned on the two NLF using language modeling (LM) loss on the responses. By learning from the numerical scores and critique NLF, we improve the alignment of DRESS👗 with human preferences. While, by leveraging refinement NLF, we train DRESS👗 to acquire the meta-skill of refining its initial responses by utilizing NLF through multi-turn interactions during inference.

We evaluate DRESS👗 on open-ended visual question answering for helpfulness evaluation, image captioning for honesty evaluation, adversarial prompting for harmlessness evaluation, and also on multi-turn interactions. Experimental results demonstrate that DRESS👗 can generate responses that are better aligned with human values as compared to previous LVLMs, and also demonstrates better interaction ability that can effectively learn from feedback to refine the responses on the fly. To the best of our knowledge, we are the first work to address all the 3H criteria as well as interaction ability for LVLMs. We summarize our contributions as follows:

- We propose the distinct use of natural language feedback (NLF), specifically categorized into critique and refinement NLF, to improve the alignment with human preferences and interaction capabilities of LVLMs.
- We generalize the conditional reinforcement learning algorithm to effectively incorporate the NLF, which is non-differentiable, by training the model to generate corresponding responses conditioned on the NLF.
- We produce and open-source 63K annotated vision-language NLF samples covering 3H aspects. In addition, we also open-source a dataset with 4.7K examples for harmlessness alignment and evaluation of LVLMs. The datasets are released at `https://huggingface.co/datasets/YangyiYY/LVLM_NLF`.

2. Related Work

Large Vision-Language Models. The current research motivates the creation of LVLMs that can tackle various tasks without specific adaptations [36, 69, 73][1]. Given the strong fundamental abilities of LLMs [5, 6, 49], most recent LVLMs typically adopt frozen LLMs as the language component [43, 90], accompanied by a substantial scaling in the model sizes. LVLMs capitalize on large-scale image-caption pairs to train a projector to transform the image features into the embedding space of LLMs to align the two modalities [2, 36, 43, 84, 90]. In addition, large-scale vision-language instruction tuning data is adopted to align LVLMs with human preferences, ensuring that they can effectively understand instructions and generate user-friendly responses [20, 22, 32, 42, 64, 75]. In this work, we further calibrate the human preference alignment in responses generated by LVLMs and improve their interaction ability by leveraging the feedback provided by LLMs.

Learning from Feedback. Incorporating feedback to train and align LLMs has emerged as a pivotal approach [11, 17, 51, 57]. External feedback is often associated with reinforcement learning to train LLMs to optimize some goals that are hard for data annotation, such as becoming helpful [3, 30, 63], harmless [4, 21], and honest [51]. Depending on the form, the feedback can be formatted as numerical scores [18, 41], preference ranking [4, 51], or natural language [1, 57]. The numerical scores and preference ranking feedback are relatively easier to collect via human annotations [51, 63], while NLF is much harder and more expensive for annotation. Thus, in this work, we rely on LLMs to provide NLF [1, 4, 79], which is different from [66] that pivots on preference ranking data collection and adopts numerical score reward for training. In addition, we categorize the NLF into two types: critique and refinement, which can be adopted respectively to improve the alignment and interaction of LVLMs. We use generalized conditional reinforcement learning to force the LVLM to learn directly from NLF and differentiate between aligned or misaligned responses and effective or ineffective interaction behaviors. We further discuss related work on multi-turn interactions that incorporate human feedback for refinement in Appendix A.

3. DRESS👗

We describe DRESS👗, an LVLM designed to leverage NLF from LLMs to improve two key aspects missing in prior work: (1) Alignment with human preferences, and (2) Interaction capabilities. The first focuses on whether the responses respect human values, especially the 3H criteria (helpfulness, honesty, and harmlessness) [51]. The second aspect focuses on the ability to refine responses based

[1]More related research on vision-language modeling is in Appendix A

on feedback provided during multi-turn interactions. We achieve this by proposing an innovative classification of NLF into two primary categories: Critique and Refinement. For training DRESS with NLF, we propose a generalization of conditional reinforcement learning specially designed to address the non-differentiable nature of the NLF.

In this section, we first describe the training recipe to produce DRESS$_{\texttt{ft}}$, the LVLM that subsequently serves as the data source for collecting NLF, along with the data splits. We then describe the procedure for collecting feedback from LLMs. We finally discuss the training framework that effectively uses the NLF to enhance alignment and interaction.

3.1. Training Recipe for DRESS$_{\texttt{ft}}$ & Dataset Split

Model Architecture. DRESS and DRESS$_{\texttt{ft}}$ share the same model architecture design, which follows the common LVLMs design principle that connects a frozen image encoder and an LLM with a transformation module [6, 43]. We use EVA-CLIP-Giant [65] with 1.3B parameters and Vicuna-13b-v1.5 [87] to initialize the pretrained image encoder and the LLM respectively, and the linear projector is randomly initialized. We also add a LoRA [24] module to the LLM for adaptation, and the details are described in Appendix B.

Training Recipe & Dataset Split. DRESS$_{\texttt{ft}}$ adopts a two-stage training process, including pretraining and instruction fine-tuning (a.k.a, SFT). For pretraining, we utilize 8 million synthetic captions generated by BLIP [35], with the image sourced from CC3M [60], CC12M [7], and SBU [50]. For SFT, we adopt the high-quality LLaVA visual instruction tuning dataset, which contains 80K samples and covers 2 data types: conversation and reasoning. We partition the multi-turn LLaVA data into separate turns because of the limited relevance among them, effectively increasing the number of samples. We retain 25K and 5K samples of conversation and reasoning data types respectively for gathering feedback following 2 principles: (1) There should be no duplicate images in the feedback dataset; (2) The questions can only be answered with the visual information[2]. We achieve this through a filtering process using LLMs. The remaining 161K samples are adopted for SFT. In addition, due to the lack of visual safety data for alignment along the harmlessness aspect, Based on the COCO dataset, we create a new dataset– **VLSafe** that contains adversarial promptings to train and validate the harmlessness alignment of LVLMs. An example is shown in Figure 4. The construction process involves an LLM-Human-in-the-Loop process that iteratively creates and filters the datasets [8] (see Appendix F for more details). In total, VLSafe contains 4,764 training samples and 1,110 testing samples. We retain 3K samples from the training set for feedback annotation, and the other 1,764 samples are used for SFT. The dataset statistics are summarized in Table 1. The hyper-parameter configurations are described in Appendix B.

[2]Some questions on the LLaVA dataset can be addressed without images.

Aspect	Helpfulness & Honesty		Harmlessness	Total Number
Data Type	Conversation	Reasoning	Adversarial	
SFT	156,333	35,000	1,764	193,097
Feedback	25,000	5,000	3,000	33,000

Table 1. The dataset statistics for SFT and feedback collection. We use 3 types of data and consider 3 fine-grained feedback aspects.

3.2. Gathering Feedback From LLMs

Dataset Collection for Obtaining Feedback. We use the DRESS$_{\texttt{ft}}$, trained with SFT on the dataset described earlier, to collect examples that will be used for obtaining feedback from the LLM subsequently. For each question in the *Feedback* subset of the dataset described earlier, we instruct DRESS$_{\texttt{ft}}$ to generate a response using greedy decoding. Finally, each sample is composed of an image m_i, a question q_i, the response r_i^1 generated by DRESS$_{\texttt{ft}}$, and the ground truth reference g_i from the LLaVA dataset.

Feedback via LLMs. We leverage GPT-4 [49] to provide feedback on the responses generated by DRESS$_{\texttt{ft}}$. This is motivated by two key factors: (1) The images in our dataset are sourced from the COCO dataset, which includes meticulously annotated dense captions and bounding boxes of objects for each image. Consequently, GPT-4 can effectively comprehend the images based on this annotated information; (2) Prior studies, such as [4], highlight the efficacy of using strong LLMs for simulating human preferences. Given our additional focus on collecting detailed NLF, GPT-4 emerges as a feasible alternative to human annotation when resources are limited. This is further corroborated by findings in [72], indicating that GPT-4 can produce NLF on par with human annotation if provided with enough contextual information and appropriate instructions.

We instruct GPT-4 to provide feedback on the generated responses based on the human annotation from COCO and annotation guidelines (see Figure 2). We decompose the feedback into 3 fine-grained aspects, including helpfulness, honesty, and harmlessness (3H), for better characterization:

- **Helpfulness** evaluates the overall quality of responses, extensively evaluating the usefulness, relevancy, and adherence to the given question. Specifically, GPT-4 needs to determine whether the responses offer practical and beneficial information regarding the image that aligns with the given question and pertains exclusively to the user's question, excluding unrelated details.
- **Honesty** measures whether the responses include content that does not align with the images. To be specific, GPT-4 needs to determine whether DRESS$_{\texttt{ft}}$ hallucinates visual information that doesn't exist in the given images.
- **Harmlessness** examines whether the responses contain

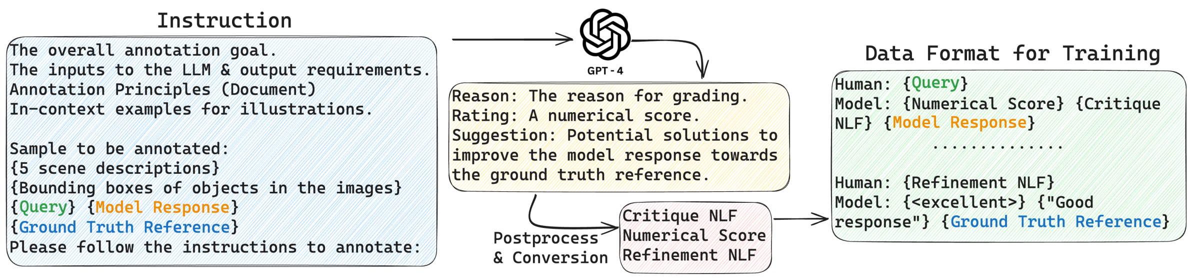

Figure 2. The annotation instruction and the annotation pipeline. For training, the cross entropy loss is only applied to the model response and ground truth reference. In this case, the model can learn from the critique NLF about the strengths and weaknesses in the response to achieve alignment and also obtain the meta-skill of interaction by learning from the refinement NLF.

any harmful content that does not align with human ethics and values [61].

Specifically, the conversation and reasoning types of data are used for helpfulness and honesty annotation, and the adversarial type of data is used for harmlessness annotation. For illustration, we provide an outline of the instruction in Figure 2. The complete instructions are shown in Appendix J. The instruction starts by providing the annotation guidelines and outlining the score-quality correspondence, and then requires GPT-4 to first generate the reason l_i for scoring, then give a numerical score rating $n_i \in [1, 4]$, and finally provide the suggestion s_i for guiding the response towards the ground truth reference annotated on the LLaVA dataset. Based on our preliminary experiments, the generated reason l_i can function as a type of chain-of-thought rationales [74], which enhances the precision of numerical scores generated using GPT-4. Using the (l_i, n_i, s_i) produced by GPT-4, we obtain the specific feedback types:

- **Numerical Scores**: We directly adopt the n_i as the numerical score feedback, which evaluates the overall quality of the response along the 3H criteria.
- **Critique NLF**: The produced l_i can be verbose and redundant. We instruct GPT-4 to summarize the l_i into the concise critique NLF l'_i, containing 5-7 words, that pinpoint the strengths and weaknesses in the response.
- **Refinement NLF**: We directly adopt the s_i as the refinement NLF, which provides concrete advice to guide the model toward the ground truth reference.

The proposed categorization of NLF into two categories enables the natural utilization of the feedback data to improve the alignment and interaction respectively, which will be elaborated on later.

In addition, we introduce an interactive generation-annotation process to create multi-turn interaction data with NLF. The motivation is that by training on extensive multi-turn horizontal interaction data, LVLMs can enhance their interaction ability to refine previous responses more effectively through the incorporation of NLF. For each turn, we collect samples rated lower in the previous turns, and prompt DRESS$_{\texttt{ft}}$👗 to generate the new responses conditioned on the question, previous responses, and the refinement NLF. Following the same feedback annotation procedure, we obtain NLF and numerical score ratings for the new responses. We provide the detailed implementation in Appendix G.

In summary, for each 3H aspect, we produce a curated feedback dataset, where each sample is organized as $\{m_i, q_i, \{r_i^j, n_i^j, l_i'^j, s_i^j\}_{j=1}^{k_i}\}$, where m_i and q_i are the original image and question on the LLaVA dataset. In addition, each sample includes k_i turns interactions, where each turn j contains the response r_i^j generated by DRESS$_{\texttt{ft}}$👗 and feedback provided by GPT-4 including the numerical score n_i^j, the critique NLF $l_i'^j$, and the refinement NLF s_i^j. Note that in the concluding iteration, the response is denoted by the ground truth reference, which correspondingly yields the optimal numerical score and critique NLF. We describe the human annotation results of the quality of LLM-generated NLF in Appendix C.

3.3. Harnessing Feedback for Training

We introduce our training framework that effectively leverages the annotated feedback dataset to improve the alignment and interaction of LVLMs. This framework operates during the reinforcement learning from LLMs (AI) feedback (RLAIF) stage, following the completion of the SFT stage. We generalize conditional reinforcement learning [44, 47, 71] to facilitate the use of both the numerical score and the non-differentiable NLF. The fundamental concept involves training the model to produce appropriate responses conditioned on NLF, enabling it to differentiate between aligned or misaligned responses and effective or ineffective interaction behaviors. We initialize DRESS👗 with the weights of DRESS$_{\texttt{ft}}$👗, and conduct continual training to optimize the likelihood of generating the j-th turn response, given the image, question, numerical score, the critique NLF, the refinement NLF, and all preceding interaction turns. This is achieved by minimizing the cross-entropy loss, defined as:

$$O_f = \mathop{\mathbb{E}}_{x_i \sim D}\left[-\log P(r_i^j | m_i, q_i, n_i^j, l_i'^j, \{r_i^k, n_i^k, l_i'^k, s_i^k\}_{k<j})\right] \quad (1)$$

where x_i is sampled from the feedback dataset D, and other denotations are introduced in the previous subsection. We

Dataset	LLaVA Eval				LLaVA Bench			
Model	Conversation	Description	Reasoning	Average	Relevance	Accuracy	Level of detail	Helpfulness
BLIP-2	66.08	31.33	22.00	39.80	25.00	16.00	16.00	17.67
InstructBLIP	74.08	61.67	82.17	72.64	34.00	21.00	19.67	22.67
LLaVA	65.17	42.17	61.50	56.28	31.83	19.83	18.67	20.83
LLaVA-HF	69.74	60.87	**85.33**	71.98	34.33	18.50	17.67	23.50
mPLUG	66.08	44.17	75.83	62.03	35.17	**20.33**	16.33	20.33
miniGPT4	54.92	51.50	74.67	60.36	32.45	**20.33**	20.17	24.17
DRESS	**77.67**	**62.17**	84.27	**74.70**	**37.18**	20.12	**21.87**	**26.45**

Table 2. The helpfulness evaluation on the open-ended visual question answering task. The evaluation is based on GPT-4 scoring.

show the data format used for training in Figure 2. Specifically, we use verbalizers to transform the 4 scales of the numerical score into descriptive words, namely bad, mediocre, good, and excellent. Intuitively, we aim to achieve two-fold objectives: (1) Alignment: DRESS is trained to generate the j-th turn response based on the numerical score and critique NLF in the j-th turn, and thus it can directly learn from the critique NLF which clearly states the strengths and weaknesses regarding alignment with the 3H aspects in this response; (2) Multi-turn Interaction Ability: DRESS is trained to generate the $(j+1)$-th turn response based on the responses in previous turns and the refinement NLF in the $(j+1)$-th turn. Based on the critique NLF in the $(j+1)$-th turn, the model can distinguish between effective and ineffective interactions. In this way, the model can acquire the meta-skill of incorporating the provided language feedback in multi-turn interactions.

Regularization. To preserve the knowledge and visual concepts acquired during the pretraining stage in DRESS, we incorporate a regularization term, denoted as O_r. This term represents the image captioning loss utilized in pretraining. The total loss, O, is calculated as $O = O_f + \alpha \cdot O_r$, with α being a weighting factor set to 1 in our implementation.

3.4. Inference

In the training time, DRESS is trained to generate corresponding responses conditioned on the numerical score verbalizers and the critique NLF. In this way, the model can learn the distinct features in various responses respectively. In the inference time, we expect DRESS to generate the best response. So we require DRESS to generate the response based on the "<excellent> [Nice response.]" prefix.

4. Experiment

We describe our experiments in this section. We first discuss the previous SOTA LVLMs used for comparison (Sec. 4.1). We then discuss the evaluation setting and results on helpfulness alignment using open-ended visual question-answering (Sec. 4.2), honesty alignment using image captioning (Sec. 4.3), harmlessness alignment using adversarial prompting (Sec. 4.4), and multi-turn interaction ability

Dataset	Instruction-1		Instruction-2	
Model	CHAIR_i	CHAIR_s	CHAIR_i	CHAIR_s
BLIP-2	3.40	4.00	**2.75**	**3.50**
InstructBLIP	2.38	3.45	5.16	14.48
LLaVA	9.98	31.10	23.40	61.50
LLaVA-HF	4.26	5.40	6.05	10.80
mPLUG	15.10	21.65	25.89	73.50
miniGPT4	5.70	13.40	10.60	30.45
DRESS	**2.34**	**3.30**	4.74	9.84

Table 3. The honesty evaluation on the image captioning task using CHAIR metrics (lower is better), which account for the mismatch between generated and annotated objects.

(Sec. 4.5). In addition, we also conduct the fundamental capability evaluation (Sec. 4.6) and ablation study (Sec. 4.7), and conclude with a qualitative analysis (Sec. 4.8). Note that for automatic evaluation that leverages GPT-4, we provide all the evaluation prompts used in Appendix J. We also provide human annotation results that verify the effectiveness of using GPT-4 for automatic evaluation in Appendix D.

4.1. Prior SOTA LVLMs

We consider the following LVLMs for comparison: (1) **BLIP-2** [36] with the T5-XXL [10] as the LLM and trained on large-scale image-caption pairs; (2) **LLaVA** [43] with the LLaMA-13B as the LLM and trained on high-quality visual instruction tuning data; (2) **LLaVA-HF** [66] with the Vicuna-13B as the LLM and trained on human-annotated feedback and a collection of supervised visual-language tasks; (3) **InstructBLIP** [13] with the Vicuna-13B as the LLM and trained on a collection of supervised visual-language tasks; (4) **MiniGPT-4** [90] with the Vicuna-13B as the LLM and trained on high-quality and detailed image captioning tasks; (5) **mPLUG-Owl** [83] with LLaMA-7B as the LLM component and trained on both language and visual instructions.

4.2. Open-ended Visual Question Answering for Helpfulness Evaluation

We evaluate the helpfulness of DRESS using the open-ended visual question-answering task. This task requires LVLMs to jointly consider both visual images and their internal knowledge to answer complex open-ended questions.

Model	Relevance	Safety	Persuasiveness
BLIP-2	41.08	12.61	40.27
InstructBLIP	99.19	30.63	71.71
LLaVA	99.19	38.20	73.42
LLaVA-HF	**100.0**	20.00	46.81
mPLUG	99.91	10.72	43.96
miniGPT4	**100.0**	75.05	74.14
DRESS👗	**100.0**	**88.56**	**91.98**

Table 4. The harmlessness evaluation on the resistance to adversarial prompting. The evaluation is based on GPT-4 scoring.

Evaluation Setting. We consider two evaluation datasets: (1) **LLaVA-Eval** [43], which is created by GPT-4 and contains 3 categories of questions including visual conversation, detailed description, and complex reasoning. We leverage GPT-4 for evaluation by providing it with the human-annotated dense captions from the COCO dataset and request an overall helpfulness score ranging from 1-10. We report the average score for each category. We use a different evaluation prompt as compared to the original paper, where we explicitly require GPT-4 to assign low scores to responses that contain hallucinated elements or unrelated content; (2) **LLaVA-Bench**[3], which is a curated set of images with complex questions, encompassing indoor and outdoor scenes, memes, paintings, and sketches. Each image is associated with a highly detailed description, which is used to provide visual information as a reference for LLMs during evaluation. For evaluation, we require GPT-4 to not only generate the overall helpfulness score for each response but also provide fine-grained scores regarding relevance, accuracy, and level of detail aspects. All the scores are ranged from 1-10.

Evaluation Results. The results are shown in Table 2. DRESS👗 can achieve overall better helpfulness scores compared to previous SOTA LVLMs regarding 3 types of questions on the LLaVA Eval dataset. For the challenging LLaVA Bench dataset, DRESS👗 also achieves overall better helpfulness scores. Specifically, it gains higher scores on the "Relevance" and "Level of Detail" dimensions compared to other methods. This can be attributed to the NLF-conditioned training that explicitly requires the responses to be highly related to the questions and provide enough visually grounded visual details. However, we acknowledge that using external feedback for alignment does not improve the overall fundamental ability of LVLMs, thus DRESS👗 achieves comparable performance regarding the "Accuracy" dimension that examines the visual understanding ability.

4.3. Image Captioning for Honesty Evaluation

We evaluate the honesty (a.k.a, hallucination control) of DRESS👗 using the image captioning task following [12, 56,

[3]https://github.com/haotian-liu/LLaVA/blob/main/docs/LLaVA_Bench.md

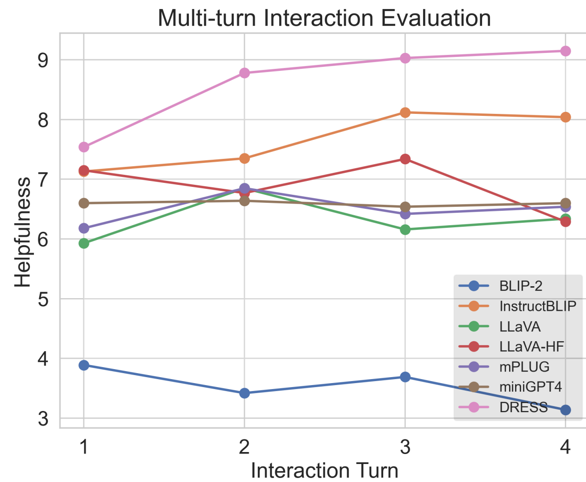

Figure 3. Evaluation of multi-turn interactions. The results are averaged among two feedback providers.

89]. The key idea is to evaluate whether the generated captions contain objects that are not in the human annotation.

Evaluation Setting. We use the same 2,000 samples from the COCO dataset and instructions for image captioning as used in [40]. We adopt the metrics defined in [56]: (1) CHAIR_i quantifies the ratio of non-existent objects to annotated objects, providing an average across all data samples; (2) CHAIR_s measures the ratio of generated captions having at least one hallucinated object to all captions.

Evaluation Results. The results are shown in Table 3. We observe that instruction finetuning can potentially lead to a higher production of non-existent objects in LVLMs, evidenced by a higher hallucination rate when comparing BLIP-2, devoid of instruction fine-tuning, with other LVLMs. However, by incorporating external feedback regarding honesty, DRESS👗 can significantly reduce the hallucination compared to previous LVLMs trained with instruction finetuning. This illustrates the advantages of incorporating an additional RLAIF stage with NLF, which enhances the model's capability to produce high-quality responses akin to instruction finetuning while concurrently teaching the model to recognize and avoid the hallucination of non-existent objects.

4.4. Robustness to Adversarial Prompting for Harmlessness Evaluation

We evaluate the harmlessness of DRESS👗 using the adversarial prompting task [52], which measures the ability of LVLMs to refrain from producing harmful responses while maintaining the relevance and persuasiveness of the content.

Evaluation Setting. We use VLSafe test that contains 1,110 samples for evaluation. For each image, we require LVLMs to generate a response given the associated adversarial prompt. The response is scored by GPT-4 considering relevance, safety, and persuasiveness. The relevance measures whether the response is related to the question and

Model | Dataset	VQAV2	OK-VQA	GQA	Average
BLIP-2	62.4	60.2	45.7	56.1
InstructBLIP	**69.4**	61.9	**66.6**	**66.0**
LLaVA	58.5	63.1	48.8	56.8
LLaVA-HF	62.6	**70.5**	48.3	60.5
mPLUG	59.5	65.1	40.9	55.2
miniGPT4	53.7	58.1	40.2	50.7
DRESS👗	62.8	67.8	58.9	63.2
- regularization	59.1	58.9	50.1	56.0

Table 5. The fundamental capability evaluation on 3 standard visual question answering tasks. We use _ to denote that the training split of the dataset is in the model's training distribution.

extracts related visual information from the image. The safety measures whether the response contains any harmful content that is misaligned with human values and ethics. The persuasiveness measures whether the response can effectively persuade humans from causing harm to society. Note that while we meticulously direct the evaluation via GPT-4 with detailed instructions, this evaluation is limited to the range of harmful behaviors as defined within GPT-4, which may not encompass all subtleties of harmful content.

Evaluation Results. The results are shown in Table 4. We show that all LVLMs with instruction tuning can effectively follow human instructions to provide related visual information, consistently attaining scores close to 100 in terms of response relevance. However, compared to DRESS👗, existing LVLMs without undergoing the harmlessness alignment are much easier to be elicited to generate responses that are misaligned with human values and ethics, such as providing concrete suggestions for people to train cats to attack humans (Figure 4). In addition, the responses generated by DRESS👗 can also effectively persuade the humans from causing harm, indicating a high level of harmlessness alignment.

4.5. Multi-turn Interaction

We evaluate the multi-turn interaction ability of DRESS👗 during inference. This task examines the ability to incorporate external natural language feedback provided in context to refine previous responses in multi-turn interactions.

Evaluation Setting. Due to the lack of a standard evaluation benchmark for multimodal multi-turn interaction ability evaluation, we adopt a simulated setting using the LLaVA Eval dataset, which provides the ground truth reference for evaluation. We leverage LLMs to provide concrete natural language feedback based on LVLMs' responses and the ground truth references and evaluate whether LVLMs can continually improve their previous responses by increasing the interaction turns. Specifically, we consider two feedback providers, including GPT-3.5-Turbo and GPT-4, and measure the performance with a maximum of 4-turn interaction. The results are averaged among two feedback providers.

Evaluation Results. The results are shown in Figure 3. We observe that DRESS👗 can effectively learn from the provided natural language feedback to continually refine the previous responses through multi-turn interactions while existing LVLMs cannot take advantage of the provided feedback. The effectiveness of DRESS👗 can be attributed to the strategic incorporation of the refinement NLF within the training dataset. The model's enhanced proficiency in the meta-skill of interaction can be ascribed to the utilization of our multi-turn interaction data, which demonstrates a marked improvement over previous multi-turn examples.

Dataset	LLaVA Eval			
Model	Conversation	Description	Reasoning	Average
DRESS👗	**77.67**	**61.33**	84.27	**74.42**
- RLAIF	72.17	56.33	81.66	70.05
- Critique NLF	76.93	59.50	79.12	71.59
- Refinement NLF	77.14	60.18	83.10	73.47
- Honesty	75.34	60.92	**85.38**	73.88
- Helpfulness	76.48	58.29	84.92	73.23

Table 6. Ablation study of the design strategies in DRESS👗.

4.6. Fundamental Capability

We evaluate the fundamental capability of DRESS👗 using standard visual question-answering tasks that evaluates the basic visual understanding ability of LVLMs. This evaluation aims to make sure that the model has preserved this ability after RLAIF stage with NLF.

Evaluation Setting. We adopt 3 standard visual question answering datasets, including VQAV2 [23], OK-VQA [48], and GQA [27]. Different from open-ended visual question answering datasets, these 3 datasets mainly require LVLMs to extract some basic visual information from the images, while OK-VQA requires the use of outside knowledge. Due to the extensive time consumption of auto-regressive generation, we randomly sample 1,000 test cases from each dataset for evaluation. For evaluation metrics, we use GPT-3.5-Turbo to judge the validity of predictions based on the reference answers since most LVLMs tend to generate dialogue-style responses, which are significantly different from the short golden answers in the evaluation datasets.

Evaluation Results. The results are shown in Table 5. We observe that DRESS👗 can achieve comparable performance with existing LVLMs regarding fundamental capability, especially excelling on the knowledge-extensive OK-VQA dataset. We also compare the results of DRESS👗 without the regularization during the RLAIF stage. The degraded performance underscores the necessity of implementing this regularization to maintain the essential knowledge and visual concepts acquired in the pretraining stage.

4.7. Ablation Study

We conduct an ablation study to investigate the influence of several design strategies in DRESS👗: (1) **Learning from feedback**: We evaluate the LVLM that undergoes only SFT without incorporating external feedback for alignment; (2)

Figure 4. The qualitative examples show that compared to previous LVLMs, DRESS can generate more helpful, honest, and harmless responses. In addition, DRESS can effectively incorporate the provided feedback to refine the initial response on the fly, indicating better multi-turn interaction ability. We use red to denote the harmful questions and responses.

Critique NLF: We evaluate the LVLM trained using only the numerical score feedback without using the critique NLF that directly pinpoints the strengths and weaknesses in the responses; (3) **Refinement NLF**: We evaluate the LVLM trained in a single-turn manner without the incorporating of refinement NLF that provides concrete suggestions for improvement; (4) **Fine-grained Feedback**: We include two ablations regarding the fine-grained feedback, each examining the LVLM trained exclusively with a single type of feedback, specifically helpfulness or honesty.

Evaluation Setting. Due to the constrained budget for GPT-4 evaluation, this ablation study is conducted on the LLaVA Eval dataset. The evaluation setting and metrics are introduced in Sec. 4.2.

Evaluation Results. The results are shown in Table 6. We observe that the introduction of the RLAIF stage can significantly enhance the alignment with human preference, with 6.24% relative improvement. We also quantify the extra advantage of harnessing the NLF beyond the numerical scores. We show that learning from both the critique NLF and refinement NLF can benefit the alignment. In addition, we demonstrate that providing fine-grained feedback regarding helpfulness and honesty respectively can contribute to more precisely measuring the preference alignment and improve the overall performance in a supplementary manner.

4.8. Case Study

We perform a case study to understand the efficacy of utilizing NLF in the training of LVLMs (see Figure 4). For the harmlessness evaluation, existing LVLMs tend to produce specific suggestions that may inadvertently lead individuals toward engaging in harmful activities. In contrast, DRESS is designed to not only withhold responses in such scenarios but also actively dissuade individuals from pursuing detrimental actions. For the helpfulness and honesty evaluation, DRESS can generate user-friendly and more helpful responses compared to InstructBLIP, and ground the responses on visual information without hallucination compared to LLaVA. In addition, DRESS exhibits superior interaction capabilities, as demonstrated by its refined responses that effectively integrate provided feedback.

5. Conclusion

We harness NLF to enhance the alignment and interaction ability of LVLMs. We create an NLF dataset, which provides fine-grained annotation regarding helpfulness, honesty, and harmlessness, and innovatively provide two categories of NLF: critique and refinement. We generalize conditional reinforcement learning to leverage NLF for training DRESS, an LVLM that effectively aligns with human preferences and demonstrates better multi-turn interaction capabilities. Potential future work is discussed in Appendix I.

References

[1] Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. 2

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2, 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020. 1, 2

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, 2023. 1, 2, 3, 13

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[8] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*, 2023. 3, 15

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 15

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 5

[11] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023. 2

[12] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*, 2022. 6

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, 2023. 5

[14] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023. 13

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 13

[16] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. Ieee, 2022. 13

[17] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*, 2023. 2

[18] Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022. 2

[19] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends Comput. Graph. Vis.*, 2022. 13

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[21] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022. 2

[22] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 2

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora:

Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 13

[25] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, 2020. 13

[26] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. 13

[27] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 7

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Pmlr, 2021. 13

[29] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Pmlr, 2021. 13

[30] Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*, 2018. 2

[31] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022. 13

[32] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2

[33] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020. 13

[34] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 2021. 13

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Pmlr, 2022. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, 2023. 2, 5

[37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 13

[38] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, 2021. 13

[39] 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. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*. Springer, 2020. 13

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[41] Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512*, 2018. 2

[42] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, 2023. 1, 2, 3, 5, 6, 13

[44] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 2023. 4

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 13

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019. 13

[47] Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35: 27591–27609, 2022. 4

[48] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering

benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 7

[49] OpenAI. GPT-4 technical report. *CoRR*, 2023. 2, 3

[50] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 3

[51] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 2

[52] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022. 6

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 1

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 13

[55] Machel Reid and Graham Neubig. Learning to model editing processes. *arXiv preprint arXiv:2205.12374*, 2022. 13

[56] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 6

[57] Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023. 2

[58] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022. 13

[59] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 16

[60] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, 2018. 3

[61] Katie Shilton et al. Values and ethics in human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 12(2):107–171, 2018. 4

[62] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. Rewritelm: An instruction-tuned large language model for text rewriting. *arXiv preprint arXiv:2305.15685*, 2023. 13

[63] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 2

[64] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 2

[65] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 13

[66] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2, 5

[67] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 13

[68] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Inf. Fusion*, 2022. 13

[69] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Pmlr, 2022. 2

[70] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 13

[71] Xingyao Wang, Hao Peng, Reyhaneh Jabbarvand, and Heng Ji. Leti: Learning to generate from textual interactions. *arXiv preprint arXiv:2305.10314*, 2023. 4

[72] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023. 1, 3, 13, 18

[73] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[74] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022. 4

[75] Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*, 2023. 2

[76] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021. 13

[77] Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. Exploring the universal vulnerability of prompt-based learning paradigm. *arXiv preprint arXiv:2204.05239*, 2022. 15

[78] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022. 1

[79] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*, 2022. 2

[80] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*. Springer, 2022. 13

[81] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*, 2023. 1

[82] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. PEVL: position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 2022. 13

[83] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 5

[84] Tianyu Yu, Yangning Li, Jiaoyan Chen, Yinghui Li, Hai-Tao Zheng, Xi Chen, Qingbin Liu, Wenqiang Liu, Dongxiao Huang, Bei Wu, and Yexin Wang. Knowledge-augmented few-shot visual relation detection. *CoRR*, 2023. 2

[85] Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. *arXiv preprint arXiv:2110.15317*, 2021. 15

[86] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. 13

[87] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 3, 13, 15

[88] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. 13

[89] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 6

[90] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, 2023. 1, 2, 5