

# Appendix - Learning to Correction: Explainable Feedback Generation for Visual Commonsense Reasoning Distractor

## A Benchmark Construction

In this section, we will describe the prompts used to generate distractors and feedback from GPT-4. Subsequently, we will present a web page we developed for manual filtering of distractors to remove those of low quality.

### A.1 Prompt for GPT-4

Given the question, answer, place, object boxes, and events depicted in the image, we utilize a language-only GPT-4 to generate distractors and feedback. Specifically, we first utilize GPT-4 to identify the educational level of the given question based on Bloom’s taxonomy, as shown in Table 6. The prompts used for distractors and feedback generation are detailed in Table 7 and 8, respectively.

### A.2 Annotation Web Page

Fig. 1 shows the web page used for manual distractor filtering.

## B More Experimental Details

In this section, we provide more implementation details and experimental results of our PEIFG model and the baseline models.

### B.1 Instruction for Q-Former

Fig. 2 shows the diverse instructions for Q-Former, which are used for alignment between visual features and learnable query tokens. The diversity of instructions demonstrates the same meaning with variations in natural language expression. Specifically, for each input sample, we randomly select one from the list of instructions for alignment.

### B.2 Refinement Details

For the refinement step, we randomly sample 800 instances from the training set and employ a top-p sampling strategy, where the temperature is set to 0.8 and  $p = 0.95$  to generate four candidate generated feedback for each input sample. Given the question, ground truth distractor, feedback, and the generated feedback, we design five diagnostic questions for GPT-4 to ascertain whether the generated feedback meets the specified criteria. These questions and a diagnostic example are shown in Table 9. Specifically, for each diagnostic question, feedback that meets the criteria is awarded 1 point, otherwise 0 point. Therefore, the final diagnostic score for each generated feedback ranges from 0 to 5. We rank the generated feedback as pairs based on the diagnostic score and utilize the direct preference optimization loss to further fine-tune the LLM.

### B.3 Experiment

We conduct experiments to investigate the sensitivity of the hyperparameters, i.e., pool size  $S$ , number of selected expert prompts  $K$ ,  $\lambda_1$  and  $\lambda_2$  of the loss function. Table 2, 1 and 3 demonstrate the experimental results. Moreover, we also investigate the influence

Table 1: Experimental results of different pool size  $S$  on feedback generation. Bold: the maximum value in the column.

$S$	B@4	METEOR	CIDEr	BERTScore
5	16.86	34.94	30.01	69.75
10	<b>17.69</b>	<b>35.77</b>	32.10	<b>71.69</b>
15	17.15	35.05	<b>32.64</b>	70.64

Table 2: Experimental results of different  $K$  settings on feedback generation. Bold: the maximum value in the column.

$K$	B@4	METEOR	CIDEr	BERTScore
1	17.32	35.26	30.12	70.62
3	<b>17.69</b>	<b>35.77</b>	<b>32.10</b>	<b>71.69</b>
5	17.12	35.03	30.25	70.62

Table 3: The sensitivity of  $\lambda_1$  and  $\lambda_2$  on feedback generation. Bold: the maximum value in the column.

$(\lambda_1, \lambda_2)$	B@4	METEOR	CIDEr	BERTScore
(0.1, 0.1)	<b>17.69</b>	<b>35.77</b>	32.10	<b>71.69</b>
(0.3, 0.3)	17.54	35.33	<b>32.65</b>	71.21
(0.5, 0.5)	17.29	35.17	32.09	71.10

Table 4: The influence of expert correlation loss  $\mathcal{L}_{cor}$  on feedback generation. Bold: the maximum value in the column.

Models	B@4	METEOR	CIDEr	BERTScore
PEIFG	<b>17.69</b>	<b>35.77</b>	<b>32.10</b>	<b>71.69</b>
PEIFG w/o Cor	16.08	33.86	29.26	70.53

of the expert correlation loss  $\mathcal{L}_{cor}$ , which aims to reduce the correlation among the expert prompts in the pool and enforce each of the expert prompt to maintain its special feature. The experiment results are shown in Table 4.

### B.4 Implementation of Baseline Models

For NLX-GPT and KICNLE, we employ full fine-tuning approach on the VCR-DF dataset. We optimize them with an initial learning rate of  $2e-5$ . For the large multimodal models (i.e., BLIP-2, InstructBLIP, VisualGLM, LLaVA-v1.5 and CogAgent), we utilize the LoRA layers into the self-attention mechanisms of their LLMs and the lora rank of BLIP-2, InstructBLIP, VisualGLM and LLaVA-v1.5 is set to 16, while it is set to 32 for CogAgent. During training, we use Adam optimizer with an initial learning rate of  $8e-5$  and cosine scheduler.

### Distractor Filtering

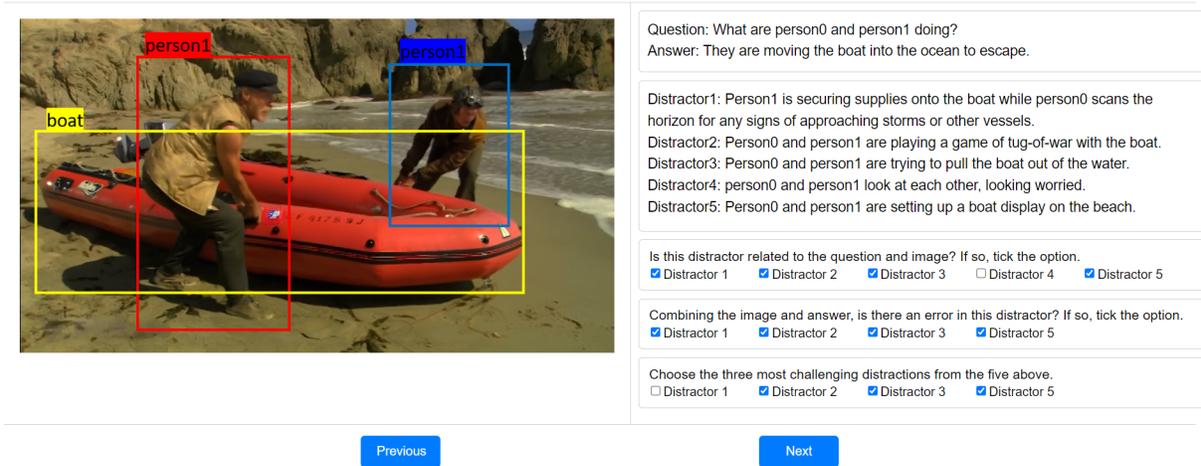


Figure 1: The web page for manual filtering.

- Instructions for Q-Former-----
- Analyze the image, question, answer, and distractor provided to pinpoint the error location. Offer a detailed explanation based on the context.
  - Identify the mistake in the context of the image, question, answer, and distractor. Elaborate on the reasons for the error with thorough rationale.
  - Locate the specific error related to the image, given the question, answer, and distractor. Provide a comprehensive explanation for the error.
  - Pinpoint the error in relation to the image with the given question, answer, and distractor. Give a detailed account of the error's cause.
  - Detect the error within the context of the image and the elements of the question, answer, and distractor. Offer a detailed rationale for the error.

Figure 2: The list of instructions for Q-Former

Table 5: Human evaluation results of baselines and our model. **Bold: the maximum value in the column.**

Method	Flu	Help	Loc	Div	Rel
InstructBLIP	1.08	0.59	0.42	0.57	0.31
VisualGLM	1.66	0.62	0.66	0.89	0.54
CogAgent	1.75	0.81	0.75	1.32	0.79
GPT-4V	<b>1.78</b>	0.77	0.89	1.76	<b>0.93</b>
PEIFG	1.74	<b>0.87</b>	<b>0.92</b>	<b>1.80</b>	0.88

### B.5 Human Evaluation

We randomly select 200 samples from the VCR-DF test set and 5 participants with good English education are asked to evaluate these samples based on the following criteria: Fluency (**Flu**) mainly reflects the grammatical correctness and fluency of the generated sentence. Helpfulness (**Help**) measures whether the generated feedback aids human in more effectively reaching to the correct answer. Logical consistency (**Loc**) refers to the logical coherence between feedback and input information. Diversity (**Div**) represents the variety of content within the three distractors generated for each

sample. Relevance (**Rel**) evaluates whether the distractor is related to the image, question and educational level. The **Flu** and **Div** are scored on a range from 0 to 2 (higher values indicate greater fluency, diversity and fewer grammatical errors), while others are binary values. Table 5 shows the human evaluation results with five metrics. We conclude that: **i)** For fluency **Flu**, the results illustrate that our PEIFG model achieves competitive results with much larger LMMs (i.e., CogAgent and GPT-4V) to generate fluent and smooth sentences. **ii)** The metrics **Help** and **Loc** are pivotal in our human evaluation. The results demonstrate PEIFG superiority of the multimodal instruction and refinement operation for LLM to generate feedback that not only aids in guiding individuals toward the correct answers but also maintains a high degree of logical coherence with the input information. Meanwhile, notice that feedback generated by GPT-4V tends to be overly verbose and may include irrelevant information, which can distract human from choosing the correct answer. **iii)** The metrics **Div** and **Rel** indicate that the distractors generated by our model outperform those from most baselines in both diversity and relevance to the input information and educational level, attributed to the specificity of selected expert prompts and integrative visual extraction strategy.

**Table 6: The prompt for GPT-4 to generate educational level.**


---

**Prompt:** You are an accurate NLP annotator. Bloom’s Taxonomy categorizes cognitive skills into six levels (i.e., Remember, Understand, Apply, Analyze, Evaluate, and Create), guiding educators in assessing and developing critical thinking abilities. Analyze the cognitive processes involved in both the question and answer. Identify the Bloom’s Taxonomy level that best corresponds to each.

Question: Why are person0 and person2 , and person3 pulling their chairs out at the same time?

Answer: Person0, person2 and person3 were waiting for person1’s signal to sit.

---

**GPT-4:**  
Educational level: Analyze

---

**Table 7: The prompt for distractor generation.**


---

**Prompt:** You are an accurate NLP annotator. You need to generate the distractors based on the provided visual content, including Event, Object, Place, Question, Answer and Educational Level. First, accurately identify the knowledge points and potential misconcepts involved in the question. Then generate three corresponding challenging distractors based on the misconcepts that easily lead to errors. The series of data I have provided are as follows:

Event: Person0 is pulling a chair out at a meeting, person1 is wearing a military uniform, person1 already appears to be in the process of sitting and is at the head of the table, indicating his status and leadership over the others

Place: In a conference room.

Object: Person0: [490,194,790,582], Person1: [1001,231,1244,632] ...

Question: Why are person0 and person2 , and person3 pulling their chairs out at the same time?

Answer: Person0, person2 and person3 were waiting for person1’s signal to sit.

Educational Level: Analyze.

---

**GPT-4:**  
Distractor1: Person0, person2, and person3 are pulling chairs out to leave the meeting.  
Distractor2: Person0, person2, and person3 are pulling chairs out to rearrange the conference room.  
Distractor3: Person0, person2, and person3 are pulling chairs out to make room for more people.  
Distractor4: Person0, person2, and person3 are pulling chairs out because they are uncomfortable.  
Distractor5: Person0, person2, and person3 are pulling chairs out to challenge person1’s authority.

---

**Table 8: The prompt for feedback generation.**


---

**Prompt:** You are an accurate NLP annotator. You need to generate the feedbacks based on the provided visual content, including Event, Place, Question, Answer, Educational level and Distractor. First, accurately identify the knowledge points and potential misconcepts involved in the question. Then analyze five corresponding challenging distractors based on the misconcepts that easily lead to errors. Given the errors in the distractor, you point out the error locations and explain the reasons for the errors as feedback. The feedback includes two aspects: the concept of confusion involved in the distractor, and an explanation. Explanation consists of simple declarative sentences, without complex structures and words. The series of data I have provided are as follows:

Event: person1 sits in a chair turning towards the people behind him, person8 walks towards the bartender at the bar, person5 is an attractive woman and the smile person2 is giving her suggests attraction,

Place: A crowded bar.

Question: Why is person2 leaning in the way that he is?

Answer: Person2 is showing attraction and flirting with person5.

Educational level: Analyze

Distractor1: Person2 is leaning to hear the bartender better.  
Distractor2: Person2 is leaning because he is tired and wants to rest.  
Distractor1: Person2 is leaning to avoid a spill from the wineglass.

---

**GPT-4:**  
Feedback1:  
Misconception: Incorrectly assuming person2’s posture is for hearing better.  
Explanation: Person2’s lean is a body language sign of attraction, not for better hearing.  
Feedback2:  
Misconception: Mistakenly thinking person2 is leaning due to tiredness.  
Explanation: Leaning is not a sign of tiredness here; it’s a flirtatious gesture towards person5.  
Feedback3:  
Misconception: Falsely believing person2 is leaning to avoid a spill.  
Explanation: There is no mentioned wineglass, so avoiding a spill is not the reason for the lean.

---

**Table 9: Five diagnostic questions for GPT-4.**

---

**Prompt:**

Question: What are person0 and person1 doing?

Answer: They are moving the boat into the ocean to escape.

Distractor: Person0 and person1 are trying to pull the boat out of the water.

Ground truth feedback:

Misconception : Incorrect direction of the boat movement.

Explanation: The distractor arises from focusing on the act of pulling without considering the context of urgency and the goal to escape, leading to the false conclusion that they are trying to pull the boat out of the water.

F1:

Educational level: Understand

Misconception: Misunderstand the direction of the boat.

Explanation: The distractor stems from concentrating solely on the action of pulling, without taking into account the urgent context and the objective to escape, which erroneously suggests that the effort is to pull the boat out of the water.

F2:

Educational level: Understand

Misconception: The distractor stems from emphasizing the action of pulling without taking into account the context of urgency, resulting in the incorrect assumption that the objective is to pull the boat out of the water.

...

Analyze the distractors and feedback F1, F2, F3, F4. Answering the following questions as output. Just give me the answer format like F1: Y N N Y N without further explanations, where Y for yes and N for No.

Q1: Is the educational level in the generated feedback consistent with the level of the ground truth feedback?

A1: Y/N

Q2: Is the explanation consistent with the given misconception?

A2: Y/N

Q3: Does the explanation provide a reasonable account of the distractor error?

A3: Y/N

Q4: Does the explanation describe why the distractor contradicts the given content?

A4: Y/N

Q5: Can this feedback help you better understand this question towards the correct answer?

A5: Y/N

---

**GPT-4:**

F1: Y Y Y Y Y

F2: Y Y Y Y N

...

---