

521 A Potential Negative Societal Impacts

522 The VLMs may generate harmful outputs based on human induction feedback, resulting in risks,
523 such as false information, discrimination, violent and pornographic content, and privacy leaks *etc.* To
524 mitigate the risks of these harmful outputs, we will strictly filter and review the model outputs based
525 on feedback in the future. In addition, users may become overly dependent on the model’s outputs
526 given feedback, neglecting the need for independent thinking and verification of information.

527 B Human Verification on FIRE

528 To evaluate the data quality of generated data in FIRE-100K, FIRE-1M, and FIRE-Bench, we conduct
529 a user study for the three splits of FIRE. Concretely, we randomly sample 100 conversations from
530 each of the three splits, and ask 10 persons to provide scores (1-5) for feedback and refined responses
531 in each turn of conversations. For the feedback, we ask the person “Please consider the quality of the
532 refined feedback, based on its correctness, relevance, clarity, and constructiveness. Give a score (1-5).
533 1 means its quality is bad, and 5 means its quality is very good”. For the refined response, we ask
534 the person “Please consider the quality of the response, based on its improvement, correctness, and
535 completeness. Given a score (1-5). 1 means its quality is bad, and 5 means its quality is very good”.
536 The interface of the user study is shown in Fig. A1. We report the average scores in Tab. A1. We
537 can find that, most users provide high scores for generated data in the three splits, showing that our
538 dataset has high-quality data.

Table A1: Average scores from humans on FIRE-100K, FIRE-1M, and FIRE-Bench, with 5 being the highest score.

FIRE-100K		FIRE-1M		FIRE-Bench	
Feedback	Response	Feedback	Response	Feedback	Response
4.87	4.66	4.84	4.73	4.88	4.74

539 C Additional Experimental Results

540 C.1 Error bar

541 We report the error bar of average turn (AT), average dialogue refinement (ADR), average turn
542 refinement (ATR), and refinement ratio (RR) in fixed dialogues. We run the model three times and
543 compute the standard deviation, as shown in Tab. A2. Comparisons among the four metrics, the
544 standard deviation is relatively small, less than 8% of the average results, showing that our method
can achieve stable feedback-refining ability.

Table A2: Results in free dialogue over all test data in FIRE.

Model	AT (↓)	ADR (↑)	ATR (↑)	RR (↑)
LLaVA-Next-8B	1	0.97	0.41	0.25
FIRE100K-LLaVA-8B	0.92 ± 0.026	1.27 ± 0.013	0.55 ± 0.042	0.34 ± 0.022
FIRE-LLaVA-8B	0.84 ± 0.015	1.56 ± 0.012	0.66 ± 0.053	0.39 ± 0.028

546 C.2 Performance Curves with respect to data number

547 We have a total of 1.1M training data in FIRE. We evaluate the performance of VLMs using different
548 number data in FIRE. In Fig. A2, we present the curves of AT, ATR, ATR, and RR using different
549 numbers of training data in FIRE. Concretely, we first use the FIRE-100K data. Then, we randomly
550 sample data from FIRE-100K, varying from 100K to 1000K, combined with FIRE-100K to train the
551 LLaVA-NEXT-8B model. Results show that more data leads to better performance. FIRE-100K data
552 brings a significant improvement, and the performance continues to grow with data increases. Then,
553 the performance slightly increases, and finally achieves obviously better feedback-refining ability
554 than the original LLaVA-Next-8B model. This experiment shows the quality of data in FIRE again.

Please evaluate the quality of the student model's response and the teacher model's feedback. Rate from 1 to 5, where a higher score indicates better quality. You can refer to the following criteria when scoring:

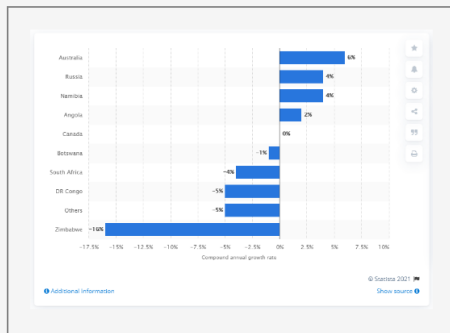
Evaluation Criteria for Student Model's Response

1. **Improvement:** Evaluate the improvement based on the teacher's feedback in the student's answer.
2. **Correctness:** Assess the accuracy of the student's answer and its alignment with known facts or ground truth.
3. **Relevance:** Evaluate if the student's response directly addresses the question and if it is free of omissions or off-topic content.
4. **Completeness:** Assess if the student's response is comprehensive, including necessary details and information.

Evaluation Criteria for Teacher Model's Feedback

1. **Correctness:** Evaluate if the feedback is based on actual errors or deficiencies in the student's answer and if it aligns with known ground truth.
2. **Relevance:** Assess if the feedback directly addresses the student's answer without introducing irrelevant information and if it relates to the specific question the student answered.
3. **Clarity:** Determine if the feedback is clear and explicit enough for the student model to understand and improve upon, using concise and easy-to-understand language.
4. **Constructiveness:** Evaluate if the feedback provides specific advice that will help the student improve and if it points out how to avoid similar mistakes or improve the answer.

Case 2



Question: What is the difference between Australia and Angola Compound annual growth rate for the top diamond producing country world wide between 2010 and 2015?

Groundtruth: 4

Student: Thought: I see various countries listed with their compound annual growth rates. I need to find the rates for Australia and Angola and calculate the difference. I think Australia's rate is higher than Angola's. Answer: The difference is 2%.

Teacher: Score: 4. Feedback: You correctly identified the need to find the compound annual growth rates for both Australia and Angola and calculate the difference. However, your calculation seems to be incorrect. Please check the rates for both countries again and ensure you are calculating the difference accurately. Remember to subtract the smaller rate from the larger rate to find the correct difference.

1 - Very Poor 2 - Poor 3 - Average 4 - Good 5 - Excellent

Student: Thought: I need to recheck the values for both countries. I see that Australia has a growth rate of 6% and Angola has a rate of 2%. To find the difference, I should subtract Angola's rate from Australia's rate. Answer: The difference is 4%.

1 - Very Poor 2 - Poor 3 - Average 4 - Good 5 - Excellent

Teacher: Score: 10. Feedback: Excellent work! You have correctly identified the compound annual growth rates for both Australia and Angola and accurately calculated the difference between them. Your answer is now correct and well-presented. Great job on adjusting your approach and achieving the correct result.

1 - Very Poor 2 - Poor 3 - Average 4 - Good 5 - Excellent

Save Scores and Next Case

Figure A1: The screenshot of the interface for the human verification.

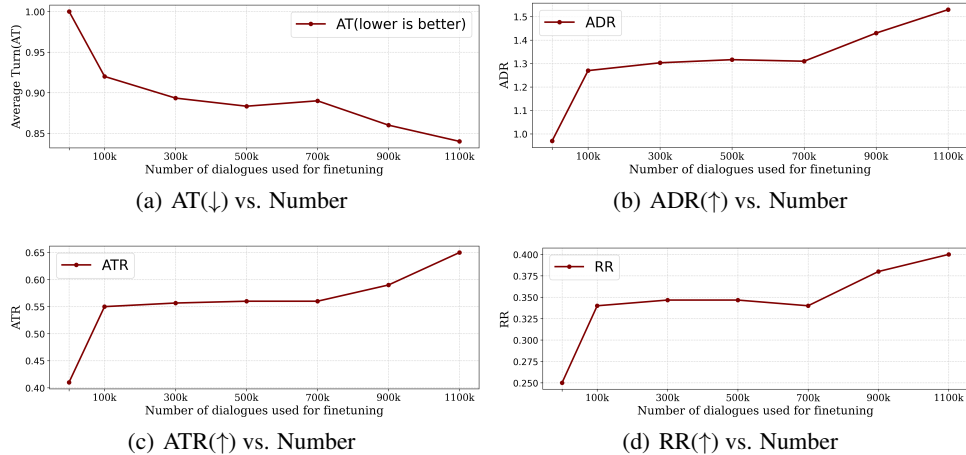


Figure A2: Performance vs. Number of dialogues for finetuning.

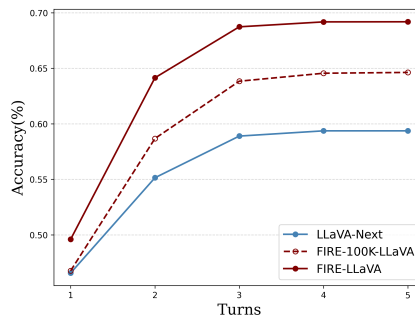


Figure A3: Accuracy (%) vs. Number of Turns.

555 C.3 Performance Curves with respect to turn number

556 In Fig. A3, we present the performance curve in FIRE-Bench with respect to the number of turns in
 557 dialogues, where LLaVA-Next, FIRE100K-LLaVA, and FIRE-LLaVA are evaluated. We measure
 558 the percentage of correctly answered samples (obtained scores larger than 8) after each turn. As the
 559 number of turns increases, the number of correctly answered samples in all three models increases.
 560 Finally, about 70% samples can be correctly answered, where 20% samples are corrected based on
 561 feedback, showing the effectiveness of FIRE-LLaVA trained on FIRE. In addition, compared with
 562 the original LLaVA-Next model, FIRE-LLaVA leads to extra 7% improvements (49% \rightarrow 69% vs
 563 46% \rightarrow 59%).

564 C.4 More VLMs

565 We further train a FIRE-LLaVA-Vicuna model that replaces LLaMA3-8B in FIRE-LLaVA with
 566 Vicuna1.5-7B. Results are shown in Tab. A3. Results of using Vicuna1.5-7B demonstrate the
 567 effectiveness of FIRE again, where FIRE-LLaVA-Vicuna has better feedback-refining ability than the
 568 original LLaVA-Next-Vicuna model on AT, ADR, ATR, and RR, showing the helpfulness for the
 569 feedback-refining ability.

Table A3: Results of FIRE-LLaVA-Vicuna in free dialogue over all test data in FIRE.

Model	AT (\downarrow)	ADR (\uparrow)	ATR (\uparrow)	RR (\uparrow)
LLaVA-Next-Vicuna	1.00	0.98	0.49	0.24
FIRE-LLaVA-Vicuna	0.94	1.11	0.57	0.27

Student: The camera next to the bottle is a Leica.

Teacher: Score: 2. Comments: The camera next to the bottle is not a Leica. Leica cameras have a distinctive design that is different from the camera in the image. Look closely at the logo and design features of the camera to identify the correct brand.

Question: What brand is the camera next to the bottle?

Student: The camera next to the bottle is a Nikon.

Ground truth Answer: Nikon.

Teacher: Score: 8. Comments: You correctly identified the brand of the camera as Nikon, which is great!

Figure A4: One case for feedback-refining ability in OCR.

Student: The image provided does not contain any identifiable landmarks or signs that would allow us to determine if this is a photo of De Bataaf, Winterswijk, or any other location.

Teacher: Score: 3. Comments: You have correctly noted the absence of identifiable landmarks or signs in the image. However, the question specifically asks if this is a photo of De Bataaf, Winterswijk. To improve your answer, consider the **possibility that the image might be of a location with a similar appearance to De Bataaf, Winterswijk**, even if there are no explicit indicators. Try to think about the **architectural style and the overall setting**, which might give you a clue about the location.

Student: Based on the architectural style and the overall setting of the image, it is possible that this could be a photo of De Bataaf, Winterswijk, or a similar location. The structure's design, **with its conical roof and the use of brick and plaster, is characteristic of windmills found in the Netherlands, particularly in the provinces of North Holland and Friesland, where such windmills are common.** However, without more specific information or a view of the surroundings that might include distinctive landmarks or signs, it cannot be confirmed with certainty that this is a photo of De Bataaf, Winterswijk.

Teacher: Score: 6. Comments: Your answer is on the right track by identifying the architectural style and the possible location based on the windmill's design. However, you need to be more confident in your identification. The structure in the image is indeed a windmill, and the setting appears to be in the Netherlands. To improve your answer, **focus on the specific features of the windmill and the surrounding environment that might indicate its location.**

Question: Is this a photo of De Bataaf, Winterswijk? Please answer yes or no.

Student: Given the architectural style and the setting of the windmill, **it is reasonable to conclude that this is indeed a photo of De Bataaf, Winterswijk**, or a similar location in the Netherlands. The conical roof, the use of brick and plaster, and the overall design are characteristic of windmills found in this region. While there are no explicit indicators such as signage or distinctive landmarks, the architectural features are sufficient to suggest that this windmill is located in the Netherlands, likely in the provinces of North Holland or Friesland.

Ground truth Answer: Yes.

Teacher: Score: 8. Comments: Your answer is quite reasonable and well-supported by the architectural features observed in the image. You correctly identified the Dutch influence in the design of the windmill.

Figure A5: One case for feedback-refining ability in visual perception.

570 C.5 More Visualization

571 We provide more cases tested on the FIRE-Bench to show the feedback-refining ability, as shown
 572 in [Figs. A4 to A11](#). We observe that FIRE-LLaVA trained on FIRE has good feedback-refining ability
 573 on different tasks, such as visual perception, visual question answering, image captioning, OCR
 574 reasoning, document understanding, math reasoning, science question answering, and chart analysis.

575 D Prompt

576 D.1 Prompt for GPT-4V

577 We show examples of prompts for GPT-4V to generate feedback-refinement conversations, as shown
 578 in [Fig. A12](#) and [Fig. A13](#).

579 D.2 Prompt for Student and Teacher models

580 We show examples of prompts for student and teacher models to simulate feedback-refinement
 581 conversations, as shown in [Fig. A14](#) and [Fig. A15](#), respectively. In [Fig. A14](#), the prompt for the
 582 student model to generate n -th response is shown. The prompt contains the last $n - 1$ rounds' student
 583 responses and the textual comments from the teacher model. The prompt for the teacher model is
 584 shown in [Fig. A15](#). Firstly, the prompt provides user instruction and ground truth. Secondly, the

The elementary substance in the image is represented by the molecule labeled "propane". "Propane is a simple organic molecule composed of three carbon atoms and one hydrogen atom. It is a gas at room temperature and is commonly used as a fuel. The other molecules, hydrazine and iodine, are more complex compounds. Hydrazine is a compound of hydrogen and nitrogen, while iodine is a non-metallic element. Therefore, the correct answer to the question is: A. propane

Score: 3. Comments: You correctly identified that propane is a simple organic molecule composed of carbon and hydrogen atoms. However, the question asks for the elementary substance, which refers to a basic chemical element. Propane is a compound, not an element. Consider the other options and think about which one is an element.

I apologize for the confusion. You are correct that propane is a compound, not an element. The elementary substances in the image are represented by the molecules labeled "hydrazine" and "iodine". Hydrazine is a compound of hydrogen and nitrogen, and iodine is a non-metallic element. Therefore, the correct answer to the question is: C. iodine

Score: 10. Comments: You correctly identified that "iodine" is an elementary substance, which is a non-metallic element. However, you should also note that "hydrazine" is a compound made up of hydrogen and nitrogen. This distinction is important for understanding the difference between elements and compounds. Great job on identifying the correct answer!

Figure A6: One case for feedback-refining ability in science question answering.

The question pertains to the relative strength of the magnetic force between the magnets in each pair. Based on the image, the magnetic force is represented by the length of the arrows. In Pair 1, the arrow is longer, indicating a greater force. In Pair 2, the arrow is shorter, indicating a lesser force. Therefore, the correct statement is A. The magnitude of the magnetic force is greater in Pair 1

Score: 4. Comments: Your observation about the lengths of the arrows is correct, but your conclusion is incorrect. The question asks which pair has a greater magnetic force, not which pair has a longer arrow. Consider the actual values given for the magnetic force in each pair to determine the correct answer.

I apologize for the misunderstanding. The question asks which pair has a greater magnetic force, not which pair has a longer arrow. According to the values provided in the image, the magnetic force in Pair 2 is greater than in Pair 1. Therefore, the correct statement is C. The magnitude of the magnetic force is greater in Pair 2

Score: 9. Comments: Your answer is correct, and you have correctly identified that the magnetic force in Pair 2 is greater than in Pair 1.

Figure A7: One case for feedback-refining ability in science question answering.

To find the length of the unknown side P, we can use the trigonometric function tangent, which is the ratio of the opposite side to the adjacent side in a right-angled triangle. The formula for tangent of an angle in a right-angled triangle is: $\tan(\theta) = \text{opposite}/\text{hypotenuse}$. In this case, the angle is 30° , and the opposite side is the marked right-angled side, which is 8.5. The hypotenuse is the side opposite the 90° angle, which is P. Therefore, we have: $\tan(30^\circ) = 8.5$

Score: 4. Comments: You correctly identified the use of the tangent function and the sides involved in the calculation. However, you need to rearrange the equation to solve for P. Remember that the tangent of an angle in a right-angled triangle is the ratio of the opposite side to the adjacent side. Try to solve for P by rearranging the equation to include both sides of the triangle.

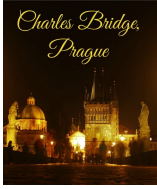
Thank you for the clarification. Let's rearrange the equation to solve for P: $\tan(30^\circ) = \text{opposite}/\text{hypotenuse}$, $\tan(30^\circ) = 8.5 / P$ To find P, we multiply both sides by P: $P * \tan(30^\circ) = 8.5$. Now, we divide both sides by $\tan(30^\circ)$: $P = 8.5 / \tan(30^\circ)$. Using a calculator to find the tangent of 30° and then dividing by that value gives us the length of P: $P = 8.5 / \tan(30^\circ)$

Score: 7. Comments: You correctly rearranged the equation and solved for P using the tangent function. However, you did not provide the final value of P. Please calculate the value of P using the given information and provide the final answer.

I apologize for the oversight. Let's calculate the value of P using the given information: $P = 8.5 / \tan(30^\circ)$. Using a calculator to find the tangent of 30° and then dividing by that value gives us the length of P: $P = 8.5 / \tan(30^\circ)$ $P = 8.5 / 0.5$ $P = 17$. Therefore, the length of the unknown side P is 17.

Score: 8. Comments: You correctly used the tangent function to find the length of P.

Figure A8: One case for feedback-refining ability in math reasoning.



Charles Bridge, Prague

Question:

Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
 Question: What is the main architectural feature of the city in the image?
 Choices: A. A castle B. A bridge C. Skyscrapers D. A mosque

Ground truth Answer:
 A. castle

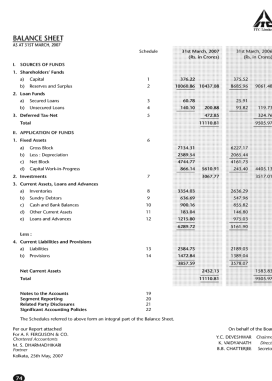
Student The main architectural feature of the city in the image is a bridge. This is evident from the title "Charles Bridge, Prague" which directly refers to the iconic bridge in Prague, Czech Republic. The image shows the bridge with its distinctive Gothic towers and statues, which are characteristic features of this historic bridge. Therefore, the correct answer to the question is: B. A bridge

Teacher Score: 4. Comments: You correctly identified the bridge as a significant architectural feature in the image. However, the question asks for the main architectural feature, and while the bridge is prominent, it is not the most iconic feature of Prague. Consider other notable architectural elements that might be more widely recognized as the main feature of the city.

Student Thank you for the feedback. Upon further consideration, the most iconic and widely recognized architectural feature of Prague is the Old Town Bridge Tower, also known as the Prague Castle, which is a significant landmark in the city. Therefore, the correct answer to the question is: A. A castle

Teacher Score: 8. Comments: You are on the right track by identifying the Prague Castle as a significant landmark.

Figure A9: One case for feedback-refining ability in visual question answering.



Schedule	2017 Actual	2016 Actual
1. Reserves and Surplus	174.57	174.57
2. Reserves	174.57	174.57
3. Surplus	0.00	0.00
4. Total	174.57	174.57

Question:
 What is the schedule no. for reserves and surplus?

Ground truth Answer:
 2

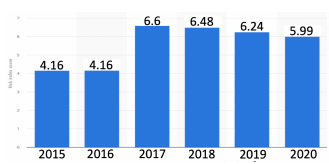
Student The schedule number for reserves and surplus in the balance sheet is 1.

Teacher Score: 3. Comments: The schedule number for reserves and surplus is not 1. Please review the balance sheet again and look for the correct schedule number listed under the "SCHEDULE NO." column next to "Reserves and Surplus."

Student I apologize for the mistake. After reviewing the balance sheet again, I found that the schedule number for reserves and surplus is actually 2.

Teacher Score: 8. Comments: You correctly identified the schedule number for reserves and surplus as 2.

Figure A10: One case for feedback-refining ability in document understanding.



Year	Risk Index
2015	4.16
2016	4.16
2017	6.6
2018	6.48
2019	6.24
2020	5.99

Question:
 Was the risk index of 2015 and 2016 together larger than 2020?

Ground truth Answer:
 yes

Student The risk index of 2015 and 2016 together is not larger than 2020.

Teacher Score: 3. Comments: There seems a mistake in adding the values for 2015 and 2016. Try adding them again and compare the sum to the score of 2020.

Student Adding scores of 2015 and 2016, 4.16 + 4.16 equals 8.32. Thus, the answer is yes, the risk index of 2015 and 2016 together is larger than 2020 because 8.32 is greater than 5.99.

Teacher Score: 10. Comments: You correctly identified number of 2015, 2016, and 2019, and finally get the correct answer.

Figure A11: One case for feedback-refining ability in chart analysis.

```
You are a helpful assistant that can generate a dialogue between a teacher and a student. The student is
trying to answer a question about an image. The student first gives a wrong answer. Based on the given
groundtruth answer, the teacher provides feedback to help the student gradually improve its answer.
Use the following template to generate the dialogue:

""
# Round 1

## Student's response
Thought: <here is the student's thought process about the question. Do NOT use the words "teacher"
or "student". >
Answer: <here is the student's answer to the question.>

## Teacher's feedback
Score: <compare the student's answer with the groundtruth answer in terms of accuracy, relevance,
helpfulness, and level of detail. Provide an overall score on a scale of 1 to 10, where a higher score
indicates better overall performance.>
Feedback: <provide feedback on the student's answer. Do not directly tell the groundtruth answer.
The feedback should identify which parts of the student's answer are incorrect, what is missing in the
student's answer, and how to improve the student's answer.>

# Round 2
...

# Round n
...

""

The number of rounds should depend on the difficulty of the question. More rounds should be used for
difficult questions, while fewer rounds should be used for easy questions.
```

Figure A12: System prompt for GPT-4V for Student-Teacher conversation generation.

```
Here are the given image, question: <question> and groundtruth answer: <groundtruth>, now generate
a dialogue:
```

Figure A13: User prompt for GPT-4V for Student-Teacher conversation generation.

585 prompt contains instructions that format the teacher model's feedback as textual comments and
586 numeric scores. Finally, the prompt incorporates the only latest student response into its context.

587 **E Datasheet for FIRE**

588 We present a Datasheet [16] for documentation and responsible usage of our internet knowledge
589 databases. The required author statement, hosting, licensing, metadata, and maintenance plan can be
590 found in the datasheet.

591 **E.1 Motivation**

592 **For what purpose was the dataset created?** We create this large-scale dataset to facilitate research
593 towards enhancing the feedback-reflecting ability of VLMs.

594 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g.,**
595 **company, institution, organization)?** This dataset was created by Pengxiang Li (Beijing Institute

You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language
 <user_instruction>

Round 1
 <student_response_round_1>
 <feedback_round_1>

...

Round n-1
 <student_response_round_n-1>
 <feedback_round_n-1>

Based on the feedback, answer the question again:

Figure A14: Prompt for student model to simulate feedback-refinement conversations.

You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language

Question: <question>
 Groundtruth: <groundtruth>

Please compare my answer with the groundtruth answer and provide helpful, detailed, and polite feedback to help me improve my answer. Formulate the feedback as:
 """"

Score: <compare the provided response with the groundtruth answer in terms of accuracy, relevance, helpfulness, and level of detail, and provide an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.>

Feedback: <provide feedback on the response. Do NOT directly tell the groundtruth answer. The feedback should identify which parts of my answer are incorrect, what is missing in the response, and how to improve the response.>
 """"

Here is the student response: <student_response>, now please provide the feedback:

Figure A15: Prompt for teacher model to simulate feedback-refinement conversations.

596 of Technology), Zhi Gao (BIGAI), Bofei Zhang (BIGAI), Tao Yuan (BIGAI), Yuwei Wu (Beijing
 597 Institute of Technology), Mehrtash Harandi (Monash University), Yunde Jia (Beijing Institute of
 598 Technology), Song-chun Zhu (BIGAI), Qing Li (BIGAI).

599 **E.2 Distribution**

600 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,
 601 organization) on behalf of which the dataset was created?** Yes, the dataset is publicly available
 602 on the internet.

603 **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** The dataset can
 604 be downloaded from <https://huggingface.co/datasets/PengxiangLi/FIRE>. We
 605 use Creative Commons Attribution 4.0 License (CC BY 4.0). The Croissant metadata can be found
 606 on the dataset hosting platform (<https://huggingface.co/>).

607 **Have any third parties imposed IP-based or other restrictions on the data associated with the**
608 **instances?** No.

609 **Do any export controls or other regulatory restrictions apply to the dataset or to individual**
610 **instances?** No.

611 E.3 Maintenance

612 **Who will be supporting/hosting/maintaining the dataset?** The authors will be supporting, host-
613 ing, and maintaining the dataset.

614 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** Please
615 contact Qing Li (liqing@bigai.ai).

616 **Is there an erratum?** No. We will make announcements if there is any.

617 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
618 Yes. New updates will be posted on <https://mm-fire.github.io/>.

619 **If the dataset relates to people, are there applicable limits on the retention of the data associated**
620 **with the instances (e.g., were the individuals in question told that their data would be retained**
621 **for a fixed period of time and then deleted)?** The images in our dataset might contain human
622 subjects, but all of them are from public datasets.

623 **Will older versions of the dataset continue to be supported/hosted/maintained?** Yes, old
624 versions will be permanently accessible on huggingface.co.

625 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for**
626 **them to do so?** Yes, please refer to <https://mm-fire.github.io/>.

627 E.4 Composition

628 **What do the instances that comprise the dataset represent?** Our data is generally stored in the
629 json files. Every instance includes the path of an image and the feedback-reflecting dialogues.

630 **How many instances are there in total (of each type, if appropriate)?** There are 1.2M samples
631 (1.1M for training set, 0.1M for test set), among which 200K are GPT-4V generated data, while the
632 rests are simulated via the FIRE-LLaVA and FIRE-LLaVA-FD.

633 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of**
634 **instances from a larger set?** We provide all instances in our Huggingface data repositories.

635 **Is there a label or target associated with each instance?** Yes.

636 **Is any information missing from individual instances?** No.

637 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social**
638 **network links)?** No.

639 **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes.
640 FIRE-100K and FIRE-1M are used for training and FIRE-Bench is used for testing.

641 **Are there any errors, sources of noise, or redundancies in the dataset?** Please refer to the
642 limitations in [Sec. 6](#).

643 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,**
644 **websites, tweets, other datasets)?** The dataset is self-contained.

645 **Does the dataset contain data that might be considered confidential?** No.

646 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening,**
647 **or might otherwise cause anxiety?** No.

648 E.5 Collection Process

649 The collection procedure, preprocessing, and cleaning are explained in detail in Section 2 of the main
650 paper.

651 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and**
652 **how were they compensated (e.g., how much were crowdworkers paid)?** All data collection,
653 curation, and filtering are done by FIRE coauthors.

654 **Over what timeframe was the data collected?** The data was collected between Jan. 2024 and
655 May 2024.

656 E.6 Uses

657 **Has the dataset been used for any tasks already?** Yes, we have used FIRE for training our VLMs,
658 including FIRE-LLaVA-Vicuna, FIRE100K-LLaVA, FIRE-LLaVA, FIRE-LLaVA-FD.

659 **What (other) tasks could the dataset be used for?** Our dataset is primarily for facilitating research
660 in enhancing the feedback-reflecting ability of VLMs. Our data might also be used to benchmark
661 existing and future VLMs.

662 **Is there anything about the composition of the dataset or the way it was collected and prepro-**
663 **cessed/cleaned/labeled that might impact future uses?** No.

664 **Are there tasks for which the dataset should not be used?** We strongly oppose any research that
665 intentionally generates harmful or toxic content using our data.

666 F Data source

667 Our dataset uses images from 27 diverse sources to provide a robust training dataset for FIRE. All
668 27 datasets are public datasets, and all the images can be downloaded via links in [Tab. A4](#). The
669 comprehensive list of the source datasets and links to their metadata are detailed below:

Table A4: Data utilized from 27 source datasets for training and test data in FIRE.

LLaVA [41]	COCO [37]	SAM [26]	VQAV2 [17]	GQA [21]
VG [27]	Web-Celebrity [8]	Web-Landmark [8]	WikiArt [54]	ALLaVA-Vflan [4]
ChartQA [46]	DocVQA [47]	DVQA [23]	GeoQA+[5]	Synthdog-EN[25]
LLaVA-in-the-Wild [41]	MMMU [63]	MME [14]	MM-Vet [62]	SEED-bench [30]
OCRvQA [49]	TextVQA [56]	Share-TextVQA [8]	A12D [24]	MathVerse [65]
MathVista [43]	ScienceQA [45]			