FIRE: A Dataset for Feedback Integration and Refinement Evaluation of Multimodal Models
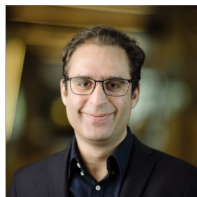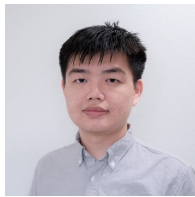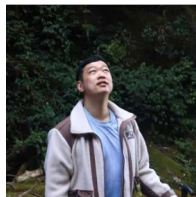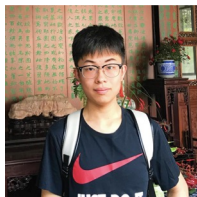
Pengxiang Li*[1,2], Zhi Gao*[1,3], Bofei Zhang*[1], Tao Yuan[1], Yuwei Wu[2],
Mehrtash Harandi[4], Yunde Jia[1], Song-Chun Zhu[1,3,5], Qing Li[1]

[1] BIGAI   [2] BIT   [3] Peking University   [4] Monash University   [5] Tsinghua University
* Equal contribution, Corresponding Author
https://mm-fire.github.io
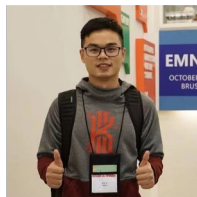
Motivation

We tackle the limitation of open-source Vision-Language Models (VLMs) in feedback-refining within a single session, aiming to enhance their interactive learning capabilities.

Ground truth: Car.

User: Which kind of vehicle is waiting for the traffic light?

LLaVA-Next-8B: The vehicles look like typical vehicles used for personal transport. The vehicles are motorcycles. *Wrong*

User: Motorcycles are not correct. Look again at the structure and features of the vehicles, such as the number of wheels and the body shape.

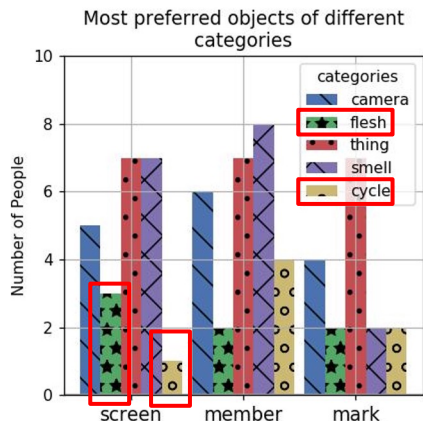**LLaVA-Next-8B:** Upon reevaluating the image, the vehicles are indeed motorcycles. *Failed in Refining*

Limited feedback-refining ability

Most preferred objects of different categories

categories: camera, flesh, thing, smell, cycle

Ground truth: No.

User: Is the object mark in the category cycle preferred by more people than the object screen in the category flesh?

Mini-Gemini-34B-HD: *Wrong* — Yes, the object mark in the category cycle is preferred by more people than the object screen in the category flesh. The bar representing the object mark in the cycle category reaches up to the number 8 on the y-axis, while the bar for the object screen in the flesh category is just above the number 2 on the y-axis.

User: Your answer is incorrect. You correctly identified the need to compare the number of people, but your conclusion is wrong. Look closely at the bar heights for the object mark in the cycle category and the object mark in the flesh category. Try again.

Mini-Gemini-34B-HD: I apologize for the mistake in my previous answer. Upon closer inspection, the object mark in the cycle category is indeed preferred by more people than the object screen in the flesh category. The bar for the object mark in the cycle category reaches up to the number 8 on the y-axis, while the bar for the object screen in the flesh category is just above the number 2 on the y-axis. Therefore, the correct answer is yes, the object mark in the cycle category is preferred by more people than the object screen in the flesh category.

Same as initial response. Failed in Refining

Limited feedback-refining ability

User
Why are the leaves turn yellow?

Overwatering ~~Wrong~~
QwenVL-72B

User
Your answer is incorrect. Please judge the possible reasons based on the texture characteristics of the yellowing leaves. Please note that overwatering will not cause the leaves to turn yellow in such pattern. Yellowing leaves may be caused by a disease in the plant.

Overwatering ~~Failed in Refining~~
QwenVL-72B

Ground truth:
Nutritional deficiency.

One reason of such limitation is that instruction-following datasets predominantly consist of data focused on direct task completion, with **little to no inclusion** of feedback-refining information.

Task setting

Data format

$$\{I, q, gt, \{r_i, f_i\}_1^n\}$$

Image   Question   Groundtruth Answer   Response   Feedback
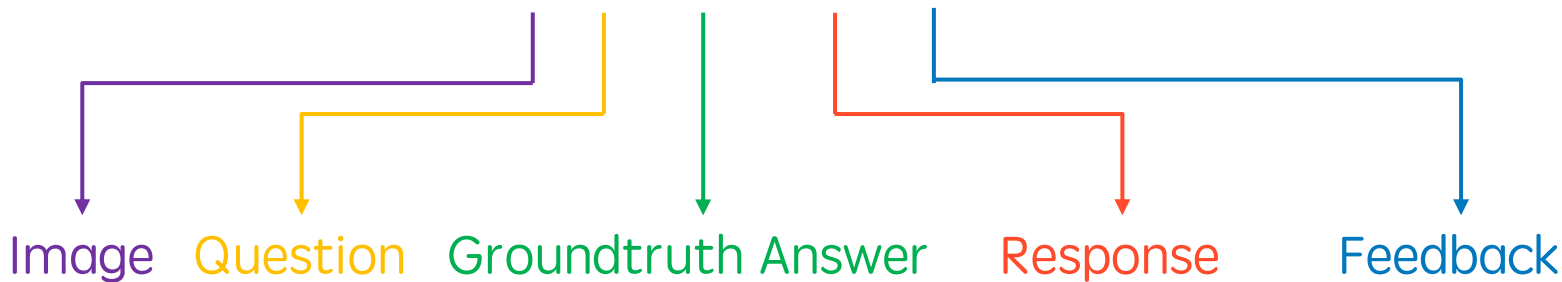
Dialogue simulation

Vision Lanuage Model
ROLE: Teacher
Playing: Scoring and feedback.

Input: I, q, gt, r_i
Output: f_i

Vision Lanuage Model
ROLE: Student
Playing: Answer the Question.

Input: I, q, f_i
Output: r_i+1

Groundtruth Answer is only visible to teacher model.

Dataset statistic

Training Set FIRE-100K & FIRE-1M

Statistic	FIRE-100K	FIRE-1M
Total dialogue	104941	1003490
Avg. Response Length	41.20	79.33
Avg. Feedback Length	41.74	54.87

FIRE offers **1.1 million** high-quality feedback-refining dialogues across **8** diverse tasks and **18** various data sources.

Dataset statistic
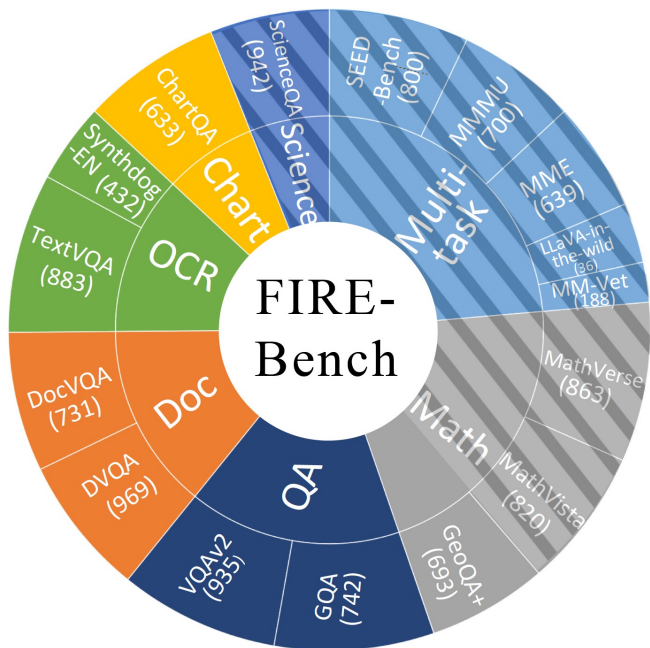
Response

Feedback

Statistic	FIRE-Bench
Total dialogue	11006
Avg. Response Length	38.70
Avg. Feedback Length	41.74

FIRE-Bench offers **11 K** high-quality feedback-refining dialogues across **8** diverse tasks and **16** various data sources, including 8 test splits from the training data sources and **8 completely new benchmarks**.

Data collection

(a) FIRE-100K Data Generation

Science	Caption	Math	Chart	OCR
Instruction: Which figure is .. GT : ...	Instruction: Describe the image. GT : ...	Instruction: compute the area.. GT : ...	Instruction : What is the lowest.. GT : ...	Instruction : What is the address ...? GT : ...

... 100K data

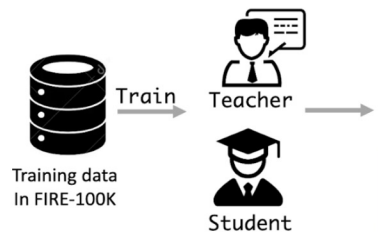You are a helpful assistant that can generate a dialogue between a teacher and a student. The student is trying to answer a question about an image. Based on the given groundtruth answer, the teacher provides feedback to help the student gradually improve its answer.

Here are the QUESTION: <question> and GROUNDTRUTH ANSWER:<answer>, now generate dialogue:"

GPT-4V

Question: <question>
GROUNDTRUTH ANSWER:<answer>

#Round 1
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}
...
#Round n
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}

FIRE-100K

We use GPT-4V to generate 100K data (FIRE-100K), with each dialogue was generated in a one-go manner.

Data collection

We use FIRE-100K to train a Teacher model and a Student model.
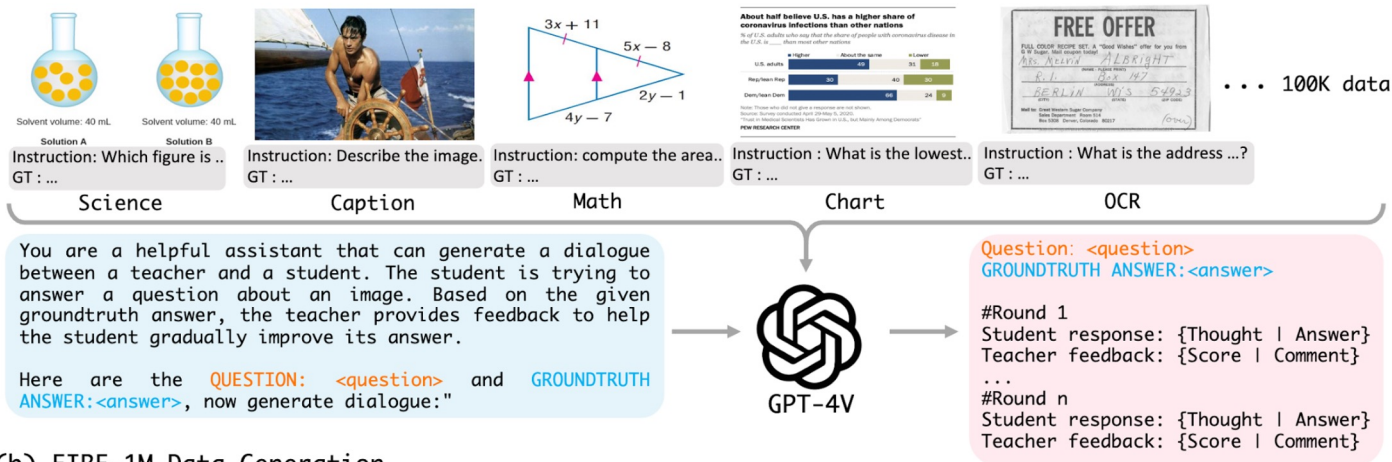
Then we simulate dialogues between the student and the teacher models and scale up the data points to ~1M, rending a split named FIRE-1M.

(b) FIRE-1M Data Generation

Train Teacher
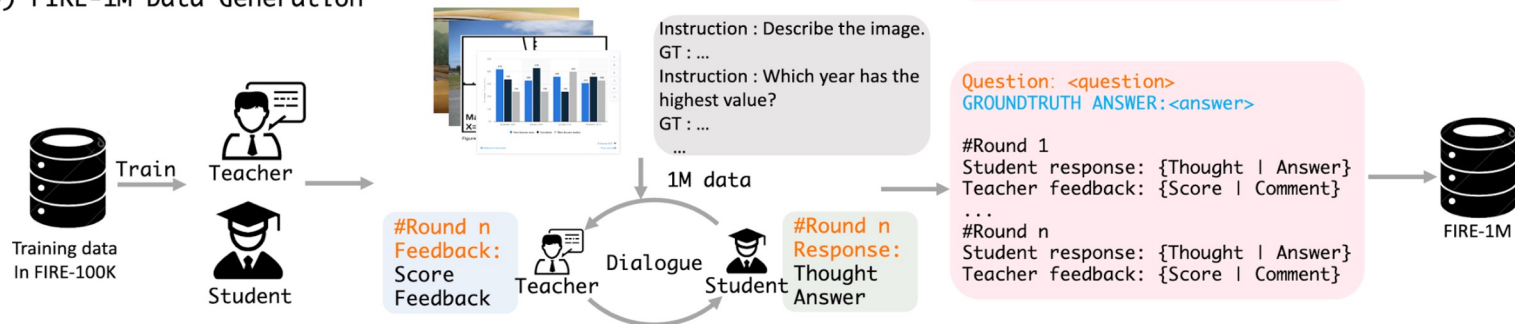
Training data
In FIRE-100K

Student

Data collection

(a) FIRE-100K Data Generation

Instruction: Which figure is .. GT : ... — Science

Instruction: Describe the image. GT : ... — Caption

Instruction: compute the area.. GT : ... — Math

Instruction : What is the lowest.. GT : ... — Chart

Instruction : What is the address ...? GT : ... — OCR

... 100K data

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. Based on the given groundtruth answer, the teacher provides feedback to help the student gradually improve its answer.

Here are the QUESTION: <question> and GROUNDTRUTH ANSWER:<answer>, now generate dialogue:"

GPT-4V

Question: <question>
GROUNDTRUTH ANSWER:<answer>

#Round 1
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}
...
#Round n
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}

FIRE-100K

(b) FIRE-1M Data Generation

Training data In FIRE-100K — Train — Teacher / Student

Instruction : Describe the image.
GT : ...
Instruction : Which year has the highest value?
GT : ...
...

1M data

#Round n Feedback:
Score Feedback — Teacher — Dialogue — Student — #Round n Response: Thought Answer

Question: <question>
GROUNDTRUTH ANSWER:<answer>

#Round 1
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}
...
#Round n
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}
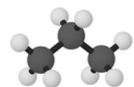
FIRE-1M

Training-Student Model

$$\{I, q, gt, \{r_i, f_i\}_1^n\}$$

Question:

Look at the models of molecules below. Select the elementary substance.
Choices:
A. propane B. hydrazine C. iodine

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.

Ground-truth Answer: 🙈

Text completion:

$$\min \mathbb{E}_{(I,q,gt,\{r_i,f_i\}_{i=1}^n)\sim\mathbb{D}} \left[-\sum_{i=2}^{n} \log P(r_i | I, q, \{r_j, f_j\}_{j=1}^{j=i-1}) \right]$$
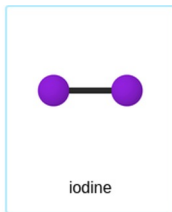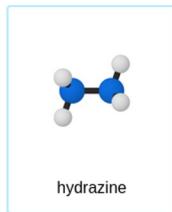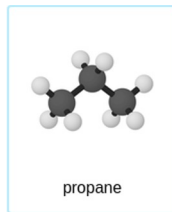
Training-Teacher Model

$$\{I, q, gt, \{r_i, f_i\}_1^n\}$$

Question:

Look at the models of molecules below. Select the elementary substance.
Choices:
A. propane B. hydrazine C. iodine

propane   hydrazine   iodine

H   C   N   I

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**
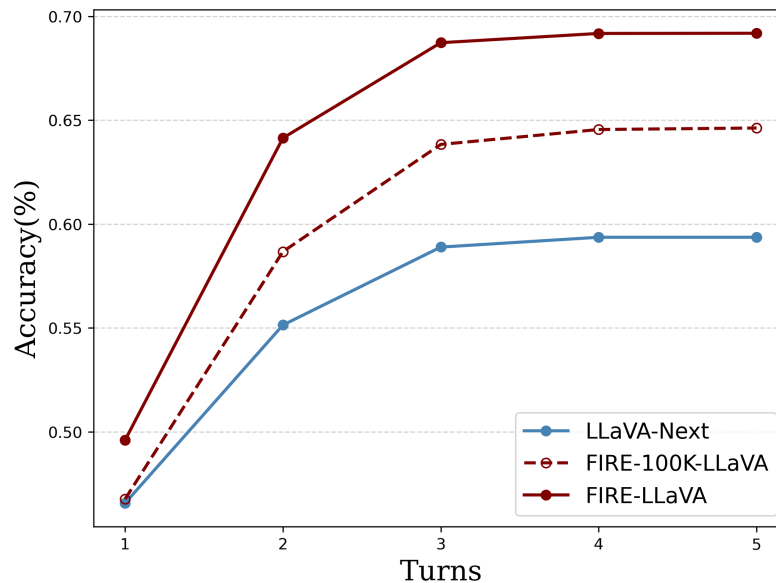
Text completion:

Ground-truth Answer:

C. iodine

$$\min \mathbb{E}_{(I,q,gt,\{r_i,f_i\}_{i=1}^n) \sim \mathbb{D}} \left[-\sum_{i=1}^{n} \log P(f_i | I, q, gt, \{r_j, f_j\}_{j=1}^{j=i-1}, r_i) \right]$$

Experiments

LLaVA-Next fine-tuned by FIRE can complete the response **faster** and has **better feedback refining capabilities** (7% more improvements) under the same feedback model.

Experiments

Average Dialogue Refinement (ADR). The ADR metric evaluates how much knowledge VLMs could learn from feedback **in a dialogue**.

Refinement Ratio (RR). RR measures the proportion of data that have a wrong initial response and a correct final response

Average Turn (AT). The AT metric evaluates **how fast** a VLM could achieve a satisfactory result based on feedback.

Average Turn Refinement (ATR). ATR evaluates how much knowledge VLMs could learn from feedback **in one turn**.

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	1.27	0.55	0.34
FIRE-LLaVA-8B	**0.84**	**1.56**	**0.66**	**0.39**

LLaVA-Next fine-tuned by FIRE can complete the response **faster** and has **better feedback refining capabilities**.

Visualization

Question:

What is the color of the gloves worn by one of the football player in the image?

A. White   B. Black
C. Green   D. Blue

Ground-truth Answer:

B. Black

The color of the gloves worn by the football player in the image is white. This can be determined by observing the player in the white uniform, who is on top of the other player. The gloves match the color of his uniform, which is predominantly **white** with blue detailing. Therefore, the correct answer to the question is: **A. White**

Score: 4.
Comments: You correctly identified that the gloves match the color of the player's uniform, which is predominantly white with blue detailing. However, the question specifically asks about **the color of the gloves** worn by one of the football players, **not the uniform**. **Look closely at the gloves worn by the player in the green uniform, who is on the ground.** Consider the options provided and re-evaluate your answer.

Upon re-evaluation, **the gloves worn by the player in the green uniform, who is on the ground, are black.** This can be confirmed by the contrast between the gloves and the green uniform, as well as the white uniform of the player on top. Therefore, the correct answer to the question is: B. Black

Score: 10.
Comments: Excellent observation! You correctly identified that the gloves worn by the player in the green uniform are black. This matches the ground truth answer. Well done on adjusting your focus and re-evaluating the image.

Visualization

Question:

How many units of the item scene were sold in the store oil?

Sales statistics for items in different stores

Ground-truth Answer:

1
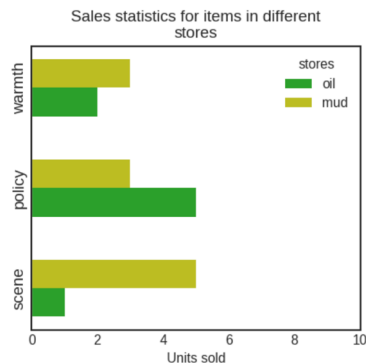
Based on the image, **it is not possible to determine the exact number of units of the item "scene" sold in the store "oil"** because the bar representing "oil" for "scene" is not fully visible. Only a portion of the bar is shown, and without the full length of the bar or a scale, the number of units sold cannot be accurately determined from the given image.

Score: 3.
Comments: You are correct that the image does not provide enough information to determine the exact number of units of the item "scene" sold in the store "oil." However, you should **consider the visible portion of the bar and any patterns or trends that might help estimate the number of units sold.** Try to use the visible information to make an educated guess.

Upon closer inspection**, the visible portion of the "oil" bar for "scene" appears to extend slightly beyond the halfway point between 0 and 2 on the x-axis.** This suggests that a small number of units were sold, possibly between 1 and 2. However, without more information or a clearer view of the entire bar, it is not possible to provide an exact number.

Score: 6.
Comments: Your observation about the visible portion of the "oil" bar for "scene" is a good start. You correctly noted that the bar extends slightly beyond the halfway point between 0 and 2. This suggests that a small number of units were sold. However, to improve your answer, **consider the possibility of the bar extending just beyond the halfway point and estimate the number of units sold more precisely**. Remember, the exact number cannot be determined without more information, but you **can make an educated guess based on the visible data**.

16

Visualization

Question:

How many units of the item scene were sold in the store oil?

Sales statistics for items in different stores
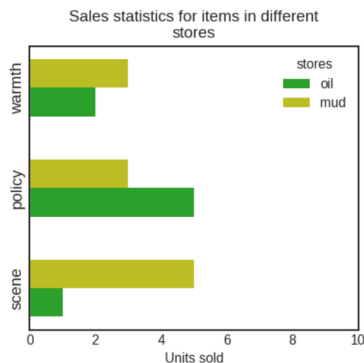
Ground-truth Answer:

1

Considering the visible portion of the "oil" bar for "scene" extends just beyond the halfway point between 0 and 2, **it is reasonable to estimate that approximately 1 unit of the item "scene" was sold in the store "oil."** This is based on the assumption that the bar extends equally beyond the halfway point on both sides, which is a common pattern for bars in such graphs. However, without more information or a clearer view of the entire bar, this remains an educated guess rather than a definitive answer.

Score: 9.
Comments: Considering the visible portion of the "oil" bar for "scene" extends just beyond the halfway point between 0 and 2, **it is reasonable to estimate that approximately 1 unit of the item "scene" was sold in the store "oil."** This is based on the assumption that the bar extends equally beyond the halfway point on both sides, which is a common pattern for bars in such graphs. However, without more information or a clearer view of the entire bar, this remains an educated guess rather than a definitive answer.

Takeaway

We present **FIRE**, a feedback-refinement dataset with **1.1M multi-turn conversations**, which empowers VLMs to refine their responses based on given feedback.

We develop **FIRE-LLaVA**, an advanced VLM that could improve its responses based on feedback, making efficient interaction between users and VLMs.

We have developed and released a **Feedback Model** designed to enhance VLM performance by providing targeted feedback.

Code & Dataset: https://mm-fire.github.io

FIRE: A Dataset for **F**eedback **I**ntegration and **R**efinement **E**valuation of Multimodal Models