

A DETAILS OF DATA

The details of the data mixture of image caption and VQA datasets used in **Fact** training are shown in Table 5. We only use a subset of each dataset’s training set. For image caption and half of the VQAv2 tasks, we do not generate the corresponding program. However, not all data transform to CoT rationales for teaching MLLMs in our process. We conduct filtering for the faithful of the program and the transferability of the rationale separately, ultimately retaining only those programs and rationales as shown in the table. This means the final dataset used for training MLLMs comprises 176K entries (including 20K image captions and 50K VQA data without rationale). This selective approach ensures that the MLLMs are provided with the most relevant and effectively distilled information, enhancing their reasoning and understanding capabilities.

Table 6: Data mixture of image caption and VQA datasets used in Fact generalist training.

Dataset	Description	labels	programs	rationales
COCO	Scene description	10.0K	-	-
Flickr 30K	Scene description	10.0K	-	-
VQAv2	General	50.0K	-	-
VQAv2	General	50.0K	28.5K	26.2K
GQA	Compositional	86.0K	47.5K	43.0K
OK-VQA	Knowledge	9.0K	5.1K	4.9K
TallyQA	Counting	48.4K	35.4K	31.9K
Total		263.4K	116.5K	106.0K

It should be noted that only in the code generation part of VQAv2, the code pre-train model we use is Code-Llama 70B [36].

Our data filtering results and ablation experiments on transferability also prove that even if a rationale is satisfied with faithfulness and conciseness, transferability is still an important part of the rationale.

B EVALUATION MATRIX

Table 7: Evaluation metrics and prompts we used in downstream tasks.

Dataset	Split	Metrics	Prompt
COCO	test	CIDEr Score	A short image caption:
Flickr 30K	test	CIDEr Score	A short image caption:
VQAv2	test-dev	VQA Acc	Based on the image, respond to this question with a short answer: {question}
GQA	test-dev	VQA Acc	Based on the image, respond to this question with a short answer: {question}
OK-VQA	val	VQA Acc	Based on the image, respond to this question with a short answer: {question}
TallyQA	test	VQA Acc	Answer the question with a number. {question}

C HUMAN EVALUATION

In this section, we explain the sources of error details of our experiments in Figure 6. We manually chose 100 responses from the GQA dataset and manually undertook a detailed error analysis. The errors identified were classified into four main categories:

- Logical errors, highlighting failures in the coherence of reasoning across sentences.
- Factual errors, denoting the presence of incorrect information in the answers.
- Format errors, observed when the model either did not address the question directly or provided answers that were irrelevant to the posed questions.
- Localization errors, identified when the model referenced parts of the image that were either incorrectly identified or non-existent.