On the Evaluation and Refinement of Vision-Language Instruction Tuning Datasets

Anonymous ACL submission

Abstract

There is an emerging line of research on multimodal instruction tuning, and a line of benchmarks has been proposed for evaluating these models recently. Instead of evaluating the models directly, in this paper, we try to evaluate the Vision-Language Instruction-Tuning (VLIT) datasets. Also, we seek the way of building a dataset for developing an all-powerful VLIT model, which we believe could also be of utility for establishing a grounded protocol for benchmarking VLIT models. For effective evaluation of VLIT datasets that remains an open question, we propose a *tune-cross-evaluation* paradigm: tuning on one dataset and evaluating on the others in turn. For each single tune-evaluation experiment set, we define the Meta Quality (MQ) to quantify the quality of a certain dataset or a sample. On this basis, we develop the Dataset Quality (DQ) covering all tune-evaluation sets to evaluate the comprehensiveness of a dataset. To build a comprehensive dataset and developing an all-powerful model for practical applications, we define the Sample Quality (SQ) to quantify the all-sided quality of each sample. Extensive experiments validate the rationality of the proposed evaluation paradigm. Based on the holistic evaluation, we build a new dataset, REVO-LION (REfining VisiOn-Language InstructiOn tuNing), by collecting samples with higher SQ from each dataset. Remarkably, even with only half of the complete data, the model trained on REVO-LION can achieve the performance comparable to simply adding all VLIT datasets up. Furthermore, REVO-LION also incorporates an evaluation set, which is designed to serve as a convenient benchmark for future research in the field.

1 Introduction

The large-scale multimodal model GPT-4 (OpenAI, 2023) has recently exhibited strong power in generating desired answers from given images and instructions. Inspired by its remarkable success, various multimodal instruction tuning models (Dai et al., 2023; Chen et al., 2024; Li et al., 2023a; Luo et al., 2024) have been proposed towards different aspects of Vision-Language (VL) understanding, such as MiniGPT-4 (Zhu et al., 2023) for detailed description and LLaVAR (Zhang et al., 2023) for text-rich image understanding. With the rapid development of Vision-Language Instruction-Tuning (VLIT), evaluating these models becomes increasingly important, for which several benchmarks (Yin et al., 2024; Xu et al., 2024; Liu et al., 2024b) have been released recently.

Different from these existing benchmarks that concentrate on evaluating VLIT models directly, our goal is one step back: **evaluating VLIT datasets**. The motivation comes from the insights into current VLIT models, including two similarities and one difference. *The first similarity is the model architecture* as shown in Fig. 1. The image feature is firstly extracted by a frozen vision encoder (Fang et al., 2023). Then, a learnable projection module, which can be simply designed either as the linear layer in LLaVA (Liu et al., 2024a) or a more sophisticated one like Q-Former in InstructBLIP (Dai et al., 2023), transforms the image feature to the text space. Finally, by feeding the transformed image feature and instruction text into the frozen Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023), the instruction-following answer is generated. *The second similarity is the multi-stage learning scheme*. During training, common large-scale image-text pairs (Ordonez et al., 2011; Schuhmann et al., 2021; Sharma et al., 2018) are leveraged for the cross-modal feature alignment in the prior stage. Then, the customized high-quality instruction data is used to train the VLIT model to generate coherent and desired output in the later stage. *The difference is exactly the high-quality instruction data targeting at different aspects of VL understanding*, as concluded in Table 1. To be more consistent with

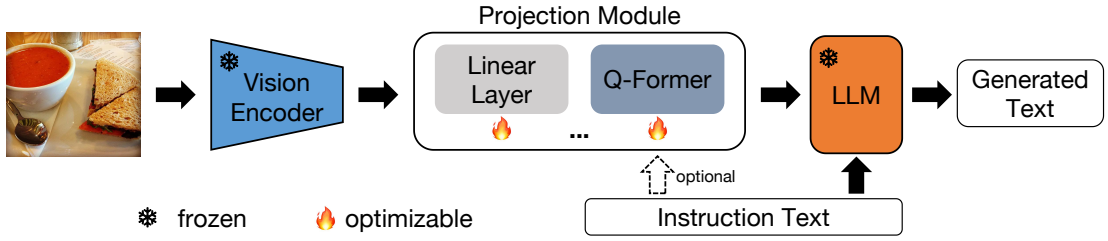


Figure 1: The popular architecture in current vision-language instruction tuning methods (Dai et al., 2023; Liu et al., 2024a). Extracting the visual feature by the image encoder, transferring the visual feature into the language space, and generating text output via a frozen Large Language Model (LLM).

LLMs, the annotations in these datasets are almost generated or augmented by GPT. It follows that curating proper instruction tuning datasets is essential in VLIT, which motivates us to evaluate VLIT datasets and look into their quality.

However, there exist limitations when using current benchmarks for evaluation. The style of annotations in benchmarks (Xu et al., 2024; Liu et al., 2024b) is quite different from the style of the open-ended texts generated by LLMs, causing possible bias for assessment. Besides, human voting (Xu et al., 2024) and ChatGPT/GPT-4 (OpenAI, 2023) are leveraged for performance evaluation. While the former is labor-intensive and liable to cause subjective evaluation, and the latter is inconvenient and unstable for widespread use because of the API availability and the changeable output. Additionally, it is worth noting that although evaluations utilizing GPT models have demonstrated the highest agreement with human evaluations (Bitton et al., 2023), both GPT models and human evaluation are not well-suited for large-scale evaluation scenarios due to practical considerations.

To conduct a comprehensive analysis of VLIT datasets, we introduce a pioneering *tune-cross-evaluation* paradigm shown in Fig. 2. This paradigm allows us to thoroughly assess the datasets. The fundamental concept is that each dataset serves a dual purpose: it can be utilized for model development and also function as a benchmark for the specific aspect it was designed to address. Our evaluation paradigm benefits from annotations consistent with LLMs, enabling us to define the Meta Quality (MQ) as the average score measured by caption metrics, including BLEU, METEOR, and ROUGE-L. This model-free and human-free evaluation strategy, utilizing MQ to measure performance in each tune-evaluation experiment set, offers greater convenience and stability than GPT-involved scoring and a more objective assessment than human voting. Building upon the proposed MQ, we devise the concepts of Dataset Quality (DQ) and Sample Quality (SQ) to measure the overall capability of each dataset and sample, combining all tune-evaluation sets.

Taking a step further, the other goal in this paper is **refining VLIT datasets** according to the holistic evaluation on a set of VLIT datasets. On one hand, existing VLIT models are only equipped with one or several abilities in VL understanding, which leads to unsatisfying performance in comprehensive evaluations. On the other hand, existing benchmarks build evaluation datasets by collecting datasets from different tasks (Krizhevsky et al., 2009; Lu et al., 2022) with annotations inconsistent with the open-ended generated texts (Xu et al., 2024; Yin et al., 2024), which causes inaccurate evaluation. As a result, a dataset encompassing multiple VLIT capabilities is critical for developing an all-powerful model and building an unbiased benchmark in a convenient way.

To this end, we build the so-called REVO-LION dataset by REfining VisiOn-Language InstructiOn tuNing datasets, which is composed of samples with higher SQ from each dataset as listed in Table 1. As a compact subset of the original datasets, REVO-LION is shown empirically to be more sample-efficient than simply merging the raw datasets together, which validates the effectiveness of the proposed SQ and the refinement strategy. We make following contributions:

(1) We propose a paradigm namely *tune-cross-evaluation* for the holistic analysis on VLIT datasets.

(2) We define a model-free and human-free evaluation metric, Meta Quality (MQ), as the mean score measured by BLEU, METEOR, and ROUGE-L. Based on MQ, Dataset Quality (DQ) and Sample

Table 1: Popular vision-language instruction tuning datasets on current VLIT methods. These datasets are used to build the proposed REVO-LION in this paper.

Datasets	Size	Purpose
DetGPT (Pi et al., 2023)	50K images and around 30K query-answer pairs	Reasoning-based object detection.
LAMM (Yin et al., 2024)	186K image-language instruction-response pairs	Daily conversation, factual knowledge reasoning, detailed description,visual task dialogue.
LLaVAR (Zhang et al., 2023)	16K high-quality instruction following data.	Text-rich image understanding
LLaVA (Liu et al., 2024a)	58K in conversations, 23K in detailed description, 77K in complex reasoning.	Conversations, detailed description, complex reasoning.
Macaw (Lyu et al., 2023)	69K image instances.	Human-written style text generation.
MiniGPT-4 (Zhu et al., 2023)	Around 3.5K image-text pairs.	Comprehensive image description.
LRV (Liu et al., 2023)	Around 120K instances.	Robust visual instruction with mitigated hallucination issue.

Quality (SQ) are devised to quantify the quality of each dataset and sample in VLIT, respectively.

(3) We collect and release a comprehensive dataset called REVO-LION, by refining public mainstream VLIT datasets. REVO-LION consists of a training set for developing a highly capable VLIT model and an evaluation set that serves as an effective benchmark.

2 Related Work

2.1 Vision-Language Instruction Tuning

With the success of ChatGPT and InstructGPT (Ouyang et al., 2022) in solving tasks aligned with human instructions, subsequent Large Language Models (LLMs) (Taori et al., 2023; Peng et al., 2023; Ding et al., 2023; Zhou et al., 2024; Du et al., 2022; Chiang et al., 2023) have been further devised by fine-tuning open-source LLMs (Touvron et al., 2023; Zeng et al., 2022) using instruction data (Wang et al., 2023) in the last two years.

Standing on the shoulder of LLMs, many Vision-Language Instruction Tuning (VLIT) models (Su et al., 2023; Ye et al., 2023; Luo et al., 2024; Li et al., 2023a) have been proposed within a year. These models are similarly constructed by using a projection module to connect the pre-trained vision model for visual perception and the language model for text generation. The projection module is firstly trained on common image-text pairs for VL alignment, then on high-quality data for instruction tuning. One of the most impactful methods is InstructBLIP (Dai et al., 2023), which is built upon the VL alignment achieved by the Q-Former in BLIP2 (Li et al., 2023b). After collecting and transforming 28 datasets from 11 tasks into instruction format, InstructBLIP (Dai et al., 2023) takes the instruction as a guidance of Q-Former to extract instruction-aware visual features for further tuning. Similar to InstructBLIP, MiniGPT-4 (Zhu et al., 2023) is firstly pre-trained on large-scale datasets (Ordonez et al., 2011; Schuhmann et al., 2021) for VL alignment, then curates around 3500 high-quality instruction data, with the assistance of ChatGPT and Vicuna (Chiang et al., 2023) targeting at comprehensive image description, for instruction tuning in the second stage. LRV (Liu et al., 2023) constructs a dataset including both positive and negative instructions for robust tuning with mitigated hallucination issues based on MiniGPT-4. Simpler than MiniGPT-4, LLaVA (Liu et al., 2024a) adopts a linear layer to bridge the gap between visual and language space in the first stage using 595K image-text pairs filtered from CC3M. Then, by using ChatGPT and GPT-4, 158K instruction samples including conversations, detailed descriptions, and complex reasoning are collected in LLaVA for instruction tuning in the second stage. Similar to LLaVA (Liu et al., 2024a), DetGPT (Pi et al., 2023) collects around 30K query-answer pairs towards reasoning-based object detection for instruction tuning in the second stage, LLaVAR (Zhang et al., 2023) enhances the text-rich image understanding ability by collecting 16K text-rich image data, Macaw (Lyu et al., 2023) builds a dataset consisting of 69K instances for human-style text generation.

To make a brief summary, existing VLIT models mostly share the similar model architecture and the two-stage learning scheme. The major difference lies in the instruction data used in the second stage. Beyond current VLIT models targeting at certain aspects, collecting a comprehensive dataset lies the foundation for developing an all-powerful VLIT model.

2.2 VLIT Benchmarks

With the rapid development of VLIT models, how to comprehensively and effectively evaluate these models becomes a concurrent significant problem. To this end, several benchmarks (Zeng et al., 2024; Yu et al., 2023; Bitton et al., 2023) have been proposed in the last few months. The pioneering benchmark is the LVLM-eHub (Xu et al., 2024), which evaluates VLIT models by quantifying the performance and human voting in the online arena platform. Immediately after LVLM-eHub, LAMM (Yin et al., 2024) is proposed for evaluation on 9 common image tasks by collecting 11 datasets. Except for task-specific metrics, LAMM adopts GPT as a judgment for performance evaluation. However, MME (Fu et al., 2023) argues that human voting and GPT scoring bring problems of subjectivity and inaccuracy. For this, MME exams perception and cognition abilities covering 14 subtasks by manually constructing instruction-answer pairs and leading the tested models to answer “yes" or “no", which is designed for objective and accurate quantitative statistics. Nevertheless, such performance evaluation that heavily relies on generating “yes" or “no" is not quite reasonable, because existing VLIT models usually target at detailed tasks instead of making decisions from “yes" or “no" strictly. For fine-grained ability assessment, MMBench (Liu et al., 2024b) curates a dataset covering 20 fine-grained skills, and all instances are transformed into multi-choice problems. For robust evaluation, it employs ChatGPT for answer extraction and judgment in the proposed circular evaluation strategy, which is unable to evaluate the models directly on the generated texts, causing inaccurate assessment.

In short, there are three aspects that are not fully satisfied in existing benchmarks: 1) collecting datasets with annotations consistent with open-ended generated texts for evaluation; 2) avoiding human subjectivity in data selection and evaluation; 3) designing stable and convenient quantification metrics. We argue that it is possible to meet these conditions via our *tune-cross-evaluation* paradigm on VLIT datasets with the proposed quality metrics at both dataset and sample levels. In particular, based on our deep dive in Sec. 2.1, we propose shifting the focus of model evaluation, which existing benchmarks are paying great efforts on, to dataset evaluation.

3 Methodology

3.1 Tune-Cross-Evaluation Paradigm

As shown in Fig. 2, we propose the *tune-cross-evaluation* paradigm to evaluate VLIT datasets, which are specifically listed in Table 1. Note that these datasets are all in English such that we do not need to handle the language bias problem, which is also not the focus of this paper. On one hand, each dataset is employed to develop a model by instruction tuning. On the other hand, because these VLIT datasets are almost annotated by leveraging GPT-4 (OpenAI, 2023) or ChatGPT for text generation or augmentation, each dataset also represents a standard on the aspect that the dataset is constructed towards, by which the proper annotations consistent with open-ended generated texts are accessible. Based on the VL alignment learned in the first stage by the model with the architecture in Fig. 1, at each time, we select one dataset from these datasets for instruction tuning, and the remaining datasets are used for test at this time. For example, when we use DetGPT (Pi et al., 2023) for instruction tuning, the tuned model equipped with great reasoning-based object detection ability will be further tested on other datasets, and they involve testing the model’s ability such as daily conversation, factual knowledge reasoning, detailed description, etc. By taking turns to cycle in this way, we finally get the comprehensive quality evaluation of each dataset and each sample. To quantify the comprehensiveness, we define the Meta Quality (MQ), Dataset Quality (DQ) and Sample Quality (SQ), and detail them in the following sections.

3.2 Meta Quality (MQ)

In LVLM-eHub (Xu et al., 2024), the authors show that metrics in caption tasks are ineffective for VLIT evaluation due to the style differences between the diverse open-ended generated texts and the ground-truths in the datasets curated prior to LLMs, which are outdated compared to LLMs. Benefiting from the proposed *tune-cross-evaluation* paradigm, when making full use of VLIT datasets as evaluation ones, the proper annotations, which are created by GPT models to be consistent with LLMs, are available. Therefore, with mitigated style differences, to perform a model-free and human-free evaluation, in which we do not rely on other models such as GPT or human for scoring, we define the Meta Quality (MQ) as the average of scores measured by caption metrics to quantify the

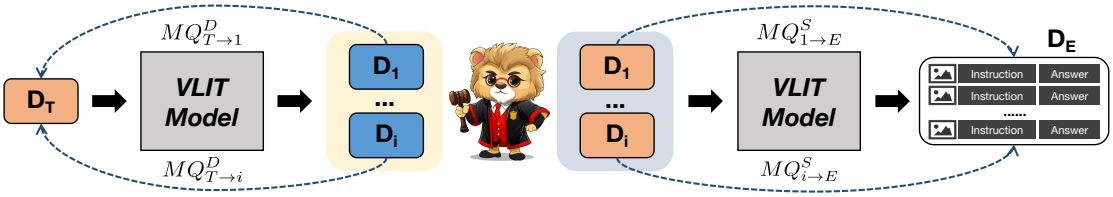


Figure 2: The overall framework of the proposed *tune-cross-evaluation* paradigm. *Left*: The diagram of Dataset Quality (DQ) evaluation. Each dataset adopted for testing measures the quality of the tuning dataset D_T on the aspect that the testing datasets are constructed towards. *Right*: The diagram of Sample Quality (SQ) evaluation. Each dataset used for tuning measures how well the samples in the testing set D_E match with the ability that the tuning dataset is constructed towards.

one-side quality of each dataset or sample within a single tune-evaluation experiment. Considering the time-consuming process in calculating sample-wise MQ if using SPICE, we use BLEU@1 (B@1), BLEU@2 (B@2), BLEU@3 (B@3), BLEU@4 (B@4), METEOR (M), and ROUGE-L (R) as the components for MQ definition. CIDEr is set as a hold-out metric in data refinement in Sec. 4.4. The MQ is formulated as:

$$MQ = mean(\sum_{i=1}^{i=4} B@i + M + R). \quad (1)$$

The ablation of the combinations is studied in Sec. 4.2. It should be noted that the MQ can be commonly used to measure on a set of samples. When the number of samples is 1, it actually measures the sample-wise quality. For distinction, we denote the MQ measured on a dataset and a sample as MQ^D and MQ^S, respectively.

3.3 Dataset Quality (DQ)

In the proposed *tune-cross-evaluation* paradigm, each time we select a dataset denoted as D_T from the set of datasets S for instruction tuning, the remaining datasets denoted as $D_i (i \in S, i \neq T)$ are then leveraged as evaluation ones for inference, thus measuring the quality of the tuning dataset on the aspect that the evaluation datasets are constructed towards one by one, as shown on the left side of Fig. 2. Note that though the datasets to be evaluated in this paradigm are in different sizes, we do not explicitly separate a metric for data size. When using a dataset as an evaluation one, its size has been implicitly integrated into the measurement. A dataset with larger size is more likely to exhibit better comprehensive ability, as compared between LLaVA-Detailed description and MiniGPT-4, which both concentrate on detailed image description in Table 2 in experiments. In a single tune-evaluation set, the one-side dataset quality is denoted as $MQ^D_{T\to i}$, in which the right arrow indicates the direction from the tuning dataset to the evaluation dataset. Specifically, we set the quality $MQ^D_{T\to T}$ that each tuning dataset exhibits on its aspect as 1, the maximum value of MQ. Therefore, when a dataset is set as the tuning one, its comprehensive quality measured by all capabilities in S is formulated as the sum of all one-side qualities:

$$\begin{aligned} DQ_T &= MQ^D_{T\to T} + \sum_{i\in S, i\neq T} MQ^D_{T\to i} \\ &= 1 + \sum_{i\in S, i\neq T} MQ^D_{T\to i}, T \in S. \end{aligned} \quad (2)$$

By setting each dataset as the tuning one and the remaining as evaluation ones in turn, the comprehensive DQ, which measures various capabilities, for all datasets can be calculated.

3.4 Sample Quality (SQ)

Because the MQ can only be calculated on the inference datasets, it is hard to measure the quality of each sample in the tuning dataset when keeping the same evaluation direction in DQ, i.e., the inference datasets are regarded as standards. In contrast, when a dataset D_E is set as the inference one, we hold that the model equipped with the ability of dataset $D_i (i \in S, i \neq E)$, after tuned on which, is supposed to be a standard. By this way, the $MQ^S_{i\to E}$ for each sample in D_E measures how close the sample matches with the ability of the tuning dataset D_i, as shown on the right side of Fig. 2. To calculate the comprehensive quality that each sample exhibits on other aspects, other than DQ having the ability corresponding to itself, we define the SQ as a weighted sum:

$$SQ_E = \sum_{i\in S, i\neq E} DQ_i \cdot MQ^S_{i\to E}. \quad (3)$$

Table 2: DQ evaluated on SPLIT1 and SPLIT2 by using the Q-Former based architecture.

D_T	DetGPT	LAMM	LLaVAR	LLaVACo	LLaVADe	LLaVARe	Macaw	MiniGPT-4	LRV
Q-Former+SPLIT1	2.55	2.63	2.49	2.68	2.40	**2.85**	2.31	2.38	**1.99**
Q-Former+SPLIT2	2.56	2.64	2.50	2.67	2.41	**2.83**	2.32	2.37	**1.99**

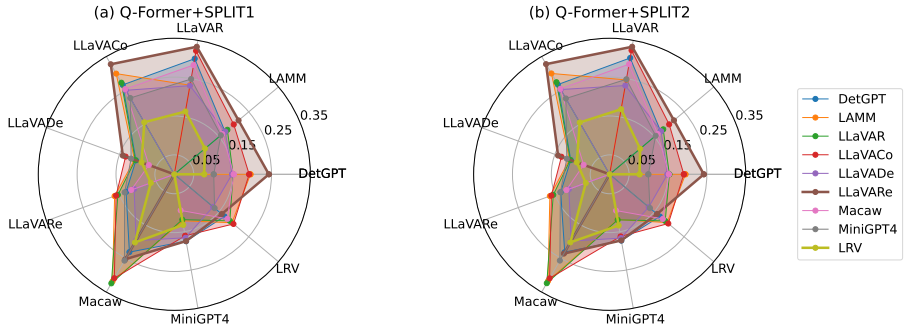


Figure 3: Visualizations of $MQ^D_{T\to i}(i \neq T)$ in dataset quality evaluation. Lines with different colors represent different datasets D_T used for instruction tuning.

We use the DQ_i as the weights for objective evaluation, the higher DQ_i represents a more confident evaluation when using dataset D_i to tune the model. By setting each dataset as the inference one and the remaining as tuning ones in turn, the comprehensive SQ for each sample exhibits on other datasets can be calculated.

3.5 REVO-LION

To build a comprehensive dataset integrating all capabilities of the evaluated datasets, a simple yet direct way is to merge these datasets into one without more operations. As suggested in the analysis (Zeng et al., 2024), data quality is more significant than data quantity. Therefore, we propose to REfine VisiOn-Language InstructiOn tuNing (REVO-LION) datasets according to the proposed SQ, which measures the comprehensive quality of each sample exhibits on other datasets. To preserve all capabilities, we collect samples with higher SQ from each dataset to compose REVO-LION. Formally, we denote the portion that the number of selected samples to the number of all samples in each dataset as P. The lower bound of SQ in dataset $D_i(i \in S)$ corresponding to the portion P is τ_i^P. For each sample $\boldsymbol{x}_i^k \in D_i$, if the SQ of it SQ_i^k is no lower than τ_i^P, the sample is collected in REVO-LION, which is formulated as:

$$S1 = \bigcup_{i \in S} \boldsymbol{x}_i^k, \quad (\boldsymbol{x}_i^k \in D_i, SQ_i^k >= \tau_i^P). \tag{4}$$

We denote this refinement strategy as $S1$, which is validated to be more effective than "Random Refinement" ($S2$) and "Gaussian Refinement" ($S3$) in Sec. 4.4. After performing the data evaluation and creating REVO-LION from the datasets in Table 1, we split it into a training set and an evaluation set. The former can serve as a common dataset for developing an all-powerful VLIT model, and the latter can serve as a convenient benchmark covering all capabilities of these datasets and equipping with ideal annotations, based on which the caption metrics can be conveniently employed for model-free and human-free evaluation.

4 Experiments

The evaluated VLIT datasets are clarified in Table 1. In our main experiments, we adopt the architecture of InstructBLIP (Dai et al., 2023). Implementation details are delivered in the Appendix A. In addition, experiments of data evaluation and refinement based on the architecture adopting the linear layer as the projection module are given in Appendix B.

4.1 DQ and SQ Evaluation

By setting each dataset as the tuning one D_T, its one-side qualities measured by other datasets $MQ^D_{T\to i}(i \neq T)$ are visualized in Fig. 3. The areas enclosed by brown and yellow lines are the largest and smallest, indicating that LLaVA-Reasoning and LRV hold the greatest and poorest comprehensive capability. It follows that LLaVA-Reasoning exhibits the highest DQ and LRV shows the lowest DQ among these datasets, as shown in Table 2 computed by Eq. 2. We infer its reason as that LLaVA-Reasoning includes various problems of which the difficulty varies from low to high. As

Table 3: Ablation study on the definition of MQ. The blue numbers after the results represent their relative rankings. The bold blue numbers indicate the inconsistent ranking relations.

D_T (Q-Former+SPLIT1)		DetGPT	LAMM	LLaVAR	LLaVACo	LLaVADe	LLaVARe	Macaw	MiniGPT-4	LRV
C1	DQ	2.55 **(4)**	2.63 **(3)**	2.49 (5)	2.68 (2)	2.40 (6)	2.85 (1)	2.31 (8)	2.38 (7)	1.99 (9)
	$MQ^D_{T\to Eval600}$	1.37 **(3)**	1.35 **(4)**	1.27 (5)	1.42 (2)	1.16 (6)	1.54 (1)	1.11 (8)	1.13 (7)	0.78 (9)
C2	DQ	2.48 (5)	2.61 (3)	2.57 (4)	2.66 **(2)**	2.38 (8)	2.73 **(1)**	2.46 (6)	2.40 (7)	2.20 (9)
	$MQ^D_{T\to Eval600}$	0.64 (5)	0.70 (3)	0.69 (4)	0.73 **(1)**	0.57 (8)	0.71 **(2)**	0.63 (6)	0.59 (7)	0.49 (9)
C3	DQ	2.79 (6)	2.95 (3)	2.94 **(4)**	3.00 **(2)**	2.72 (8)	3.05 **(1)**	2.83 (5)	2.72 (7)	2.59 (9)
	$MQ^D_{T\to Eval600}$	0.50 (6)	0.56 (3)	0.57 **(2)**	0.59 **(1)**	0.47 (8)	0.54 **(4)**	0.53 (5)	0.48 (7)	0.44 (9)

shown in Fig. 6 in the Appendix C, easy reasoning problems may be similar to description problems, while hard reasoning problems may require logical thoughts. As a result, LLaVA-Reasoning exhibits the greatest comprehensive capability. Besides, the results achieved on SPLIT1 and SPLIT2 demonstrate a high degree of consistency, indicating that the DQ evaluation can provide common and objective data analysis.

In addition, detailed evaluation cases of SQ are delivered in Appendix E.

4.2 Ablation Study on MQ and DQ

To validate the rationality of the definition of MQ, based on which DQ is devised, we perform ablation studies on 3 combinations of MQ, in which C1 refers Eq. 1, C2 and C3 are defined as:

$$\begin{aligned} C2: MQ &= mean(B@4 + M + R); \\ C3: MQ &= mean(M + R). \end{aligned} \tag{5}$$

Given a dataset, its DQ quantified by a reasonable evaluation criteria should be consistent with its performance in the comprehensive evaluation. By setting *Eval600* as the comprehensive evaluation set and $MQ^D_{T\to Eval600}$ as the performance quantification, according to the three definitions, the results and relative orders of DQ and $MQ^D_{T\to Eval600}$ achieved on SPLIT1 are shown in Table 3. Compared with C3, C1 and C2 make more consistent results between DQ evaluation and $MQ^D_{T\to Eval600}$. Because a dataset owning higher DQ should exhibit better all-sided ability, and perform better in the comprehensive evaluation, C1 and C2 are more rational than C3. To preserve a more general evaluation covering as many metrics as possible, we choose C1 as the final definition of MQ, based on which DQ and SQ are devised.

4.3 Single Dataset VS. Merged Dataset

To build a comprehensive dataset integrating all capabilities, a simple yet direct way is to add all these single datasets together into one, denoted as "Merge". By setting *Eval600* as the evaluation dataset, the $MQ^D_{T\to Eval600}$ achieved by setting each single dataset and the merged dataset as tuning one D_T is compared in Table 4. The simply merged dataset achieves the greatest result, showing adding all datasets together can contribute to an all-powerful model that exhibits the best performance on the comprehensive evaluation set covering all capabilities.

4.4 REVO-LION and Ablation Study on Refinement Strategy

It has been validated that combining all datasets together can develop an all-powerful model in a comprehensive evaluation compared with single datasets in Sec. 4.3. Considering that data quality is more significant than data quantity (Zeng et al., 2024), we further perform data refinement based on the above holistic evaluation. Specifically, we collect part of the samples from each dataset to build a comprehensive dataset. In addition to the refinement strategy defined in Eq. 4, denoted as $S1$, we design another two strategies for comparisons. The second strategy, namely $S2$, collects the samples from each dataset randomly with the same amount as in $S1$. The third strategy $S3$ adopts the Gaussian distribution for sample selection. Specifically, for each dataset $D_i (i \in S)$, we calculate the mean value μ_i and the standard deviation σ_i of SQ of the samples in D_i. The sample whose SQ exists within an interval of λ times the standard deviation σ_i around the mean value μ_i will be selected. $S3$ is formulated as:

$$\begin{aligned} S3 &= \bigcup_{i \in S} \boldsymbol{x}_i^k, \\ (\boldsymbol{x}_i^k \in D_i, SQ_i^k &\in [\mu_i - \lambda \cdot \sigma_i, \mu_i + \lambda \cdot \sigma_i]). \end{aligned} \tag{6}$$

We adopt CIDEr, the hold-out metric in defining MQ, to measure the comprehensive performance of the model tuned on the refined dataset, thus making

Table 4: $MQ^{D}_{T\rightarrow Eval600}$ on SPLIT1 and SPLIT2 by using the Q-Former based architecture.

D_T	DetGPT	LAMM	LLaVAR	LLaVACo	LLaVADe	LLaVARe	Macaw	MiniGPT-4	LRV	Merge
Q-Former+SPLIT1	1.37	1.35	1.27	1.42	1.16	1.54	1.11	1.13	0.78	**1.64**
Q-Former+SPLIT2	1.38	1.36	1.29	1.43	1.18	1.55	1.12	1.12	0.79	**1.64**

Table 5: Evaluation on *Eval600* measured by CIDEr using refinement strategies $S1$ and $S2$ with the portion P ranging from 10% to 100%. "Nums" refers to the number of samples in the refined dataset for tuning.

Portion (P)		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Q-Former+SPLIT1	Nums	92828	185650	278473	371293	464115	556940	649760	742584	835406	928225
	$S1$-CIDEr	163.43	168.56	171.64	174.54	175.13	177.21	194.70	177.16	176.64	175.49
	$S2$-CIDEr	165.21	168.63	170.18	171.91	172.82	174.37	174.22	175.62	176.03	175.49
Q-Former+SPLIT2	Nums	92807	185608	278410	371211	464012	556815	649616	742418	835219	928017
	$S1$-CIDEr	165.03	170.87	173.56	174.89	175.99	178.33	178.87	178.23	178.80	175.49
	$S2$-CIDEr	165.81	169.49	172.40	173.89	175.78	176.23	177.16	178.23	179.17	175.49

an objective evaluation for the data refinement. By setting *Eval600* as the comprehensive evaluation set, and selecting a portion P of samples in each dataset, the result comparisons between $S1$ and $S2$ are given in Table 5. The setting when $P = 100\%$ refers to simply adding all datasets together.

For the refinement strategy S1, the results when $P \in [50\%, 90\%]$ are all competitive and even better than those when using the simply merged dataset. It shows that S1 successfully collects high-quality samples in the refined dataset. Specifically, the CIDEr rises with the increase of P from 10% to 70%. When we select the top 50% samples with higher SQ from each dataset, we can already achieve competitive performance comparable to those using the entire data. Then, the CIDEr achieves the highest when $P = 70\%$. When P increases from 70% to 100%, the CIDEr results decrease, which is caused by the involvement of samples with lower SQ. Besides, comparing S1 with $S2$, when keeping the number of collected samples from each dataset the same, results achieved by selecting samples with higher SQ are almost better than those achieved by random selection, which validates that S1 is more effective than $S2$. Moreover, for the refinement strategy $S2$, the CIDEr rises with the increase of P from 10% to 90%. It demonstrates that with the lack of effective data evaluation and refinement strategies, a direct way for improving the performance is just expanding the scale of datasets.

In addition, results from the refinement strategy $S3$ when setting the times λ within [1.0, 1.5, 2.0] are given in Table 6. Comparing the results achieved by setting $\lambda = 1.0$ in $S3$ with those achieved by setting $P = 70\%$ in S1 in Table 5, though more samples are collected in $S3$, the performance achieved by S1 with fewer samples is better. The same phenomenon also occurs in the comparison between setting $\lambda \in [1.5, 2.0]$ in $S3$ and setting $P = 90\%$ in S1 in Table 5. The comparisons prove that S1 is more effective than $S3$. According to the above, the effectiveness of the proposed *tune-cross-evaluation* paradigm and the refinement strategy is systematically validated.

Table 6: Evaluation on *Eval600* measured by CIDEr using the refinement strategy $S3$ by setting $\lambda \in [1.0, 1.5, 2.0]$. "Nums" refers to the number of image-instruction-answer triplets.

Times (λ)		1.0	1.5	2.0
Q-Former+SPLIT1	Nums	697374	838771	880426
	CIDEr	173.94	175.07	176.52
Q-Former+SPLIT2	Nums	697346	838650	880206
	CIDEr	175.88	178.45	179.32

On this basis, we aim to release the REVO-LION dataset by performing the evaluation and refinement on the original datasets without partitions. Details are given in Appendix D.

5 Conclusions and Outlook

In this paper, we pioneer the analysis of VLIT datasets and propose the *tune-cross-evaluation* paradigm. A model-free and human-free metric, namely Meta Quality (MQ), is defined for meta evaluation. It has been extended to Dataset Quality (DQ) and Sample Quality (SQ) for quantitative evaluation. Based on the holistic evaluation, we build a refined dataset REVO-LION by collecting samples with higher SQ, which is proved to be sample efficient with great performance. In the released version, REVO-LION includes a train set for developing an all-powerful model, and an evaluation set to serve as a convenient yet stable benchmark.

Limitations

The evaluation paradigm are only limited to the datasets analyzed in this paper. The more datasets with various capabilities are involved in the evaluation, the more comprehensive analysis is achieved. As a result, the refined dataset can be used to develop a VLIT model performing well in more aspects, and also as a more comprehensive evaluation benchmark. Incorporating more datasets into the proposed evaluation paradigm is flexibly allowed.

References

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: a benchmark for vision-language instruction following inspired by real-world use. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26898–26922.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. 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*, pages 3558–3568.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024. Visual instruction tuning with polite flamingo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17745–17753.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky and 1 others. 2009. Learning multiple layers of features from tiny images.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2024. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and 1 others. 2023. Detgpt: Detect what you need via reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14172–14189.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, FZJ-2022-00923. Jülich Supercomputing Center.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*, pages 2556–2565.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, and 1 others. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, and 1 others. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. 2024. What matters in training a gpt4-style language model with multimodal inputs? In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7930–7957.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Implementation Details

Data Preparation. To ensure that each dataset is independent of each other and has no overlapping samples, in DetGPT (Pi et al., 2023), we remove samples generated from MiniGPT-4 (Zhu et al., 2023); in LLaVAR (Zhang et al., 2023), we remove samples generated from LLaVA (Liu et al., 2024a). Concentrating on the vision-language field, in LAMM (Yin et al., 2024) and Macaw (Lyu et al., 2023), we only use the released image-text data. In addition, the data in LLaVA (Liu et al., 2024a) is divided into three independent ones: LLaVA-Conversation (LLaVACo), LLaVA-Detailed description (LLaVADe), LLaVA-Reasoning (LLaVARe) for their clear difference. In our main experiments, we adopt the architecture of InstructBLIP (Dai et al., 2023).

To validate the effectiveness of the proposed data refinement strategy, we need to design an evaluation set covering all capabilities of these datasets. For this, we collect 80% samples from each dataset to build independent training sets, on which the *tune-cross-evaluation* paradigm and refinement are performed, and collect 600 samples from the remaining 20% samples to build a balanced and comprehensive evaluation set, namely *Eval600*, as shown in Fig. 4. We choose 600 samples from each dataset for testing because the smallest dataset MiniGPT-4 (Zhu et al., 2023) includes about 3,500 samples, and the 20% includes no more than 700 samples. To build a balanced and comprehensive evaluation set, we finally set the number of selected samples from each dataset for evaluation as 600. To verify the effectiveness of the proposed evaluation paradigm and data refinement on different data partitions, and ensure the universality of experimental effect verification, we perform such data split twice and get two sets, denoted as SPLIT1 and SPLIT2.

Table 7: The hyperparameters for instruction tuning using the architecture of InstructBLIP (Dai et al., 2023), which adopts the Q-Former as the projection module.

Hyperparameters	
Epochs	5
Warmup Epochs	1
Warmup initial learning rate	1e-8
Warmup end learning rate	1e-5
Warmup Schedule	Linear
Learning rate decay	Cosine
End (Minimum) learning rate	0
Batch size	128
Optimizer	AdamW
AdamW β	(0.9, 0.999)
Weight decay	0.05

Table 8: The hyperparameters for instruction tuning using the architecture of LLaVA (Liu et al., 2024a), which adopts the linear layer as the projection module.

Hyperparameters	
Epochs	3
Learning rate	2e-5
Learning rate decay	Cosine
Batch size	128
Optimizer	AdamW
Weight decay	0.0

Instruction Tuning. The learnable projection module is the Q-Former in BLIP2 (Li et al., 2023b), the vision encoder is the pre-trained ViT-G/14 from EVA-CLIP (Fang et al., 2023), and the language model is Vicuna-7B (Chiang et al., 2023). Specifically, based on the selected vision encoder and language model, the Q-Former used for instruction tuning has been pre-trained on 129M images (Li et al., 2023b), including COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011) and LAION400M (Schuhmann et al., 2021). Based on the official code of InstructBLIP (Dai et al., 2023), the learning hyperparameters during instruction tuning are listed in Table 7. Each dataset has been adopted for tuning on 8 Nvidia A100 (80G) GPUs with the vision encoder and language model kept frozen, only parameters in the Q-Former are optimized.

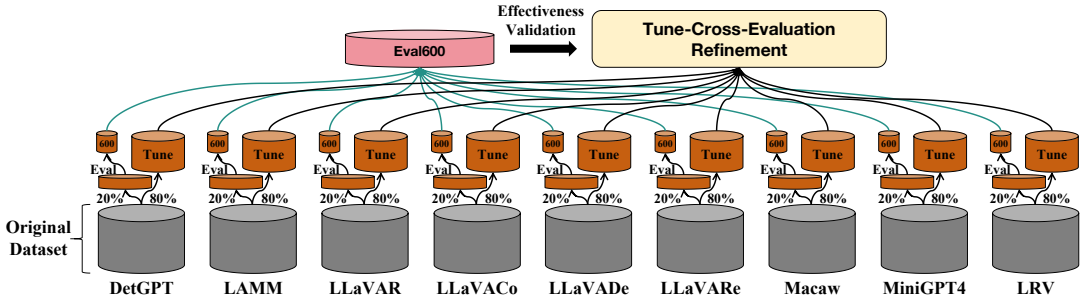


Figure 4: The diagram of the data split process. It is designed to validate the effectiveness of the proposed *tune-cross-evaluation* paradigm and the data refinement strategy in main experiments. Each original dataset is divided into two parts: 80% samples are collected as a tuning set for data evaluation and refinement, and 600 samples from the remaining 20% are collected into a balanced and comprehensive evaluation set. For robust validation, we perform such partitions twice, thus creating SPLIT1 and SPLIT2 that are used in the main experiments.

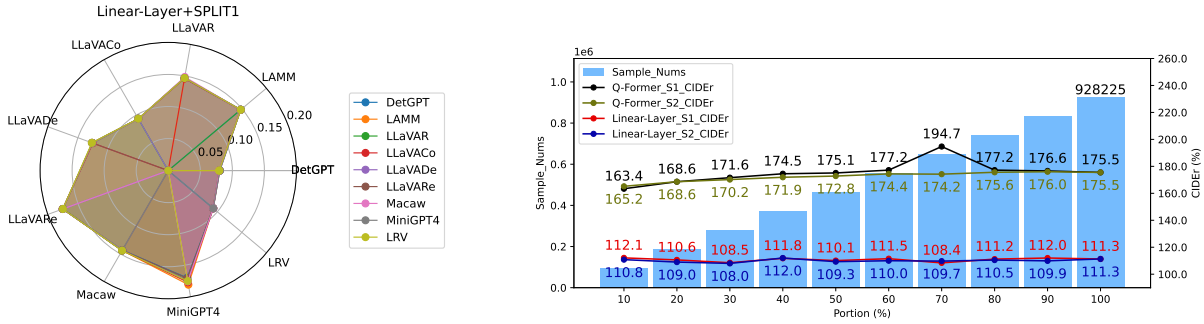


Figure 5: Results of data evaluation and refinement using the linear projection-based architecture. (Left) Visualizations of $MQ^D_{T\to i}(i \neq T)$ in DQ evaluation using the linear projection module. (Right) Result comparisons between using the Q-Former andthe linear layer for projection using strategies $S1$ and $S2$.

B Data Evaluation and Refinement using the Linear Projection Module

For a supplementary, we perform the data evaluation and refinement using SPLIT1 based on the architecture adopting the linear layer as the projection module for VL alignment. Specifically, we take the architecture of LLaVA (Liu et al., 2024a). The vision encoder is the pre-trained ViT-L/14 in CLIP (Radford et al., 2021), and the language model is Vicuna-7B (Chiang et al., 2023). The linear layer used for instruction tuning has been pre-trained on 558K image-text pairs from LAION (Schuhmann et al., 2021), CC (Sharma et al., 2018) and SBU (Ordonez et al., 2011). We adopt the official code of LLaVA (Liu et al., 2024a) for instruction tuning with their default learning hyperparameters, which are given in Table 8. Each dataset has been adopted for tuning on 8 Nvidia A100 (80G) GPUs with the vision encoder and language model kept frozen, and only parameters in the linear layer are optimized.

When setting each dataset as the tuning one D_T, its one-side qualities $MQ^D_{T\to i}(i \neq T)$ measured by other datasets are given in Fig. 5 (Left). It shows that each dataset exhibits extremely high similarity in the dataset-wise evaluation, leading to almost equal DQ for each dataset, compared with the results in Fig. 3. Consequently, using the linear projection-based VLIT model cannot effectively distinguish differences among datasets, resulting in invalid data evaluation. In addition, based on the evaluation, we perform the data refinement using strategies $S1$ and $S2$. The refinement results achieved by using the Q-Former and the linear layer for projection are shown in Fig. 5 (Right). Obviously, when keeping both the vision encoder and the language model frozen, using the linear projection module results in a much more unsatisfying performance than using the Q-Former. Then, taking a deep comparison between the results achieved by S1 and $S2$ using the linear layer-based architecture, the CIDEr results vary within a small range when the portion P of selected samples in each dataset changes. The highest result in S1 refinement, which is higher than using all the data, is achieved when only collecting 10% of samples

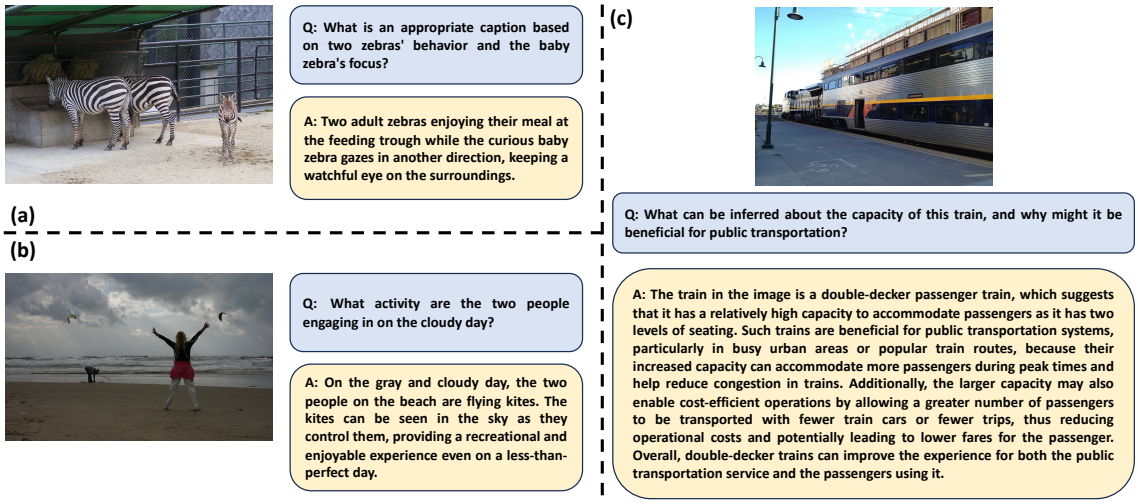


Figure 6: Three samples in LLaVA-Reasoning. (a) and (b) are easy reasoning problems, and similar to problems of describing images. (c) is a hard reasoning problem requiring logical thoughts. Q: question. A: answer.

with higher SQ from each dataset. It shows that as a much simpler projection module, the linear layer does not need as much high-quality instruction data as the Q-Former. The simplicity of the projection module limits the greatest performance that can be improved by expanding the data scale. Besides, compared with $S2$, the strategy $S1$ is almost better with different portions.

Except for the effectiveness of $S1$, which has been validated compared with $S2$, other results are inconsistent with ones using the architecture of InstructBLIP, and the linear projection module is not as good as the Q-Former. We make the deep analysis as follows. (1) From the perspective of the architecture, linear projection is quite simple in transferring the visual feature to the language space. While Q-Former adopts the pre-trained BERT (Kenton and Toutanova, 2019) as initialization, and extracts the desired visual feature according to the texts using a more sophisticated cross-attention mechanism. (2) From the perspective of the pre-trained dataset, both the linear layer and the Q-Former have been pre-trained on large-scale image-text pairs for VL alignment before instruction tuning. As demonstrated in Sec A, Q-Former in BLIP2 (Li et al., 2023b) has been pre-trained on 129M images (Li et al., 2023b) from COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011) and LAION400M (Schuhmann et al., 2021). While the linear layer in LLaVA (Liu et al., 2024a) has been pre-trained only on 558K image-text pairs from LAION (Schuhmann et al., 2021), CC (Sharma et al., 2018) and SBU (Ordonez et al., 2011). The significant difference between the amount of pre-training dataset results in a much poorer VL alignment of the linear projection than the Q-Former.

C Supplemented Figures

The figures that are referenced in the main manuscript are presented in this section.

Three samples form LLaVA-Reasoning are visualized in Fig. 6, demonstrating its great diversity covering problems vary from easy to hard.

D REVO-LION Release

According to the validated effectiveness on specific data preparation in above experiments, to release the REVO-LION, the evaluation and refinement are performed on the original datasets without partitions. As analyzed in Appendix B, using the linear layer as the projection module for VL alignment is inferior to using the Q-Former. Therefore, we adopt the architecture of InstructBLIP, the detailed setting of which is delivered in Appendix A, for data evaluation and refinement.

In the released dataset REVO-LION, the *tune-cross-evaluation* paradigm is directly performed on each original dataset without partition, as shown in Fig 7. According to the results in Table 5, setting the portion $P = 70\%$ can achieve the best performance. Therefore, we release the dataset with setting $P = 70\%$. After refining each dataset, we divide it into a train set for instruction tuning and an evaluation set as a convenient yet stable benchmark. To keep a balanced dataset for eval-

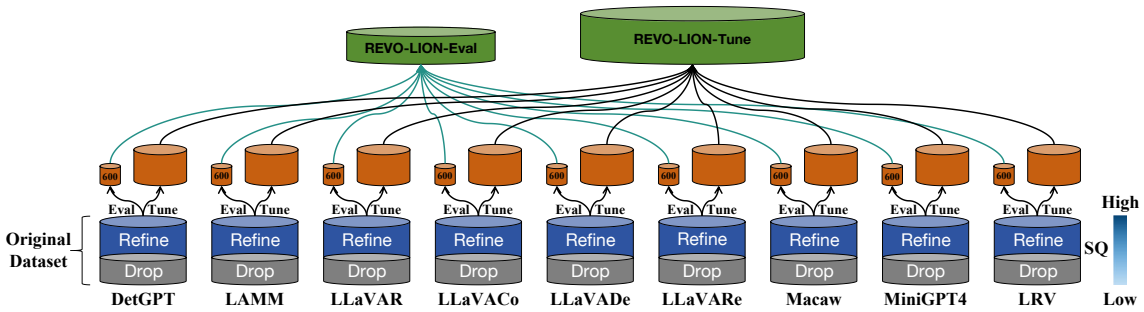


Figure 7: The refining process of buiding REVO-LION from existing VLIT datasets. The proposed *tune-cross-evaluation* paradigm is directly performed on each original dataset without partition. After the holistic evaluation, the top 70% samples with higher SQ in each dataset are collected, in which 600 samples are collected into the balanced and comprehensive evaluation benchmark, namely REVO-LION-Eval, and the remaining are collected into the refined tuning dataset, namely REVO-LION-Tune, for developing an all-powerful model.

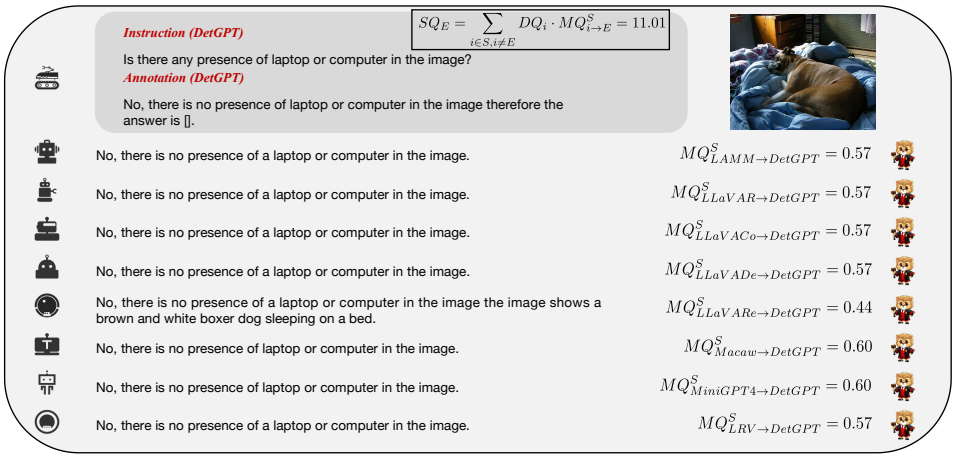


Figure 8: A sample in DetGPT with high SQ measured by other datasets.

uation, we select 600 samples from each refined dataset to build the evaluation set, namely REVO-LION-Eval. The remaining samples in each refined dataset are combined into the instruction tuning dataset, namely REVO-LION-Tune. Statistics of number of image-instruction-answer triplets in the REVO-LION-Tune is given in the Table 9. Moreover, as the annotations in REVO-LION share the same style of open-ended texts generated by LLMs, the caption metrics can be directly adopted for a model-free and human-free performance evaluation when using the REVO-LION-Eval as the benchmark.

Table 9: Statistics of number of image-instruction-answer triplets in the released REVO-LION-Tune dataset.

Datasets	Size
DetGPT (Pi et al., 2023)	21486
LAMM (Yin et al., 2024)	352035
LLaVAR (Zhang et al., 2023)	23084
LLaVACo (Liu et al., 2024a)	177807
LLaVADe (Liu et al., 2024a)	15668
LLaVARe (Liu et al., 2024a)	53051
Macaw (Lyu et al., 2023)	45187
MiniGPT-4 (Zhu et al., 2023)	1808
LRV (Liu et al., 2023)	101588
Sum	791714

E SQ Evaluation Cases

Based on the proposed *tune-cross-evaluation* paradigm, the sample-level data quality are evaluated by the proposed SQ. Taking the Fig. 8 as an example, by calculating $MQ^S_{i \to E}(i \neq E)$ between the generated answers and the annotation, the comprehensive quality of this sample can be obtained. Due to the high similarity among the generated answers, the calculated SQ of this sample is quite high.

More evaluation samples with high SQ and low SQ are presented from Fig. 9 to Fig. 25

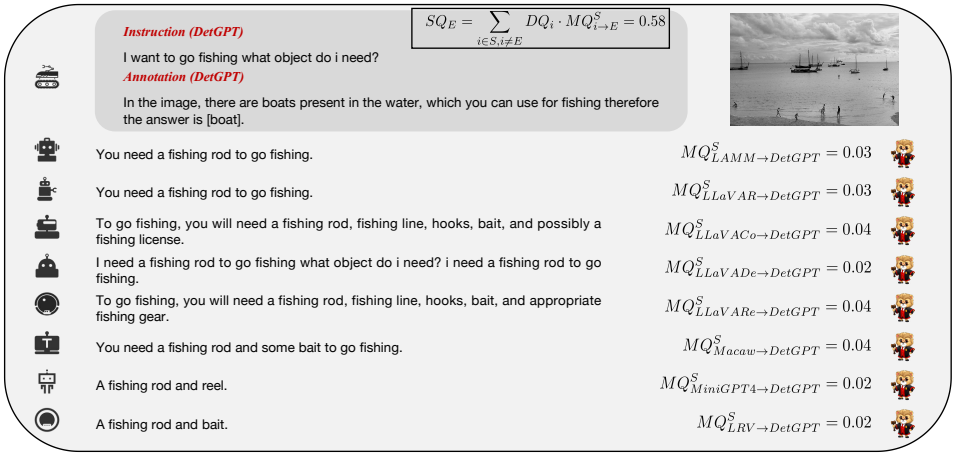
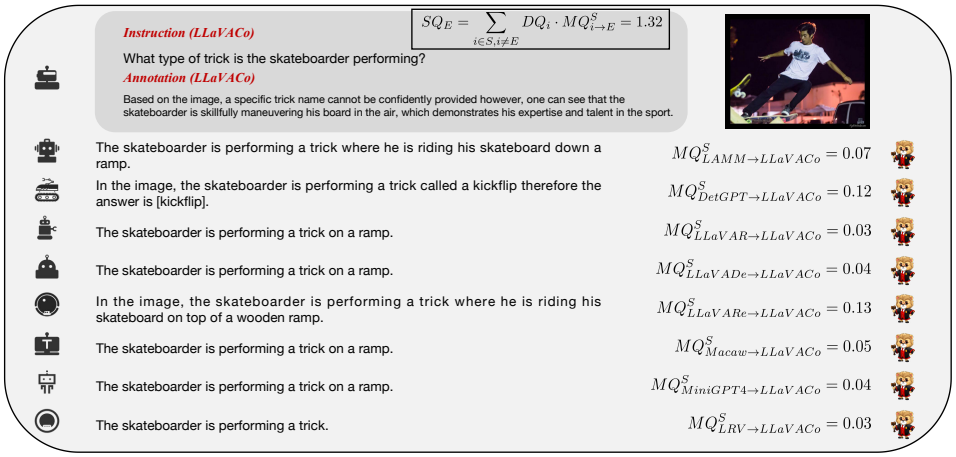


Figure 9: A sample in DetGPT with low SQ measured by other datasets.



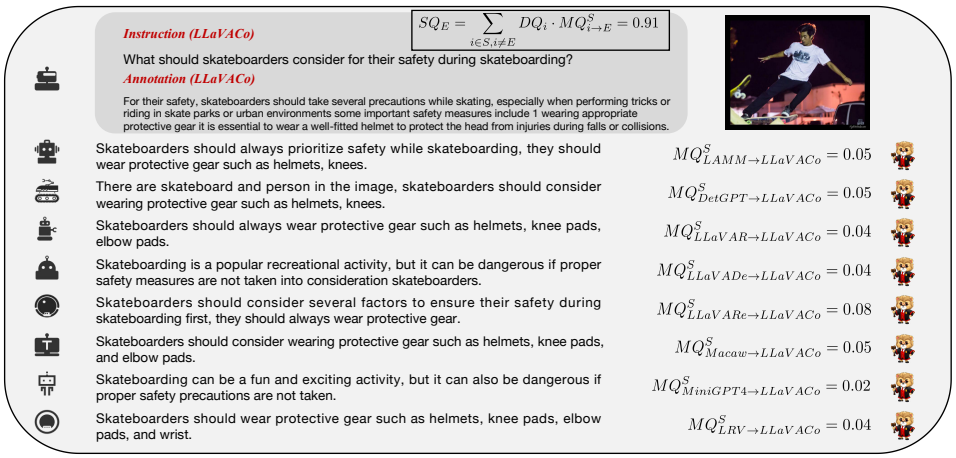
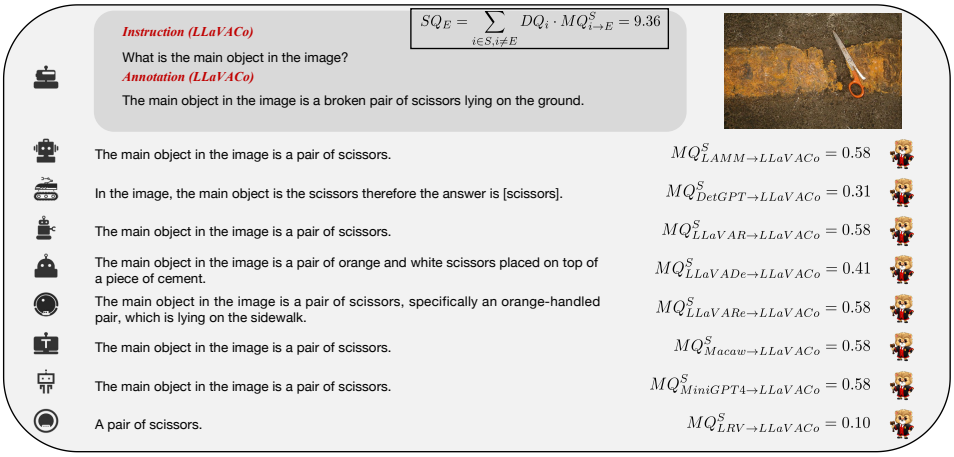
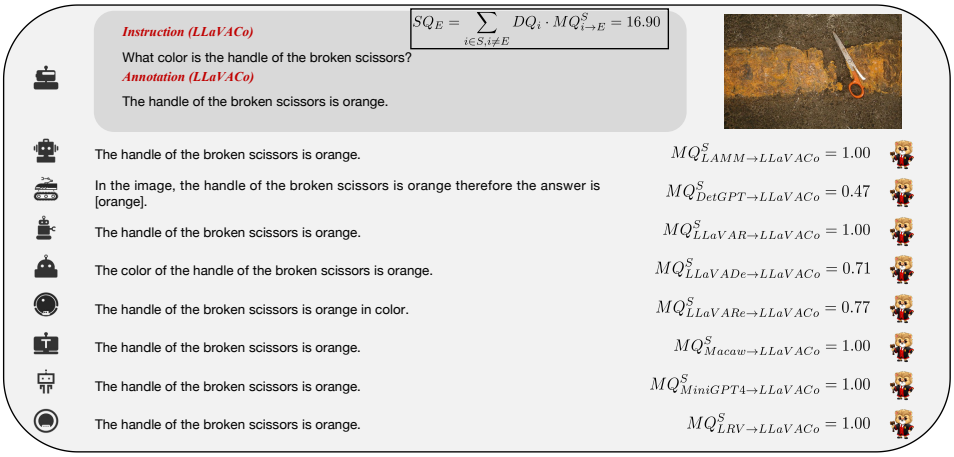


Figure 10: A conversation sample in LLaVA-Conversation with low SQ measured by other datasets.





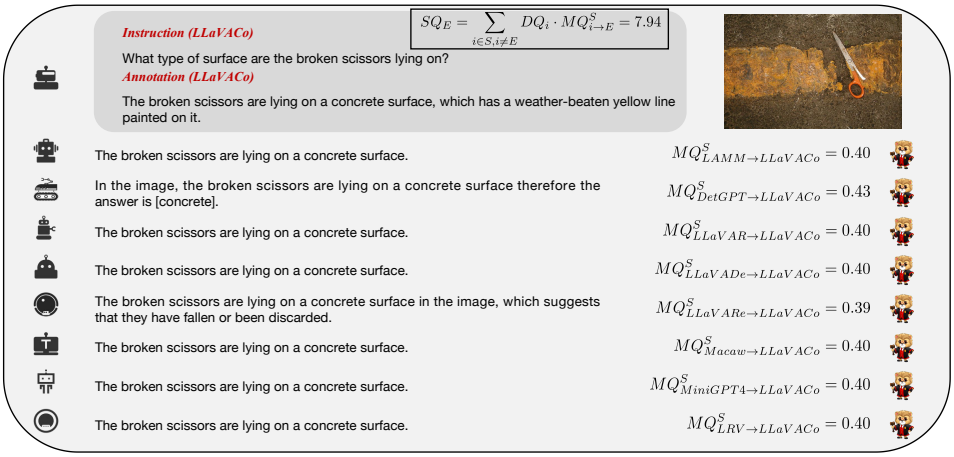


Figure 11: A conversation sample in LLaVA-Conversation with high SQ measured by other datasets.

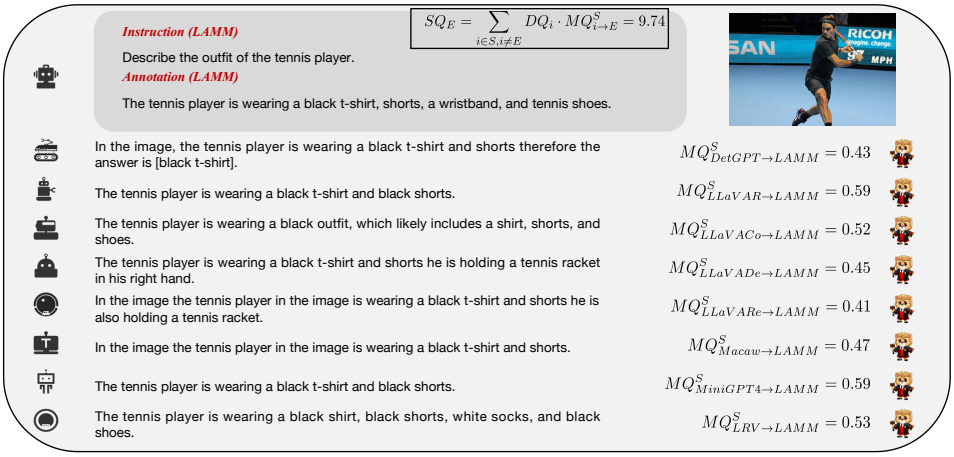


Figure 12: A sample in LAMM with high SQ measured by other datasets.

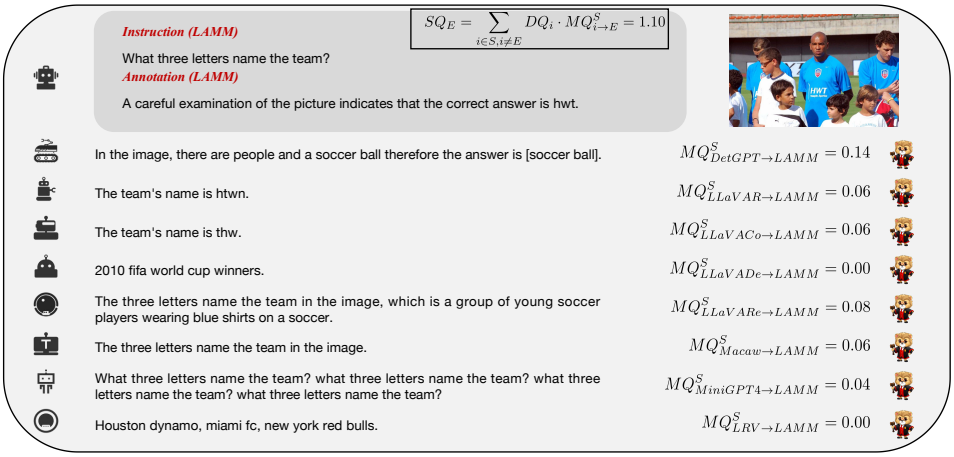


Figure 13: A sample in LAMM with low SQ measured by other datasets.

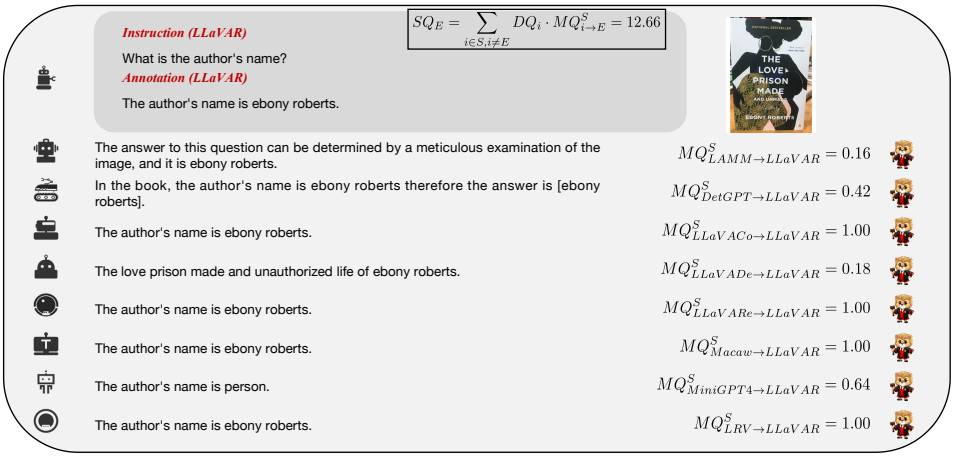


Figure 14: A sample in LLaVAR with high SQ measured by other datasets.

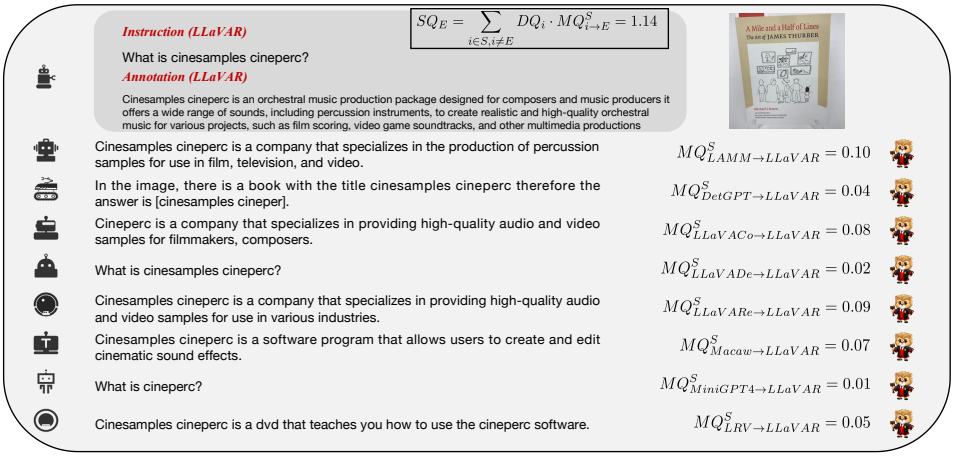


Figure 15: A sample in LLaVAR with low SQ measured by other datasets.

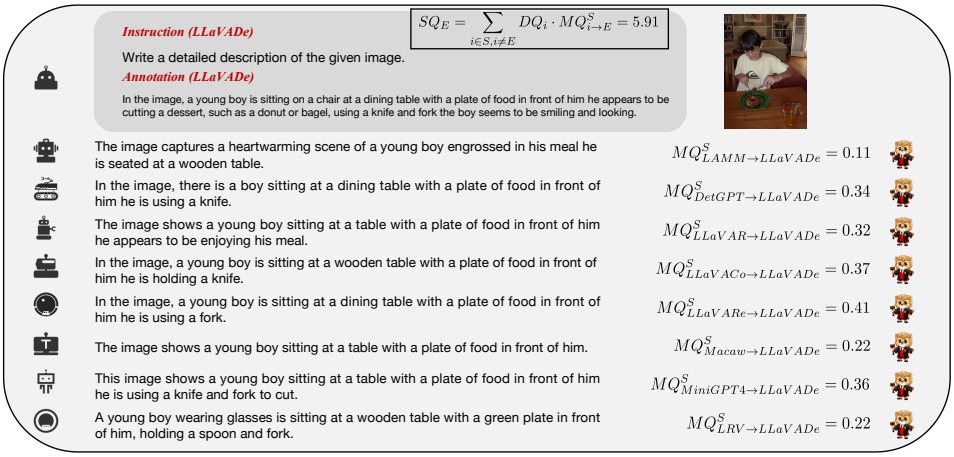


Figure 16: A sample in LLaVA-Detailed description with high SQ measured by other datasets.

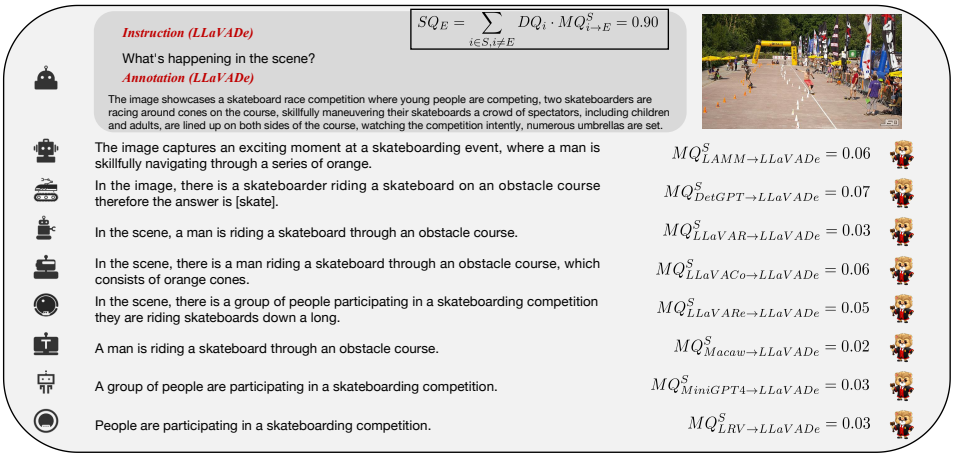


Figure 17: A sample in LLaVA-Detailed description with low SQ measured by other datasets.

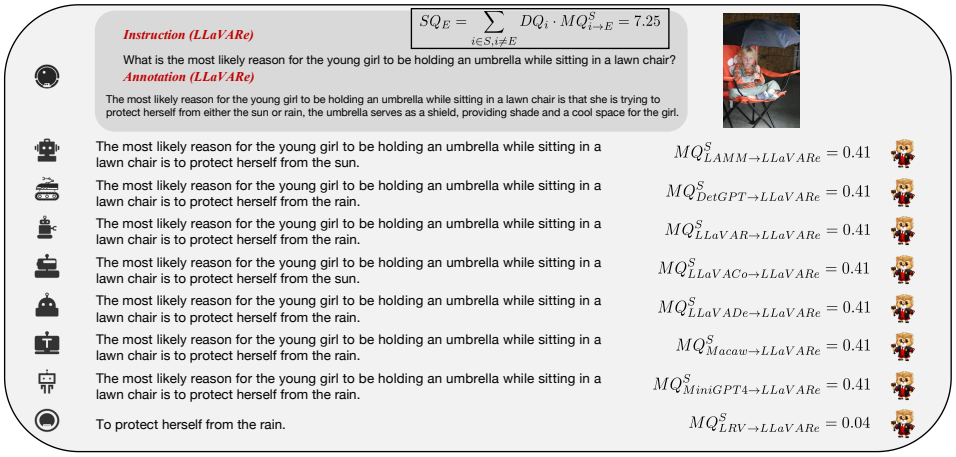


Figure 18: A sample in LLaVA-Reasoning with high SQ measured by other datasets.

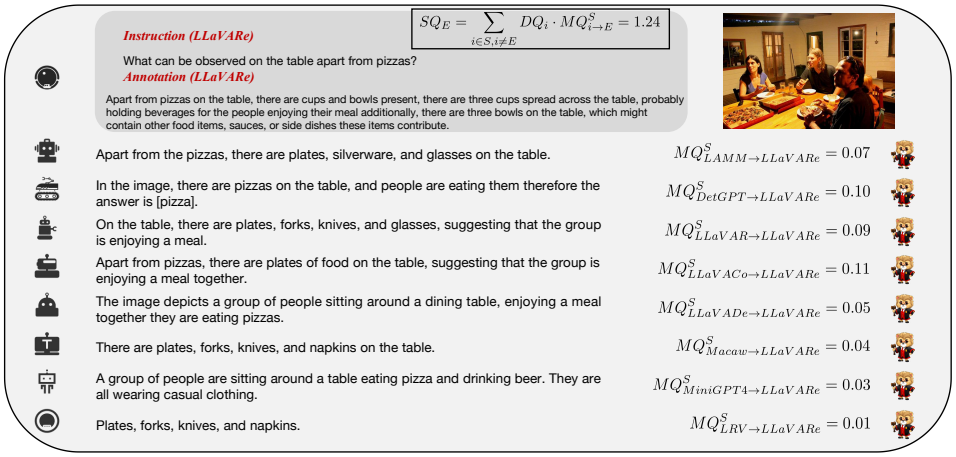


Figure 19: A sample in LLaVA-Reasoning with low SQ measured by other datasets.

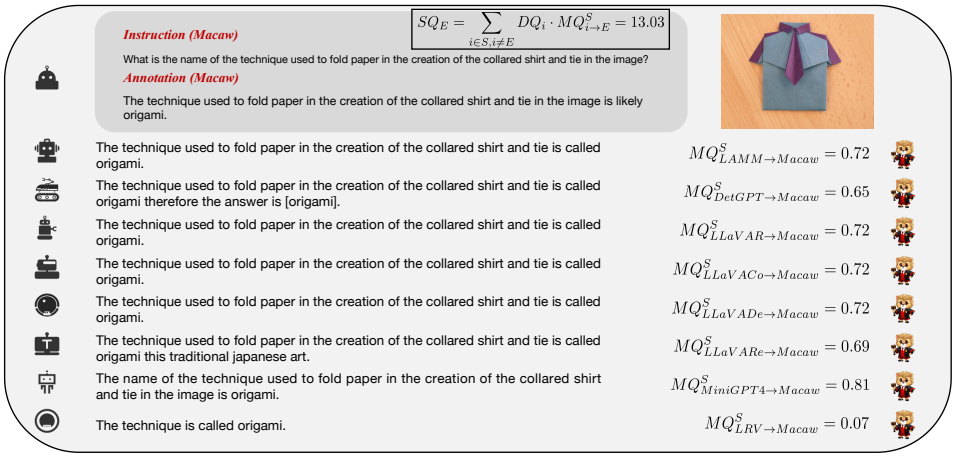


Figure 20: A sample in Macaw with high SQ measured by other datasets.

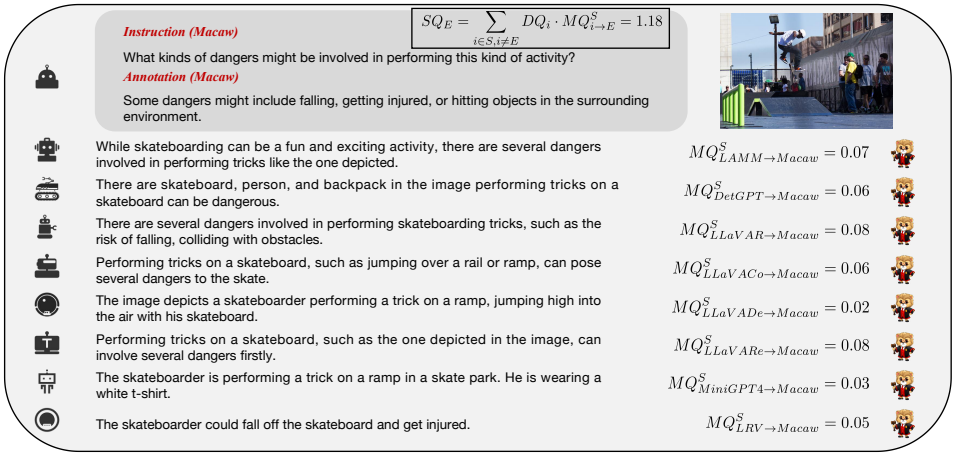
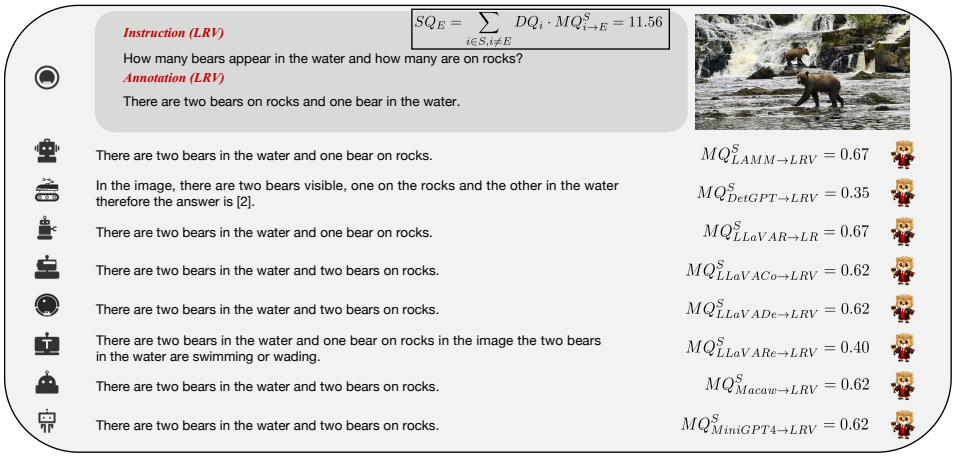


Figure 21: A sample in Macaw with low SQ measured by other datasets.



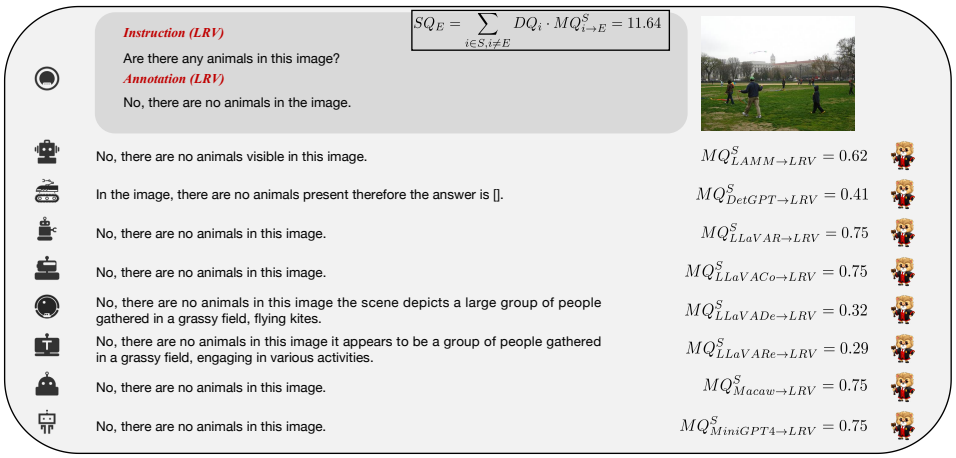


Figure 22: Two samples in LRV with high SQ measured by other datasets.

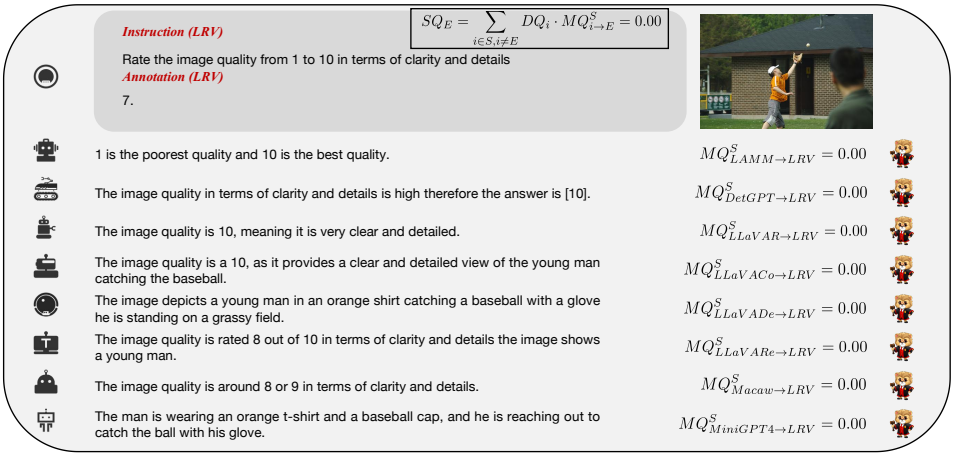


Figure 23: A sample in LRV with low SQ measured by other datasets.

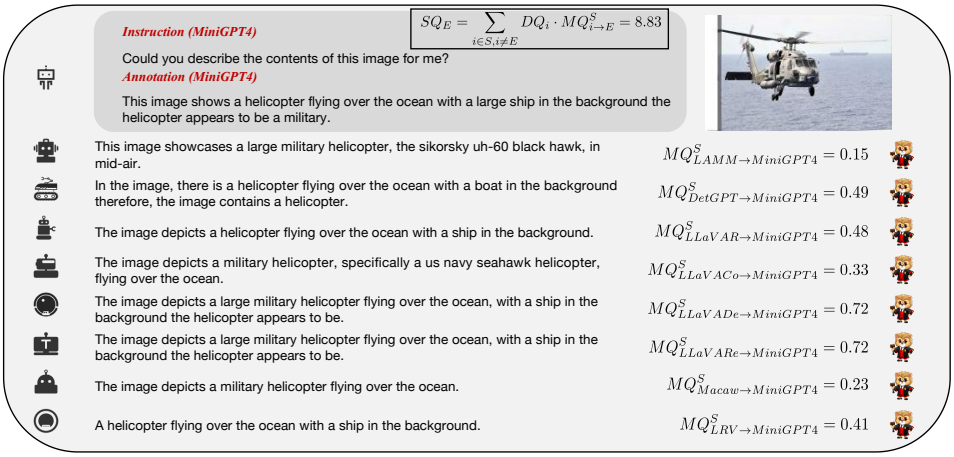


Figure 24: A sample in MiniGPT-4 with high SQ measured by other datasets.

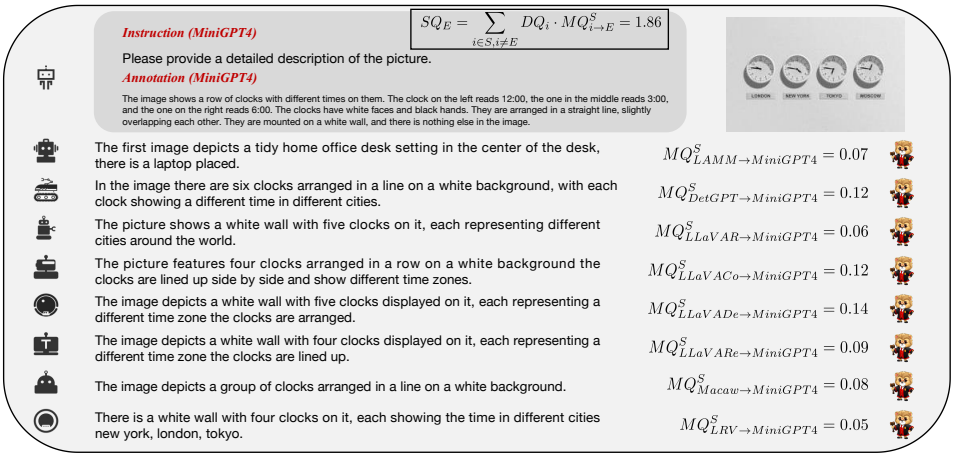


Figure 25: A sample in MiniGPT-4 with low SQ measured by other datasets.