Picking the Cream of the Crop: Visual-Centric Data Selection with Collaborative Agents

Anonymous ACL submission

Abstract

To improve Multimodal Large Language Models' (MLLMs) ability to process images and complex instructions, researchers predominantly curate large-scale visual instruction tuning datasets, which are either sourced from existing vision tasks or synthetically generated using LLMs and image descriptions. However, they often suffer from critical flaws, including misaligned instruction-image pairs and low-quality images. Such issues hinder training efficiency and limit performance improvements, as models waste resources on noisy or irrelevant data with minimal benefit to overall capability. To address this issue, we propose a **Vi**sual-Centric **S**election approach via **A**gents Collaboration (ViSA), which centers on image quality assessment and image-instruction relevance evaluation. Specifically, our approach consists of 1) an image information quantification method via visual agents collaboration to select images with rich visual information, and 2) a visual-centric instruction quality assessment method to select high-quality instruction data related to high-quality images. Finally, we reorganize 80K instruction data from large open-source datasets. Extensive experiments demonstrate that ViSA outperforms or is comparable to current state-of-the-art models on seven benchmarks, using only 2.5% of the original data, highlighting the efficiency of our data selection approach. Moreover, we conduct ablation studies to validate the effectiveness of each component of our method.

1 Introduction

Recent advancements in multimodal large language models (MLLMs) have led to significant progress in general visual understanding (OpenAI, 2023; Dai et al., 2023a; Liu et al., 2023). By integrating state-of-the-art large vision models (LVMs) with large language models (LLMs) and training on vast amounts of visual data, MLLMs have demonstrated the ability to comprehend complex images, achieving remarkable performance across a variety of visual tasks, including object recognition (Zhang et al., 2024b), visual question answering (Yu et al., 2024), and image captioning (Levinboim et al., 2021). The training of MLLMs generally involves two main steps: first, pretraining on image-caption pairs to align the large vision model with the language model; second, instruction tuning on downstream visual tasks to enable the model to better understand complex images and perform sophisticated visual tasks. To improve the multimodal understanding and task-solving abilities of MLLMs, existing approaches typically collect large-scale visual task data from a variety of sources for constructing instruction tuning datasets (Li et al., 2024a; Tong et al., 2024).

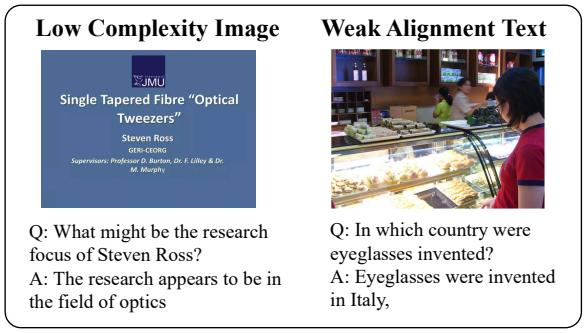


Figure 1: Two common issues in visual instruction tuning datasets. (Left) Low-Complexity Image: The image lacks meaningful visual information. (Right) Low-Quality Text: The instruction is weakly aligned with the image.

Specifically, existing works have primarily focused on collecting open-source training datasets from existing visual tasks (Li et al., 2024a; Liu et al., 2024a; Tong et al., 2024), or synthetically generating various instructions using GPT-4 based on the collected images and description (Liu et al., 2023, 2024b). Such approaches enable the collection of large amounts of instructional data at a relatively low cost, without the need for extensive manual annotation. However, they often introduce

significant noise into the training datasets. For example, images with low complexity or irrelevant instruction-image pairs can be problematic. As illustrated in Figure 1, the left subfigure shows an image of a presentation slide, which contains only a simple icon and basic presentation information. Such images provide little useful visual content or knowledge, making it difficult for models trained on such images to generalize to more complex real-world applications, where informative visuals are essential for learning robust visual representations. Another common issue is when the instruction is weakly aligned with the image, as shown in the right subfigure, where the instruction is a commonsense question that can be answered without referencing the image. This type of data fails to effectively link the visual model's perception with the textual knowledge stored in the language model, reducing the efficiency of multimodal learning.

To address this issue, we present a novel approach called **Vi**sual-Centric **S**election via **A**gents Collaboration (ViSA). This method leverages collaboration among multiple visual agents to assess the quality of visual data, selecting high-quality images and instruction pairs to enhance the training efficiency of MLLMs. The approach consists of two key components: 1) **Visual Information Quantification**, where agents collaborate to quantify visual elements and assess the richness of diverse image perspectives, selecting informative images; and 2) **Image-Centric Instruction Quality Quantification**, which focuses on selecting instructions tightly coupled with the images. This is achieved by leveraging multiple agents to calculate metrics such as prior token perplexity and image-text mutual information, enabling the evaluation of response quality and its relevance to the visual content. Additionally, to improve the collaboration across multiple visual agents, we introduce a Shapley value based on the Pearson correlation coefficient to weigh the reliability of each agent's assessment.

To thoroughly assess the effectiveness of our method, we filter out 80K high-quality instruction data from the LLaVA-OneVision dataset (Li et al., 2024a). We then conduct extensive experiments to evaluate the training effectiveness of the filtered data. The results on seven visual benchmarks show that a model trained with less than 5% of the data used by baselines achieves performance comparable to, or even surpassing, state-of-the-art models, highlighting the importance of high-quality data for understanding intricate images. Furthermore, we observe significant improvements when continuing the training of a 72B model with a small set of high-information-density data, demonstrating that larger models can still benefit from informative images and well-aligned instructions. Overall, our agent collaboration-based data selection method effectively reduces noise in training datasets, leading to a significant increase in MLLM training efficiency.

Our main contributions can be summarized as follows:

- We propose a novel multi-agent framework for quantifying and selecting high-quality visual instruction data. This framework evaluates both image informativeness and instruction relevance, leveraging a diverse set of visual agents to assess general visual elements as well as diverse image-specific features. To the best of our knowledge, this is the first work to explicitly introduce image informativeness evaluation in the context of visual data selection.

- We introduce a new instruction quality quantification method based on prior token perplexity and image-text mutual information, which measure the response quality for selecting instructions tightly coupled with the images, respectively. We propose the Shapley value approach based on the Pearson correlation coefficient to effectively combine the evaluations from multiple agents.

- We conduct extensive experiments and ablation studies, demonstrating that our approach significantly improves the training efficiency of MLLMs. Using only 80K selected samples, our method enables both 2B-scale and 7B-scale vision models to achieve performance on par with state-of-the-art models. Moreover, with a larger 72B-scale model, our high-quality selected data consistently enhances the model's performance in visual understanding and reasoning tasks.

2 Related Work

Vision Large Models : The development of Vision Large Models (VLMs) has led to significant advances in visual comprehension and reasoning (Li et al., 2022, 2023b; Zhu et al., 2024; Liu et al., 2023, 2024a; Li et al., 2025). The training of multimodal large models (MLLMs) typically consists of two key stages. First, vision-language alignment is achieved by mapping the visual representations obtained from a pretrained Large Visual Model (LVMs) (Sun et al., 2023; Radford

et al., 2021) into the LLM representation space through a learnable projection network (Merullo et al., 2023; Li et al., 2023b). Subsequently, the aligned models are fine-tuned using instruction data to enhance their ability to process complex visual instructions (Chen et al., 2023; Bai et al., 2023; Li et al., 2024c; Dai et al., 2023b). Recently, there has been a growing trend in high-resolution MLLMs (Guo et al., 2024; Li et al., 2024b; Bai et al., 2023; Wang et al., 2024). Consequently, the ability to benefit from complex images has become increasingly important as informative images that can maximize these models' potential. Hence, our work aims to develop a quantitative assessment of image informativeness to select high-quality data.

Vision Instruction Datasets : Instruction tuning aims to enhance a model's ability to understand and reason about images through diverse vision-language tasks, such as visual question answering (Antol et al., 2015), image captioning (Radford et al., 2021), and object detection (Ravi et al., 2024). MultiInstruct (Xu et al., 2023) introduced the first human-annotated multimodal instruction tuning dataset, designed to enhance the zero-shot capabilities of pretrained VLMs. Building upon this foundation, subsequent studies have increasingly focused on curating diverse image datasets and harnessing the power of large multimodal models, such as GPT-4V, to automatically generate large-scale visual instruction data (Li et al., 2024a; Xu et al., 2024). Several recent studies have explored different strategies to improve visual instruction tuning. Some works (Zhang et al., 2023; Hu et al., 2024) employ OCR tools and GPT-4 to generate instruction-following tasks that help VLMs better understand text within images. Meanwhile, recent works (Wang et al., 2023; Chen et al., 2024) directly prompt GPT-4V with images as inputs to generate visual instruction tuning data. Unlike previous works that focus on instruction generation strategies, our approach emphasizes high-quality data selection by systematically assessing image informativeness and image-instruction relevance.

3 Method

In this section, we introduce our visual-centric selection via agents collaboration. First, to identify informative images, we comprehensively consider various aspects of image information density assessment, including evaluation based on general visual elements §3.1 and assessment from other diversity visual information perspectives §3.2. Subsequently, we refine multimodal data selection on the textual level, incorporating text quality evaluation based on prior token perplexity and text relevance assessment based on mutual information §3.3. Finally, using our proposed method, we curated a high-quality instruction fine-tuning dataset §3.4.

3.1 Visual Elements Quantification

Evaluating image complexity from the perspective of visual elements is crucial for understanding image informativeness. However, existing methods primarily focus on object-level recognition or coarse-grained overall image assessment, lacking a fine-grain quantification of visual complexity. To address this, we assess image informativeness based on general visual elements (e.g., visual objects, graphical elements, characters, etc.), which are inherently linked to image complexity. Specifically, we introduce two evaluation metrics: **Segmentation Complexity Score (SC Score)** and **Object Alignment Score (OA Score)**. The SC Score is designed to assess the richness of graphical elements and characters in an image, leveraging a visual segmentation method. In contrast, the OA Score evaluates the richness of objects in an image using the object detection approach based on a predefined category set.

Specifically, for the SC Score, we predefine 512 anchor points and employ the SAM2 model (Ravi et al., 2024), which segments target regions based on the provided points, to generate a set of segmentation boxes denoted as M. The SC Score is then computed using the following formula:

$$\text{Score}(SC) = \text{count}(\text{IoU}(M, 0.75)). \quad (1)$$

To evaluate the richness of objects, motivated by prior grounding work (Ren et al., 2024), we first predefine a set of over 1800 common image object categories. Using the DINO model (Liu et al., 2024c), which identifies the presence and location of predefined tags within images, we detect the visual objects corresponding to each category. The OA Score is then computed by calculating the frequency of each category and normalizing it using the term frequency-inverse document frequency weight algorithm (TF-IDF), as shown in the following:

$$\text{Score}(OA) = \sum_i \text{TF-IDF}(C_i), \quad (2)$$

where C_i is the predicted frequency of the i-th

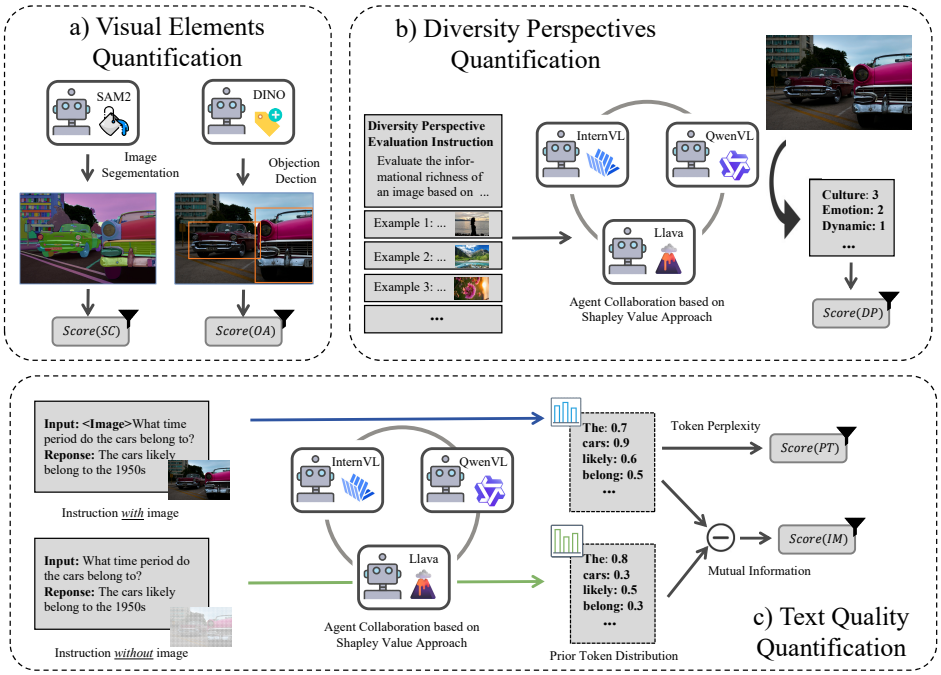


Figure 2: Overview of our agent collaboration for visual data selection. The $Score(SC)$ denotes the Segmentation Complexity Score. The $Score(OA)$ shows the Object Alignment Score. The $Score(DP)$ indicates the Diversity Perspective Score. The $Score(PT)$ denotes the Prior Token Perplexity Score. The $Score(IM)$ denotes the Image-Text Mutual Information Score.

category, which emphasizes the importance of rare objects within the image set.

3.2 Diversity Perspectives Quantification

We found that while the aforementioned visual element metrics provide a estimate of image informativeness with expectations, they are insufficient for fully capturing the richness of other types of image information. To address this limitation, we propose a multi-agent-based **Diversity Perspective Score (DP Score)**, designed to evaluate images from multiple perspectives and dimensions, including events, emotions, culture, composition, and dynamics. These diverse perspectives offer a more comprehensive view of image informativeness, accounting for the varied origins and functional purposes of image creation. Specifically, we design a detailed evaluation guideline, complete with multiple examples, prompting various state-of-the-art visual agents to assess the image from each perspective and produce a final integrated score. For the specific evaluation guideline used, please refer to Appendix A.

To integrate the final score of each agent, we use a Shapley value approach (Lundberg and Lee, 2017) based on the Pearson correlation coefficient to compute the weight of each agent, motivated by the idea of determining model weights based on score consistency. Specifically, the weight W_j of each agent is calculated as follows:

$$
\begin{aligned}
W_j &= \sum_{S \subseteq N \setminus \{j\}} \omega(S) \Delta\rho(S, j), \\
\omega(S) &= \frac{|S|!(|N| - |S| - 1)!}{|N|!}, \\
\Delta\rho(S, j) &= \rho(S \cup \{j\}) - \rho(S)
\end{aligned}
\tag{3}
$$

where N represents the set of all agents, and S is any subset of agents excluding the j-th agent. Pearsonr(S) is the Pearson correlation coefficient of the score distribution over the agents set S, measuring the linear relationship between the scores. The overall diversity perspective score for each image is then computed as the weighted average of the diversity perspective scores E_j from multiple agents.

$$
\text{Score}(DP) = \sum_{|N|}^{j} W_j E_j. \tag{4}
$$

3.3 Text Quality Quantification

To fully capture the visual and textual aspects of image informativeness, the quality of the associated text is also important. In this regard, we propose a multi-agent-based text quality quantification method, which includes two key metrics: the **Prior Token Perplexity Score (PT Score)**, which measures the quality of the textual content, and

the **Image-Text Mutual Information Score (IM Score)**, which assesses the degree of association between the text and the image. Similar to the approach in Section 3.1, we utilize three different agents to compute the PT Score and the IM Score separately, and then aggregate these scores using the Shapley value approach to obtain their comprehensive quality score.

Specifically, the PT Score is calculated by evaluating the perplexity of the response, which is defined as follows. Due to the significant variability in the length distribution of the data instructions, we adopt a more length robust approach to prior token perplexity (Liu et al., 2024e):

$$P_j = \exp\left(-\frac{1}{N}\sum_{i=1}^{K}\log P(w_i|w_{<i})\right), \quad (5)$$

$$\text{Score}(TP) = \sum_{N}^{j}\sum_{S\subseteq N\setminus\{j\}}\omega(S)\Delta\rho(S,j)P_j,$$

where K denotes the number of prior tokens. This score reflects how predictable or coherent the sequence of tokens is based on preceding words.

Next, we compute the IM Score, which measures the mutual dependence between the image and the associated text, formulated as following:

$$I_j = H(\text{Text}) - H(\text{Text}|\text{Image}), \quad (6)$$

$$\text{Score}(IM) = \sum_{N}^{j}\sum_{S\subseteq N\setminus\{j\}}\omega(S)\Delta\rho(S,j)I_j,$$

where $H(\text{Text})$ is the entropy of the text without the image input, and $H(\text{Text}|\text{Image})$ is the conditional entropy of the text given the image. A higher mutual information value indicates a stronger association between the text and the image.

3.4 Data Selection

To evaluate the effectiveness of our data selection approach, random sample 175K instruction data points from the open-source LLaVA-OneVision dataset (Li et al., 2024a). We then applied the SC score, OA score, and DP score to filter out 15%, 20%, and 13% of the data, respectively, resulting in 100K high-quality image samples. Subsequently, we further refined the dataset by applying the TP score and IM score to select 10% of the data, yielding a final set of 80K high-quality instruction fine-tuning samples.

4 Experiment

4.1 Evaluation

To evaluate the performance of our method, we introduce three well-established image understanding benchmarks, including VQAv2 (Antol et al., 2015), OKVQA (Marino et al., 2019), TextVQA (Singh et al., 2019), along with four cutting-edge datasets designed for complex visual comprehension and reasoning, including MMBench (Liu et al., 2024d), MME-RealWorld (Zhang et al., 2024a), SEED-Bench (Li et al., 2023a), MMMU (Yue et al., 2024). We provide more detail in the Appendix D.

4.2 Baselines

We compare our approach against SOTA MLLMs, including Qwen2-VL (Wang et al., 2024), Llava-OneVision (Li et al., 2024a), Llama-3.2-Vision (Dubey et al., 2024), as well as InternVL2 and InternVL2.5 (Chen et al., 2023). all of which have been trained on massive-scale datasets with extensive computational resources. To comprehensively evaluate the effectiveness of our method across different model scales, we conduct experiments using models of three representative parameter sizes: small-scale (2B parameters), medium-scale (7B–11B), and large-scale (70B–90B). Due to the computational constraints, we employ the INT4 quantized versions of all large-scale models.

4.3 Implementation Detail

We conduct full fine-tuning on Qwen2-VL-2B and Qwen2-VL-7B as our base models. Due to computational constraints, we adopt Qwen2-VL-72B-instruction as the base model for large-scale experiments and apply LoRA for efficient continued fine-tuning. All models are optimized using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 for full fine-tuning and 1e-4 for LoRA-based fine-tuning, along with a weight decay of 0.01. We employ a linear warmup strategy with a 0.1 warmup ratio and set the LoRA rank to 16. The maximum sequence length is 2048 tokens, and all models are trained and evaluated using the BFloat16 floating-point format. For training efficiency, we use a batch size of 32 with four-step gradient accumulation on a single GPU. Additionally, we follow the original chat template of Qwen2-VL for model interactions. We used beam search with the beam size set to 4 for generation. We run all the experiments on the NVIDIA H800 GPU.

Models	Data Scale	OKVQA	VQAv2	TextVQA	MMBench	MME-RW	SEED	MMMU
InternVL2-2B	>5M	41.2	75.9	72.0	70.9	37.8	71.8	33.0
InternVL2.5-2B	>8M	39.2	**77.8**	71.9	**79.0**	34.6	**72.8**	40.3
Qwen2-VL-2B	>5M	44.3	71.5	77.6	74.6	31.2	68.2	33.6
Qwen2-VL-2B-AC (ours)	80K	**48.4**	73.3	**81.4**	73.0	**39.1**	69.4	**42.2**
Llama3.2-vision-11B	>3M	23.6	72.2	56.5	65.8	38.5	72.7	48.1
Llava-Onevision-7B	>1.6M	**61.6**	82.5	80.3	80.9	48.6	75.4	47.9
InternVL2-8B	>5M	46.4	79.2	77.9	79.4	37.4	75.4	51.2
InternVL2.5-8B	>8M	54.4	80.7	80.9	**82.5**	46.4	**76.8**	**56.2**
Qwen2-VL-7B	>5M	58.0	83.0	**85.4**	81.0	48.3	75.0	50.6
Qwen2-VL-7B-AC (ours)	80K	54.2	**83.6**	84.3	80.7	**49.4**	75.2	49.9
Llama3.2-vision-90B	>3M	46.7	73.1	71.0	84.3	45.4	75.8	38.2
Llava-Onevision-72B	>1.6M	**66.3**	84.8	82.6	86.4	48.8	77.8	50.6
InternVL2-76B	>5M	60.9	84.6	87.1	87.3	**49.1**	76.5	52.2
InternVL2.5-78B	>8M	57.4	85.2	86.3	**91.9**	46.4	77.0	**62.5**
Qwen2-VL-72B	>5M	52.9	83.2	90.8	90.6	40.8	78.3	59.1
Qwen2-VL-72B + AC-lora (ours)	80K	58.0	**90.5**	**91.3**	89.8	47.4	**79.8**	60.6

Table 1: Main result of visual understanding task. The "Data Scale" refers to the instruction data scale reported in each respective paper. Models with more than 72B parameters use InT4 quantification. The best results in each task are highlighted in **bold**.

Models	Data Scale	OKVQA	VQAv2	TextVQA	MMBench	MME-RW	SEED	MMMU
Qwen2-VL-2B-AC	80K	48.4	73.3	81.4	73.0	39.1	69.4	42.2
- Instruction Quality	100K	43.5	72.2	75.8	70.8	36.0	68.1	38.3
- Diversity Perspective	115K	42.4	72.6	74.9	69.7	35.4	67.7	37.5
- Visual Element	175K	41.0	70.3	72.1	68.5	33.9	67.1	35.8
Qwen2-VL-7B-AC	80K	54.2	83.6	84.3	80.7	49.4	75.2	49.9
- Instruction Quality	100K	50.4	81.9	82.4	77.3	47.9	73.2	48.1
- Diversity Perspective	115K	49.3	81.7	81,9	76.5	47.0	72.8	48.2
- Visual Element	175K	47.2	79.3	79.1	75.7	45.3	71.8	44.1

Table 2: Ablation study of small-scale and middle-scale on visual understanding task. The "Data Scale" refers to the training data size before filtering.

4.4 Main Result

In this section, we explore the two following research questions: (1) Q1: How does our data compare with current mainstream training datasets in terms of efficiency? and (2) Q2: Can a small amount of high-quality data continuously improve a well-trained, large-scale model?

Q1: A small amount of high-quality data significantly efficiency improves the instruction-following ability of pretrained MLLMs. As shown in Table 1, with less than 5% of the baseline training data, our method achieves comparable or superior performance across all datasets. Specifically, our Qwen2-VL-2B-AC model achieves an average score of 61.0% on 7 benchmarks, outperforming Qwen2-VL-2B (57.3%), InternVL2.5-2B (59.4%), and InternVL2-2B (57.5%). For middle-scale models, our Qwen2-VL-7B-AC model reaches an average score of 68.2%, surpassing Llama3.2-vision-11B (53.9%), InternVL2-8B (63.8%), and matching the performance of Llava-Onevision-7B (68.1%), Intern2.5-8B (68.3%), and Qwen2-VL-7B (68.7%). These results suggest that the key factor for training pretrained MLLMs is not solely the amount of data, but rather the quality of the data. A small amount of high-quality data can significantly boost the multimodal instruction-following capability of pretrained MLLMs, enabling the model to activate the visual knowledge learned during pretraining and perform a variety of visual tasks. Moreover, we observe that on the MME-RealWorld dataset, a complex real-world visual benchmarks, our model consistently outperforms the baseline, highlighting the crucial role of complex visual data in real-world applications. However, we also see that compared to the small-scale model, the advantages of high-quality data are less pronounced in the middle-scale model. This may be because larger models require more high-quality data to fully fine-tune the deeper parameters, a finding also shown in the work of training scaling law (Kaplan et al., 2020).

Q2: A small amount of high-quality data can continuously boost large-scale MLLMs. As shown in Table 1, continued training on the Qwen2-VL-72B model leads to noticeable improvements in its image understanding capabilities. The Qwen2-VL-72B-AC model achieves 73.9% on 7 datasets, surpassing other baseline models. Specifically, compared to the original Qwen2-VL-72B, our fine-

tuned model shows significant gains on the two datasets that focus on complex image perception and understanding: OKVQA and MME-RealWorld (4.1% and 6.6% improvement, separately). These results suggest that even large models trained on vast amounts of data can still be hungry for high-quality data, and a small amount of high-quality data can rapidly boost their performance. Moreover, this also highlights the importance of high-quality data for effective continued instruction tuning in very large models.

5 Ablation Study

In this section, we conduct ablation experiments to assess the effectiveness of the proposed filtering methods, as presented in Table 2. Our findings reveal that even when the total amount of data is increased by removing the specific filter, the performance quickly degrades. This suggests that fine-tuning with a small amount of high-quality data is significantly more efficient than using large amounts of noisy data. Our method effectively filters out noisy data, thereby improving the fine-tuning efficiency of the model. Notably, the performance shows the most significant decline when the instruction quality filter is removed, with average performance drops of 3.1% and 2.5% for the small and middle-scale models, respectively. This confirms our observation that weakly aligned instructions in the current visual instruction data have a detrimental effect. Poorly aligned instruction data negatively impacts the perception and knowledge alignment between the visual encoder and the large language model, ultimately reducing the overall performance. Moreover, the removal of the visual element filter also leads to a noticeable decrease in performance (1.6% and 2.1% for the small and middle-scale models, respectively). This emphasizes the critical role of image quality in data selection; low-quality or noisy images hinder the efficiency of visual representation learning by the vision encoder.

6 Complex Visual Understanding

To fully explore the role of high-quality data in complex image perception and understanding, we selected six sub-tasks of image comprehension from the practical related benchmark, including sentiment classification, quality assessment, scene perception, and style classification from MMBench, as well as scene understanding from SEED Bench and visual perception from MME-RealWorld. We report the results of our method on a 7B model setup, as shown in Figure 4.

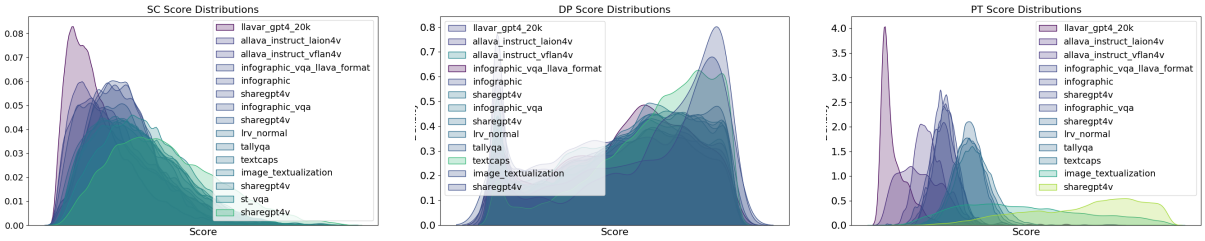
Our findings reveal that, unlike the similar overall performance with baselines in the main results, our Qwen2-VL-7B-AC model consistently outperforms the baseline across all complex image understanding tasks. This demonstrates that informative image data is crucial for improving a model's ability to perceive and understand complex images. While recent works (Li et al., 2024d; Guo et al., 2024) in visual model architectures often focus on increasing image resolution to enhance complex image understanding, higher resolution images do not necessarily equate to greater complexity. For example, compared to a high-resolution document scan with more content, a lower-resolution yet content-rich street scene image proves to be far more beneficial for a model's visual representation learning. Images containing dense visual elements effectively activate the model's ability to recognize and learn the relationships between these elements, thus bridging the gap between the visual model's low-level visual perception and the world knowledge stored in the pretrained language model.

7 Quantification Distribution Analysis

In this section, we present a visualization and detailed analysis of the proposed image informativeness metrics, focusing on three metrics, with the additional metrics provided in Appendix B.

Segmentation complexity score distribution As shown in Figure 3, we observed that the distribution of the segmentation complexity score varies across different datasets. For example, the score of the "llava_gpt4_20k" dataset is relatively low, as it typically contains text-image pairs with limited visual content, resulting in fewer visual elements within the dataset. In contrast, the "sharegpt4v" dataset primarily includes scene images from real-life scenarios, which are generally richer in visual elements, leading to higher SC scores. This difference indicates that the "sharegpt4v" dataset contains more complex images. This distribution demonstrates that our proposed segmentation complexity score effectively reflects the number of visual elements in the data, indirectly capturing the complexity of the images.

Diversity perspectives score distribution Unlike the distribution of other scores, the diversity



(a) (b) (c)

Figure 3: Score distributions across different datasets: (a) segmentation complexity, (b) diversity perspectives and (c) prior token perplexity.

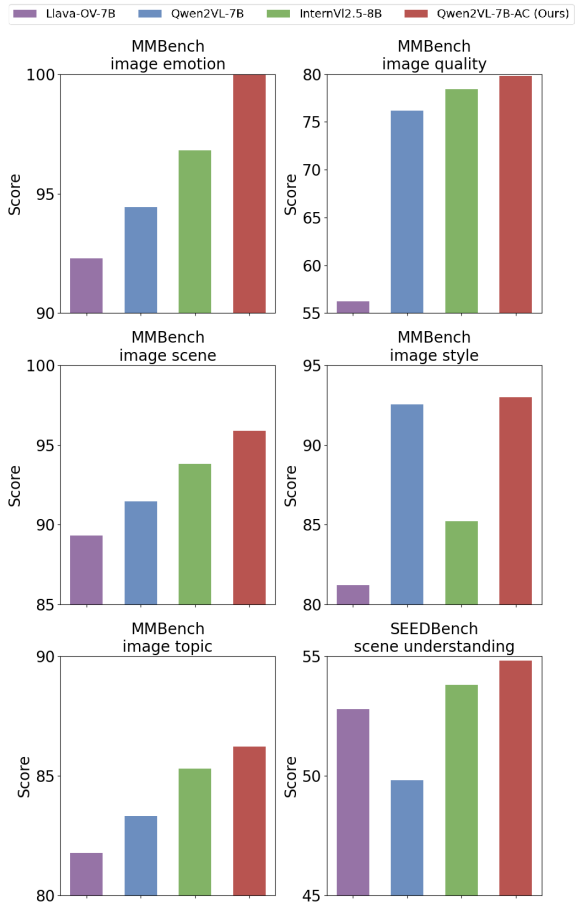


Figure 4: Results of the medium-scale model on six curated complex image understanding and perception datasets.

perspectives score exhibits a notably extreme trend, with the visual agent tending to assign either the highest or the lowest ratings. Our analysis suggests that this is primarily due to the multifaceted nature of the image evaluation process, where images are assessed with multiple dimensions. The diversity in the types and sources of images further complicates the task of providing a comprehensive score. Additionally, we intentionally designed the system to prevent the agent from assigning low scores based on a few weak dimensions of an image. For example, we cannot devalue the quality of a close-up photograph of still life simply due to the lacking of visual events. As a result, the agent tends to classify images as either extremely poor or exceedingly rich in terms of visual content. This underscores the substantial challenge in evaluating images across multiple dimensions of quality. The diverse characteristics of images make it difficult to achieve a balanced and nuanced assessment, which remains an open issue in multimodal evaluation tasks.

Prior token perplexity score distribution For analytical convenience, we apply the negative transformation to the prior token perplexity scores when visualization. Upon examining the distribution of the prior token perplexity scores, we observe significant variability across datasets. This can likely be attributed to the greater disparity in the response text across different tasks. While higher scores (shifted towards the right) indicate that the token probability distribution is more certain. However, excessively high scores may also reflect overly simplistic responses. Therefore, when setting the threshold for filtering prior token perplexity scores, we exclude a small number of outlines values to enhance training efficiency.

8 Conclusion

In this paper, we present a novel approach called visual-centric selection approach via agent collaboration, which improve the training efficiency and effectiveness of MLLMs through reduce the noise data. Extensive experiments indicate that , with less than 5% of the data, models trained with our method achieve comparable or even superior results to state-of-the-art approaches. Our paper demonstration that even a small number of high quality data could surpass a large scale of data. Future work will focus on developing more diverse and comprehensive image quality evaluation metrics to capture a broader range of visual complexities and enhance the robustness of model training.

Limitations

Despite our discoveries and improvements, we must acknowledge certain limitations in our work:

Data Size: Due to resource constraints, our experiments were conducted on a dataset of 200K samples, from which 80K instruction data were finally selected. This limitation restricts our exploration of the application of ViSA on larger datasets. In future work, we aim to extend the use of ViSA to a broader range of open-source datasets.

Theoretical Foundation: While our method for quantifying image information richness has yielded some promising results, it currently lacks a solid theoretical foundation, limiting the generalizability of our approach. Given the diverse sources and purposes of image data, a comprehensive evaluation of image quality remains a challenging task. We plan to investigate additional theories and methods for image information metrics in future.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 11198–11201. ACM.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, volume 15141 of *Lecture Notes in Computer Science*, pages 390–406. Springer.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen,

and Zhuowen Tu. 2024. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 2256–2264. AAAI Press.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Tomer Levinboim, Ashish V. Thapliyal, Piyush Sharma, and Radu Soricut. 2021. Quality estimation for image captions based on large-scale human evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3157–3166. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. 2024c. Lmeye: An interactive perception network for large language models. *IEEE Trans. Multim.*, 26:10952–10964.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024d. Monkey: Image resolution and text label are important things for large multi-modal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26753–26763. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jihao Liu, Xin Huang, Jinliang Zheng, Boxiao Liu, Jia Wang, Osamu Yoshie, Yu Liu, and Hongsheng Li. 2024b. Mm-instruct: Generated visual instructions for large multimodal model alignment. *CoRR*, abs/2406.19736.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024c. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pages 38–55. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Zhenyu Liu, Dongfang Li, Xinshuo Hu, Xinping Zhao, Yibin Chen, Baotian Hu, and Min Zhang. 2024e. Take off the training wheels! progressive in-context learning for effective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2743–2757. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International*

Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment anything in images and videos. *CoRR*, abs/2408.00714.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: assembling open-world models for diverse visual tasks. *CoRR*, abs/2401.14159.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15271–15342. Association for Computational Linguistics.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

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. *CoRR*, abs/2306.17107.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024a. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? *CoRR*, abs/2408.13257.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. 2024b. Recognize anything: A strong image tagging model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 1724–1732. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Diversity Perspectives Prompt

We present the default version prompt for diversity perspectives quantification used in our experiment in the Table 3.

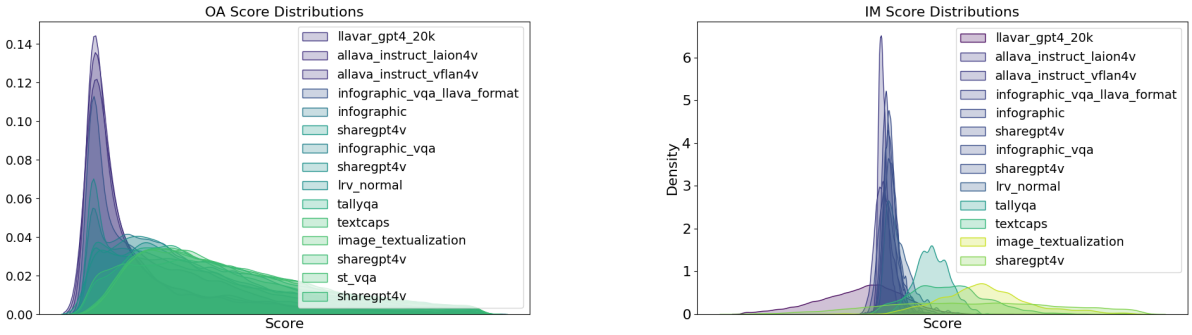
B Quantification Distribution Analysis

Object alignment score distribution As shown in Figure 5, our object alignment score exhibit similar trends with previous segmentation complexity score. However, while the SC score tends to reflect small visual elements such as shapes and color blocks, the OA score is more aligned with visual entities. As a result, the distribution of the OA score is more concentrated. Higher OA scores indicate the presence of more rare and distinctive visual entities, highlighting the uniqueness of the image. In contrast, images with lower OA scores contain more common and less distinctive entities, resulting in lower informational richness. The presence of too many common entities may introduce bias into the visual model's representation learning, thereby affecting its efficiency.

Image-text mutual information score distribution The distribution of the image-text mutual information score reveals a concerning trend: certain open-source datasets primarily focus on evaluating instruction-following ability, often overlooking the critical relationship between the instruction and the corresponding image. For example, we observed that the llava_gpt4_20k dataset frequently contains textual explanations or discussions that are largely unrelated to the image content. Such data not only fails to effectively guide the visual model in learning meaningful visual representations but also introduces significant noise into the visual instruction learning process. This highlights the need for careful curation of datasets to mitigate such mismatches and enhance the quality of visual instruction learning. These findings emphasize the importance of aligning image-instruction pairs in multimodal datasets, ensuring that both components are relevant and complementary to the task at hand. Such principles should guide future efforts in constructing more effective and coherent multimodal datasets.

C Case Study

To provide a more intuitive demonstration of the proposed image information richness metrics, we present a case study, shown in Figure 6. In this



(a) (b)

Figure 5: Score distributions across different datasets: (a) object alignment, and (b) image-text mutual information.

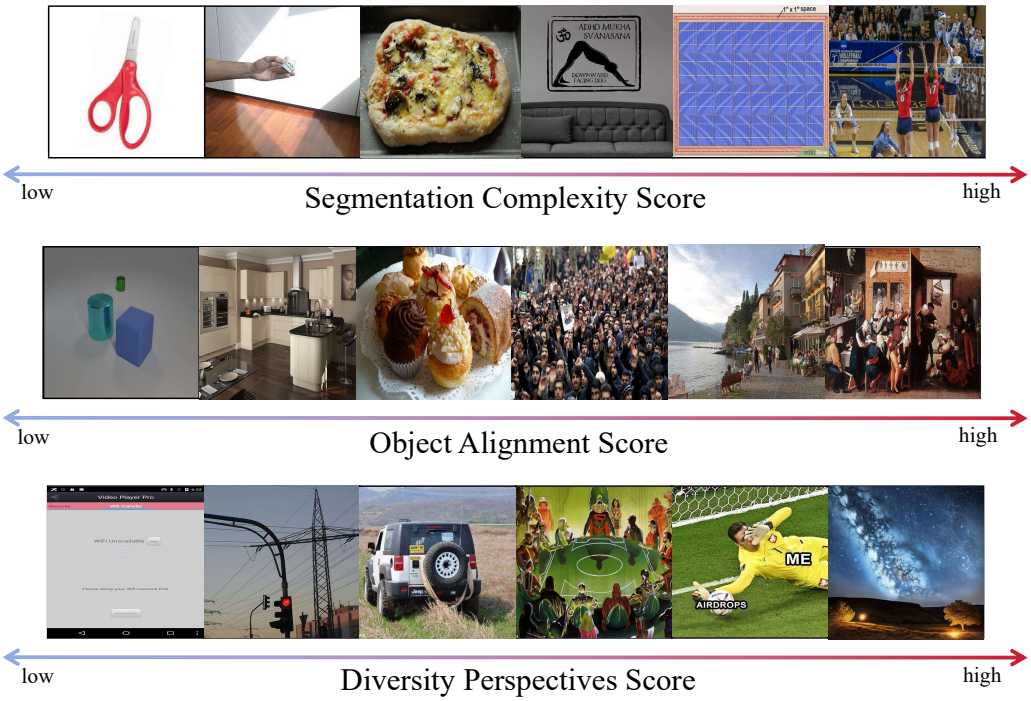


Figure 6: Case study of our proposed image informative metrics. Images on the right exhibit higher scores and richer content, while images on the left correspond to lower scores and simpler content.

figure, images on the right correspond to higher scores, while images on the left are associated with lower scores. Upon observation, we find that all three proposed metrics effectively filter for images with richer content. The segmentation complexity score favours simple visual elements, such as shapes and color blocks. For instance, in the SC score examples, the image on the far left—featuring only a single object, a pair of scissors—receives a low SC score, whereas the volleyball match image on the far right, with richer content, achieves a higher score. However, we also present an example of misjudgment: the second image on the right, a fabric with repeated textures, has a relatively high SC score despite its overall simplicity. This highlights the limitation of using basic image segmentation as a sole measure of image complexity.

In contrast, the object alignment score does not suffer from this issue. For example, in the third image on the right, depicting a crowd, the presence of numerous people does not result in an inflated OA score, as the frequent repetition of entities in the image reduces their individual weight. Thus, the image is correctly assigned a higher OA score without any misjudgment.

Regarding the diversity perspectives score, we focus on the multidimensional evaluation of images, which allows us to prioritize "interesting" images. For instance, the third image on the right,

a comic, and the second image, a meme, receive high scores for emphasizing specific cultural elements. The photo on the far right earns the highest score due to its aesthetic value, while the screenshot on the far left, primarily composed of text, is assigned a low score due to its lack of visual information.

D Evaluation Detail

In this section, we provide additional information and details regarding the evaluation benchmarks. VQAv2 is a widely adopted open-domain visual question answering dataset, consisting of over one million question-answer pairs. For our evaluation, we sample a subset of 10,000 questions. OKVQA is designed to assess models' ability to answer visual questions by leveraging external knowledge sources. Likewise, TextVQA includes over 45,000 text-centric questions based on more than 28,000 images extracted from selected categories of the OpenImages dataset. To evaluate advanced multimodal reasoning capabilities, MMBench (V1.1) provides a diverse set of 4876 multiple-choice questions compiled from various sources, spanning 20 distinct capability dimensions. MME-RealWorld (MME-RW) serves as a benchmark tailored to real-world applications, focusing on practical scenarios; we utilize its lite version, which consists of 1919 of the most challenging questions. SEED-Bench (SEED) comprises 14232 carefully annotated multiple-choice questions, covering 12 key evaluation dimensions. Lastly, MMMU consists of 1050 challenging questions across a wide range of interdisciplinary tasks, requiring college-level subject knowledge and advanced reasoning skills. For each dataset, we follow the answer-matching strategy recommended in its original paper and compute accuracy with a widely used open-source evaluation toolkit (Duan et al., 2024).

The default version of PICA prompt with an example

 # Image Complexity Assessment Task

In this task, you will evaluate the informational richness of an image based on multiple dimensions. Each image may contain multiple layers of information, including various visual elements, contextual settings, emotional expressions, and cultural backgrounds. Please provide a detailed evaluation based on the following rating criteria.

Rating Criteria:

The overall score will be based on all valid indicators. Images with particularly strong performance in certain dimensions (e.g., complex context or strong emotional expression) will receive higher scores, and weaker performance in other dimensions will not significantly lower the final score. It is recommended to focus on the image's overall informational layers, element rarity, context building, and uniqueness rather than the absence of a single dimension.

1 Point: Very low information, typically a single subject with a simple background or no clear context. The image lacks depth and multi-dimensional understanding, with few visual elements and no notable uniqueness.
2 Point: Low information, with a few simple elements. It may have basic context or structure but lacks detail and uniqueness. The image is plain and suitable for basic understanding tasks, with weak emotional or contextual expression.
3 Point: Moderate information, with a reasonable number of elements or layers. Context, emotion, or uniqueness are more noticeable, and the image provides moderate multi-dimensional understanding. It has some depth in visual elements and context but doesn't meet all high-level requirements.
4 Point: High information, with rich elements and clear context, showing uniqueness and diverse layers. The image is detailed, with strong emotional or contextual expression, but may lack some rare or background features. Suitable for in-depth analysis, particularly in context, emotion, culture, or history.
5 Point: Very rich information, with diverse elements, complex scenes, unique perspectives, and abundant detail. The image features rare qualities and deep context, emotion, and cultural background. Ideal for advanced analysis with multi-dimensional depth.

Evaluation Dimensions

1. **Details and Materiality**
Detail Density: Evaluate the number and complexity of details in the image. The denser the details, the greater the informational content.
Material and Texture: Evaluate whether the image showcases multiple materials (e.g., wood, metal, fabric) and whether the textures are clearly visible. A variety of material layers adds to the visual richness.
Detail Layers and Spatial Perception: Evaluate whether the image presents multiple layers of details, such as clear details in the foreground and a blurred background, creating a sense of depth.

2. **Context and Narrative**
Context Constructio: Evaluate whether the image constructs a clear context or scene that conveys rich information through visual elements (e.g., a family gathering, office work, festive events).
Narrative Complexity: Evaluate whether the image depicts one or more complex events or actions, increasing the narrative depth and informational content.
Implied Background Story: Evaluate whether the image implies a potential story or context through visual elements, such as through the environment, actions of people, or objects that hint at social, historical, or cultural contexts.

3. **Emotion and Atmosphere**
Emotional Expression: Evaluate whether the image conveys a specific emotion or atmosphere through elements like lighting, color tones, and composition. Emotional expression increases the image's complexity.
Emotional Layering: Evaluate whether the image expresses multiple emotional layers or emotional shifts through different elements. Images with rich emotional layers often have greater depth.

4. **Cultural and Historical Context**
Cultural Characteristics: Evaluate whether the image contains elements that reflect a specific cultural, historical, or social context (e.g., distinctive architectural styles, clothing, festivals).
Historical Background: Evaluate whether the image reflects historical events, periods, or characteristics. Historical elements enhance the image's depth.

5. **Camera Angle and Composition**
Unique Perspective: Evaluate whether the image features a unique or creative angle, using uncommon perspectives.
Composition Complexity: Evaluate whether the image's composition is complex and varied, using techniques like contrast, symmetry, and spatial distribution to enhance the image's depth. Complex compositions effectively convey more information.

6. **Dynamics and Interaction**
Dynamic Elements: Evaluate whether the image includes dynamic elements (e.g., movement of people or objects, hints of time progression).
Interactivity: Evaluate whether the image depict interaction between elements (e.g., people talking, animals hunting)? Interaction enhances the appeal and complexity of the image.

Table 3: The default version of PICA prompt with an example