

Mastering Collaborative Multi-modal Data Selection: A Focus on Informativeness, Uniqueness, and Representativeness

Qifan Yu^{1*} Zhebei Shen^{1*} Zhongqi Yue^{2*} Yang Wu³ Bosheng Qin¹ Wenqiao Zhang¹
Yunfei Li³ Juncheng Li¹✉ Siliang Tang¹ Yueting Zhuang¹✉

¹Zhejiang University, ²Nanyang Technological University, ³Ant Group

{yuqifan, shenzhebei, junchengli, siliang, yzhuang}@zju.edu.cn

nickyuezhongqi@gmail.com, wy306396@antgroup.com

Abstract

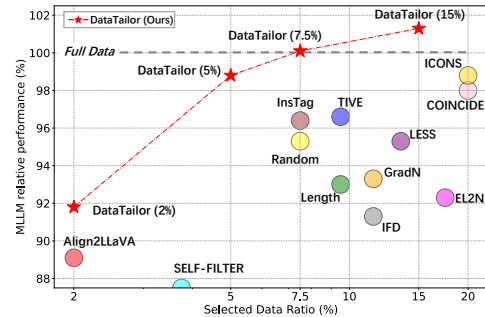
Instruction tuning fine-tunes pre-trained Multi-modal Large Language Models (MLLMs) to handle real-world tasks. However, the rapid expansion of visual instruction datasets introduces data redundancy, leading to excessive computational costs. We propose a collaborative framework, **DataTailor**, which leverages three key principles—*informativeness, uniqueness, and representativeness*—for effective data selection. We argue that a valuable sample should be informative of the task, non-redundant, and represent the sample distribution (i.e., not an outlier). We further propose practical ways to score against each principle, which automatically adapts to a given dataset without tedious hyperparameter tuning. Comprehensive experiments on various benchmarks demonstrate that **DataTailor** achieves 101.3% of the performance of full-data fine-tuning with only 15% of the data, significantly reducing computational costs while maintaining superior results. This exemplifies the “Less is More” philosophy in MLLM development. The code and data is available in this [URL](#).

1. Introduction

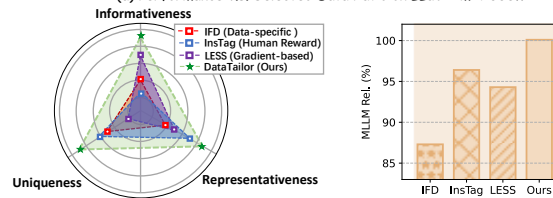
The rapid development of Multi-modal Large Language Models (MLLMs) has made promising progress on various multi-modal tasks [2, 55, 61]. A typical MLLM is developed through two main training stages: pre-training on vast image-text pairs and fine-tuning on task-specific multi-modal instructions. Notably, the fine-tuning stage is critical for enhancing the instruction-following capabilities of MLLMs. Yet, this stage can become exceedingly time-consuming due to the large-scale but low-quality instruction data. Hence, the community is interested in fine-tuning data

*Equal contribution.

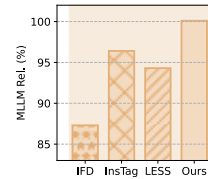
✉Corresponding author.



(a) Performance v.s. Selected Data Ratio on LLaVA-mix-665k



(b) Metric triangle among informativeness, uniqueness, and representativeness



(c) MLLM relative performance of DataTailor compared with other data selection methods

Figure 1. (a) The Performance v.s. Selected Data Ratio on LLaVA-mix-665k of DataTailor compared with SOTA data selection methods. (b) Metric triangle among informativeness, uniqueness, and representativeness when applying IFD [28] (data-specific methods), InsTag [36] (human-reward methods), LESS [51] (gradient-based methods), and DataTailor. (c) The corresponding MLLM performance on LLaVA-mix-665k [33] of different methods.

selection methods, such that an MLLM trained on the selected subset yields comparable or even better performance.

Existing MLLM data selection methods [9, 19, 35, 49] largely follow similar ideas from the NLP community [28, 36, 45, 51, 60]. They can be divided into three main categories: (1) *Data-specific methods* [6, 10, 28] leverage vast hand-designed rules to select data in specific datasets, which are not flexible and robust for diverse tasks. (2) *Human-reward methods* [36, 60] utilize human feedback to select data, which are both time-consuming and expensive. (3) *Gradient-based methods* [31, 45, 51] select sam-

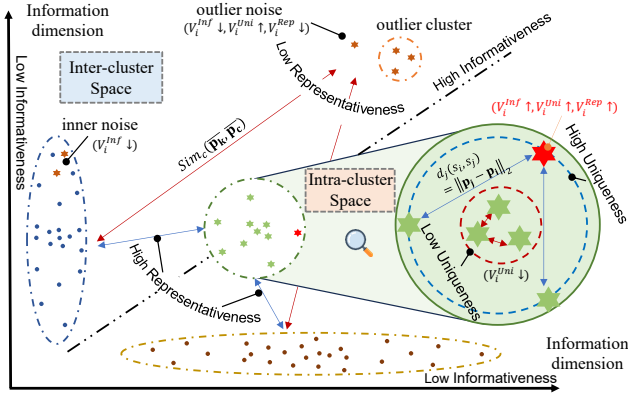


Figure 2. Illustration of the proposed method based on informativeness, uniqueness, and representativeness, where the x and y axes show information dimensions in latent space. The red star denotes a high-quality sample that satisfies the three principles.

ples whose gradients are similar to the average training gradient of the dataset. However, they require additional training on downstream tasks, making total computation costly.

To address these deficiencies, we build a systematic data selection method for MLLM called **DataTailor**. We evaluate each sample with three principles and select the most valuable samples, leading to state-of-the-art MLLM performance with a fraction of data (c.f. Fig. 1). The three principles are: (1) **Informativeness**: a valuable sample should be informative of the hard task, *e.g.*, If the task is reasoning, describing the movement differences between skiing and ice skating is more informative and complex than simply describing someone skiing. In Fig. 2 where each axis (heuristically) represents an orthogonal dimension of task information, points along the diagonal carry more information about the task. (2) **Uniqueness**: A valuable sample should be distinct from others, offering unique insights rather than prevalent commonsense knowledge (c.f. Fig. 2 near the blue dashed region in the intra-cluster space demonstrates high uniqueness). (3) **Representativeness**: it should be a typical sample in the data distribution. This prevents selecting noisy outliers or mislabeled samples (c.f. Fig. 2 the clusters connected by blue lines in the inter-cluster space exhibit high representativeness for the overall dataset).

We further propose a practical method to measure the value of each sample against each principle. For informativeness, we take motivations from information theory [7, 8]. For each sample, we analyze the singular value distribution of its features and use the entropy of the singular values to determine if it is informative of the task. To compute uniqueness and representativeness, we first cluster the samples based on their visual and textual features. This allows efficient calculation for uniqueness, as we can simply measure the average distances of a sample to its neighbors in the same cluster, and mark those with a large dis-

tance as unique. Then we find connected clusters and mark samples in those clusters as representative. Hence, noisy or mislabeled data from a far-away cluster can be filtered.

Moreover, as multi-modal samples exhibit varying structures and complexities across diverse tasks, we propose an adaptive weight to combine the values, which removes the need for expensive hyper-parameter tuning. We also adaptively determine the proportion of selected data for each task by using the average largest singular value of samples in the task, which empirically reflects task difficulty and correlates with training robustness. Combining these techniques, DataTailor synergizes the three principles for data selection and achieves an optimal balance between data volume and model performance (as shown Fig. 1(a) red line).

To our knowledge, we are the first to explore sample relationships between multi-modal instructions systematically. Through extensive experiments, we demonstrate that DataTailor exhibits significant effectiveness in data selection for MLLMs (with less than 5% data but achieving over 95% performance). This effectiveness is further demonstrated through quantitative metrics designed for the three principles (c.f. Fig. 1(b)), proving that DataTailor improves data selection by addressing the essential aspects of each principle (detailed analyses is in Sec. 4.4 and Appendix D.1 & D.2). In contrast, other methods lack a systematic evaluation, particularly of sample relationships, leading to weaknesses in uniqueness and representativeness and resulting in suboptimal MLLM performance (c.f. Fig. 1(c)). Remarkably, when DataTailor increases the data selection ratio, multi-modal data selection can even outperform full data fine-tuning (achieve 101.3% performance with 15% data), truly exemplifying the concept of “Less is More”. Overall, our main contributions are summarized as follows:

- We identify three key principles (*i.e.*, informativeness, uniqueness, and representativeness) from a systematic perspective to master multi-modal data selection.
- We propose a unified framework, **DataTailor**, to adaptively integrate these principles for value evaluation to optimize multi-modal data selection in a collaborative way.
- Extensive results show DataTailor’s effectiveness in optimizing all three principles during selection and achieving new SOTA performance on various benchmarks.

2. Related Work

2.1. Multi-modal Large Language Model

With the outstanding performance of LLMs in zero-shot settings, early work combining LLMs with visual modalities has demonstrated impressive visual language comprehension abilities [9, 15, 23, 25–27, 55]. Recently, more powerful MLLMs have emerged [5, 17, 33, 38, 39, 53, 54, 57, 61], which possess perceptual abilities for visual-language tasks and excellent reasoning abilities. Generally, the training

process of MLLMs primarily includes two stages: the pre-training stage and the instruction tuning stage. Recent studies [13, 14, 32, 42, 48, 56] have focused on the second stage to enhance instruction-following abilities. However, this stage is gradually facing inevitable computational overhead due to the growing volume of multi-modal data [35, 41]. It is critical to identify a small subset of high-quality instructions to improve MLLM fine-tuning efficiency.

2.2. Instruction-based Data Selection

Although MLLMs have demonstrated remarkable performance, data redundancy is becoming apparent with the rapid growth of visual instruction datasets, similar to challenges in LLMs [6, 11, 60]. Previous LLM-based works mainly focus on using pre-defined specific rules [6, 28], human feedback [36], or gradient-based approximation [3, 51] to select high-quality data. INSTRUCTMINING [6] uses 9 indicators to fit specific rules for data selection. In contrast, DataTailor automatically assesses values without hand-designed rules, offering greater robustness and flexibility. Moreover, these methods only focus on individual sample values while neglecting similar or noisy data when applied to more complex multi-modal instructions. For MLLM-based data selection, TIVE [35] identifies redundancy in multi-modal instructions and selects valuable data at the task and instance level through gradient similarity, while ICONS [50] extends this by integrating specialist influence estimation. However, they both require extra training on downstream tasks for data selection. SELF-FILTER [49] attaches evaluation models and updates its parameters during training to select high-value samples. InstructionGPT-4 [47] selects 200 instructions for MiniGPT4 [61], but it is unscalable for other settings. COINCIDE [22] roughly clusters data by conceptual representations, while overlooking the value differences between clusters. These methods largely follow prior approaches and overlook the complex relationships between samples, which limits their generalization. To mitigate these limitations, we are the first to adopt a collaborative perspective for multi-modal data selection, focusing on three core principles: informativeness, uniqueness, and representativeness.

3. Method

As illustrated in Figure 3, our DataTailor framework consists of four primary steps: (1) The *informative value* captures the information density in latent space, directly reflecting informativeness to enhance MLLM generalization. (2) The *unique value* identifies distinct samples within the intra-cluster space, reflecting the uniqueness of sample relationships to reduce redundancy effectively. (3) The *representative value* captures samples that align closely with the overall dataset distribution in the inter-cluster space, ensuring representativeness and preventing compromise by noisy

outliers. (4) Finally, DataTailor adaptively integrates these three values to enable collaborative multi-modal data selection. Next, we will elaborate on the details of each step.

3.1. Problem Formulation

Given a visual input X_v and a textual query X_q , the objective of the MLLM is to predict the correct answer X_a for downstream tasks. We formulate multi-modal data selection as selecting fewest samples $S^* = \{s_1, \dots, s_k\}$ from the candidate dataset S for MLLM training that achieves the best inference performance, where $s_i = (X_v, X_q, X_a)$ and k is the total data selection proportion for S . Typically, the sample value is measured by the performance perturbations to the MLLM when removing the sample [21], but this is not practical for large-scale datasets. Therefore, we propose three core principles (*i.e.*, informativeness, uniqueness, and representativeness) that practically identify the most valuable samples to improve MLLM’s performance on downstream tasks. More concrete instantiations of each principle are in Appendix D.2. Then, we introduce each principle.

3.2. Informative Value Estimation in Latent Space

Although several approaches have been proposed for multi-modal data selection [35, 47, 49, 50], they are either infeasible at scale due to computational overhead or are limited in generalizability by pre-defined evaluation rules. Previous studies [30, 34, 51] have shown that difficult samples contribute more to improving the performance of downstream tasks in MLLMs. Thus, DataTailor directly estimates samples’ essential difficulty for multi-modal data selection.

However, it is non-trivial to quantify sample difficulty for multi-modal instructions due to varying information representations in different modalities. Drawing upon previous spectral analysis [7, 8, 52], we perform singular value decomposition (SVD) on the unified feature matrix at the token level for image and text modalities to reflect the difficulty of samples, as it consists of the variations across different tokens in latent space. Formally, given a multi-modal sample s_i , we obtain its unified feature matrix $\mathbf{M}_i \in \mathbb{R}^{L_i \times d}$ from the penultimate layer, where L_i is the token length and d is the feature dimension. We adopt this layer as the lower layers struggle to capture complex features, while the final layer suffers biases of pre-training datasets [34]. Besides, the penultimate layer retains earlier layer information, providing richer features. Subsequently, we perform SVD on the matrix in latent space $\mathbf{M}_i = \mathbf{U}_i \hat{\Sigma}_i \mathbf{V}_i^\top$ and its corresponding diagonal singular matrix is defined as follows:

$$\hat{\Sigma}_i = \{\sigma_1, \dots, \sigma_{L_i}\} \quad (1)$$

where we assume $L_i \leq d$ [33] and all singular values $\{\sigma_j\}_{j=1}^{L_i}$ are listed in order. Building on this, we compute the entropy of normalized singular values as the informative

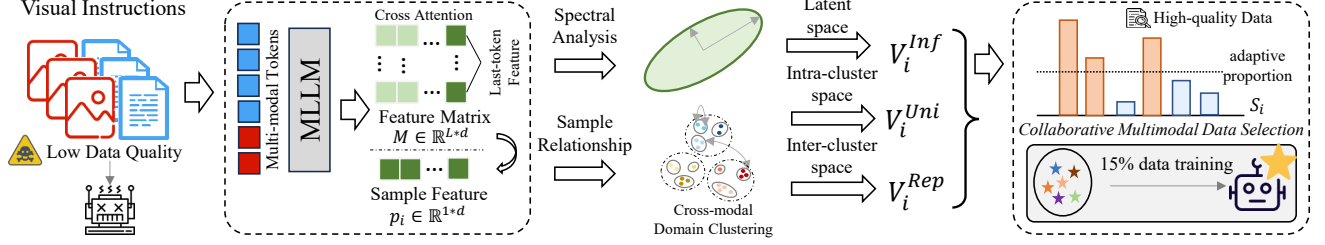


Figure 3. Overview of our proposed **DataTailor** for automatically selecting high-quality multi-modal data through the collaboration of three principled values (*i.e.*, informative value, unique value, and representative value) from a systematic perspective.

value to assess the information density of each data:

$$V_i^{Inf} = - \sum_{j=1}^{L_i} \frac{\sigma_j}{\sum_{k=1}^{L_i} \sigma_k} \log \frac{\sigma_j}{\sum_{k=1}^{L_i} \sigma_k} \quad (2)$$

Intuitively, simple samples contain redundant information in images or text, and only part of information is necessary to answer the query (see Appendix D.2 for examples). As a result, the columns of their feature matrices exhibit strong linear dependence, which drive down the smaller singular values, causing the top ones to be much bigger by comparison, leading lower singular value entropy. In contrast, more challenging samples are richer in information, with feature matrices that are closer to full rank, resulting in more uniform singular values and higher entropy. In this way, SVE can as a practical approximation to measure the difficulty of samples, enabling the selection of hard samples to enhance the performance of MLLM in downstream tasks.

3.3. Unique Value in Intra-cluster Space

To enrich the value of samples for multi-modal data selection, it is crucial to emphasize the unique values of samples to select unique samples from similar clusters. The unique value estimation in the intra-cluster space consists of two steps: *cross-modal domain clustering* and *unique value calculation*. Next, we introduce each step in detail.

Cross-modal Domain Clustering. To efficiently distinguish unique samples, we propose an optimized Cross-modal Domain Clustering that aggregate sufficiently similar samples into cluster in each task. This process begins with quantifying the ℓ_2 -norm distance between samples for cluster variance computation. It then utilizes the Ward criterion [37] to merge clusters with the minimum increase in sum of squared errors (SSE) progressively, computed as:

$$\Delta \text{SSE} = \frac{n_A \cdot n_B}{n_A + n_B} \cdot \|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|_2 \quad (3)$$

where n_A, n_B denote cluster cardinalities and $\boldsymbol{\mu}_A, \boldsymbol{\mu}_B$ represent cluster centroids. The clustering terminates when the ΔSSE exceeds the threshold $\lambda \Delta \text{SSE}_{\max}$, where ΔSSE_{\max} denotes the maximum increase in SSE. Here, λ controls

the number of clusters and is set to 0.1 by default to match the total data selection proportion for capturing sample relationships. Moreover, we use parallel computation and memory optimization for efficiency. More implementation details and analyses are shown in Appendix B.3.

Unique Value Calculation. To quantify the uniqueness of samples, we focus on identifying discriminative samples in the intra-cluster space that contribute uniquely to training. The key intuition is that unique samples exhibit larger distance from similar samples (as shown in Fig. 2) to contain more information in the cluster for generalization. Therefore, we introduce a distance coefficient to assign high unique values to these distinctive instructions based on their distance from surrounding samples as follows:

$$V_i^{Uni} = \sum_{s_j \in \mathbf{C}, j \neq i} \|\mathbf{p}_j - \mathbf{p}_i\|_2 \cdot \frac{V_j^{Inf}}{\sum_{k \in \mathbf{C}} V_k^{Inf}} \quad (4)$$

where $\|\mathbf{p}_j - \mathbf{p}_i\|_2$ is the Euclidean distance of two samples s_i, s_j in the cluster \mathbf{C} in latent space. We assign higher reference weights to samples with greater informative value when computing their unique and representative values, reflecting their stronger influence on other samples. In this manner, these challenging instructions are more likely to be selected due to their enhanced unique values.

3.4. Representative Value in Inter-cluster Space

Empirically, evaluating samples solely with their informativeness and uniqueness within clusters may introduce outlier noisy or mislabeled data. These samples, despite higher uniqueness, have weaker associations with other clusters, limiting their ability to represent the overall dataset. Therefore, we introduce an inter-cluster representative coefficient to measure relationships across clusters, ensuring that selected samples capture representative features from the overall dataset, thereby avoiding the selection of noisy data:

$$\tau_i^c = \frac{1}{K-1} \sum_{k \neq c}^K \exp(\text{sim}(\overline{\mathbf{p}}_k, \overline{\mathbf{p}}_c)) \quad (5)$$

where $\overline{\mathbf{p}}_c$ is the average sample feature in the target cluster \mathbf{C} that contains sample s_i , $\{\overline{\mathbf{p}}_k\}_{k \neq c}^K$ is the average sample

feature of other clusters and K is the number of clusters in each task. We use the feature of the last token to represent the sample feature as it aggregates all visual and textual features by cross-attention. Here, $\text{sim}(\cdot, \cdot)$ is the cosine similarity and $\exp(\cdot)$ is used to amplify the effect of clusters. Based on this coefficient, we then assign the weighted representative value to the instruction s_i as follows:

$$V_i^{Rep} = \tau_i^c \cdot \frac{V_i^{Inf}}{\sum_{k \in \mathbf{C}} V_k^{Inf}} \quad (6)$$

In this way, the representative value uses the association coefficient to ensure that selected samples align with the overall distribution. When the value is high, it indicates that the selected samples can effectively represent other samples, reducing the impact of noisy data and enhancing the overall representativeness in conjunction with uniqueness. More implementation details are shown in Appendix C.1.

3.5. Adaptively Collaborative Data Selection

Although we obtain multi-scale values from three complementary perspectives, combining them to select ideal samples is challenging since these multi-modal samples exhibit varying instruction rounds. Multi-IF [18] points out that multi-turn instructions prioritize informative value due to their weak interrelationships while single-turn ones emphasize unique and representative value due to their limited information. Inspired by this, we introduce an influence factor based on the number of response rounds of each multi-modal instruction for adaptively collaborative data selection to enable adaptive, collaborative data selection, enhancing the synergy among these three values as follows:

$$V_i = \frac{r_i}{r_i + 2} \cdot V_i^{Inf} + \frac{1}{r_i + 2} \cdot (V_i^{Uni} + V_i^{Rep}) \quad (7)$$

where r_i denotes the conversation round of each multi-modal instruction. We employ a 1:1 ratio to simplify the balance between uniqueness and representativeness values, as their trade-off remains stable. With this synergistic value for data selection, we can identify informative and unique instructions while adequately representative (c.f., Fig. 4). We also show detailed demonstration of DataTailor addressing each of three principles in Sec. 4.4 and Appendix D.1.

In addition, we observe that standardizing the data selection proportion across all tasks limits selection diversity due to differences in task difficulty. Since spectral analysis shows that samples with the higher largest singular value exhibit less directional diversity for generalization, they require more selected data to improve the model’s training robustness. Thus, we propose adaptive data selection proportion k_p for each task S_p based on largest singular values, which is computed as follows:

$$k_p = \frac{x_p^2 \cdot |S_p|}{\sum_q x_q^2 \cdot |S_q|} \cdot k, \quad x_p = \text{avg}\left(\frac{\sigma_{\max}}{\sum_{j=1}^{L_i} \sigma_j}\right) \quad (8)$$

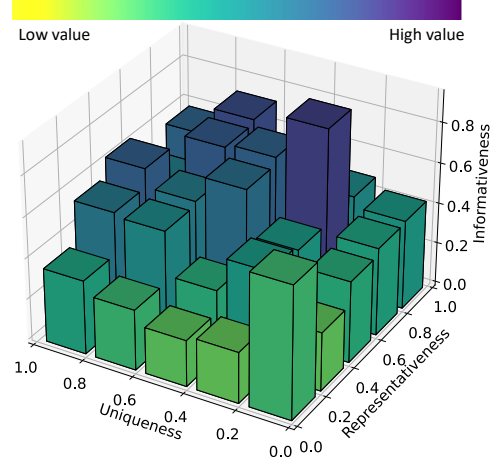


Figure 4. Visualization of the collaboration among informativeness, uniqueness, and representativeness for data selection, where each bar represents a subset defined by a specific value interval.

where x_q is the average ratio of the largest singular value σ_{\max} to the sum of all singular values for all samples in the task S_q , $|S_q|$ is the number of samples in the task, and k is the whole data selection proportion. Through collaborative value assessment with task-adaptive proportions, DataTailor promotes more diversity during MLLM data selection. More details and analyses are shown in Appendix B.2.

4. Experiments

We first evaluate DataTailor on the standard data selection of MLLM on various benchmarks (§ 4.2). Additionally, we examine its transferability to other datasets (§ 4.3) and conduct an in-depth analysis (§ 4.4) for further evaluation.

4.1. Experimental Setup

Multi-modal Instruction Data & Backbone. As ideal data selections should be adaptable to diverse MLLM instruction datasets, we integrate DataTailor with two widely-used datasets to conduct experiments for its effectiveness evaluation: 1) MiniGPT4-Instruction [61] includes about 3.5K instances refined by ChatGPT from detailed descriptions. 2) LLaVA-1.5-mix-665k [33] is a wider collection with 665K instructions, which encompass a wide range of task categories, including dialogue-based Q&A pairs, multiple-choice short Q&A, detailed descriptions, and text-only reasoning tasks. For the general setting, we conduct experiments on MiniGPT-4-7B and LLaVA-v1.5-7B.

Benchmarks. We assess our methods using a mix of MLLM-specific benchmarks and more general downstream tasks. Aligned with SOTA MLLM methods, we include MLLM benchmarks for comprehensiveness: MME [12] is used to evaluate MLLM’s ability of perception and cognition; SEED-Bench [24] involves multi-modal tasks across 12 perspectives with the assistance of GPT-4; POPE [29]

Methods	Valid Data	MLLM Benchmarks						VQA Benchmarks				Captioning		Rel.	
		MME-P \uparrow	MME-C \uparrow	SEED-1 \uparrow	POPE \uparrow	MM-Vet \uparrow	LLaVA-Wild \uparrow	MMMU(val) \uparrow	VizWiz \uparrow	SciQA \uparrow	GQA \uparrow	VQA-v2 \uparrow	TextVQA \uparrow		NoCaps (val) \uparrow
MiniGPT4-Instruction														MiniGPT4-7B	
MiniGPT4-7B	3.4k	717.4	259.6	23.8	68.3	19.0	25.1	23.4	36.0	36.3	32.2	32.1	21.4	111.5	100.0%
Random	0.2k	698.4	227.7	25.1	69.7	17.2	27.9	19.1	18.3	34.0	19.2	33.2	17.2	105.1	89.1%
Length	0.2k	683.4	209.6	26.7	69.8	18.3	26.7	17.4	29.9	35.6	32.5	33.7	17.4	106.5	94.7%
EL2N [40]	0.2k	668.6	207.6	26.5	72.0	17.4	27.3	23.1	41.9	36.1	32.9	36.3	23.7	108.3	102.2%
IFD [28]	0.2k	678.6	213.8	29.1	47.4	18.9	28.4	20.6	42.7	38.1	28.3	36.0	23.4	106.6	99.8%
InsTag [36]	0.2k	715.6	237.9	26.8	70.4	15.8	28.1	21.7	40.0	38.1	30.1	34.5	22.2	105.9	100.8%
LESS [51]	0.2k	698.5	191.4	22.4	71.8	19.8	25.8	22.2	38.4	35.4	26.0	34.4	16.6	109.7	95.4%
InstructionGPT-4 [47]	0.2k	716.9	229.6	17.4	71.6	21.3	25.3	14.9	29.9	35.1	26.8	34.8	22.1	106.8	93.3%
SELF-FILTER [49]	0.5k	438.7	128.6	21.7	71.4	19.4	24.3	20.4	41.3	35.7	30.4	35.0	22.0	105.6	92.8%
TIVE [35]	0.2k	707.0	200.9	23.6	72.3	17.5	25.8	18.2	31.4	33.8	26.4	35.1	17.5	108.9	92.7%
DataTailor (Ours)	0.2k	720.6	263.9	27.3	69.8	21.4	28.4	23.6	40.8	37.7	30.7	34.7	21.0	106.9	104.6%
LLaVA-1.5-mix-665k														LLaVA-7B	
LLaVA-v1.5-7B (LoRA)	665k	1476.9	267.9	67.4	86.4	30.9	67.9	32.8	47.8	70.0	63.0	79.1	58.2	106.5	100.0%
Random	50k	1387.5	287.5	59.7	85.7	29.5	64.5	32.2	42.3	70.0	55.0	73.7	53.1	107.7	95.3%
Length	50k	1357.0	265.7	47.0	82.6	29.7	67.6	33.9	49.2	60.9	55.5	70.7	45.2	88.2	91.0%
EL2N [40]	50k	1077.3	252.5	59.3	80.8	21.1	40.1	33.6	44.4	71.0	41.7	61.0	41.7	86.9	82.3%
GradN [40]	50k	1275.4	303.6	58.3	75.7	24.8	68.2	32.4	37.8	70.9	44.9	64.0	46.0	101.9	89.3%
IFD [28]	50k	1113.4	301.8	55.1	76.7	27.6	63.1	33.0	48.7	48.2	41.9	64.2	43.6	106.8	87.3%
InsTag [36]	50k	1317.1	345.0	57.4	82.1	29.6	68.1	34.0	47.4	69.3	52.5	63.2	53.3	108.3	96.4%
LESS [51]	50k	1344.8	281.8	61.2	79.4	28.3	65.5	33.0	44.4	71.0	53.4	71.8	52.0	106.2	94.3%
SELF-FILTER [49]	25k	955.7	262.5	47.5	76.0	26.6	60.3	30.6	40.8	59.4	3.6	2.1	5.6	82.3	65.8%
TIVE [35]	50k	1334.8	248.6	62.2	85.9	30.2	67.9	33.1	45.1	71.4	56.2	73.8	51.1	96.0	94.6%
COINCIDE [22]	133k	1495.6	-	-	86.1	-	67.3	-	46.8	69.2	59.8	76.5	55.6	-	98.0%
ICONS [50]	133k	1485.7	-	-	87.5	29.7	66.1	-	50.1	70.8	60.7	76.3	55.6	-	98.8%
DataTailor (Ours)	50k	1461.2	362.5	61.7	82.1	30.4	69.3	-	46.3	70.9	57.7	75.0	53.1	107.2	100.1%
DataTailor (Ours)	100k	1476.2	319.3	63.6	85.3	31.8	71.1	33.2	49.5	71.0	60.5	76.7	55.7	108.7	101.3%

Table 1. Comprehensive comparison between DataTailor and other baselines for multi-modal data selection on MLLM and downstream general benchmarks. Our results are shown in the gray block. Due to limited resources, we all use the LoRA model for fair comparisons.

mainly evaluates the MLLM’s hallucination problems; MM-Vet [58] and LLaVA-Wild [34] assess the model’s open-ended conversational capabilities; The MMMU [59] consists of more challenging scientific problems to assess reasoning ability. For general VQA tasks, VizWiz [16] and ScienceQA [43] contain unseen visual queries and multiple-choice questions to evaluate the zero-shot generalization of MLLMs. VQA-v2 [4] and GQA [20] access the model’s visual perception abilities with open-ended questions while TextVQA [44] focuses on text-rich questions. For captioning, we transfer MLLMs to the NoCaps [1] validation set. We also show the amount of valid data selected to demonstrate the effectiveness of data selection. Note that **Rel.** in all tables represents the relative boost on all benchmarks

Baselines. We use the following baselines: 1) **Traditional data selection:** it includes traditional random selection; length-based selection; GradN [40] and EL2N [40] use the L2-norm of the gradient and the error vector for selection, respectively. 2) **LLM data selection:** it directly transfers the selection methods from LLMs to MLLMs, including data-specific methods [6], human-reward methods [36], and gradient-based method [51]. 3) **MLLM-specialized selection:** it involves methods specifically designed for data selection in MLLM, including InstructionGPT-4 [47], SELF-FILTER [49], TIVE [35], COINCIDE [22], ICONS [50].

4.2. Main Results on Multi-modal Data Selection

We report the results of our DataTailor and other diverse data selection methods for the MiniGPT4 and LLaVA shown in Table 1. Based on the observation of experimental results, we have summarized the following conclusions:

Multi-modal instruction data suffers from serious redundancy, resulting in overall poor data quality. We can observe that in most benchmarks, even randomly selecting a

small amount of instruction data does not result in a performance drop. Moreover, in some cases, simply selecting part of the data outperforms utilizing the entire dataset (33.2 v.s. 32.1 of VQA-v2 on MiniGPT-4), suggesting that excessive low-quality data hinder several MLLMs’ capabilities on the contrary. Qualitatively, as shown in Fig. 1, most methods achieve 80% performance with less than 20% of the data. This confirms our analysis of the data redundancy in multi-modal datasets and the necessity of data selection.

For LLM data selection approaches (i.e., IFD [28], InsTag [36], and LESS [51]), the performances across several benchmarks overall remain unsatisfactory. Although these baselines explicitly distinguish tasks and features, they still underperform DataTailor by 12.8%, 3.7%, and 5.8% on overall relative performance due to their rough information modeling and disregard for sample relationships. Moreover, all LLM data selection methods demonstrate severe shortcomings in representativeness, which leads to a decline in the performance of general VQA tasks (average 49.6 in TextVQA and 49.3 in GQA). In contrast, we observe that DataTailor obtains overall improvement on various tasks. For an intuitive illustration, we visually compare the metrics of three principles of DataTailor and those LLM data selection methods, as shown in Figure 1(b). Similarly, those LLM data selection methods exhibit deficiencies in at least one dimension, whereas DataTailor consistently achieves promising results. This result demonstrates DataTailor’s capability to effectively promote these principles for multi-modal data selection, rather than roughly selecting samples based on individual values.

Our DataTailor can be flexibly applied to different MLLMs. We incorporate our DataTailor into the two most popular MLLMs for evaluation. Despite the diversity in data and model structures, our DataTailor consistently im-

Methods	Feature Backbone	MME \uparrow	SEED-1 \uparrow	MMMU(val) \uparrow	SciQA \uparrow	Rel.
mPLUG-264k [54]						
Full Data (100%)	-	1243.4	34.3	26.9	41.1	100.0%
Random (5%)	-	1183.2	33.9	26.6	40.6	97.9%
TIVE [35] (5%)	LLaVA-vicuna-7B	1177.1	34.0	26.9	40.2	97.9%
DataTailor (5%)	LLaVA-vicuna-7B	1260.0	34.5	27.8	41.9	101.8%
	mPLUG-OWL-7B	1217.2	34.3	28.0	41.3	100.6%
Bunny-695k [17]						
Full Data (100%)	-	1778.1	70.7	38.7	70.9	100.0%
Random (5%)	-	1578.8	64.4	36.7	70.1	93.4%
TIVE [35] (5%)	LLaVA-vicuna-7B	1542.4	61.7	34.7	68.4	90.0%
DataTailor (5%)	LLaVA-vicuna-7B	1582.8	65.3	37.1	75.8	96.0%
	Bunny-Phi-3B	1599.8	63.3	37.3	74.4	95.2%

Table 2. Transferability analysis of multi-modal data selection.

proves relative performance of data selection across all benchmarks compared to random selection (e.g., +15.5% on MiniGPT4-7B and +4.8% on LLaVA-v1.5-7B). Notably, despite comprising only 7.5% data, DataTailor consistently outperforms full fine-tuning on the challenging MMMU benchmark (23.6 v.s. 23.4 on MiniGPT4-7B and 33.9 v.s. 32.8 on LLaVA-7B). These results indicate that our proposed method consistently addresses data redundancy across different datasets and architectures.

Compared with MLLM-specialized selection methods, DataTailor exceeds SOTA for overall benchmarks. Specifically, DataTailor achieves an average of 100.1% relative performance, outperforming TIVE [35] (94.6%) and the latest ICONS [50] (98.8%). Notably, when increasing data selection ratio to 15%, DataTailor surpasses the performance of the full tuning of LLaVA-v1.5-7B (103.0% for MLLM benchmarks and 101.3% for total benchmarks). This indicates that a small amount of high-quality data is more crucial than a large volume of low-quality data for enhancing MLLMs, which truly exemplifies “Less is More”.

4.3. Transferability of Multi-modal Data Selection

Previous MLLM-specialized methods [35, 51] relied on the same MLLMs as in the training phase for data selection. The transferability analysis aims to investigate whether multi-modal data selected by surrogate models can be efficiently transferred into the target MLLM. Here, we use LLaVA-v1.5-7B as the surrogate model to select valuable data for the target MLLMs mPLUG-Owl-7B [54] and Bunny-3B [17], whose corresponding target datasets are mPLUG-264k and Bunny-695k, respectively. Moreover, we explore DataTailor’s robustness among other feature backbones with different structures and parameters. Table 2 presents the results of our DataTailor and other baselines

We observe that, (a) despite inconsistencies between the data selection model and the target MLLMs, DataTailor still consistently achieves over 95% relative performance of full fine-tuning with only 5% data (101.8% in mPLUG-264k and 96.0% in Bunny-695k). This demonstrates the powerful transferability of DataTailor and its potential for surrogate data selection. In contrast, TIVE [35] performs similarly to or worse than random selection (e.g., 90.0% v.s. 93.4% in Bunny-695k) though it outperforms it in Tab. 1. This discrepancy may stem from TIVE’s strong correlation with

Methods	Principled Values			Benchmarks			
	V_i^{Inf}	V_i^{Uni}	V_i^{Rep}	MME \uparrow	MMMU(val) \uparrow	SciQA \uparrow	Rel.
1 Full Data	-	-	-	1744.8	32.8	70.0	100.0%
2 Random	✓	✓	✓	1675.0	32.2	70.0	95.3%
3 $+V_i^{Inf}$	✓	✓	✗	1759.3	34.9	70.2	98.0%
4 $+V_i^{Uni}$	✓	✓	✓	1716.2	33.5	69.8	97.3%
5 $+V_i^{Rep}$	✓	✓	✓	1771.4	33.8	68.5	97.5%
6 DataTailor	✓	✓	✓	1823.7	33.9	70.9	100.1%
7 w/o adaptive collaboration	✓	✓	✓	1770.2	34.0	70.2	98.8%
8 w/o adaptive proportion	✓	✓	✓	1753.4	33.3	70.1	97.7%
9 w/o adaptive collaboration & proportion	✓	✓	✓	1730.0	32.4	69.3	97.2%

Table 3. Ablation study of each module in DataTailor. All experiments are with 7.5 % selection proportion on LLaVA-mix-665k and Rel. denotes the average boost on all 13 benchmarks.

training gradients of the specific model. (b) The stable performance transition (nearly 1%) indicates that DataTailor emphasizes the inherent value of samples rather than relying on the model features as in previous feature-based baselines, which highlights its transferability for data selection.

4.4. In-depth Analysis

Analysis of Instruction Selection Factors. To investigate our DataTailor deeply, we study the ablation variants of different factors in Table 3. Specifically, we analyze their independence using the following ablation strategy: 1) $+V_i^{Inf}$: we only use the informative value. 2) $+V_i^{Uni}$: we only include the unique value. 3) $+V_i^{Rep}$: we only consider the representative value. 4) w/o adaptive collaboration & proportion: we gradually remove adaptive weights for collaboration and adaptive proportions for selection. Note that we adopt adaptive proportion in all Row 2-5 for fair comparison. The results of Row 3 indicate that informative value is the most crucial principle. Also, Row 4 and 5 suggest the importance of unique and representative values, as unique values support MLLMs’ discriminative capabilities, while representative values enhance their generative capabilities. Furthermore, decreasing performance in rows 7-9 suggests that the adaptively collaborative strategy effectively ensures diversity between tasks in multi-modal data selection.

Instantiation of Three Principles. To explore DataTailor addressing three principles for data selection, we clarify three principles based on the following insights: (a) **informativeness** is crucial for generalization when initializing MLLMs; (b) **uniqueness** is guaranteed to promote MLLMs after partial training by novel contributions; (c) **representativeness** improves relevant samples’ performance during training. Specifically, we instantiate three settings to evaluate each of them essentially: (a) the improvement with 1% data added to initial MLLMs reflects informativeness; (b) the improvement with extra 1% data added to MLLMs based on 15% random data training reflects uniqueness; (c) the average reduced loss of relevant samples with 1% data added reflects representativeness. We normalize each metric and compare it to three types of baselines in Fig. 1(b). The selected data from baselines show limitations in at least one principle. In contrast, DataTailor effectively selects valuable data for all three aspects of MLLM. Please refer to Appendix D.1 for more details of instantiation analysis.

Robustness Analysis of DataTailor. (a) *Model scale ro-*

Methods	MMMU(val)	LLaVA-Wild	SciQA	GQA	Rel.
LLaVA-v1.5-13B					
Full Data (100%)	35.2	69.5	72.6	63.3	100.0%
Random (15%)	33.0	67.3	71.8	60.7	96.3%
DataTailor (15%)	36.4	75.0	73.9	61.2	102.4%

Table 4. Model scale robustness analysis of DataTailor.

Methods	Redundancy Perturbation			Noise Perturbation		
	POPE	SciQA	GQA	POPE	SciQA	GQA
LLaVA-v1.5-7B						
Full Data (100%)	84.7	68.2	56.1	83.0	65.3	51.2
Random (15%)	82.1	64.0	40.7	81.5	63.9	43.2
TIVE (5%)	81.1	54.1	42.5	80.9	62.4	45.4
DataTailor (5%)	81.4	65.9	46.9	84.2	63.7	48.5

Table 5. Dataset robustness analysis of DataTailor with redundancy perturbation and noise perturbation.

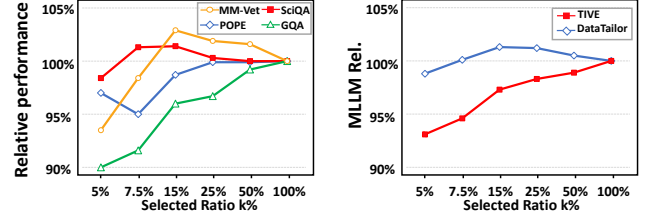
business. In Table 4, we verify the robustness of DataTailor with larger model scales. When keeping the same architecture and using LLaVA-v1.5-13B with higher parameters, our method demonstrates enhanced performance with larger scale MLLMs, especially +5.5 points on the LLaVA-Wild benchmark. This shows the stability of DataTailor at model scales. (b) *Dataset robustness*. In Table 5, we explore the dataset robustness of DataTailor with two challenging perturbations of candidate datasets: **redundancy perturbation** and **noise perturbation**. Specifically, we construct these perturbations with 50k redundant data and 50k noisy data by resampling and wrong answer combination. Note that we also sample 50k normal data from LLaVA-665k for balance. With redundancy and noise perturbations, we notice that DataTailor consistently demonstrates superior performance with limited data, whereas TIVE experiences a significant performance drop (65.9 v.s. 54.1 of SciQA on redundancy disturbance and 48.5 v.s. 45.4 of GQA on noise disturbance). It indicates DataTailor can bring out more distinctive and representative samples to identify the truly valuable samples from the redundant and noisy data for better robustness, which is crucial for discrimination tasks.

Influence of Selection Proportion $k\%$ in DataTailor. As shown in Figure 5 (a), when the selected data volume is relatively small, the model’s performance improves as the data scale increases. However, due to the limited valuable data, further increasing the data volume introduces redundancy and noise, which degrades data-sensitive visual recognition performance (*i.e.* MM-Vet). This reveals the necessity of selecting data to ensure efficiency and maintain performance. Moreover, DataTailor quickly obtains over 100% overall performance (100.1% with only 7.5% data) while TIVE exhibits only limited performance while growing slowly, as shown in Figure 5 (b). It indicates that DataTailor considers the uniqueness of the selected samples to avoid repetition and achieve continuous improvement.

Computation Cost Analysis. Since the overhead of data selection is crucial for effective pruning methods, we analyze the computational cost of DataTailor when selecting 7.5% data. We find that TIVE even exceeds the original training cost, which has certain limitations. In contrast,

	Warmup		Data Selection (7.5%)		Training	
	Complexity	Actual	Complexity	Actual	Complexity	Actual
Full Model	-	-	-	-	$\mathcal{O}(D \cdot S)$	100 H
TIVE [35]	$\mathcal{O}(D \cdot S_{\text{warmup}})$	8 H	$\mathcal{O}(D \cdot S)$	100 H	$\mathcal{O}(D \cdot S^*)$	7.5 H
DataTailor (Ours)	-	-	$\mathcal{O}(S)$	15 H	$\mathcal{O}(D \cdot S^*)$	7.5 H

Table 6. Asymptotic complexity and wall-clock runtime (measured with 4×3090 for LLaVA-v1.5-7B experiments on LLaVA-mix-665k dataset). $|D|$ is the complexity of gradient computation.



(a) Data Scaling in LLaVA-1.5-mix-665k

(b) Data Scaling Comparison

Figure 5. Ablation study of selection ratio $k\%$ in DataTailor.

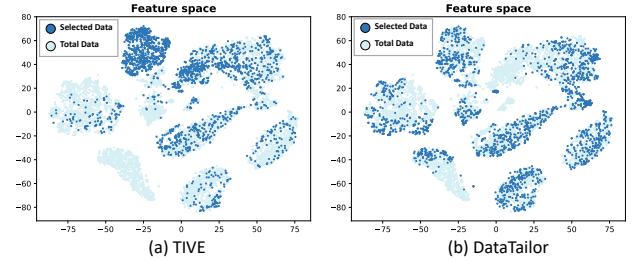


Figure 6. Visualization using t-SNE on feature space.

DataTailor saves nearly 80% of the overall time while outperforming the full model’s performance. This confirms the efficiency of data selection with DataTailor for MLLMs.

Distribution of Selected Data. To give an intuitive perspective on the selected data, we employ t-SNE [46] on the feature space of data chosen by DataTailor in Fig. 6. Notably, DataTailor selects informative samples without redundancy or deviation, while TIVE, despite high informativeness, focuses solely on gradient similarity, leading to redundancy and outlier noise. This visualization confirms the effectiveness of our method adhering to three principles.

5. Conclusion and Future Work

In this paper, we reveal the drawbacks of existing data selection methods and identify three systematic principles of informativeness, uniqueness, and representativeness as fundamental to optimizing multi-modal data selection. Building on this, we propose a unified framework, DataTailor, to synergistically integrate these principles for value evaluation and adaptively address the varying structure and complexity of samples across diverse tasks, thereby mastering collaborative multi-modal data selection. Comprehensive experiments on the challenging MLLM and general VQA benchmarks show that DataTailor significantly improves the performance of MLLM data selection. In the future, we aim to extend DataTailor to more challenging interleaved datasets with extra modalities such as video and audio.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (62436007), the Key R&D Projects in Zhejiang Province (No. 2024C01106, 2025C01030), the Zhejiang NSF (LRG25F020001) and Wallenberg-NTU Presidential Postdoctoral Fellowship. We thank all the reviewers for their valuable comments.

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 6
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1
- [3] Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*, 2024. 3
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 6
- [5] Wendong Bu, Yang Wu, Qifan Yu, Minghe Gao, Bingchen Miao, Zhenkui Zhang, Kaihang Pan, Yunfei Li, Mengze Li, Wei Ji, et al. What limits virtual agent application? omnibench: A scalable multi-dimensional benchmark for essential virtual agent capabilities. *arXiv preprint arXiv:2506.08933*, 2025. 2
- [6] Yihan Cao, Yanbin Kang, and Lichao Sun. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*, 2023. 1, 3, 6
- [7] Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. *arXiv preprint arXiv:2309.17002*, 2023. 2, 3
- [8] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019. 2, 3
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2
- [10] Qianlong Du, Chengqing Zong, and Jiajun Zhang. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*, 2023. 1
- [11] Dante Everaert and Christopher Potts. Gio: Gradient information optimization for training dataset selection. *arXiv preprint arXiv:2306.11670*, 2023. 3
- [12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 5
- [13] Minghe Gao, Juncheng Li, Hao Fei, Liang Pang, Wei Ji, Guoming Wang, Zheqi Lv, Wenqiao Zhang, Siliang Tang, and Yueting Zhuang. De-fine: De composing and re fin ing visual programs with auto-feedback. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7649–7657, 2024. 3
- [14] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355, 2024. 3
- [15] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023. 2
- [16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6
- [17] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 2, 7
- [18] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. 5
- [19] Hongzhe Huang, Zhewen Yu, Jiang Liu, Li Cai, Dian Jiao, Wenqiao Zhang, Siliang Tang, Juncheng Li, Hao Jiang, Haoyuan Li, et al. Align²llava: Cascaded human and large language model preference alignment for multi-modal instruction curation. *arXiv preprint arXiv:2409.18541*, 2024. 1
- [20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6
- [21] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. 3
- [22] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*, 2024. 3, 6

- [23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2
- [24] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 5
- [25] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2
- [26] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023.
- [27] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu, and Yueting Zhuang. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12601–12617, 2023. 2
- [28] Ming Li, Yong Zhang, Zhitao Li, Jiu-hai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023. 1, 3, 6
- [29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5
- [30] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 3
- [31] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for llm-based recommendation. *arXiv preprint arXiv:2401.17197*, 2024. 1
- [32] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 3, 5
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 6
- [35] Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024. 1, 3, 6, 7, 8
- [36] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3, 6
- [37] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011. 4
- [38] Kaihang Pan, Zhaoyu Fan, Juncheng Li, Qifan Yu, Hao Fei, Siliang Tang, Richang Hong, Hanwang Zhang, and Qianru Sun. Towards unified multimodal editing with enhanced knowledge collaboration. *arXiv preprint arXiv:2409.19872*, 2024. 2
- [39] Kaihang Pan, Siliang Tang, Juncheng Li, Zhaoyu Fan, Wei Chow, Shuicheng Yan, Tat-Seng Chua, Yueting Zhuang, and Hanwang Zhang. Auto-encoding morph-tokens for multimodal llm. *arXiv preprint arXiv:2405.01926*, 2024. 2
- [40] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021. 6
- [41] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momen-tor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024. 3
- [42] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms’ compositional reasoning by spatio-temporal graph-guided self-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3284–3294, 2025. 3
- [43] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 6
- [44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6
- [45] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8
- [47] Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigt-4. *arXiv preprint arXiv:2308.12067*, 2023. 3, 6
- [48] Biao Wu, Fang Meng, and Ling Chen. Curriculum learning with quality-driven data selection. *arXiv preprint arXiv:2407.00102*, 2024. 3

- [49] Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*, 2023. 1, 3, 6
- [50] Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024. 3, 6, 7
- [51] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024. 1, 3, 6, 7
- [52] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022. 3
- [53] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2
- [54] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 2, 7
- [55] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21560–21571, 2023. 1, 2
- [56] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 3
- [57] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025. 2
- [58] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6
- [59] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 6
- [60] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3
- [61] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 3, 5