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

Supplementary Material

A. Overview

In this supplementary material, we present:

- More detailed analyses of DataTailor (Section B).
- More experimental details (Section C).
- Additional Experiment Analyses (Section D).

B. DataTailor Framework

B.1. Principle Definition

To avoid the computational infeasibility of the impact of samples on downstream task performance, we define three principles based on the geometric statistical properties of the samples to serve as practical approximation for assessing the value of multi-modal instructions:

Definition B.1 (Informativeness). It quantifies the samples' difficulty for downstream tasks within latent space. The sample difficulty is related to its feature distribution diversity. A sample $s \in S, S \subseteq D$ is high-informativeness if its token-level feature distribution is more diverse, indicating it contains a wider variety of information at the token level. By prioritizing such samples during training, models can learn more robust representations, thereby improving inference accuracy on downstream tasks.

Intuitively, simple samples often contain redundant information (*e.g.*, images with large areas of meaningless background or answers with repetitive descriptions). In such cases, the model can "trickily" focus on a subset of information to complete the task. Due to the high similarity of tokens in the sample, the feature matrix columns (or rows) exhibit strong linear dependence, meaning the matrix has a low rank and contains significant redundancy. According to Singular Value Decomposition (SVD), smaller singular values decrease significantly, and larger ones become disproportionately larger, leading to lower singular value entropy. Mathematically, if the feature matrix $\mathbf{M}$ has low rank, its singular value matrix Σ can be represented as:

$$\mathbf{M} = U\Sigma V \tag{1}$$

where the smaller singular values in Σ approach zero, while the larger singular values dominate. The singular value entropy (SVE) is defined as the entropy of normalized singular values [5], which is computed as:

$$\mathbf{SVE} = -\sum_{i=1}^{r} p(\sigma_i) \log p(\sigma_i) \tag{2}$$

where σ_i is a singular value, $p(\sigma_i)$ is the normalized probability distribution of the singular values, and r is the rank of the singular value matrix. Since many of the singular values are close to zero, the entropy is low, reflecting the simplicity of the sample. In contrast, when the feature matrix of a sample is more information-rich and closer to full rank, singular values contribute more evenly, leading to higher entropy, which indicates a more complex sample.

Therefore, singular value entropy serves as a practical approximation of sample informativeness, enabling the selection of more challenging samples that encapsulate a diverse range of information for downstream tasks.

Definition B.2 (Uniqueness). It measures deviation from local data density. A sample s is high-uniqueness in neighborhood $\mathcal{N}(s)$ if:

$$\min_{s' \in \mathcal{N}(s)} \|s - s'\| \geq \delta \cdot \text{diam}(\mathcal{N}(s)) \tag{3}$$

where $\delta \in (0, 1)$ thresholds the relative margin and $\text{diam}(\cdot)$ is the diameter of the subset within the intra-cluster space. Due to local density constraints, models are forced to learn non-degenerate decision boundaries [6], which ensures the distinction of sample distributions and improves adversarial robustness. According to Lemma 2.4 [3], maximizing the Euclidean intra-cluster margins achieves this through geometric packing in ℓ_2-space, which is equivalent to sampling along the data manifold boundary to reduce redundancy between samples thereby improving training robustness [11]. It is worth noting that the dynamic computation based on the selected sample may be influenced by the greedy side effects during the selection process, making it difficult to achieve a global optimal solution.

Therefore, the static Euclidean distance between all neighborhood samples in the intra-cluster space can approximately capture data density of sample sets, allowing the selection of unique samples for continual performance improvement. Given a sample s and its neighbor samples $x_j \in \mathbf{C}$, the Euclidean distance $d_{i,j}$ can be calculated as follows:

$$d_{i,j} = \|\mathbf{p_j} - \mathbf{p_i}\|_2 \tag{4}$$

where a smaller Euclidean distance of two samples in the latent space indicates that these samples are highly similar, leading to a lower uniqueness value.

Definition B.3 (Representativeness). It ensures samples adhere to population-level statistics. A sample s is representative if its feature vector $\phi(s)$ satisfies:

$$W_1(\mathcal{P}_S, \mathcal{P}_D) \leq \gamma \tag{5}$$

where W_1 is Wasserstein-1 distance, $\mathcal{P}_S$ is average feature of the subset, and $\mathcal{P}_D$ is average feature of the global distribution. By incorporating the Wasserstein-1 constraint, it minimizes domain shift and stabilizes the dynamics of gradient descent [4]. According to Proposition 1 [16], cosine similarity to cluster centroids approximates W_1-optimal transport under spherical normalization. Therefore, the similarity distribution from the inter-cluster space can approximately capture the alignment between samples and the overall data distribution for minimizing domain shift. The representativeness value can be measured by the cluster density of the sample, which is commonly estimated using the similarity distribution between cluster centroids:

$$\tau_i^c = \frac{1}{K-1} \sum_{k \neq c}^{K} \exp(\text{sim}(\overline{\mathbf{p_k}}, \overline{\mathbf{p_c}})) \tag{6}$$

where $\overline{\mathbf{p_c}}$ is the feature of the cluster centroid that is calculated by the average feature of samples in the cluster. Specifically, high similarity indicates that the cluster containing the sample is well-aligned with other clusters, making it a strong representative of the overall data distribution.

B.2. Adaptive Data Proportion for Data Selection

To adapt to various task complexities within multi-modal datasets, we introduce adaptive proportion of selected data for each task. The challenge stems from the fact that task difficulty is inherently difficult to assess directly. We observe that data lacking directional diversity causes the generated trajectories to collapse into a limited subspace dominated by a few principal components. This reflects the complexity of the task, highlighting the need for more training data to enhance the robustness of MLLMs. Specifically, given the feature matrix $\mathbf{M}$ of a sample s_i, the largest singular value calculation relies on Singular Value Decomposition (SVD):

$$\mathbf{M} = U\Sigma V^T, \Sigma = \{\sigma_1, ..., \sigma_r\} \tag{7}$$

where Σ is the singular value matrix and r represents the singular value matrix rank of each sample. Here $r = L_i$ since the sample feature dimension d is much larger than the sample token length L_i and σ_i represents each singular value.

Definition B.4 (Largest Singular Value Ratio). The largest singular value ratio (LSVR) is defined as the ratio of the maximum singular value to the sum of singular values. LSVR captures the distribution of significant eigenvectors in a particular direction and reflects the task difficulty for training robustness.

$$\textbf{LSVR} = \frac{\sigma_{\text{max}}}{\sum_{j=1}^{r} \sigma_i} \tag{8}$$

where σ_{max} denotes the dominant singular value. When the LSVR becomes significantly large, it indicates insufficient variation in eigenvectors along the principal direction as follows:

$$\frac{\sigma_{\text{max}}}{\sum_{i=1}^{r} \sigma_i} \gg \frac{1}{r} \tag{9}$$

This indicates that information is concentrated in specific singular value directions in the latent space. Such spectral imbalance in singular values reduces the effective dimensionality of learned representations, meaning that less knowledge can be extracted from individual samples. Consequently, such difficult tasks with larger LSVR require higher data selection proportions to ensure sufficient learning. Specifically, we compute the average of the LSVR for all samples in the task as follows:

$$x_p = avg(\frac{\sigma_{max}}{\sum_{j=1}^{r} \sigma_j}) \tag{10}$$

To amplify the contribution of task difficulty to data selection, we square the average largest singular value ratio and normalize it based on the number of samples corresponding to each task, yielding the data selection ratio as follows:

$$k_p = \frac{x_p^2 \cdot |S_p|}{\sum_q x_q^2 \cdot |S_q|} \cdot k \tag{11}$$

where $|S_q|$ is the corresponding sample number of each task. Then, we adjust the data selection rate of each task from k to k_p to achieve task-adaptive proportions. Once the data selection ratio for each task is determined, we utilize the synergistic sample value from DataTailor to perform collaborative multi-modal data selection for each task.

B.3. Cross-Modal Domain Clustering

B.3.1. Algorithmic Formulation

The clustering pipeline commences with inter-sample affinity quantification via ℓ_2-norm distance measurements within each initial category. Starting with all nodes as individual clusters, we iteratively merge the pair (A,B) exhibiting the minimal increase in Sum of Squared Errors (SSE) to construct a dendrogram:

$$A, B = \arg\min_{A,B \in \mathcal{P}} \Delta\text{SSE}(A, B) \tag{12}$$

where $\mathcal{P}$ denotes the partition of nodes into distinct clusters. The SSE increment from merging clusters A and B is defined as:

$$\Delta\text{SSE}(A, B) = \frac{n_A n_B}{n_A + n_B} \|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|_2 \tag{13}$$

with n_A, n_B representing cluster sizes and $\boldsymbol{\mu}_A, \boldsymbol{\mu}_B$ their centroids. This criterion minimizes intra-cluster variance growth. To theoretically characterize the merging behavior, we formalize a key monotonicity property inherent to Ward's algorithm.

Theorem 1 (SSE Monotonicity in Ward's Method)**.** *For a hierarchical merging process $\{\mathcal{P}_k\}_{k=0}^{n-1}$ under Ward's algorithm (where $\mathcal{P}_0$ contains singleton clusters), the total SSE increment satisfies the non-decreasing property:*

$$\Delta\mathrm{SSE}(\mathcal{P}_k) \geq \Delta\mathrm{SSE}(\mathcal{P}_{k-1}), \quad \forall k \geq 1,$$

where $\Delta\mathrm{SSE}(\mathcal{P}_k)$ denotes the incremental SSE from merging $\mathcal{P}_{k-1}$ to $\mathcal{P}_k$.

This monotonic progression ensures that earlier merges correspond to more natural cluster unions, while later merges sacrifice increasing amounts of variance. Leveraging this property, we define a threshold-based partitioning rule to extract clusters from the dendrogram hierarchy. A dendrogram $T = (V, E)$ consists of leaf nodes $V_{\text{leaf}} = \{s_1, ..., s_n\}$ and internal nodes, which serve as merge points annotated with ΔSSE values. Given a threshold $T = \lambda \cdot \Delta\mathrm{SSE}(\mathcal{P}_{n-1})$, the optimal partition is determined by:

$$\mathcal{P}^* = \max\{k \mid \Delta\mathrm{SSE}(\mathcal{P}_k) \leq T\}$$

B.3.2. GPU-Accelerated Distance Computation

For efficient pairwise distance measurement between n samples in $\mathbb{R}^d$, we implement a parallelized ℓ_2-norm computation framework using CUDA-optimized matrix operations. The distance matrix $D \in \mathbb{R}^{n\times n}$ can be derived through the algebraic identity, which:

$$D_{x,y} = \sqrt{S_{x,x} + S_{y,y} - 2S_{x,y}} \tag{14}$$

where $X \in \mathbb{R}^{n\times d}$ represents the feature matrix containing n samples, and $S = XX^T \in \mathbb{R}^{n\times n}$ represents the similarity matrix. Our kernel-based implementation achieves $\mathcal{O}(n^2d/m)$ theoretical speedup through massive parallelization across GPU cores, effectively transforming an originally $\mathcal{O}(n^2d)$ complexity operation into a highly parallelizable matrix multiplication task. In addition, we use the parallel pipeline strategy to extract feature vectors of the feature matrix while calculating the ℓ_2-norm distance between sample features for efficiency.

B.3.3. Optimized Hierarchical Clustering Merge

To address the quadratic complexity inherent in conventional hierarchical clustering, we implement a memory-efficient variant of the nearest-neighbor chain algorithm. This optimization framework features:

- Randomized stack initialization with cluster prototypes
- Iterative nearest-pair identification via stack
- In-stack merging with $\mathcal{O}(n)$ per-operation cost, where each node undergoes $\mathcal{O}(1)$ amortized operations.

The proposed acceleration strategy reduces computational overhead to $\mathcal{O}(n^2)$ while preserving the theoretical guarantees of Ward's method. By optimizing the clustering process, the time consumption of the clustering was accelerated by $90\times$, resulting in significantly improved performance.

Threshold λ	w/o clustering	0.05	0.1	0.25	0.5
MLLM Rel.	98.1%	99.8%	101.3%	100.0%	98.5%

Table 1. The analysis of different similarity thresholds for cross-modal domain clustering in extrinsic value estimation.

B.3.4. Threshold Parameter Analysis

Since the quality of clustering is critical for the domain-based adaptive data proportion in DataTailor, we further explore the effect of dynamic threshold to cross-modal domain clustering on the multi-modal data selection in Table 1. The clustering threshold determines the cluster size based on the dataset's sample distribution. Therefore, setting it close to the overall data selection proportion ensures an appropriate size of clusters that effectively captures sample relationships. Our experiments reveal that low or high thresholds compromise the constraints on the uniqueness or representativeness of high-quality samples, leading to lower performance of DataTailor. Thus, we set the appropriate threshold λ as 0.1 for cross-modal domain clustering, which is close to the total data selection proportion.

B.4. Balance between Three Principles

Since multi-modal samples exhibit varying structures, we propose an adaptive weight to combine the three principal values. We restate the underlying principles behind the three properties to show their effectiveness: (a) Informativeness. It determines external relationships due to its core training contributions. Because MLLMs rely on token-level inputs, the SVD of token feature space reveals the samples' contribution to the MLLM. That's why we use singular value entropy to reflect the value of samples for generalization. (b) Uniqueness. For repeated samples, their uniqueness is adjusted during selection based on the distribution of chosen points by normalization, ensuring duplicates are treated differently. (c) Representativeness. It aims to isolate undesired noisy samples that may have a high uniqueness score, while general noisy data can be identified by its lower informativeness (i.e., average 0.297 in clusters).

In general, it is important to balance the above three principles to select data collaboratively. On the one hand, we explore various ratios of these two external values (*i.e.*, uniqueness and representativeness) in Figure 2. The smooth performance transition (less than 2%) near the 1:1 ratio suggests that the trade-off between them remains stable. Therefore, we select the 1:1 ratio as the optimal value for collaboration. On the other hand, we use adaptive weights between the information values and the two values for the varying instruction rounds of the samples in the dataset.

$V^{Uni}:V^{Rep}$	0:1	0.5:1	1:1	1:0.5	1:0
MLLM Rel.	99.6%	100.5%	101.3%	100.3%	99.5%

Table 2. Balance between uniqueness and representation in DataTailor for data selection of MLLMs.

C. More Experimental Details

C.1. Implementation Details

Following prior research [14, 18] and each dataset scale, we keep 5% as the data proportion (0.2k) for data selection on MiniGPT4-Instruction [23] and 7.5% as the data proportion (50.0k) for data selection on on LLaVA-1.5-mix-665k [12] for the standard setting. In the transferability analysis, we uniformly set 5% as the data proportion (12.3k) for data selection on mPLUG-Owl-7B-264k-Instructions [21] and 5% as the data proportion (34.7k) for data selection on Bunny-695k [8]. During the data selection process, we retain all parameters from the original model but freeze all gradients. DataTailor evaluates the values of the three principles for multi-modal samples using the initialized features of the pre-trained model. This allows DataTailor to select high-quality samples while efficiently maintaining strong transferability.

During data selection in DataTailor, we balance the uniqueness and representativeness of different clusters. First, we normalize uniqueness and representativeness values across clusters by removing the influence of spatial distribution and cluster size. Next, we standardize the impact of the average sample value across clusters. Specifically, within the same task, we uniformly scale informativeness, uniqueness, and representativeness metrics to a $[0, 1]$ range to ensure consistent distribution alignment. By harmonizing these normalized values across samples, we enable collaboration among Informativeness, Uniqueness, and Representativeness values. This methodology fosters balanced metric collaboration in our adaptive data selection framework, ensuring proportional consideration of all three criteria.

During fine-tuning, we apply the LoRA strategy [9] to fine-tune each dataset and its subsets from various data selection methods due to the limited GPU resources. For LLaVA-v1.5-7B, we use 4*3090 GPUs for fine-tuning, where the batch size of each device is set to 12 and the training epoch is set to one epoch. For MiniGPT-4-7B, we use 1*A6000 GPU for fine-tuning, where the batch size of each device is set to 12 and the training epoch is set to 5 epoch. During fine-tuning, we only distinguish the dataset scale through various data selection methods and keep all other training parameters consistent for a fair comparison.

C.2. Candidate Datasets Details

MiniGPT4-Instruction. It contains approximately 3,500 instruction pairs, each consisting of an image and a corresponding detailed description. The correctness of each image description is manually verified to ensure high quality.

LLaVA-v1.5-mix-665k. This is currently the most extensive multimodal instruction dataset, encompassing instruction data across a wide range of tasks. It contains a variety of datasets: VQA [2], OCR [15], region-level VQA [10], visual conversation [13] and language conversation [1] data. For all datasets, QA pairs from the same training image are merged into a single conversation, and excessively long data is filtered out to improve training efficiency. As a result, this process yields 665k instruction pairs across 13 tasks.

mPLUG-Owl-7B-264k-Instructions. It gathers pure text instruction data from two distinct sources: 52k data from the Alpaca [17] and 54k from the Baize [19]. Additionally, it involves 158k multi-modal instruction data from visual conversations in the LLaVA dataset [13]. In this way, it incorporates both pure text instruction data and multimodal instruction data, demonstrating that DataTailor is well-suited for diverse data selection tasks.

Bunny-695k. It primarily utilizes SVIT-mix-665k [22], replacing ShareGPT-40k [1] with WizardLM-evol-instruct-70k [20] to create Bunny-695k. Compared to LLaVA-665K, this dataset contains more complex multi-modal instructions, enabling the evaluation of DataTailor's ability to transfer to more intricate multi-modal data selection.

Methods	Selected Data Evaluation			MLLM Rel.
	Informativeness	Uniqueness	Representativeness	
IFD (7.5%)	32.3	0.341	30.3	87.3%
INSTAG (7.5%)	30.9	0.347	34.4	96.4%
LESS (7.5%)	34.0	0.314	33.3	94.3%
$+V_i^{Inf}$ (7.5%)	34.5	0.348	34.8	98.0%
$+V_i^{Uni}$ (7.5%)	33.4	**0.364**	34.5	97.3%
$+V_i^{Rep}$ (7.5%)	33.9	0.343	**35.0**	97.5%
DataTailor (7.5%)	**34.8**	0.358	34.9	100.1%

Table 3. Quantitative Valuation of Three Principles. All setups are on the model of LLaVA-v1.5-7B and the dataset of LLaVA-mix-665K for both data selection and MLLM training.

D. Additional Experiment Analyses

D.1. Quantitative Valuation of Three Principles

To quantitatively investigate how DataTailor addresses each of the three principles for data selection, we designed three experimental settings to assess them. Here we show the detailed metric values for DataTailor and other baselines on three experimental setups in Table 3.

D.2. More scalable MLLMs Results

Further, we apply DataTailor to **larger and newer backbones** (*i.e.*, Qwen-2-VL-7B & 72B) for more robust evaluation in the Table 4. Since Qwen's data is closed-source, we fine-tune on the open Virgo dataset [7]. Similarly, we observe that DataTailor exhibits competitive performance with

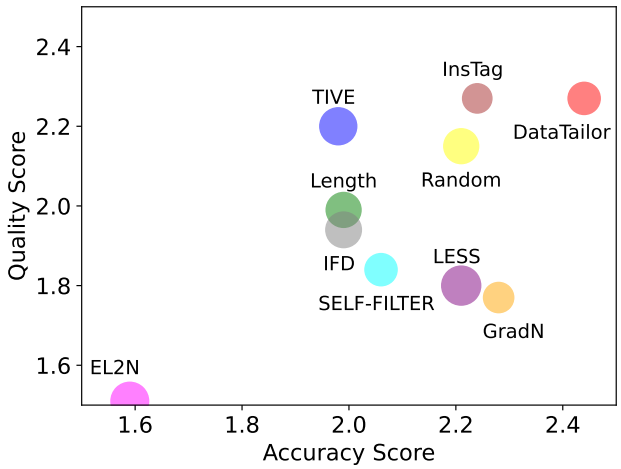

Figure 1. Quality score (y-axis, higher is better), accuracy score (x-axis, higher is better), and the stability (circle sizes, smaller is better) of MLLMs' responses on OwlEval benchmark. We set the data selection ratio for each method to 7.5%.

Methods	Dataset	MMMU	MathVerse	MathVision	MLLM Avg.
QVQ-72B-preview	-	66.0	41.5	38.2	48.6
InternVL2.5-78B	-	70.0	39.2	32.2	47.1
Qwen2-VL-7B	Virgo(100%)	46.7	36.7	24.0	35.8
w/ random	Virgo(15%)	46.7	32.5	23.4	34.2
w/ DataTailor	Virgo(15%)	50.1	34.8	23.8	**36.2**
Qwen2-VL-72B	Virgo(100%)	65.0	48.1	38.6	50.6
w/ random	Virgo(15%)	59.4	43.7	36.8	46.6
w/ DataTailor	Virgo(15%)	63.0	46.2	40.3	**49.8**

Table 4. The scalable results on math and reasoning benchmarks of DataTailor with more scalable MLLMs.

only 15% data (36.2 v.s. 35.8 of full data), outperforming other **open-source MLLMs** baselines.

D.3. Instantiation Comparisons of Three Principles

To fully explore the characteristics of valuable samples that are meaningful for downstream tasks, we further present a few instantiation comparisons of sample characteristics from the three principal perspectives in Figure 2. The left side shows the data preferred by DataTailor, and the right side shows the remaining data.

From the perspective of **informativeness**, we can observe that the samples selected by DataTailor contain richer information and various description, whereas other samples suffer from excessive redundancy in responses and numerous blank pixels in images (*e.g.*, the right image shows only snow slope and the text repeating snowboard). This indicates that selecting samples according to informativeness can select samples as complex as possible to improve inference accuracy in downstream tasks while also facilitating the generation of richer and higher-quality content.

From the perspective of **uniqueness**, we can observe that the samples selected by DataTailor in each cluster contain

Methods	Informativeness	Uniqueness	Representativeness	MLLM Rel.
$+V^{Inf}$ (15%)	34.3(33.1)	-	-	99.2%(97.7%)
$+V^{Uni}$ (15%)	-	0.363(0.347)	-	98.5%(97.4%)
$+V^{Rep}$ (15%)	-	-	34.6(32.5)	98.6%(97.6%)

Table 5. Deep Analysis for Calculation of Three Values

unique insights (*e.g.*, critically mention artistic installation) and in-depth novel analysis (*e.g.*, analyze diversity of the ornamental tree). However, other examples in the cluster show similar information and primarily describe basic, pre-learned commonsense knowledge, which limits MLLMs' ability to enhance generalization on downstream tasks continuously. This suggests that selecting samples based on uniqueness enables the inclusion of distinctive samples, which provide deeper and more novel insights to continuously enhance the reasoning capability during the fine-tuning phase of MLLM.

From the perspective of **representativeness**, we observe that the samples selected by DataTailor exhibit accurate descriptions and aligned answers. This is because the method ensures that the selected samples represent the overall distribution, avoiding noise and mislabeled data. Samples with low representativeness tend to exhibit mislabeled answers and outlier features, which can lead to incorrect optimization directions, ultimately hindering the performance improvement of downstream tasks. This suggests the need for assessing the representativeness of samples to filter out noisy or mislabeled data.

Based on the above, we propose three essential principles (*i.e.*, informativeness, uniqueness, and representativeness) for the practical selection of valuable samples that truly contribute to the downstream inference performance of MLLMs. To further illustrate the superiority of these three quantitative metrics for three principles, we explore the insights underlying three values and conduct more **ablation** of potential alternatives (results are of red color in Table 5) For **informativeness**, we adopt the SVE based on the theory that more informative samples have feature matrices closer to full rank, leading to more uniform singular values and thus higher SVE. We explore pair-wisely calculating mutual information between token-level features as an alternative. As SVE captures the overall distribution of feature directions, it achieves better informativeness and MLLM performance for data selection (99.2% v.s. 97.7%). For **Uniqueness**, we adopt the Euclidean distance based on the theory that larger distances in the feature space separate more unique samples and samples containing pre-learned common knowledge are clustered near the center. We explore directly selecting one sample from each cluster as an alternative. As Euclidean distance measures samples' deviation from dense regions, it obtains better MLLM performance with uniqueness (98.5% v.s. 97.4%). For **Representativeness**, we adopt cosine similarity based on the theory that it reflects the true directional alignment of the

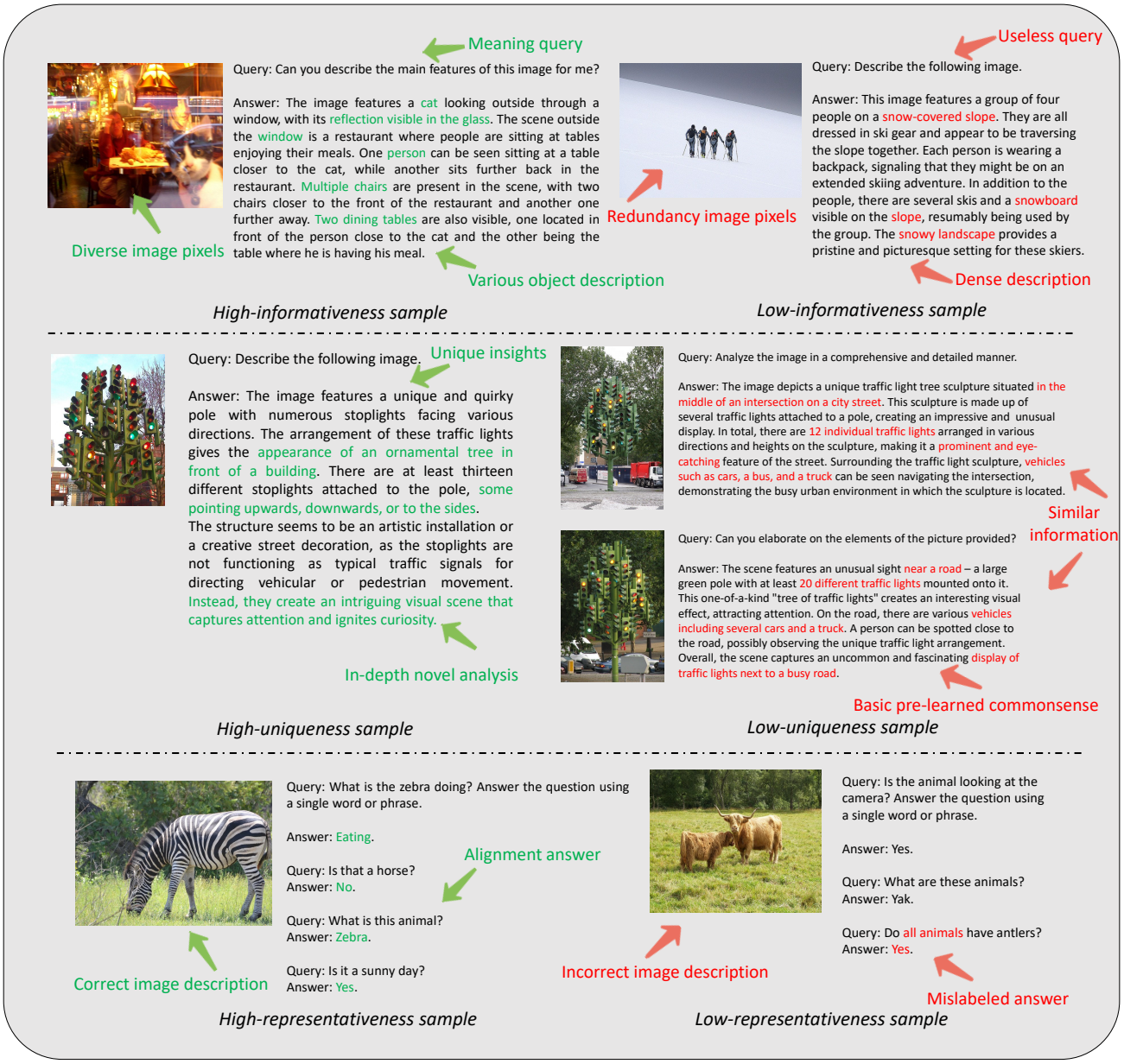

Figure 2. Instantiation Comparisons of DataTailor in addressing three core principles (*i.e.* informativeness, uniqueness, and representativeness) for selecting valuable multi-modal instruction samples.

overall distribution. We explore using the overall Euclidean distance between samples as an alternative. As cosine similarity emphasizes overall directional alignment instead of being influenced by the magnitude of outlier samples like distance, it performs better in MLLM with true representativeness (98.6% v.s. 97.6%).

D.4. Limitations

We observe some failure cases that DataTailor cannot discriminate long-term instructions when the reasoning process is omitted, *e.g.*, math reasoning and code program.

D.5. Human Evaluation

To comprehensively evaluate whether data selection ensures the open-ended capabilities of MLLMs, we conduct further

human evaluations using the OwlEval benchmark. OwlEval [21] is an open-ended evaluation set comprising 82 artificially constructed questions. We evaluated responses from all models on a 3-0 scale (aligned with option A-D in the official setting), assessing quality based on informativeness and alignment with the question, and accuracy based on consistency with image content. Furthermore, we calculate the score variance for all responses of the MLLMs using different data selection methods to assess model stability. We visualize the human-evaluation results in Figure 1. We observe that using DataTailor for data selection best preserves the response capabilities of MLLMs, enabling them to provide both informative answers and maintain the highest level of accuracy. This demonstrates that DataTailor effectively selects representative samples to support the overall capabilities of MLLMs, addressing the challenge of collaborative multimodal data selection without overemphasizing specific abilities.

References

[1] Sharegpt, 2023. https://sharegpt.com/. 4

[2] 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. 4

[3] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680, 2008. 1

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018. 2

[5] 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. 1

[6] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013. 1

[7] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025. 4

[8] 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. 4

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[11] Semi-Supervised Learning. Semi-supervised learning. *CSZ2006. html*, 5(2):1, 2006. 1

[12] 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. 4

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[14] 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. 4

[15] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4

[16] François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019. 2

[17] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 4

[18] Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*, 2023. 4

[19] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023. 4

[20] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[21] 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. 4, 7

[22] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 4

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