VisNec: Measuring and Leveraging Visual Necessity for Multimodal Instruction Tuning

Mingkang Dong[1,2]*, Hongyi Cai[2]*, Jie Li[1], Sifan Zhou[3], Bin Ren[4], Kunyu Peng[5], and Yuqian Fu[1,6]†

[1]SJTU, [2]Universiti Malaya, [3]CMU, [4]MBZUAI, [5]KIT, [6]KAUST

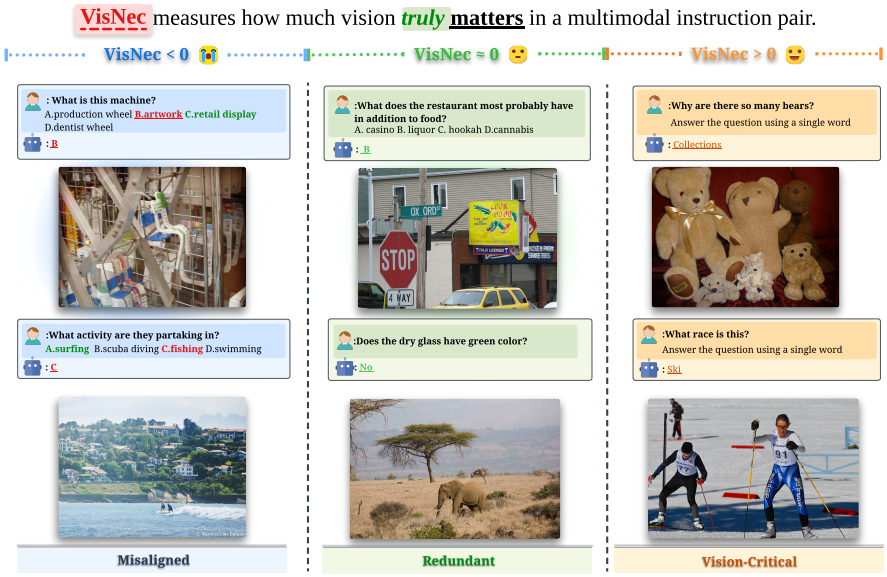

Fig. 1: Illustration of the VisNec score, which measures the contribution of visual input in multimodal instruction tuning. Samples are categorized as *Misaligned* (VisNec < 0, visual input degrades prediction), *Redundant* (≈ 0, visual input provides no additional benefit), or *Vision-Critical* (> 0, visual input substantially improves prediction).

Abstract. The effectiveness of multimodal instruction tuning depends not only on dataset scale, but also critically on whether training samples genuinely require visual reasoning. However, existing instruction datasets often contain a substantial portion of visually redundant samples (solvable from text alone), as well as multimodally misaligned supervision that can degrade learning. To address this, we propose **VisNec (Visual Necessity Score)**, a principled data selection framework that measures

* Equal contribution.
† Corresponding author: Yuqian Fu was a visiting scholar at SJTU during this work.

the marginal contribution of visual input during instruction tuning. By comparing predictive loss with and without visual context, VisNec identifies whether a training instance is vision-critical, redundant, or misaligned. To preserve task diversity, we combine VisNec with semantic clustering and select high-necessity samples within each cluster. Across 10 downstream benchmarks, training on only 15% of the LLaVA-665K dataset selected by VisNec achieves 100.2% of full-data performance. On the smaller Vision-Flan-186K dataset, our selection not only further reduces data size but surpasses full-data training by 15.8%. These results demonstrate that measuring and leveraging visual necessity provides an effective solution for both efficient and robust multimodal instruction tuning. **Project Page:** **https://dmk041218.github.io/VisNec/**.

Keywords: Multimodal Large Language Model · Multimodal Instruction Tuning · Data Selection · Visual Necessity

1 Introduction

Multimodal instruction tuning [21, 23, 33, 44] plays a central role in training multimodal large language models (MLLMs), enabling them to follow complex vision-language instructions and generalize across diverse tasks [10, 13, 17, 19, 20, 24, 28, 30, 34–38, 42, 45, 46]. Beyond architectural design, the quality of instruction data fundamentally determines whether models develop genuine cross-modal reasoning or rely on superficial correlations. Consequently, curating high-quality multimodal instruction data is critical for robust and reliable performance.

While large-scale instruction datasets [15] collected from diverse sources have successfully trained strong MLLMs, directly using all available data presents two notable limitations. First, it is inefficient. As datasets scale to hundreds of thousands or millions of samples, training costs grow substantially, yet not all samples contribute equally to learning. Second, large-scale instruction data inevitably contains a significant proportion of redundant or misaligned samples. Visually redundant instances can often be answered through linguistic shortcuts alone (e.g., predicting "green" for "the color of grass"), providing little incentive for visual grounding. When such samples dominate training, models tend to exploit textual correlations rather than integrate visual evidence, weakening cross-modal dependence. In contrast, multimodal misalignment, arising from annotation errors or noisy data pipelines, produces image-text pairs with inconsistent supervision. Training on such samples can actively degrade visual reasoning and amplify hallucinations at inference time. These observations raise a fundamental question: *Can we identify a small yet truly valuable subset of multimodal instruction data that is both efficient and effective?*

Existing data selection approaches primarily rely on generic importance or diversity signals, such as gradient influence [39], transferability [26], or clustering-based coverage [7, 14]. However, these methods treat multimodal samples holistically and do not explicitly distinguish the independent contribution of the visual modality. As a result, selected subsets may still favor linguistically easy samples

or retain harmful misaligned supervision. In contrast, we posit that the value of a multimodal instruction sample largely depends on whether visual input is necessary for resolving predictive uncertainty. This perspective shifts the focus from generic sample importance to **visual necessity** as the guiding principle for multimodal data selection.

To implement this principle, we draw inspiration from the theory of V-usable information [8], which formalizes how much a variable reduces predictive uncertainty beyond what is already available from other contexts. While V-usable information has been used to assess textual data quality, extending this principle to multimodal instruction data remains underexplored. We argue that this perspective provides the right lens for evaluating visual supervision: a training instance is only as valuable as the uncertainty that can be uniquely resolved by visual input. Grounded in this insight, we introduce **Visual Necessity Score (VisNec)**, a lightweight data selection framework that explicitly quantifies the marginal contribution of the visual modality. By contrasting model losses with and without visual input, VisNec produces a principled per-sample measure of visual necessity. As shown in Fig. 1, VisNec successfully identifies instances that are vision-critical, redundant, or misaligned. Moreover, considering that visual dependency varies across task types, e.g., geometric reasoning naturally demands stronger visual grounding than knowledge-based QA, we integrate stratified semantic clustering with intra-cluster VisNec ranking. This ensures that the selected subset maintains both visual indispensability and broad task coverage.

Extensive experiments across ten benchmarks demonstrate that VisNec recovers or even surpasses full-data performance using only 15% of the training data, achieving 100.2% on LLaVA-665K [21] and 115.8% on Vision-Flan-186K [40]. Besides, our method generalizes across model scales from 3B to 32B. Our main contributions are summarized as follows:

1. We identify a critical yet overlooked limitation in existing multimodal data selection: the neglect of the visual modality's independent contribution, which leads to widespread "pseudo-multimodal" samples that reinforce linguistic shortcuts and hinder true cross-modal reasoning.
2. We propose VisNec, a lightweight and model-relative data selection framework that quantifies the marginal contribution of visual input, ensuring the selected subset is both visually indispensable and task-diverse.
3. We achieve state-of-the-art performance across diverse training datasets and benchmarks, demonstrating that measuring visual necessity enables both data efficiency and improved multimodal reasoning robustness.

2 Related Work

Multimodal Instruction Tuning. To enable Multimodal Large Language Models (MLLMs) to effectively handle diverse downstream tasks, instruction tuning has emerged as a fundamental stage in the training pipeline. Established frameworks such as LLaVA [21], InstructBLIP [6], and Qwen-VL [1] demonstrate that fine-tuning on a broad range of instruction-following pairs is critical

for aligning multimodal representations with human intent. However, the effectiveness of this alignment typically relies on the availability of large-scale, high-quality datasets. As these collections scale to hundreds of thousands of samples, the computational resources and training time required for fine-tuning become significant bottlenecks. Furthermore, large-scale instruction datasets, whether automatically collected or model-generated, inevitably include redundant and misaligned samples, which can weaken visual supervision and potentially degrade multimodal reasoning.

Information-Theoretic Data Quality. A fundamental question in data-centric learning is how to quantify the value of individual training samples. V-usable information [8] formalizes this as the reduction in predictive uncertainty contributed by a variable beyond existing context, providing a principled view of feature-level informativeness. Building on this perspective, dataset difficulty measures [29] and loss-based scoring methods [16] treat predictive loss as an indicator of sample quality in LLM fine-tuning. These works demonstrate that loss-based signals are effective in single-modal settings. However, by operating on text alone, existing approaches cannot disentangle visual informativeness from linguistic ambiguity. A sample may appear difficult due to annotation noise or textual complexity rather than genuine visual contribution. By contrast, our work extends this information-theoretic perspective to multimodal instruction tuning by explicitly isolating the marginal contribution of visual input through counterfactual loss comparison, thus providing a modality-aware quality signal.

Data Selection for Instruction Tuning. Existing data selection methods suffer from a range of practical limitations. Complexity-driven approaches like Self-Filter [5] rely on multi-scoring mechanisms that are themselves computationally expensive. Methods such as PreSel [32] and CoIDO [41] offload quality judgment to closed-source APIs, incurring both financial costs and privacy concerns. COINCIDE [14] and XMAS [26] improve scalability through clustering and SVD-based alignment trajectories, respectively, yet still impose notable overhead on large-scale datasets. Gradient-based methods like ICONS [39] offer precise importance estimation at the cost of per-sample backpropagation, making them prohibitively slow in practice. Others are directly adapted from the NLP paradigm [3, 4], applying language-centric quality metrics that inherently favor samples with rich textual content but weak visual grounding. IFD [16], most closely related to our work, selects samples by their response prediction loss under text-only input. However, this conflates genuine visual complexity with textual ambiguity or annotation noise. A misaligned sample may score high on IFD precisely because the image misleads the model, yet such samples are detrimental to training. VisNec resolves this by computing the differential between blind and multimodal losses, explicitly rewarding visual informativeness while penalizing misalignment. Overall, existing data selection methods typically suffer from limited computational efficiency, dependence on external resources, or insufficient sensitivity to cross-modal quality, making truly efficient multimodal instruction tuning difficult in resource-constrained settings.

3 Methodology

Problem Formulation. Let $\mathcal{D} = \{(v_i, t_i, y_i)\}_{i=1}^{N}$ denote a multimodal instruction tuning dataset, where v_i represents the input image, t_i the textual instruction, and y_i the target response. The objective for fine-tuning a Multimodal Large Language Model (MLLM), parameterized by θ, is to minimize the multimodal training loss $\mathcal{L}_{\text{MM}}$, defined as the negative log-likelihood of the response y given the multimodal context:

$$\mathcal{L}_{\text{MM}}(\theta; v, t, y) = -\sum_{j=1}^{L} \log P(y_j \mid y_{<j}, v, t; \theta), \tag{1}$$

where L is the length of the response tokens and $y_{<j}$ denotes the tokens preceding position j. While effective, blindly optimizing this objective over massive datasets often leads to inefficiency, as the model may overfit to linguistic priors in samples where visual information is redundant. A fundamental question arises: which samples genuinely require visual reasoning, and which are solvable by linguistic priors alone? Equation (1) does not differentiate between these cases, treating all samples equally, regardless of their visual necessity.

3.1 Visual Necessity Score (VisNec)

Towards Visual-Centric Data Selection. Grounded in the V-usable information framework [8], which quantifies how much a variable X reduces predictive uncertainty about Y via:

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y \mid X), \tag{2}$$

where $H_{\mathcal{V}}(Y)$ and $H_{\mathcal{V}}(Y \mid X)$ denote the V-entropy without and with input X respectively, we argue that the value of a multimodal training sample should be measured by how much the visual modality uniquely reduces predictive uncertainty beyond what text alone provides. This motivates a direct approximation within an MLLM: we quantify visual necessity as the marginal loss reduction attributable to the visual input:

$$\Delta\mathcal{L}(v, t, y) = \mathcal{L}(y \mid t) - \mathcal{L}(y \mid t, v). \tag{3}$$

A larger $\Delta\mathcal{L}$ indicates higher visual necessity, meaning the visual input substantially reduces predictive uncertainty beyond what text alone provides. Conversely, a smaller or negative $\Delta\mathcal{L}$ suggests the sample is either linguistically solvable or multimodally misaligned. We address the concrete computation of $\mathcal{L}(y \mid t)$ within an MLLM in the following section.

The Blind Forward Pass. To instantiate $\mathcal{L}(y \mid t)$ within a multimodal model, we perform a counterfactual inference called the *Blind Forward Pass*. For each sample (v, t, y), we replace image token positions with padding tokens and suppress their attention contributions by setting the corresponding attention mask

to zero, ensuring that no visual information can propagate into response token representations through any attention computation. This keeps the model architecture and parameter space identical to the standard forward pass, with the presence of visual input as the sole controlled variable between the two passes. To ensure that both passes compute loss over identical response token positions, image token positions are excluded from loss computation via `ignore_index`, such that $\mathcal{L}_{\text{Blind}}$ and $\mathcal{L}_{\text{MM}}$ are both averaged exclusively over response tokens:

$$\mathcal{L}_{\text{Blind}}(\theta; t, y) = -\sum_{j=1}^{L} \log P(y_j \mid y_{<j}, t; \theta). \tag{4}$$

$\mathcal{L}_{\text{Blind}}$ thus quantifies the model's residual uncertainty in predicting y from linguistic context t alone. Although the blind input lies outside the MLLM's training distribution, this does not undermine the validity of VisNec as a selection signal: since data selection operates via intra-cluster ranking rather than absolute score thresholds, any systematic bias introduced by the out-of-distribution input cancels out in the relative ordering.

Visual Necessity Score. We define the VisNec score S_{VisNec} as the discrepancy between the blind loss and the multimodal loss, representing the marginal contribution of the visual modality to the model's prediction:

$$S_{\text{VisNec}}(v, t, y) = \mathcal{L}_{\text{Blind}}(\theta; t, y) - \mathcal{L}_{\text{MM}}(\theta; v, t, y). \tag{5}$$

This formulation offers a clear interpretation of sample utility:

- $S_{\text{VisNec}} > 0$: **High Visual Gain.** The image significantly reduces prediction error, indicating the sample requires true cross-modal reasoning.
- $S_{\text{VisNec}} \approx 0$: **Redundancy.** The model predicts the answer equally well without the image, suggesting the sample relies heavily on linguistic priors.
- $S_{\text{VisNec}} < 0$: **Misalignment/Noise.** The presence of the image increases the loss, implying the visual content contradicts the text or introduces noise.

3.2 Semantic-Aware Stratified Sampling

Relying solely on VisNec scores for selection carries a risk: the distribution of visual dependency is non-uniform across tasks. For instance, geometric reasoning tasks naturally have higher VisNec scores than OCR tasks. A naive top-k selection might bias the dataset toward specific domains. To preserve broad task diversity, we adopt a coarse-to-fine paradigm that first groups samples by semantic intent and then performs intra-cluster selection.

Instruction Clustering. To capture the true semantic structure of the dataset, we focus on the core task definition encapsulated in the user's query. Specifically, we extract the question component q_i from each textual instruction t_i (filtering out system prompts) and map it into a semantic embedding space. We then perform K-Means clustering to learn a partition function $\pi : q \to \{1, \ldots, K\}$, which assigns each question to one of K semantic clusters. This explicitly groups samples based on *task intent*, isolating distinct capabilities such as geometric reasoning, OCR, and creative generation.

Intra-Cluster Selection. We construct the k-th semantic cluster $\mathcal{C}_k = \{(v_i, t_i, y_i) \in \mathcal{D} \mid \pi(q_i) = k\}$. Within each cluster, we first remove samples with $S_{\text{VisNec}} \leq 0$, as these indicate either misaligned image-text pairs where the visual input actively misleads the model or redundant samples where the visual context contributes nearly nothing to the response. The remaining samples are then ranked by their VisNec scores. Given a budget ratio r, we select the top-$r\%$ instances to form the final subset:

$$\mathcal{D}_{\text{select}} = \bigcup_{k=1}^{K} \text{Top-}r\% \left(\{(v_i, t_i, y_i) \in \mathcal{D} \mid \pi(q_i) = k\}; S_{\text{VisNec}} \right). \tag{6}$$

Overall Framework. As illustrated in Fig. 2, our framework operates in two stages. In the first stage, each sample is scored by computing the difference between its text-only loss and multimodal loss within the target MLLM, yielding a per-sample Visual Necessity Score that measures how much the visual input actually contributes to the response. Samples with non-positive scores are discarded as they are redundant or misaligned, while the rest are ranked by their VisNec scores. In the second stage, we perform semantic clustering over the instructions and apply intra-cluster selection based on these scores, ensuring that the final subset is both visually indispensable and task-diverse.

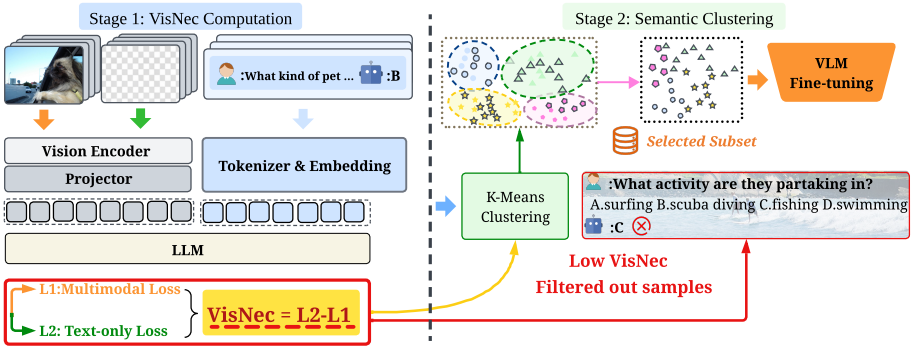

Fig. 2: Overview of the VisNec Data Selection Framework. Each sample undergoes two forward passes to compute its VisNec score. Samples with non-positive scores are filtered out, and the remainder are grouped via K-Means clustering, from which the top-$r\%$ are selected within each cluster for fine-tuning.

3.3 Case Study

To illustrate how VisNec identified problematic and high-value samples, we show three representative cases in Fig. 3.

Fig. 3: Qualitative comparison of Low (−1.2087), Medium(0.2058) and High VisNec (3.4727) samples, illustrating misaligned, redundant and visual dependent samples.

Low VisNec (Misalignment). The left example receives a negative score (-1.21), revealing multiple annotation errors. First, the question "Is it outdoors?" is answered "Yes," but the image clearly shows an indoor scene with furniture and walls. Second, when asked "In which part of the picture is the happy woman?" the answer is "Left"—yet no woman is visible in the image at all. These contradictions mean the model performs better by ignoring the image entirely. Without visual input, it can at least avoid being misled by annotations that do not match the actual content. VisNec correctly assigns a negative score to such samples and discards them outright, as their visual content actively contradicts the textual supervision rather than complementing it.

Medium VisNec (Redundant). The middle example scores 0.21, indicating that the visual input contributes little beyond what linguistic priors already provide. Questions such as "Is the cow's horn large or small?" and "Which place is it?" can largely be answered through general world knowledge alone, without any visual grounding. Such samples are not harmful but offer minimal cross-modal supervision, as the model learns nothing it could not have inferred from text alone. VisNec correctly assigns them a near-zero score and deprioritizes them during selection, effectively treating them as visually redundant.

High VisNec (Valuable). The right example scores 3.47, showing strong visual dependency. The questions request descriptions of specific regions defined by coordinates [0.0, 0.25, 1.0, 0.66] and [0.0, 0.57, 0.34, 0.86]. Without the image, these questions are impossible to answer since no amount of linguistic prior knowledge can determine what objects occupy those spatial locations. Such samples require

leveraging visual grounding to identify the objects at these specific coordinates. By filtering out negative scores and prioritizing high scores within each semantic cluster, VisNec removes noisy annotations and text-solvable shortcuts while retaining samples that teach cross-modal reasoning.

4 Experiments and Results

4.1 Experimental Setup

Datasets. To validate the effectiveness and scalability of our proposed approach across various magnitudes of instruction-tuning data, we conduct experiments on two datasets: LLaVA-665K and Vision-Flan-186K. LLaVA-665K is a widely used dataset, comprising more than 600K samples across about 10 different tasks. Vision-Flan-186K includes 191 distinct tasks, which allows us to assess the method's performance when dealing with a high number of tasks and a relatively smaller quantity of data per task. For both experiments, we employ a 15% sampling ratio to evaluate performance under budget-constrained scenarios.
Baselines. We evaluate our approach against a diverse set of competitive baselines, ranging from traditional sampling to recent state-of-the-art (SOTA) methods, including Random, Self-Filter, EL2N [29], TypiClust [12], IFD [16], PreSel [32], XMAS [26], COINCIDE [14], CLIP-Score [31], and CoIDO [41]. Specifically, XMAS, PreSel, and CoIDO represent recent SOTA techniques in multimodal data selection. In contrast, IFD and EL2N are representative methods originally developed for instruction tuning or data selection in the NLP / supervised-learning setting.
Evaluation Benchmarks. We evaluate the performance of different fine-tuned MLLMs across 10 multimodal evaluation benchmarks that test various capabilities of MLLMs. The benchmarks include: 1) Visual Question Answering: VQAV2 [11], GQA [13]; 2) Knowledge-Grounded QA: SQA-I [25]; 3) Multiple-choice understanding: MM-Bench (EN/CN) [22] and MME-Perception [9]; 4) Text understanding in images: TextVQA [35]; 5) Open-ended generation: LLaVA-Wild Bench [21]; 6) Hallucination and factual tests: POPE [18], MM-Vet [43].
Training Details. To ensure a fair and rigorous comparison, we strictly adhere to the official training configurations across all experiments, including model architectures, hyperparameters, and optimization schedules. Specifically, starting from the LLaVA-v1.5 [21] Stage-1 pre-trained checkpoint (feature alignment), we conduct LoRA fine-tuning for 1 epoch with a default learning rate of 2×10^{-4}, following the standard LLaVA training pipeline. For the K-means clustering involved in our approach, we set the number of clusters to $K = 20$. The main experimental results are summarized in Tab. 1 and Tab. 2, which report the performance comparison between our approach and several state-of-the-art baselines on the LLaVA-665K and Vision-Flan-186K datasets, respectively. Given the varying scales across different benchmarks, we report the Average Relative Performance (Rel.) to facilitate a unified comparison. For each benchmark, relative

Table 1: Overall performance and efficiency comparison of selection approaches for fine-tuning LLaVA-v1.5-7b on the LLaVA-665K dataset using a 15% (98K samples) sampling ratio across various multimodal evaluation benchmarks. "Rel" indicates relative performance. The best result is **bolded** and the runner-up is underlined.

	Method	VQAv2	GQA	LLaVA Wild	SQA-I	TextVQA	MME-P	MMBench en	MMBench cn	POPE	MM-Vet	Rel.
	LLaVA-v1.5-7B on LLaVA-665K											
0	Full-Data(665K)	79.1	63.0	67.9	68.4	57.9	1476.9	64.3	58.3	86.4	30.0	100%
1	Random	75.3	55.1	58.8	67.8	54.3	1397.5	61.0	53.5	84.9	30.2	94.2%
2	Self-Filter [ACL-2024]	74.0	56.3	60.6	62.3	51.4	1356.5	48.1	45.4	86.3	29.0	89.3%
3	EL2N [NeurIPS-2021]	76.1	54.7	59.0	66.5	50.2	1405.2	58.2	48.5	83.3	30.0	91.9%
4	TypiClust[ICML-2022]	76.0	59.8	65.2	68.2	53.3	1396.2	64.3	57.1	85.6	29.7	96.9%
5	IFD [NAACL-2024]	74.0	57.8	62.3	66.5	51.8	1307.2	57.2	50.6	**86.6**	28.1	92.1%
6	PreSel [CVPR-2025]	76.5	57.9	65.6	**70.1**	55.2	1457.7	64.8	56.5	85.4	29.6	97.7%
7	XMAS [arxiv-2025]	75.1	**61.2**	62.4	67.0	55.2	**1485.7**	64.3	54.1	85.8	31.0	97.3%
8	COINCIDE[EMNLP-2024]	76.1	60.2	64.9	67.7	54.8	1414.9	60.5	53.9	86.4	28.5	95.8%
9	CLIP-Score[ICML-2021]	71.7	56.9	61.3	64.5	53.4	1380.3	51.0	48.0	84.0	29.9	92.0%
10	CoIDO[NeurIPS-2025]	75.8	61.0	66.7	68.2	56.0	1419.5	63.3	55.5	85.6	31.4	97.7%
11	ICONS[arxiv-2025]	76.3	60.7	65.3	65.3	55.2	1435.6	63.1	55.8	85.7	30.4	97.1%
12	VisNec(ours)	**78.0**	60.8	**69.8**	67.9	**56.2**	1457.2	**64.9**	**59.1**	86.0	**32.1**	**100.2%**

performance is defined as:

$$\text{Rel} = \frac{\text{The Model's Performance}}{\text{Full Fine-tuned Model's Performance}} \times 100\%. \tag{7}$$

This metric represents the percentage of the full-data baseline performance achieved by the model trained on the selected data subset.

4.2 Main Results on Multi-modal Data Selection

Superiority Over State-of-the-Art Baselines. Tab. 1 presents a quantitative comparison of VisNec against competitive baselines for fine-tuning LLaVA-v1.5-7B. Under a strict 15% data budget on the LLaVA-665K dataset, VisNec achieves a remarkable relative performance of 100.2%, consistently surpassing all baseline approaches. Specifically, VisNec outperforms the Random baseline by 5.8% and the second-best method by 2.3%. We attribute this to a fundamental distinction in how VisNec defines sample value: rather than relying on generic importance signals or external quality proxies, VisNec explicitly computes the difference between blind and multimodal losses, directly measuring how much the visual input reduces predictive uncertainty. This allows VisNec to prioritize vision-critical samples while discarding those where the image actively misleads the model—a dimension that existing methods do not explicitly capture. Notably, VisNec even exceeds the full-data (100%) fine-tuning baseline on LLaVA-Wild, MMBench-EN/CN, and MM-Vet, benchmarks that demand open-ended reasoning and hallucination resistance. These benchmarks are precisely where noisy and visually redundant training samples are most damaging, and where removing them yields the greatest benefit.

Generalization Across Datasets. Given that the LLaVA-1.5 dataset covers a relatively limited range of tasks, we evaluate the cross-dataset generalization

Table 2: Overall performance and efficiency comparison of selection approaches for fine-tuning LLaVA-v1.5-7b on the Vision-Flan-186K dataset using a 15% (28K samples) sampling ratio across various multimodal evaluation benchmarks. "Rel" indicates relative performance. The best result is **bolded** and the runner-up is underlined.

	Method	VQAv2	GQA	LLaVA Wild	SQA-I	TextVQA	MME-P	MMBench en	MMBench cn	POPE	MM-Vet	Rel.
	LLaVA-v1.5-7B on Vision-Flan-186K											
0	Full-Data(186K)	69.6	46.0	35.7	55.6	38.3	1238.1	53.4	48.2	85.7	27.7	100%
1	Random	66.5	43.8	33.5	62.1	38.7	1238.6	43.6	43.1	83.0	28.1	96.7%
2	Self-Filter [ACL-2024]	64.9	42.5	32.1	59.3	42.6	1262.2	42.1	43.8	80.9	25.1	95.0%
3	EL2N [NeurIPS-2021]	63.6	42.2	32.8	62.7	42.4	1253.8	44.7	37.7	79.5	27.1	95.2%
4	TypiClust[ICML-2022]	65.8	43.1	30.4	59.2	37.7	1194.1	32.4	45.1	81.6	28.2	92.0%
5	IFD [NAACL-2024]	65.0	42.4	29.8	57.8	42.0	1210.9	30.4	40.8	82.6	26.9	91.6%
6	PreSel [CVPR-2025]	64.1	41.9	39.4	66.2	39.7	1218.8	50.4	45.4	84.1	29.1	100.6%
7	XMAS [arxiv-2025.10]	65.5	49.4	34.2	58.1	42.4	1151.2	51.1	44.0	76.2	24.3	96.9%
8	COINCIDE[EMNLP-2024]	66.0	44.5	34.6	63.9	33.0	1184.4	49.6	48.2	84.3	26.1	97.1%
9	CLIP-Score[ICML-2021]	63.0	41.6	31.2	62.3	39.7	1058.0	37.5	44.2	82.0	28.0	92.2%
10	CoIDO[NeurIPS-2025]	66.7	46.8	37.6	66.2	43.2	1298.8	51.4	47.3	**85.6**	28.3	103.6%
11	ICONS[arxiv-2025]	67.2	48.8	35.4	60.2	49.9	1252.5	51.3	45.4	83.0	28.6	103.2%
12	VisNec(ours)	65.0	**57.1**	**39.5**	**74.5**	**51.6**	**1505.5**	**62.6**	**53.1**	82.0	**32.1**	**115.8%**

of our approach on the more challenging Vision-Flan-186K dataset, which spans 191 distinct tasks. As reported in Tab. 2, VisNec achieves an exceptional relative performance of 115.8%, outperforming the Random baseline by 18.4% and the full-data fine-tuned model by 15.8%. This result reveals a critical insight: a substantial portion of Vision-Flan-186K is not merely redundant but **harmful** to model training. Samples with low or negative VisNec scores, where visual content contradicts textual annotations, actively degrade cross-modal reasoning. By explicitly filtering such misaligned samples, VisNec confirms that in multimodal instruction tuning, data quality is far more critical than data quantity.

4.3 Transferability across Architectures and Scales

To verify whether VisNec captures marginal data value rather than model-specific biases, we evaluate cross-architecture transferability by applying VisNec scoring directly on Qwen2.5-VL [2] family across three scales (3B, 7B and 32B) on the LLaVA-665K dataset. Despite significant architectural differences from LLaVA-v1.5, the subsets selected by Qwen2.5-VL-3B, 7B, and 32B all achieve remarkable performance. As shown in Tab. 3, VisNec using only 15% of the full data achieves 103.8%, 104.0%, and 102.4% of their respective full data performance. This demonstrates that VisNec captures the intrinsic visual necessity of the data rather than model-specific preferences, making it a robust and architecture-agnostic selection strategy for scaling next-generation MLLMs with minimal data overhead.

4.4 Parameter Sensitivity Study

To evaluate the sensitivity of our proposed method, we conduct an ablation study on two key components: the number of clusters k in K-means and the

Table 3: Performance comparison of Qwen2.5-VL (3B, 7B and 32B) trained LLaVA-665k subset versus full data. **Rel.** indicates relative performance compared to the Full method.

Model	Data	Method	VQAv2	GQA	LLaVA Wild	SQA-I	TextVQA	MME-P	MMB (en)	MMB (cn)	POPE	MM-Vet	Rel.
Qwen2.5VL-3B	665k	Full	79.7	59.4	73.5	70.3	73.6	1475.6	72.0	69.0	87.6	50.2	100%
	98k	Random	78.3	58.4	72.1	71.3	72.1	1439.2	71.9	68.3	87.0	49.1	98.8%
		VisNec	**80.7**	**59.5**	**75.4**	**76.0**	**76.8**	**1552.5**	**76.2**	**75.4**	**88.3**	**50.5**	**103.8%**
Qwen2.5VL-7B	665K	Full	83.4	**62.9**	77.8	75.6	77.8	1592.5	81.3	78.1	87.5	50.7	100%
	98k	Random	82.1	61.3	77.5	75.2	79.2	1549.3	80.2	76.2	88.1	48.5	98.7%
		VisNec	**83.3**	62.1	**80.9**	**87.7**	**83.7**	**1603.6**	**82.8**	**80.2**	**88.5**	**54.3**	**104.0%**
Qwen2.5VL-32B	665K	Full	**82.1**	60.2	75.5	81.9	77.8	1684.5	80.8	78.6	**88.5**	52.3	100%
	98k	Random	80.3	58.1	74.2	80.6	75.1	1648.2	79.5	78.2	86.3	51.6	97.9%
		VisNec	81.5	**60.6**	**78.0**	**86.0**	**78.2**	**1687.2**	**82.2**	**81.4**	88.1	**57.7**	**102.4%**

selection ratio ρ.

Analysis of the Number of Clusters (k). We investigate the impact of the number of clusters k in our stratified sampling strategy. As shown in Fig. 4, our method remains highly stable across a range of k from 10 to 30. The relative performance (Rel) fluctuates within a narrow margin of 1.3%, consistently maintaining over 98.9% of the full-data performance and outperforming other methods. This suggests VisNec is robust to different levels of semantic granularity. We use $k = 20$ as the default setting in our experiments.

Analysis of Selection Ratio. We investigate the impact of the selection ratio

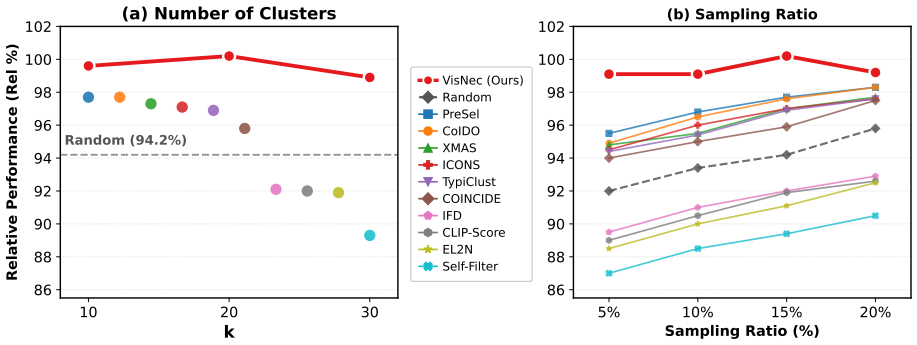

Fig. 4: Impact of (a) the number of clusters k in K-means; (b) the sampling ratio ρ.

(ρ) on model performance, sweeping from 5% to 20%. As illustrated in Fig. 4, VisNec exhibits a remarkable ability for strategic data concentration. Specifically, using only 5% of the training data, our method restores 99.1% of the full-data baseline performance. The performance reaches its peak at $\rho = 15\%$, achieving 100.2% relative performance, which effectively outperforms the model trained on the entire LLaVA-665K dataset. This result indicates that VisNec goes beyond mere data reduction; it performs a semantic-aware refinement of

the instruction set. By leveraging VisNec scores to identify samples with the highest multi-modal information gain within each semantic cluster, VisNec isolates a high-fidelity core-set that maximizes learning efficiency. The slight performance dip at 20% (99.2% Rel.) reinforces the "less-is-more" paradigm in visual instruction tuning, confirming that excessive data may introduce gradient noise or redundancy. Consequently, we utilize 15% as our optimal selection ratio to maintain peak accuracy with minimal computational cost.

4.5 Contribution of Loss Components

The core premise of VisNec is that sample value is best captured by the marginal contribution of the visual modality over linguistic context. To validate this, we disaggregate our scoring mechanism and evaluate the individual contributions of each loss component against a Random sampling baseline, as detailed in Tab. 4. Our analysis reveals a clear hierarchy in selection efficacy. The Text-only strategy (95.6% Rel.) outperforms Random sampling (94.2%), showing that linguistic difficulty is a reasonable proxy for sample quality. However, the Multimodal-only variant surprisingly falls below the random baseline at 94.0%. Without the contrastive signal from the text-only loss, the selection process tends to favor samples with high visual complexity but low instructional clarity, effectively introducing noise into the training subset.

In contrast, the full VisNec framework achieves 100.2% relative performance, a 4.6% margin over the best single-modality baseline. This confirms that VisNec captures a synergy between textual and visual signals that neither modality alone can provide—identifying samples where the two are mutually informative rather than redundant or conflicting. Consequently, training on just 15% of VisNec-selected data surpasses full-data supervision, indicating that the differential loss formulation is key to isolating genuinely cross-modal samples.

Table 4: Ablation study on loss components. We report results on LLaVA-v1.5-7B as the base model. Our default setting is k=20, $\rho = 15\%$.

Model	VQAv2	GQA	LLaVA Wild	SQA-I	TextVQA	MME-P	MMBench (en)	MMBench (cn)	POPE	MM-Vet	Rel.
LLaVA-7B 100%	79.1	63.0	67.9	68.4	57.9	1476.9	64.3	58.3	86.4	30.0	100%
Random	75.3	55.1	58.8	67.8	54.3	1397.5	61.0	53.5	84.9	30.2	94.2%
Text	75.0	58.3	66.0	64.7	54.7	1436.2	61.4	55.2	86.6	28.4	95.6%
Multimodal	76.2	57.5	59.7	65.4	54.1	1443.5	60.2	51.6	85.3	29.2	94.0%
VisNec	**78.0**	**60.8**	**69.8**	**67.9**	**56.2**	**1457.2**	**64.9**	**59.1**	**86.0**	**32.1**	**100.2%**

4.6 Contribution of Clustering

We further investigate whether semantic clustering is truly necessary or whether selecting top-r% samples by VisNec score alone (TopVisNec) suffices. As shown in Tab. 5, TopVisNec outperforms Random sampling yet still lags behind the full

framework by 3.2%. This gap suggests that VisNec scores alone, without cluster-level constraints, are insufficient to maintain a balanced task distribution across the selected subset.

Table 5: Ablation study on clustering. We report results on LLaVA-v1.5-7B as the base model. Our default setting is k=20, $\rho = 15\%$.

Model	VQAv2	GQA	LLaVA Wild	SQA-I	TextVQA	MME-P	MMBench (en)	MMBench (cn)	POPE	MM-Vet	Rel.
LLaVA-7B 100%	79.1	63.0	67.9	68.4	57.9	1476.9	64.3	58.3	86.4	30.0	100%
Random	75.3	55.1	58.8	67.8	54.3	1397.5	61.0	53.5	84.9	30.2	94.2%
Top-VisNec	75.9	58.5	66.8	66.2	55.6	1383.3	63.9	56.9	**86.2**	30.0	97.0%
VisNec	**78.0**	**60.8**	**69.8**	**67.9**	**56.2**	**1457.2**	**64.9**	**59.1**	86.0	**32.1**	**100.2%**

4.7 Cost Analysis

We evaluate the computational efficiency of VisNec by comparing it with several state-of-the-art data selection baselines on the LLaVA-665K dataset using LLaVA-v1.5-7B. As shown in Tab. 6, while full-data fine-tuning requires 76.0 GPU-hours, our method VisNec achieves superior performance (100.2% Rel.) with a total cost of only 23.0 GPU-hours. The data selection process of VisNec takes only 12.0 GPU-hours, which is significantly faster than standard filtering methods like Self-Filter (73.5 GPU-hr) and COINCIDE (55.5 GPU-hr). Unlike PreSel or CoIDO, VisNec does not rely on any external LLM APIs [27].

Table 6: Comparison of computational cost and performance on LLaVA-665K. Time metrics are reported in H100 GPU-hours. "API Cost" refers to extra expenditures for external LLM services (e.g., GPT-4). Rel. indicates overall relative performance of the full-data fine-tuned model.

Methods	Selection Time	Finetuning Time	API Cost	Total Cost	Rel.
Full Finetune	–	76.0 GPU-hr	×	76.0 GPU-hr	100%
PreSel	9.0 GPU-hr	11.0 GPU-hr	✓	20.0 GPU-hr + API Cost	97.7%
CoIDO	25.0 GPU-hr	11.0 GPU-hr	✓	36.0 GPU-hr + API Cost	97.7%
Self-Filter	73.5 GPU-hr	11.0 GPU-hr	×	84.5 GPU-hr	89.3%
XMAS	36.0 GPU-hr	11.0 GPU-hr	×	47.0 GPU-hr	97.3%
COINCIDE	55.5 GPU-hr	11.0 GPU-hr	×	66.5 GPU-hr	95.8%
VisNec (Ours)	12.0 GPU-hr	11.0 GPU-hr	×	23.0 GPU-hr	**100.2%**

5 Conclusions

In this paper, we introduce VisNec (Visual Necessity Score), a data-centric framework designed to address the redundancy and misalignment prevalent in current multimodal instruction tuning datasets. By explicitly quantifying the marginal contribution of visual context relative to linguistic priors, VisNec effectively filters out misaligned and redundant samples while preserving task diversity. Extensive experiments demonstrate that VisNec consistently identifies high-value, visually indispensable samples that outperform full-data baselines at a fraction of the computational cost, with strong robustness to hyperparameters and cross-architecture transferability.

6 Acknowledgment

This work is supported by the computing resources provided by TeleAI.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Cai, H., Fu, Y., Fu, H., Zhao, B.: Mergeit: From selection to merging for efficient instruction tuning (2025), https://arxiv.org/abs/2503.00034
4. Cai, H., Li, J., Rahman, M.M., Dong, W.: Low-confidence gold: Refining low-confidence samples for efficient instruction tuning (2025), https://arxiv.org/abs/2502.18978
5. Chen, R., Wu, Y., Chen, L., Liu, G., He, Q., Xiong, T., Liu, C., Guo, J., Huang, H.: Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 4156–4172. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.246, https://aclanthology.org/2024.findings-acl.246/
6. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023), https://arxiv.org/abs/2305.06500
7. Dong, M., Cai, H., Lei, X., Li, J., Zhang, T., Pu, M.: Once-for-all: A train-once and select-anytime framework for multimodal instruction tuning (2026), https://arxiv.org/abs/2605.26761
8. Ethayarajh, K., Choi, Y., Swayamdipta, S.: Understanding dataset difficulty with $\mathcal{V}$-usable information (2025), https://arxiv.org/abs/2110.08420

9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394
10. Fu, Y., Wang, R., Ren, B., Sun, G., Gong, B., Fu, Y., Paudel, D.P., Huang, X., Van Gool, L.: Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In: ICCV (2025)
11. Goyal, Y., Khurana, T., Gokhale, D., Agrawal, A., et al.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
12. Hacohen, G., Dekel, A., Weinshall, D.: Active learning on a budget: Opposite strategies suit high and low budgets (2022), https://arxiv.org/abs/2202.02794
13. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
14. Lee, J., Li, B., Hwang, S.J.: Concept-skill transferability-based data selection for large vision-language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
15. Li, J., Cai, H., Dong, M., Pu, M., You, S., Wang, F., Huang, T.: Pistachio: Towards synthetic, balanced, and long-form video anomaly benchmarks (2026), https://arxiv.org/abs/2511.19474
16. Li, M., Zhang, Y., Li, Z., Chen, J., Chen, L., Cheng, N., Wang, J., Zhou, T., Xiao, J.: From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning (2024), https://arxiv.org/abs/2308.12032
17. Li, Y., Fu, Y., Qian, T., Xu, Q., Dai, S., Paudel, D.P., Van Gool, L., Wang, X.: Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. arXiv preprint arXiv:2508.10729 (2025)
18. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.20, https://aclanthology.org/2023.emnlp-main.20/
19. Li, Z., Chen, Z., Fu, Z., Wang, W., Hu, Y., Guan, W., Nie, L.: Omniego-r^2: A routed reasoning framework for the 1st cross-domain egocross challenge at cvpr 2026. arXiv preprint arXiv:2605.24481 (2026)
20. Liang, G., Wang, Z., Hu, J., Zhou, H., Xue, Z., Zhang, J., Xu, D., Yu, Q.: Render-in-the-loop: Vector graphics generation via visual self-feedback. arXiv preprint arXiv:2604.20730 (2026)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
22. Liu, Y., Duan, H., Zhang, Y., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
23. Liu, Z., Zhou, K., Zhao, W.X., Gao, D., Li, Y., Wen, J.R.: ViFT: Towards visual instruction-free fine-tuning for large vision-language models. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 10341–10366. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.547, https://aclanthology.org/2025.findings-emnlp.547/

24. Lu, K., Zhou, S., Xu, H., Xu, G., Yang, Z., Wang, Y., Xiao, Z., Long, J., Li, M.: Yo'city: Personalized and boundless 3d realistic city scene generation via self-critic expansion. arXiv preprint arXiv:2511.18734 (2025)
25. Lu, P., Mishra, S., Xia, T., et al.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
26. Naharas, N., Nguyen, D., Bulut, N., Bateni, M., Mirrokni, V., Mirzasoleiman, B.: Data selection for fine-tuning vision language models via cross modal alignment trajectories (2025), https://arxiv.org/abs/2510.01454
27. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., etc.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
28. Pan, J., Wang, R., Qian, T., Mahdi, M., Fu, Y., Xue, X., Huang, X., Gool, L.V., Paudel, D.P., Fu, Y.: V^2-sam: Marrying sam2 with multi-prompt experts for cross-view object correspondence (2025), https://arxiv.org/abs/2511.20886
29. Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. CoRR **abs/2107.07075** (2021), https://arxiv.org/abs/2107.07075
30. Pu, M., Lim, M.K., Chong, C.Y., Loy, C.C.: Sigma: Semantically informative pre-training for skeleton-based sign language understanding. arXiv preprint arXiv:2509.21223 (2025)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
32. Safaei, B., Siddiqui, F., Xu, J., Patel, V.M., Lo, S.Y.: Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning (2025), https://arxiv.org/abs/2503.07591
33. Shen, Y., Xu, Z., Wang, Q., Cheng, Y., Yin, W., Huang, L.: Multimodal instruction tuning with conditional mixture of lora (2024), https://arxiv.org/abs/2402.15896
34. Shu, Y., Ren, B., Xiong, Z., Paudel, D.P., Van Gool, L., Demir, B., Sebe, N., Rota, P.: Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. arXiv preprint arXiv:2506.01667 (2025)
35. Singh, A., Natarajan, V., Shah, M., et al.: Towards vqa models that can read. In: CVPR (2019)
36. Tan, Y., Wu, Z., Fu, Y., Zhou, Z., Sun, G., Zamfir, E., Ma, C., Paudel, D., Van Gool, L., Timofte, R.: Xtrack: Multimodal training boosts rgb-x video object trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5734–5744 (2025)
37. Wen, D., Peng, K., Zheng, J., Chen, Y., Shi, Y., Wei, J., Liu, R., Yang, K., Stiefelhagen, R.: Mica: Multi-agent industrial coordination assistant (2025), https://arxiv.org/abs/2509.15237
38. Wen, D., Qi, L., Peng, K., Yang, K., Teng, F., Luo, A., Fu, J., Chen, Y., Liu, R., Shi, Y., Sarfraz, M.S., Stiefelhagen, R.: Go beyond earth: Understanding human actions and scenes in microgravity environments (2025)
39. Wu, X., Xia, M., Shao, R., Deng, Z., Koh, P.W., Russakovsky, O.: Icons: Influence consensus for vision-language data selection (2025), https://arxiv.org/abs/2501.00654
40. Xu, Z., Feng, C., Shao, R., Ashby, T., Shen, Y., Jin, D., Cheng, Y., Wang, Q., Huang, L.: Vision-flan: Scaling human-labeled tasks in visual instruction tuning.

In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 15271–15342. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.905, https://aclanthology.org/2024.findings-acl.905/

41. Yan, Y., Zhong, M., Zhu, Q., Gu, X., Chen, J., Li, H.: Coido: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization (2025), https://arxiv.org/abs/2510.17847
42. Yu, J., Zhou, S., Yang, D., Li, S., Wang, S., Hu, X., Xu, C., Xu, Z., Shu, C., Yuan, Z.: Mquant: Unleashing the inference potential of multimodal large language models via static quantization. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1783–1792 (2025)
43. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2024), https://arxiv.org/abs/2308.02490
44. Zha, Y., Zhou, K., Wu, Y., Wang, Y., Feng, J., Xu, Z., Hao, S., Liu, Z., Xing, E.P., Hu, Z.: Vision-g1: Towards general vision language reasoning with multi-domain data curation (2025), https://arxiv.org/abs/2508.12680
45. Zhang, D., Fu, Y., Yang, R., Miao, Y., Qian, T., Zheng, X., Sun, G., Chhatkuli, A., Huang, X., Jiang, Y.G., et al.: Egonight: Towards egocentric vision understanding at night with a challenging benchmark. arXiv preprint arXiv:2510.06218 (2025)
46. Zheng, X., Dongfang, Z., Jiang, L., Zheng, B., Guo, Y., Zhang, Z., Albanese, G., Yang, R., Ma, M., Zhang, Z., et al.: Multimodal spatial reasoning in the large model era: A survey and benchmarks. arXiv preprint arXiv:2510.25760 (2025)