From Factors to Methods: A Comprehensive Survey on Visual Instruction Tuning Data Selection

Anonymous ACL submission

Abstract

The progress of visual–language models has made visual instruction tuning central to multimodal alignment, yet its effectiveness depends heavily on the composition of training data. Visual instruction datasets are often heterogeneous and redundant, necessitating principled data selection to ensure downstream performance. Despite growing interest, prior studies remain fragmented, relying on disparate evaluation criteria and inconsistent terminology that obscure the underlying design space. To address this, we present the first comprehensive survey of visual instruction data selection, providing a unified perspective on both evaluation and selection mechanisms. We introduce a factor-based analytical framework that organizes core data properties into a structured Data Evaluation Factor Library. Furthermore, we categorize existing methods into feature-based, prediction-based, gradient-based, and hybrid paradigms, analyzing how they operationalize scoring and filtering. By bridging evaluation factors with selection mechanisms, this survey consolidates fragmented insights and outlines future directions for data-efficient multimodal learning.

1 Introduction

Visual-Language Models (VLMs) (Liu et al., 2024a; Bai et al., 2023; Chen et al., 2024c; Li et al., 2025b) have established a central paradigm for multimodal intelligence, driven largely by Visual Instruction Tuning (VIT). This process aligns pretrained models with human intent through large-scale image–instruction–response data. Traditionally, VIT has been treated as a data-scaling problem, wherein performance improvements are pursued primarily by expanding the dataset size. However, recent findings challenge this assumption, supporting the "Less Is More" philosophy (Liu et al., 2025b; Ye et al., 2025; Muennighoff et al., 2025). These studies demonstrate that reasoning ability is highly sensitive to data composition, and that within redundant or heterogeneous datasets (Xu et al., 2025a; Yu et al., 2025), indiscriminate scaling can dilute critical learning signals.

Motivated by this paradigm shift, visual instruction data selection has emerged as an active research direction. Various methods have been proposed to identify high-value samples, leveraging signals such as model predictions (Chen et al., 2024b; Lyu et al., 2025), representation similarity (Bi et al., 2025; Lee et al., 2024), training dynamics (Lee et al., 2025), and gradients (Wu et al., 2025; Liu et al., 2025b). Although these approaches yield substantial gains in data efficiency, the field remains fragmented. Existing studies often adopt ad-hoc evaluation criteria and focus on isolated aspects of data quality, which impedes the generalization of insights and the comparison of effectiveness across different settings.

A core challenge underlying this fragmentation is the lack of a unifying abstraction that separates data evaluation from selection mechanisms. Prior work frequently entangles the criteria used to assess data value with the algorithms employed to exploit them, thereby obscuring the broader design space and limiting systematic analysis. Consequently, it remains unclear how different methods interrelate, what implicit assumptions they share, and where their fundamental limitations lie.

To address these issues, this survey presents the first comprehensive and systematic review of visual instruction data selection. We adopt a factor-based perspective that abstracts the common data properties implicitly considered across prior work, providing a principled lens for analyzing data quality and utility independent of specific algorithms. Furthermore, we organize existing methods into feature-based, prediction-based, gradient-based, and hybrid paradigms according to their selection mechanisms, and analyze how these paradigms operationalize data scoring, filtering, and distribution

balancing. By bridging data evaluation principles with selection strategies, this survey clarifies the design space of visual instruction data selection, consolidates fragmented insights, and highlights open challenges and future research directions toward more data-efficient multimodal learning.

Our main contributions can be summarized as:

- We present the first comprehensive review of visual instruction data selection, synthesizing fragmented literature to formalize the paradigm shift from traditional data scaling to data efficiency.
- We propose a unified factor-based perspective for data evaluation that decouples evaluation criteria from selection mechanisms, providing a principled framework to analyze data quality independent of specific algorithms.
- We establish a structured taxonomy and highlight future frontiers, categorizing methods into feature-based, prediction-based, gradient-based, and hybrid paradigms to clarify the design space and identify key open challenges.

The rest of this paper is structured as follows: §2 reviews preliminaries; §3 introduces our factor library for data evaluation; and §4 presents a taxonomy of selection methods. We then outline a unified evaluation pipeline in §5, identify open challenges in §6, and conclude in §7. The structural organization is visualized in Fig. 1.

Related Survey. Existing surveys (Albalak et al., 2024; Zhang et al., 2025a) predominantly address data selection within general machine learning or large language models, limiting their scope to unimodal text. Despite the rapid rise of Visual-Language Models, a dedicated systematic review of visual instruction data selection remains absent. Furthermore, prior surveys typically organize work solely by methodological categories, often overlooking the underlying mechanisms. We address these gaps by introducing a unified "from factors to methods" framework, providing fine-grained theoretical insights and principled guidance beyond standard taxonomies.

2 Preliminaries

2.1 Visual Instruction Turning

VIT aims to align VLMs with human intent by optimizing the conditional likelihood of target responses given multimodal inputs. Formally, let f_θ denote a VLM with parameters θ and a training dataset $\mathcal{D} = \{(x_i^v, x_i^t, y_i)\}_{i=1}^N$, where x_i^v, x_i^t, and y_i represent the visual input, textual instruction, and target response, respectively. The objective of VIT is to minimize the cross-entropy loss:

$$\mathcal{L}_{\text{train}}(\theta) = -\sum_{i=1}^{N} \log p_\theta(y_i \mid x_i^v, x_i^t) \tag{1}$$

This promotes coherent, instruction-following responses grounded in multimodal contexts. However, persistent challenges in instruction diversity, alignment, and scalability motivate research on principled data selection strategies.

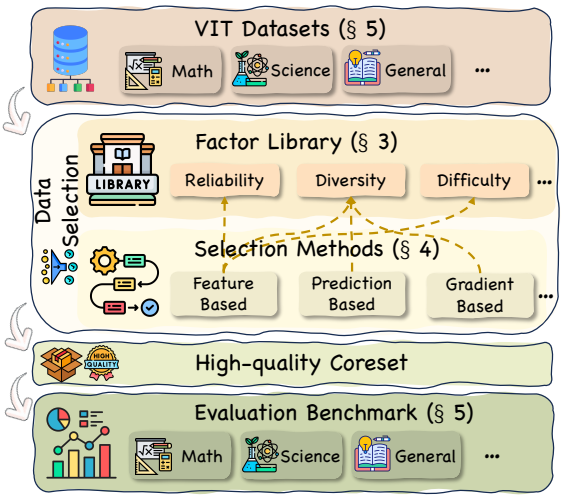

Figure 1: Pipeline for coreset construction and evaluation. Samples are selected from VIT datasets using feature-, prediction-, gradient-based, or hybrid methods built upon the Factor Library, followed by domain-specific benchmarking.

2.2 VIT Data Selection

VIT dataset selection optimizes the trade-off between data quantity and performance by identifying a high-quality subset $\mathcal{S}$ from corpus $\mathcal{D}$ that maximizes generalization within a strict budget. We formulate this as a bi-level optimization problem, where the outer loop maximizes the expected metric $\mathcal{M}$ over target distribution $\mathcal{P}$ subject to the inner loop's training minimization:

$$\begin{aligned} &\max_{\mathcal{S}\subseteq\mathcal{D}} \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\mathcal{M}(f_{\theta^*(\mathcal{S})}(x), y)\right] \\ &\text{s.t. } |\mathcal{S}| \le b,\ \theta^*(\mathcal{S}) = \arg\min_\theta \mathcal{L}_{\text{train}}(\mathcal{S};\theta) \end{aligned} \tag{2}$$

where $b \ll |\mathcal{D}|$ denotes the budget. As a direct solution is computationally intractable, effective strategies approximate the gradient contribution of data points using proxies such as difficulty and diversity. This process targets a Pareto-optimal frontier where the curated subset $\mathcal{S}^*$ significantly reduces training costs while preserving multimodal reasoning capabilities by filtering redundancy and noise.

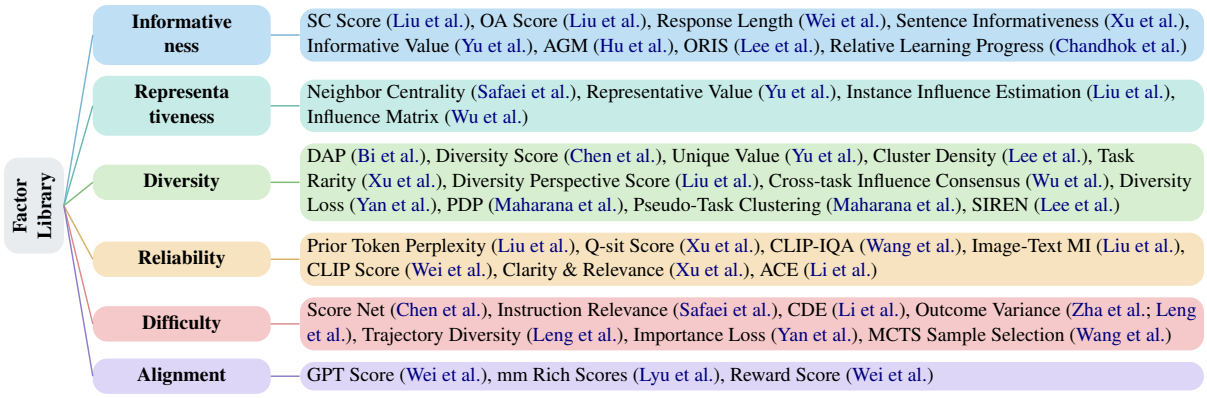

Figure 2: Comprehensive VIT Data Evaluation Factor Library.

3 Data Evaluation Factor Library

In this section, we introduce the Data Evaluation Factor Library, a taxonomy designed to assess the utility of VIT data through decoupled evaluation dimensions. In contrast to prior ad-hoc or task-specific metrics, this framework facilitates the fine-grained analysis and comparison of data selection methods. By standardizing these dimensions, the library establishes a unified foundation for diagnosing and enhancing multimodal data quality.

3.1 Informativeness

Informativeness quantifies the density of effective signals within a sample. It distinguishes data exhibiting rich structural and semantic patterns from trivial or redundant inputs, thereby ensuring that only samples with high intrinsic salience contribute to the learning process.

In the visual domain, informativeness encompasses structural and semantic complexity. Metrics like Segmentation Complexity (SC) and Object Alignment (OA) (Liu et al., 2025a) employ SAM2 (Ravi et al., 2024) and DINO (Liu et al., 2023b) to measure object salience and visual density. The Image Grounding (IG) Score (Maharana et al., 2025) isolates visual utility by comparing textual perplexity with and without image input.

Textual informativeness pertains to signal quality and density. While Response Length (Wei et al., 2023) serves as a heuristic for balancing reasoning depth with conciseness, granular metrics such as Sentence Informativeness and Complexity (Xu et al., 2025b) filter trivial content to preserve syntactic and semantic diversity.

Regarding multimodal integration, approaches such as Informative Value (Yu et al., 2025) quantify joint information density using the entropy of normalized singular values derived from fused representations. Complementing this static analysis with a perturbation-based perspective, Attention-Guided Masking (AGM) (Hu et al., 2025) evaluates effective mutual information by quantifying the loss divergence incurred when high-attention hidden states are suppressed. Finally, in continuous data streams, Online Relative Informativeness Selection (ORIS) (Lee et al., 2025) dynamically prioritizes samples exhibiting high Fisher Information relative to the model's current training state, normalized against historical statistics. Similarly, Relative Learning Progress (RLP) (Chandhok et al., 2025) serves as a dynamic proxy for learnability; it tracks the velocity of accuracy improvement across skill clusters to explicitly target samples within the model's zone of proximal development.

3.2 Representativeness

Representativeness ensures that selected samples accurately reflect the global characteristics of the dataset, thereby preventing the model from learning biased or incomplete patterns.

Recent methods employ diverse strategies to identify samples that best encapsulate the underlying data distribution. One prevalent strategy involves cluster-based analysis. For instance, Neighbor Centrality (Safaei et al., 2025) utilizes cluster centrality to quantify sample representativeness, determined via the mean cosine similarity between a sample's feature vector and those of its k nearest neighbors in the feature space. Similarly, Representative Value (Yu et al., 2025) computes an inter-cluster association coefficient based on the cosine similarity between cluster centroids; a higher coefficient indicates that the sample's cluster is central to the overall distribution, thus mitigating the selection of noisy or outlier data. A distinct, gradient-based approach is adopted by Instance Influence Estimation (Liu et al., 2025b), which lever-

ages influence functions to compute the effect of one sample on others, thereby identifying highly influential samples as representative. Extending this gradient-based paradigm, the Influence Matrix (Wu et al., 2025) redefines representativeness through downstream transfer. It computes gradient inner products with multiple external validation sets and selects samples that consistently exert positive influence across tasks, yielding a subset with broadly reusable capabilities.

3.3 Diversity

Diversity quantifies the semantic and structural variance of data across domains, styles, or concepts. Its primary objective is to ensure broad informational coverage, thereby facilitating the learning of richer distributions and enhancing downstream performance. Existing diversity factors can be broadly classified into two categories: intrinsic data-centric evaluation and consensus-based evaluation.

Intrinsic data-centric factors quantify redundancy based on feature space topology or explicit semantic content. In the context of static embeddings, metrics such as Diversity-Aware Pruning (DAP) (Bi et al., 2025) utilize Pearson correlation coefficients to filter high-redundancy subsets. Similarly, the Diversity Score (Chen et al., 2024b) and Permanent Data Pruning (PDP) (Maharana et al., 2025) employ cosine similarity to penalize candidates that resemble the already-selected set or existing cluster members. To capture intra-cluster distinctiveness, Unique Value (Yu et al., 2025) measures the Euclidean distance between instruction embeddings, prioritizing samples with high uniqueness. Beyond pairwise comparisons, Cluster Density (Lee et al., 2024) introduces a distribution-aware factor, which minimizes Maximum Mean Discrepancy by sampling inversely to density estimated via Gaussian kernels. Complementing these latent metrics, Task Rarity (Xu et al., 2025a) targets explicit semantic balance by prioritizing samples with rare predicted task labels to prevent distribution skew. Extending into the optimization landscape, gradient-based representation factors capture dynamic redundancy. Pseudo-Task Clustering (Maharana et al., 2025) uses gradient vectors to define distinct pseudo-tasks for coverage sampling. Meanwhile, Similarity-aware Information Redundancy Elimination (SIREN) (Lee et al., 2025) introduces an iterative penalty factor, dynamically down-weighting candidates based on their gradient cosine similarity to selected samples to minimize informational overlap without full re-evaluation.

Finally, consensus-based factors derive robustness from multiple evaluative sources. The Diversity Perspective Score (Liu et al., 2025a) aggregates assessments from diverse "visual agents", applying Shapley values (Lundberg and Lee, 2017) to compute a reliability-weighted consensus. In a cross-task context, Cross-task Influence Consensus (Wu et al., 2025) evaluates utility by computing gradient similarities against multiple validation sets, using majority voting to identify samples with consistent positive influence across diverse tasks.

3.4 Reliability

Reliability characterizes the stability and trustworthiness of multimodal data, defined by two dimensions: intrinsic quality and cross-modal alignment.

Intrinsic factors evaluate the purity of individual modalities. In the textual domain, the Prior Token Perplexity Score (Liu et al., 2025a) assesses fluency and coherence by measuring the probability of preceding tokens. For visual data, the Q-sit Score (Xu et al., 2025a) identifies low-level distortions (e.g., blur, noise) using a fine-tuned VLM, while the CLIP-IQA Score (Wang et al., 2022) evaluates image embeddings against positive and negative prompt anchors to derive a quality metric.

Cross-modal factors ensure semantic consistency between visual and textual signals. The Image-Text Mutual Information Score (Liu et al., 2025a) quantifies this interdependence through entropy analysis, whereas the CLIP Score (Wei et al., 2023) computes the direct cosine similarity between image and textual feature vectors. Alternatively, Clarity and Relevance (Xu et al., 2025a) employs a generative approach, validating alignment by measuring the comprehensibility and relevance of VLM-generated textual descriptions relative to the source image. Finally, the Attention Confidence Estimator (ACE) (Li et al., 2025a) targets internal reasoning stability, analyzing self-attention patterns to reject samples driven by irrelevant tokens rather than semantic context.

3.5 Difficulty

Difficulty quantifies the cognitive load imposed by a data sample, reflecting the complexity required to bridge the modality gap relative to the model's current capabilities. Estimation methods generally fall into three categories: auxiliary estimation, optimization dynamics, and stochastic sampling.

Auxiliary estimation relies on external scoring mechanisms. ScoreNet (Chen et al., 2024b) employs a dedicated network trained alongside the VLM to predict instruction failure, prioritizing samples that drive gradient updates.

Optimization dynamics estimate difficulty based on internal loss signals. The Instruction Relevance Score (Safaei et al., 2025) measures the informational value of an instruction by computing the loss difference with and without it; a larger gap indicates a more challenging sample that requires strong prompt adherence. The Causal Discrepancy Estimator (CDE) (Li et al., 2025a) formalizes this using the Potential Outcome Model (Rubin, 2005), isolating "cognitive samples" by quantifying the causal gain in accuracy provided by visual information over text-only counterfactual.

Stochastic sampling evaluates difficulty dynamically through uncertainty estimation. The Outcome Variance Score (Zha et al., 2025; Leng et al., 2025) utilizes multiple rollouts to estimate uncertainty, targeting samples with high entropy that are neither trivial nor excessively difficult. The Trajectory Diversity Score (Leng et al., 2025) prioritizes questions with diverse reasoning paths to ensure that the data promote robust problem-solving rather than rigid pattern matching. Complementing these, Monte Carlo Tree Search (MCTS)-based sample selection (Wang et al., 2025b) repurposes tree search to quantify difficulty via the iterations required to reach a solution, efficiently isolating samples at the model's reasoning frontier.

3.6 Alignment

Alignment measures the extent to which multimodal data accurately reflects human intent and preferences. Unlike objective reliability or statistical informativeness, alignment captures the overall quality and adherence to instructions through learned proxies of human judgment.

Key factors include LLM-as-a-Judge frameworks (Wei et al., 2023; Lyu et al., 2025) and Reward Model-based evaluations (Wei et al., 2023). For instance, the GPT Score (Wei et al., 2023) employs VLMs such as GPT-4 to assess response quality from a comprehensive perspective. To capture finer granularity, Multimodal Rich Scorers and Styler (Lyu et al., 2025) utilize GPT-4o to evaluate fourteen perception-based dimensions and interaction styles. Additionally, the Reward Score (Wei et al., 2023) quantifies the degree to which data align with human preferences, providing a model-based proxy for user satisfaction.

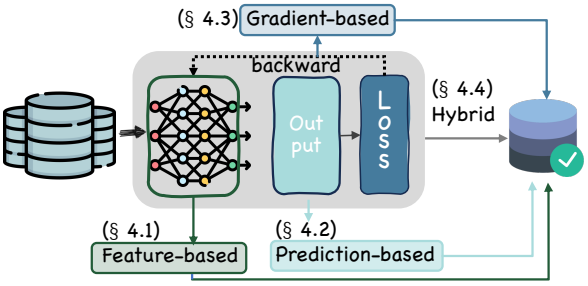

Figure 3: VIT data selection methods classified by intervention points: Feature-based (§ 4.1), Prediction-based (§ 4.2), Gradient-based (§ 4.3), and Hybrid (§ 4.4).

3.7 Discussion: Interactions and Implications

While we define these factors in isolation, practical data selection necessitates their joint optimization and the management of their inevitable interactions. A primary tension exists between *Informativeness* and *Diversity*: maximizing information density often concentrates selection within high-complexity clusters, thereby inducing redundancy, whereas prioritizing diversity risks allocating the budget to distinct yet low-utility outliers. Similarly, *Representativeness* inherently favors high-frequency modes, directly opposing *Difficulty*-based criteria that target the long tail—a conflict that can skew the data distribution if left unbalanced. Furthermore, *Reliability* serves as a critical gating constraint, as label noise frequently manifests as artificial difficulty (e.g., high loss), while *Alignment* proxies instruction adherence but guarantees neither factual correctness nor cross-modal consistency. Given the complexity of these interactions, effective synthesis is essential. We next review representative selection methods, detailing how they integrate these factors through specific designs.

4 Data Selection Methods

In this section, we systematically categorize existing VIT data selection methods, grounded in the aforementioned factor library, into four primary paradigms, distinguished by the stage at which selection is performed (Fig. 3). We provide an in-depth analysis of the characteristics, advantages, and limitations of each paradigm. We summarize these methods in Tab. 1, with further details provided in Tab. 2.

4.1 Feature-based Methods

Feature-based methods quantify sample utility by applying analytical heuristics to learned intrinsic data representations, such as hidden states or em-

beddings. These approaches are generally categorized into two distinct paradigms: sample-level scoring, which evaluates instances in isolation, and dataset-level analysis, which leverages the global topology of the feature space.

Sample-level methods derive quality metrics for individual instances by analyzing their intrinsic feature properties. For instance, DataTailor (Yu et al., 2025) aggregates multiple dimensions of data quality, measuring *Diversity* via Euclidean distances and estimating *Informativeness* through Singular Value Decomposition. Complementing this geometric perspective, PRISM (Bi et al., 2025) addresses the fundamental pathology of representation anisotropy in pre-trained features. It employs an implicit re-centering strategy to rectify global semantic drift, utilizing Diversity-Aware Pruning to filter out high-redundancy samples.

Dataset-level methods leverage global structure, typically through clustering. COINCIDE (Lee et al., 2024) extracts features using a lightweight VLM and clusters them to analyze transferability and density, jointly optimizing for *Diversity* and *Representativeness*.

Discussion. The primary advantage of feature-based methods lies in their computational efficiency. By operating on pre-computed representations, often without the need for auxiliary training, they provide a scalable evaluation framework. However, their reliance on hand-crafted heuristics poses a significant limitation; designing effective scoring functions demands substantial domain expertise, and heuristics optimized for specific distributions often generalize poorly to novel tasks.

4.2 Prediction-based Methods

In contrast to Feature-based methods, which analyze internal representations, Prediction-based methods leverage the outputs of auxiliary or target models to assess sample utility. This paradigm replaces manual heuristic design with learned evaluation functions, shifting the dependency from domain expertise to proxy model capability. Broadly, these approaches employ proxy models in two primary capacities: as direct zero-shot judges or as supervisors for training bespoke scoring networks.

The first strategy utilizes large-scale, pre-trained models as direct evaluators. SCALE (Xu et al., 2025a) employs specialized models such as Q-sit (Zhang et al., 2025b) and Qwen (Yang et al., 2025) to independently assess image and text components for *Informativeness* and *Reliability*, while gauging *Diversity* by comparing generated descriptions against the ground truth. Similarly, SoTA with Less (Wang et al., 2025b) integrates Qwen-VL (Bai et al., 2025) within a Monte Carlo Tree Search (Browne et al., 2012) framework. It identifies samples that are either structurally complex (requiring >5 reasoning steps) or inherently challenging (remaining unsolved after 50 steps), thereby effectively selecting for *Difficulty*.

The second strategy trains dedicated scoring modules, supervised either by empirical performance feedback or stronger teacher models. InstructionGPT-4 (Wei et al., 2023) pioneered this direction by introducing a learnable data selector. Rather than relying on static heuristics, it trains a self-attention network to predict "genuine quality labels" derived from the validation performance of selected data subsets, thereby automating the discovery of high-value samples. Building upon this concept of learned quality mapping, mmSSR (Lyu et al., 2025) bootstraps the process by employing GPT-4o to annotate a subset across 14 multimodal capabilities; this data fine-tunes the target VLM into a specialized reward model capable of scoring the entire dataset. Similarly, Self-Filter (Chen et al., 2024b) trains a scoring network to predict the *Difficulty* of instructions.

Discussion. While prediction-based methods reduce the reliance on manual feature engineering, they introduce distinct challenges. First, selection efficacy is bounded by the capabilities of the proxy model; biases or reasoning flaws in the judge can propagate into the selected subset, a phenomenon known as proxy model bias. Second, methods necessitating bespoke scoring networks incur substantial computational overhead, including training and inference costs that may prove prohibitive for large-scale datasets compared to lightweight feature-based heuristics.

4.3 Gradient-based Methods

Gradient-based methods estimate sample utility by analyzing backpropagation derivatives to capture the dynamic interaction between data and the model. Unlike static features, these approaches leverage gradient properties like magnitude, direction, and inner products to quantify the influence of samples on optimization and generalization.

Recent approaches often employ hierarchical strategies. TIVE (Liu et al., 2025b) introduces a

Table 1: Summary of representative data selection methods. Tr-Free: Training-free status (✓: Yes, ✗: No). Cost: ○/◐/● denote Low/Medium/High selection costs. Key Factors: Selection Factors. Eval. Dim.: Evaluation dimensions (3-letter abbr.).

Cat.	Method	Venue	Tr-Free	Cost	Key Factors	Eval. Dim.
Feature	DataTailor (2025)	ICCV	✓	○	Informative, Representative, Unique Value	*Inf*, *Rep*, *Div*
	PRISM (2025)	ArXiv	✓	○	Diversity-Aware Pruning	*Div*
	COINCIDE (2024)	EMNLP	✓	○	Cluster Density	*Ref*, *Div*
Prediction	InstructionGPT-4 (2023)	ArXiv	✗	◐	Length, CLIP, GPT-4, Reward Score	*Inf*, *Rel*, *Ali*
	Self-Filter (2024b)	ACL	✗	●	ScoreNet, Diversity Score	*Dif*, *Div*
	mmSSR (2025)	ArXiv	✓	◐	mm Rich Scorers	*Ali*
	SCALE (2025b)	EMNLP	✓	◐	Sentence Informativeness, Task Rarity, Clarity & Relevance	*Inf*, *Div*, *Rel*
	SOTA with Less (2025b)	NeurIPS	✓	●	MTCS Sample Selection	*Dif*
Gradient	TIVE (2025b)	MM	✗	●	Instance Influence Estimation	*Rep*
	ICONS (2025)	ArXiv	✗	●	Influence Matrix, Cross-task Influence Consensus	*Rep*, *Div*
	OASIS (2025)	ArXiv	✗	◐	ORIS, SIREN	*Inf*, *Div*
Hybrid	PreSel (2025)	CVPR	✓	○	Neighbor Centrality, Instruction Relevance	*Inf*, *Dif*
	ViSA (2025a)	ArXiv	✓	◐	Image-Text MI, Prior Token PPL, Diversity Perspective, SA, OA Score	*Inf*, *Div*, *Rel*
	Δ-AttnMask (2025)	ArXiv	✓	◐	Attn-Guided Masking	*Inf*
	RAP (2025a)	ArXiv	✓	◐	Causal Discrepancy Estimator, Attention Confidence Estimator	*Dif*, *Rel*
	PROGRESS (2025)	ArXiv	✗	◐	Relative Learning Progress	*Inf*, *Div*
	CoIDO (2025)	NeurIPS	✗	◐	Diversity, Importance Loss	*Div*, *Dif*

two-level framework. At the task level, it quantifies *Difficulty* via average gradient magnitudes to adjust data ratios. At the sample level, it ensures *Representativeness* by computing gradient inner products between intra- and inter-task samples to calibrate sampling probabilities. Similarly, ICONS (Wu et al., 2025) utilizes a two-stage consensus mechanism to ensure robustness. In the specialist stage, it derives per-task influence scores, which are then aggregated in the generalist stage via an Influence Consensus algorithm. Final selection relies on majority voting to retain samples that exhibit consistent positive influence across diverse tasks.

Addressing the computational bottlenecks of continuous data streams, OASIS (Lee et al., 2025) approximates sample utility via last-layer gradients. It combines Online Relative Informativeness Selection (ORIS), which evaluates importance against global historical statistics, with Similarity-aware Information Redundancy Elimination (SIREN), an iterative mechanism that minimizes informational overlap based on gradient similarity.

Discussion. Gradient-based methods offer superior fidelity by directly quantifying the model's learning dynamics, thereby eliminating the reliance on proxy models. However, this precision typically entails significant computational overhead for large-scale models. While efficient approximations like OASIS alleviate this burden by restricting computation to the last layer, standard full-gradient approaches remain resource-intensive compared to feature-based alternatives. Furthermore, the sensitivity of gradients to the specific loss landscape risks selecting subsets that are overly specialized to target tasks, potentially compromising transferability.

4.4 Hybrid Methods

Hybrid methods integrate signals from feature-based, prediction-based, and gradient-based paradigms to identify high-value samples. This synergy is particularly vital in complex multimodal settings, where accurately assessing sample quality necessitates capturing both intrinsic data structure and rigorous model-specific dynamics.

Early approaches often employ robust heuristic aggregation. PreSel (Safaei et al., 2025) combines prediction-based relevance, derived by contrasting model outputs with and without visual inputs, with DINOv2 (Liu et al., 2023b) feature clustering to select central samples. ViSA (Liu et al., 2025a) adopts a broader strategy, aggregating diverse indicators such as SAM2-based (Ravi et al., 2024) segmentation complexity, TF-IDF features, and token perplexity. Moving away from such external dependencies, Δ-AttnMask (Hu et al., 2025) proposes a self-contained hybrid framework. It integrates attention-based feature analysis with prediction dynamics by measuring the loss discrepancy between original states and those masked at high-attention regions, thereby directly quantifying visual-textual alignment without reliance on auxiliary models. Similarly, RAP (Li et al., 2025a) targets complex reasoning capabilities by combining causal inference with internal feature analysis; it employs a Causal Discrepancy Estimator to filter language-prior shortcuts via counterfactual predictions, coupled with an

Attention Confidence Estimator that validates the semantic reliability of the reasoning process.

However, these methods typically exhibit a loose coupling between importance and diversity and entail substantial computational costs due to the necessity of exhaustive dataset processing. To address these inefficiencies, CoIDO (Yan et al., 2025) reformulates data selection as a unified optimization problem. By leveraging training loss to proxy difficulty and spectral clustering on multimodal features to ensure diversity, it introduces a cohesive framework governed by a lightweight scorer trained on a minimal subset. This facilitates efficient joint optimization without requiring exhaustive dataset traversal. Similarly targeting efficiency, PROGRESS (Chandhok et al., 2025) circumvents the need for auxiliary scoring networks or heavy gradient computations. It employs unsupervised multimodal clustering to partition skills and dynamically prioritizes samples based on relative learning progress , thereby constructing a self-paced curriculum that targets the model's zone of proximal development.

Extending these principles to the domain of lifelong learning, Adapt-∞ (Maharana et al., 2025) introduces a strategy tailored for evolving data streams. It integrates gradient-based clustering for skill identification with dynamic prediction-based scoring, thereby enabling robust adaptation to continuous distributional shifts.

Discussion. Hybrid methods effectively mitigate the limitations of individual paradigms by cross-referencing semantic features with model-derived feedback. Nevertheless, the complexity inherent in aggregating disparate signals often results in significant computational overhead. While CoIDO alleviates this burden through unified optimization, the transition to continuous data streams presents distinct challenges. Approaches such as Adapt-∞ represent a critical advancement toward scalable, autonomous tuning; by applying hybrid principles to real-time environments, they dynamically manage redundancy and skill acquisition without reliance on static reference datasets.

5 Evaluation

To evaluate the efficacy of data selection strategies, prevalent methods (Yan et al., 2025; Bi et al., 2025; Yu et al., 2025) typically adopt large-scale instruction tuning datasets, such as LLaVA-1.5 (Liu et al., 2024a), SVIT-Mix (Zhao et al., 2023), Cambrian-7M (Tong et al., 2024), and Vision-Flan (Xu et al., 2024), as the candidate pool. The standard evaluation protocol involves curating a 10–20% subset of these datasets to fine-tune a model, aiming to achieve performance comparable to full-dataset training on downstream benchmarks. These benchmarks typically include Visual Question Answering (e.g., VQAv2, GQA, TextVQA), hallucination evaluation (e.g., POPE), and comprehensive multimodal assessments (e.g., MMBench, MME). More details of visual instruction datasets and benchmarks can be found in our Appendix.

6 Future Prospects

Existing data selection relies heavily on static, isolated heuristics, often neglecting factor dependencies and the non-stationarity of training. Future research should develop unified frameworks that explicitly model these interactions while dynamically adapting strategies as model capabilities evolve. This adaptability is particularly critical for the Reinforcement Learning stage, where metrics effective for SFT often degrade or fail during preference optimization. Furthermore, selection paradigms must expand beyond image–text data to encompass pure vision, video, and audio tasks. Ultimately, establishing stage-aware criteria backed by theoretical guarantees will be essential for robust and scalable multimodal alignment.

7 Conclusion

In this survey, we provided a comprehensive review of data selection for Visual Instruction Tuning, underscoring the paradigm shift from data quantity to data quality. To structure this fragmented field, we introduced a unified "Factors to Methods" framework. First, we established the Data Evaluation Factor Library, delineating six dimensions of data utility: Informativeness, Representativeness, Diversity, Reliability, Difficulty, and Alignment. Second, we systematically categorized selection strategies into Feature-based, Prediction-based, Gradient-based, and Hybrid paradigms, analyzing how each operationalizes these factors for scoring and filtering. By bridging theoretical evaluation criteria with practical selection mechanisms, this survey clarifies the design space of VIT data selection and establishes a foundation for future research into more robust, unified, and theoretically grounded optimization frameworks.

8 Limitations

While this survey provides a comprehensive taxonomy and conceptual framework for visual instruction data selection, it acknowledges several limitations. First, we do not re-evaluate the reviewed methods under a unified experimental protocol. Given that prior studies employ disparate backbones, data mixtures, training budgets, and evaluation suites, their reported results are not directly comparable, thereby precluding head-to-head comparisons. Second, our scope is confined to visual instruction tuning; we do not extend our analysis to the broader data curation and selection strategies for multimodal pre-training, which constitutes a complementary yet distinct avenue of research.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Preprint*, arXiv:2402.16827.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jinhe Bi, Yifan Wang, Danqi Yan, Aniri, Wenke Huang, Zengjie Jin, Xiaowen Ma, Artur Hecker, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *Preprint*, arXiv:2502.12119.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.

Shivam Chandhok, Qian Yang, Oscar Manas, Kanishk Jain, Leonid Sigal, and Aishwarya Agrawal. 2025. Learning what matters: Prioritized concept learning via relative error-driven sample selection. *arXiv preprint arXiv:2506.01085*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024b. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *Preprint*, arXiv:2402.12501.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

Jucheng Hu, Suorong Yang, and Dongzhan Zhou. 2025. δ-attnmask: Attention-guided masked hidden states for efficient data selection and augmentation. *Preprint*, arXiv:2508.09199.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. *Preprint*, arXiv:2406.10995.

Minjae Lee, Minhyuk Seo, Tingyu Qu, Tinne Tuytelaars, and Jonghyun Choi. 2025. Oasis: Online sample selection for continual visual instruction tuning. *arXiv preprint arXiv:2506.02011*.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, and 1 others. 2025. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*.

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

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. 2019. Visual question answering as reading comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shenshen Li, Kaiyuan Deng, Lei Wang, Hao Yang, Chong Peng, Peng Yan, Fumin Shen, Heng Tao Shen, and Xing Xu. 2025a. Truth in the few: High-value data selection for efficient multi-modal reasoning. *Preprint*, arXiv:2506.04755.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025b. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Zhenyu Liu, Yunxin Li, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. 2025a. Picking the cream of the crop: Visual-centric data selection with collaborative agents. *Preprint*, arXiv:2502.19917.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2025b. Less is more: High-value data selection for visual instruction tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 3712–3721, New York, NY, USA. Association for Computing Machinery.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Scott Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. In *Nips*.

Mengyao Lyu, Yan Li, Huasong Zhong, Wenhao Yang, Hui Chen, Jungong Han, Guiguang Ding, and Zhenheng Yang. 2025. Cream of the crop: Harvesting rich, scalable and transferable multi-modal data for instruction fine-tuning. *Preprint*, arXiv:2503.13383.

Adyasha Maharana, Jaehong Yoon, Tianlong Chen, and Mohit Bansal. 2025. Adapt-∞: Scalable continual multimodal instruction tuning via dynamic data selection. In *The Thirteenth International Conference on Learning Representations*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

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

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. 2025. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. *Preprint*, arXiv:2503.07591.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. 2022. Exploring clip for assessing the look and feel of images. *Preprint*, arXiv:2207.12396.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025a. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. *Preprint*, arXiv:2505.10557.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. SoTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

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

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2025. Icons: Influence consensus for vision-language data selection. *Preprint*, arXiv:2501.00654.

Mingjie Xu, Andrew Estornell, Hongzheng Yang, Yuzhi Zhao, Zhaowei Zhu, Qi Xuan, and Jiaheng Wei. 2025a. Better reasoning with less data: Enhancing vlms through unified modality scoring. *Preprint*, arXiv:2506.08429.

Mingjie Xu, Andrew Estornell, Hongzheng Yang, Yuzhi Zhao, Zhaowei Zhu, Qi Xuan, and Jiaheng Wei. 2025b. Better reasoning with less data: Enhancing vlms through unified modality scoring. *Preprint*, arXiv:2506.08429.

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

Yichen Yan, Ming Zhong, Qi Zhu, Xiaoling Gu, Jinpeng Chen, and Huan Li. 2025. Coido: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization. *arXiv preprint arXiv:2510.17847*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2025. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. *Preprint*, arXiv:2412.06293.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2025. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *Preprint*, arXiv:2508.12680.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. 2025a. A survey on data selection for llm instruction tuning. *Journal of Artificial Intelligence Research*, 83.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Zicheng Zhang, Haoning Wu, Ziheng Jia, Weisi Lin, and Guangtao Zhai. 2025b. Teaching lmms for image quality scoring and interpreting. *Preprint*, arXiv:2503.09197.

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

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models.

A VIT Selection Methods

Table 2 presents a systematic comparison of representative data selection methodologies. We categorize existing works into four paradigms: Feature-based, Prediction-based, Gradient-based, and Hybrid. To quantify computational overhead and assist researchers in balancing selection efficacy against training costs, the Cost column categorizes efficiency relative to TIVE. Given that TIVE represents a typical high-cost baseline in current literature, we use it to establish a relative scale. We reference efficiency metrics reported in original papers or provide estimates based on our empirical experience. Methods are classified as Low ($0 \sim 50\%$ of TIVE, suitable for rapid iteration), Medium ($50 \sim 75\%$), or High (others).

Several key trends emerge from this comparative analysis:

Overall landscape. Table 2 summarizes 17 representative visual instruction data selection methods spanning four paradigms. Feature-based methods (3/17) rely on embedding-space structure and are uniformly training-free with low cost. Prediction-based methods (5/17) introduce learned or model-driven scorers, yielding a wider cost range and mixed reliance on auxiliary training. Gradient-based methods (3/17) consistently require training and incur high computational overhead due to influence or gradient estimation. Hybrid methods (6/17) are the most diverse in design, typically combining lightweight heuristics with model-dependent signals to balance efficiency and selectivity.

Training-free and cost trade-offs. A majority of methods are training-free (10/17), concentrated in feature-based and hybrid categories, suggesting a strong community preference for selectors that can be plugged into existing pipelines without additional optimization loops. Cost is dominated by medium-compute approaches (9/17), while low-cost methods remain fewer (4/17) and high-cost methods (4/17) are primarily those that require explicit scorer training or gradient-based influence estimation. Empirically, training-free does not imply weaker performance. Several training-free selectors match or exceed the full-data baseline at aggressive retention ratios, such as SOTA with Less (101.7% at 10%), DataTailor (101.3% at 15%), and hybrid approaches including Δ-AttnMask (104.4% at 20%) and RAP (103.2% at 9.3%).

Subset efficiency and relative performance. The retention ratios in parentheses reveal a broad operating regime from 5% to 37.6%. The strongest efficiency-per-sample signals come from hybrid selectors: RAP improves performance while keeping less than one-tenth of the data (9.3%), and Δ-ATTNMASK delivers the highest reported relative gain (104.4%) at 20%. In contrast, several methods underperform despite moderate retention, including SELF-FILTER (90.9% at 20%) and OASIS (95.6% at 25%). This pattern suggests that selection quality is highly sensitive to the fidelity of the scoring proxy and its robustness across heterogeneous instruction distributions, rather than being determined by retention ratio alone.

Factor coverage by paradigm. Across methods, *Diversity* and *Informativeness* are the most frequently operationalized dimensions, appearing in nearly all paradigms, while *Alignment* appears mainly in prediction-based approaches that use LLM-as-a-judge style signals (e.g., InstructionGPT-4, mmSSR). *Difficulty* is often targeted by selectors that explicitly pursue long-tail training benefits (Self-Filter, SOTA with Less, RAP, CoIDO), whereas *Representativeness* is emphasized by feature clustering or influence-based estimators (DataTailor, TIVE, ICONS). Overall, the table indicates a shift from single-factor heuristics (e.g., diversity-aware pruning) toward multi-signal designs that combine complementary proxies, especially in hybrid methods.

Evaluation concentration and generalization concerns. Most studies benchmark on a shared core suite (frequent use of MME, MMMU, POPE, MMBench, MMVet, SEEDBench), enabling partial comparability but also concentrating conclusions around a narrow evaluation distribution. Moreover, the dominant target model is LLaVA-1.5-7B, with fewer results on newer families such as Qwen2.5/Qwen3-VL. This skew suggests that cross-backbone transferability remains underexplored, and reported gains may partially reflect selector-model compatibility rather than universally improved data utility.

Takeaways. Feature-based methods offer the best simplicity-cost profile but can be sensitive to embedding quality and may miss task-specific utility. Gradient-based methods provide principled attribution signals yet are expensive and do not consistently outperform cheaper alternatives. Hybrid approaches currently appear most promising, achieving strong relative improvements under tight budgets by mixing lightweight structural cues with model-aware signals. Future benchmarks would benefit from unified protocols that (i) control for retention ratio and training compute, (ii) evaluate across multiple backbones and scales, and (iii) report stability across diverse benchmark clusters (reasoning, OCR, hallucination, and domain-specific tasks).

B VIT Datasets

The datasets listed in Table 3 constitute the primary candidate pools for modern visual instruction tuning. As the volume of available data explodes, ranging from the curated LLaVA-158k to the massive PlotQA (28.9M), identifying the most effective samples from these pools has become a central research challenge.

We categorize these candidate pools by domain to highlight the diverse sources available for data mixture construction. This structured taxonomy benefits researchers by providing a clear roadmap for data curation: it enables the systematic selection of domain-specific sources (e.g., prioritizing Math or OCR capabilities) to construct balanced training mixtures. Furthermore, by contrasting data scales across domains, this overview aids in determining where aggressive filtering strategies are most needed, thereby transforming raw, uneven candidate pools into efficient, high-performing instruction tuning datasets. For a more comprehensive collection of open datasets, we refer readers to FineVision [1].

C Evaluation Benchmarks

In the landscape of visual instruction tuning, the efficacy of a data selection algorithm is typically validated across a hierarchy of benchmarks, as outlined in Table 4.

Current research largely bifurcates evaluation into two streams. The first focuses on specialized capability enhancement, where researchers select math or document-centric data to drive performance on specific leaderboards like MathVista (Math & Science) or DocVQA (Infographic). The second stream addresses generalist data efficiency, aiming to reduce training costs while maintaining

[1] https://huggingface.co/datasets/HuggingFaceM4/FineVision

competitive performance on comprehensive benchmarks like MMMU and MMBench.

Crucially, recent studies also incorporate safety-oriented benchmarks like POPE and HallusionBench into the selection loop. This trend reflects a growing recognition that high-quality data selection must not only maximize positive knowledge transfer but also actively filter out samples that contribute to visual hallucinations.

Table 2: Summary of representative data selection methods. Tr-Free indicates training-free status (✓: Yes, ✗: No). Cost represents selection expense (○: Low, ◐: Medium, ●: High). Key Factors lists the primary selection criteria, while Eval. Dim. refers to evaluation dimensions (abbreviated). Full Data identifies the source dataset used for filtering. Target Model specifies the model trained on the coreset. Rel. Perf. (Ratio) shows the performance relative to full-data training at the indicated data retention ratio.

Method	Venue	Tr-Free	Cost	Key Factors	Eval. Dim.	Full Data	Benchmarks	Target Model	Rel. Perf. (Ratio)
Feature-based Methods									
DataTailor (2025)	ICCV	✓	○	Informative, Representative, Unique Value	*Inf*, *Rep*, *Div*	LLaVA-665K	MME, POPE, MMMU, MM-Vet, LLaVA-Wild, SEED, VizWiz, SciQA, GQA, VQA, TextVQA	LLaVA-1.5-7B	101.3%(15%)
PRISM (2025)	ArXiv	✓	○	Diversity-Aware Pruning	*Div*	LLaVA-665K	SQA, VizWiz, MMVet, POPE, MME, MMMU	LLaVA-1.5-7B	101.7% (37.6%)
COINCIDE (2024)	EMNLP	✓	○	Cluster Density	*Ref*, *Div*	LLaVA-665K	VQA, GQA, VizWiz, SQA, TextQA, POPE, MME, MMBench, LLaVABench	LLaVA-1.5-7B	97.4% (20%)
Prediction-based Methods									
InstructionGPT-4 (2023)	ArXiv	✗	◐	Length, CLIP, GPT-4, Reward Score	*Inf*, *Rel*, *Ali*	MiniGPT4-Instruction	GQA, IconQA, ScienceQA, OKVQA, DocVQA, TextVQA, STVQA, VizWiz, MMBench, MME, LLaVA-Bench	MiniGPT-4-7B	93.3%(5.8%)
Self-Filter (2024b)	ACL	✗	●	ScoreNet, Diversity Score	*Dif*, *Div*	LLaVA Instruction 158K	MMBench, MME, SEEDBench, HallusionBench, MathVista, ScienceQA, OK-VQA, TextVQA, VisDial, VCR, MSCOCO, Pope	LLaVA-1.0-7B	90.9% (20%)
mmSSR (2025)	ArXiv	✓	◐	mm Rich Scorers	*Ali*	LLaVA-OV	MMStar, MMMU, MMVet, BLINK, MMT-Bench, MME, AI2D, ScienceQA, MathVista(MINI)	LLaVA-OneVision-7B	99.11% (30%)
SCALE (2025b)	EMNLP	✓	◐	Sentence Informativeness, Task Rarity	*Inf*, *Div*, *Rel*	LLaVA-665K, ShareGPT-4V, Geometry-3K, ChartQA	A-OKVQA, CRPE, Exist, CRPE, Relation, LLaVA, Wild, MMBench EN, MME, ScienceQA, SeedBench	Qwen2.5-VL-7B	100.3% (10%)
SOTA with Less (2025b)	NeurIPS	✓	●	MTCS Sample Selection	*Dif*	Open-Source 70k	MME, MMMU, MathVista	Qwen2.5-VL-7B	101.7% (10%)
Gradient-based Methods									
TIVE (2025b)	MM	✗	●	Instance Influence Estimation	*Rep*	MC-VQA, OE-VQA, REC, VC, Caption, TC	MME, SEED, MMB, SQA, SQA-I, POPE	LLaVA-1.5-7B	101.2%(15%)
ICONS (2025)	ArXiv	✗	●	Influence Matrix	*Rep*, *Div*	LLaVA-665K	MME, SQA-I, POPE, VQAv2, LLaVA-Bench, TextVQA, MMBench, GQA, VizWiz	LLaVA-1.5-7B	98.6%(20%)
OASIS (2025)	ArXiv	✗	◐	ORIS, SIREN	*Inf*, *Div*	LLaVA-665K	Long Sequence, TRACE, COAST	LLaVA-1.5-7B	95.6%(25%)
Hybrid Methods									
PreSel (2025)	CVPR	✓	○	Neighbor Centrality, Instruction Relevance	*Inf*, *Dif*	LLaVA-665K	VQAv2, SQA, TextVQA, MME, MMBench, SEED-Bench, MM-Vet, POPE	LLaVA-1.5-7B	97.9%(15%)
ViSA (2025a)	ArXiv	✓	◐	Image-Text MI, Prior Token PPL	*Inf*, *Div*, *Rel*	LLaVA-OneVision	VQAv2, OKVQA, TextVQA, MMBench, MME-RealWorld, SEED-Bench, MMMU	Qwen2-VL-7B	99.3%(5%)
Δ-AttnMask (2025)	ArXiv	✓	◐	Attn-Guided Masking	*Inf*	LLaVA Instruction 158K	HallusionBench, MMBench, MME, POPE, ScienceQA, SEEDBench	Qwen2-VL-2B	104.4%(20%)
RAP (2025a)	ArXiv	✓	◐	Causal Discrepancy Estimator	*Dif*, *Rel*	MM-Eureka	MathVista, MMStar, MathVerse, Wemath, MMVet, LogicVista	Qwen2.5-VL-7B	103.2%(9.3%)
PROGRESS (2025)	ArXiv	✗	◐	Relative Learning Progress	*Inf*, *Div*	LLaVA-665K	MME, MMBench	LLaVA-v1.5-7B	98.8%(20%)
CoIDO (2025)	NeurIPS	✗	◐	Diversity, Importance Loss	*Div*, *Dif*	LLaVA-665K	VQAv2, GQA, VizWiz, SQA, TextQA, POPE, MME, MMBench, LLaVABench	LLaVA-1.5-7B	98.2%(20%)

Table 3: Overview of public visual instruction / VQA datasets classified by domain.

Dataset	Year	Modality	#Samples
Math & Science Datasets			
MATH-V / MATH-Vision	2024	Visual Math Problems	3K
MathVista	2024	Multi (Geo, Func, Puzzles)	6K
Geometry3K	2021	Geometry Diagrams	3K
CASIA-PGPS9K	2023	Geometry Diagrams	9K
MathV360K	2024	Multi (Geo, Plots, Word Prob)	360K
MM-MathInstruct	2024	Mixed (Geo, Charts, Synthetic)	3M
GeoQA+	2022	Geometry Diagrams	6,027
Geo170K	2025	Geometry Diagrams	177K
GeomVerse	2024	Synthetic Geometric Patterns	9K
MMK12	2025	K-12 Subject Images	17K
We-Math2.0-SFT	2025	Visual Math Problems	1,516
We-Math2.0-Standard	2025	Visual Math Problems	5,843
We-Math2.0-Pro	2025	Visual Math Problems	4,552
ScienceQA	2022	Multi (Natural, Charts, Algo)	21K
AI2D / AI2D-RST	2016	Science Diagrams	4.9K
PaperQA	2024	Scientific Paper Figures	0.1K
CoSyn-400K	2025	Synthetic Science/Reasoning	400K
ViRL39K	2025	Abstract/Reasoning Images	39K
Infograph / Data Visualization Datasets			
ChartQA	2022	Charts (Real & Synthetic)	2.5K
PlotQA	2019	Scientific Plots	28.9M
DVQA	2018	Synthetic Bar Charts	300K
FigureQA	2017	Synthetic Plots	100K
DocVQA	2021	Document Images (Scanned)	16K
MMTab	2024	Tables	82K
TabMWP	2023	Tables + Text	38K
PixmoDoc	2024	Document Screenshots	255K
ECD-10k-Images	2025	Charts / Plots	10K
MMC-Instruction	2024	Charts + Text	662K
General / Broad Visual Instruction Datasets			
TextVQA	2019	Natural Images (Scene Text)	45K
VQA v2	2017	Natural Images	265K
VizWiz	2018	User-Taken Images (Blind)	31K
VisualGenome	2017	Natural Images (Dense Cap)	1.7M
ChineseMeme	2024	Internet Memes	39K
MMEvol	2024	Mixed (Natural, Math, Web)	163K
LNQA	2024	Natural Images (Long Context)	302K
VizWiz-VQA	2024	Natural Images	12K
VQAv2 (Train)	2017	Natural Images	769K
VSR	2023	Natural Images (Spatial)	10K
WebSight	2024	Webpage Screenshots	2.7M
WildVision	2024	Natural Images (In-the-wild)	154K
Mixed / Multi-domain Datasets			
IconQA	2021	Abstract Icons / Diagrams	107K
LLaVA-CoT-100k	2025	Mixed (with CoT reasoning)	100K
LLaVA-158k	2023	Mixed	158K
LLaVA-665k	2023	Mixed	665K
LLaVA-OneVision	2024	Mixed	4.8M
VisionFlan	2024	Mixed	186K

Table 4: Multimodal benchmarks for evaluating Vision-Language Models (VLMs).

Benchmark	Venue	Main Contribution
Math & Science		
MathVista (2024)	ICLR'24	Tests mathematical reasoning under diverse visual contexts (charts, plots, geometry).
MathVision (2025a)	NeurIPS'24	Evaluates reasoning with mathematical formulas and image assistance.
ScienceQA (2022)	NeurIPS'22	Multi-modal science questions with annotated lectures and chain-of-thought explanations.
MathVerse (2024)	ECCV'24	Tests reasoning in combined mathematical formula and image problems.
DynaMath (2024)	ICLR'25	Dynamic visual mathematical benchmarks to test generalization.
WeMath (2024)	ACL'25	Multi-step reasoning benchmark for visual mathematics.
OlympiadBench (2024)	ACL'24	High-difficulty benchmark featuring Olympiad-level math and physics problems.
Infographic & Document		
ChartQA (2022)	ACL'22	Evaluates visual reasoning and data extraction from charts.
OCRBench (2024c)	SCIS'24	Comprehensive evaluation for OCR capabilities and text-centric VQA.
InfoVQA (2021a)	WACV'22	Information extraction from complex infographics.
DocVQA (2021b)	WACV'21	Visual question answering on scanned document images.
TextVQA (2019)	CVPR'19	VQA requiring reading and reasoning about text in natural scenes.
CharXiv (2024)	NeurIPS'24	Comprehensive benchmark for descriptive and reasoning tasks on charts.
General Perception & Understanding		
MMMU (2024)	CVPR'24	Expert-level benchmark covering 183 subfields, requiring college-level subject knowledge.
MMBench (2024b)	ECCV'24	Robust multi-choice evaluation using CircularEval to mitigate guessing.
MME (2025)	CVPR'24	Comprehensive suite evaluating 14 perception and cognition subtasks.
VQA v2 (2017)	CVPR'17	The standard benchmark for open-ended questions on natural images.
GQA (2019)	CVPR'19	Focuses on compositional reasoning and scene graph understanding.
VizWiz (2018)	CVPR'18	VQA on images taken by blind users; specifically tests unanswerable questions.
SEED-Bench (2023a)	CVPR'24	Large-scale benchmark evaluating both generation and comprehension across image/video.
MMStar (2024a)	CVPR'24	A curated "vision-essential" benchmark filtered to avoid blind guessing.
HallusionBench (2024)	CVPR'24	Tests visual illusions and hallucination consistency.
POPE (2023b)	EMNLP'23	Evaluates object hallucination using Yes/No questions.
RealWorldQA	-	Tests spatial understanding and object recognition in real-world driving/egocentric views.
Knowledge, Reasoning & Dialogue		
OK-VQA (2019)	CVPR'19	VQA requiring outside knowledge beyond the image content.
A-OKVQA (2022)	ECCV'22	Augmented outside knowledge VQA requiring reasoning rationales.
VCR (2019)	CVPR'19	Visual Commonsense Reasoning (Answer + Rationale selection).
MM-Vet (2023)	ICML'24	Evaluates integrated capabilities (recognition, OCR, math) using GPT-4 based scoring.
VisDial (2017)	CVPR'17	Evaluates multi-turn visual dialogue capabilities.
LLaVA-Bench (2023a)	NeurIPS'23	Small but high-quality set for checking conversational alignment.