Evian: Towards Explainable Visual Instruction-tuning Data Auditing

Zimu Jia[1], Mingjie Xu[1], Andrew Estornell[2], Jiaheng Wei[1]*

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]ByteDance Seed

jiahengwei@hkust-gz.edu

Abstract

The efficacy of Large Vision-Language Models (LVLMs) is critically dependent on the quality of their training data, requiring a precise balance between visual fidelity and instruction-following capability. Existing datasets, however, are plagued by inconsistent quality, and current data filtering methods rely on coarse-grained scores that lack the granularity to identify nuanced semantic flaws like logical fallacies or factual errors. This creates a fundamental bottleneck in developing more reliable models. To address this, we make three core contributions. First, we construct a large-scale, 300K-sample benchmark by systematically injecting diverse, subtle defects to provide a challenging testbed for data auditing. Second, we introduce a novel "Decomposition-then-Evaluation" paradigm that breaks model responses into constituent cognitive components: visual description, subjective inference, and factual claim, enabling targeted analysis. Third, we instantiate this paradigm via **EVIAN** (**E**xplainable **V**isual **I**nstruction-tuning Data **A**uditi**N**g), a pipeline that evaluates these components along the orthogonal axes of Image-Text Consistency, Logical Coherence, and Factual Accuracy. Our empirical findings challenge the prevailing scale-centric paradigm: a model fine-tuned on a compact, high-quality subset curated by EVIAN consistently surpassed models trained on orders-of-magnitude larger datasets. We also reveal that dividing complex auditing into verifiable subtasks enables robust curation, and that Logical Coherence is the most critical factor in data quality evaluation.

1 Introduction

Large Vision-Language Models (LVLMs) (Chen et al., 2024e) have recently demonstrated remarkable progress in aligning visual perception with natural language understanding, enabling a wide range of applications from medical assistance to robotic control (Yin et al., 2024; Li et al., 2025; Pang et al., 2025; Yue et al., 2024). An important factor of this success is *Visual Instruction Tuning* (VIT), which aligns visual representations with language instructions to enhance instruction-following capability (Liu et al., 2023). However, the effectiveness of VIT hinges on the quality of the underlying training data, which must strike a delicate balance between adhering to user commands and maintaining fidelity to visual inputs.

Existing datasets and filtering methods fall short of this requirement. Large-scale data synthesis (e.g., LLaVA-Instruct-150K) improves instruction following but often introduces noise (Liu et al., 2024c; Tang et al., 2024), while similarity-based filtering methods (e.g., CLIP score) promote visual grounding but lack the granularity to detect subtle semantic flaws (Wang et al., 2024a). As a result, current LVLMs frequently suffer from fine-grained errors, including object hallucination, attribute misattribution, factual inconsistency, and flawed reasoning (Liu et al., 2024a; Bai et al., 2024; Chen et al., 2024d). These deficiencies reveal a fundamental bottleneck: prevailing approaches rely on coarse, uni-dimensional quality measures that collapse diverse error types into a single opaque score.

In this work, we argue that evaluating model-generated responses requires moving beyond monolithic scoring toward structured verification. Our core insight is that a response is not an indivisible block of text but a composite of distinct, verifiable components. Building on this principle, we propose the *Decomposition-then-Evaluation* paradigm, which reframes the task of auditing complex responses into targeted sub-tasks. Specifically, we isolate and validate *pure visual descriptions* to address visual misrepresentation, *external factual claims* to correct factual inaccuracies, and *subjective inferences* to mitigate flawed reasoning.

To operationalize this paradigm, we introduce **EVIAN** (**E**xplainable **V**isual **I**nstruction-tuning

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 6272–6291
July 2-7, 2026

Data **A**uditi**N**g), an automated and interpretable framework that systematically evaluates responses along three orthogonal axes: Image-Text Consistency, Logical Coherence, and Factual Accuracy. Complementing this framework, we construct a large-scale, 300K-sample benchmark by injecting diverse, subtle defects, providing a challenging testbed for fine-grained data auditing. Our empirical findings show that models fine-tuned on compact, high-quality subsets curated by EVIAN consistently outperform models trained on orders-of-magnitude larger datasets, highlighting that interpretable data curation, rather than sheer scale, is the key to advancing LVLMs.

Our main contributions are as follows:

- To spur research in LVLM visual instruction tuning data quality and facilitate rigorous evaluation, we introduce a 300K-sample benchmark for visual instruction data selection, built by systematically injecting diverse semantic defects to support fine-grained auditing.
- We propose the *Decomposition-then-Evaluation* paradigm and instantiate it in **EVIAN**, a fully automated and interpretable framework that decomposes responses into visual descriptions, subjective inferences, and factual claims, and evaluates them along three orthogonal dimensions.
- We conduct extensive experiments showing that for LVLMs, the logical integrity of training data is a more decisive factor for downstream performance than its informational richness, establishing the critical need to prioritize reasoning and factual correctness in data curation.

2 Related Work

Vision-language data curation has progressed from coarse pre-training filters to instruction-tuning strategies, yet scalable and fine-grained evaluation remains largely missing. As a result, most existing methods still rely on shallow quality proxies, limiting their ability to diagnose subtle semantic, logical, or factual defects.

Data Selection for Vision-Language Pre-training. A central challenge in vision-language learning is selecting high-quality subsets from noisy web-scale corpora such as LAION (Schuhmann et al., 2021). Early approaches rely on similarity-based filtering with pre-trained models, including CLIP (Radford et al., 2021), ALBEF (Li et al., 2021), and BLIP (Li et al., 2022, 2023), using holistic similarity scores (Hessel et al., 2021; Xu et al., 2025a; Wang et al., 2024d) or mixture modeling (Shi et al., 2024a). More recent work adopts fine-tuned multimodal language models as learned data filters, scoring image-text pairs along multiple semantic dimensions (Wang et al., 2024c), or employs generative models for dataset sanitization via re-captioning and label correction (Vasa et al., 2025; Mahjourian and Nguyen, 2025; Zhang et al., 2024b; Zhu et al., 2023; Zhang et al., 2025). Despite their effectiveness, these methods predominantly depend on coarse proxies and offer limited insight into complex reasoning or factual errors.

Data Curation for Visual Instruction Tuning. As vision-language models shift from representation learning to instruction following (Safaei et al., 2025), data quality has become increasingly critical (Chen et al., 2024c). The "quality over quantity" principle was first demonstrated in the text domain by AlpaGasus (Chen et al., 2023) and later extended to multimodal settings by InstructionGPT-4 (Wei et al., 2023), which combines CLIP scores with GPT-4 judgments to curate compact datasets. Other approaches generate synthetic instruction data (Liu et al., 2024d; Chen et al., 2024a) or augment supervision with reasoning traces, such as Reflective Instruction Tuning (Zhang et al., 2024a). A widely adopted alternative is the LLM-as-a-Judge paradigm (Gu et al., 2024; Li et al., 2024; Pu et al., 2025), which has been shown to suffer from bias, instability, and reasoning shortcuts, particularly without ground-truth references (Shi et al., 2024b; Hwang et al., 2025; Ye et al., 2024; Guerdan et al., 2025; Wei et al., 2024; Zhang et al., 2026). Consequently, scalable and reliable instruction-tuning data curation remains an open problem.

The Gap in Fine-Grained Evaluation. The reliance on coarse filtering reflects a broader absence of scalable fine-grained evaluation. While prior work has explored alternatives to single holistic scores (Adlakha et al., 2024), systematic diagnosis of semantic errors remains limited. Early automated methods rely on fixed criteria that struggle with open-ended errors (Zhao et al., 2024), and recent pipelines such as SCALE (Xu et al., 2025b) lack explicit modeling of compositional reasoning. Task-specific benchmarks for logical reasoning (Xiao et al., 2024; Xu et al., 2025c) provide deeper analysis but are too narrow for general data auditing, while conceptual discussions of holistic

evaluation (Tu et al., 2025) stop short of actionable frameworks. This gap motivates structured, component-level evaluation that disentangles visual grounding, reasoning, and factual correctness in multimodal data.

3 The Method: Evian

We propose **EVIAN**, an automated pipeline for auditing visual instruction data. As illustrated in Figure 1, EVIAN follows a two-phase process: (i) response decomposition, which disentangles complex answers into verifiable components, and (ii) multi-faceted evaluation, which scores these components across orthogonal quality dimensions.

3.1 Problem Definition and Data Quality Metrics

We define *visual instruction data auditing* as the task of assigning interpretable quality scores to image-instruction-response triples. Formally, given $x_i = (I_i, P_i, R_i)$ from dataset D, our auditing function Φ maps each sample to a three-dimensional score vector:

$$\boldsymbol{S}_i = \Phi(x_i) = (S_{L,i}, S_{K,i}, S_{V,i}), \qquad (1)$$

where each score ranges from 1 (low) to 5 (high). The three metrics are:

- **Logical Coherence** (S_L): soundness of reasoning relative to the instruction and visual evidence.
- **Factual Accuracy** (S_K): correctness of knowledge claims against external facts.
- **Image-Text Consistency** (S_V): fidelity of the textual response to the visual input.

Together, these axes provide a comprehensive measure of data quality, capturing both semantic integrity and visual fidelity.

3.2 Phase 1: Response Decomposition via Chain-of-Thought

The first phase disentangles raw responses into verifiable components, separating visual descriptions from subjective inferences and factual claims. This is achieved through a three-step chain-of-thought (CoT) process, $\Psi_{\text{deconstruct}}$, implemented with the Qwen3-235B-A22B-Instruct-2507-FP8 model (Team, 2025). The result is an annotated response with explicit tags and a purified visual summary, which together form the basis for systematic auditing.

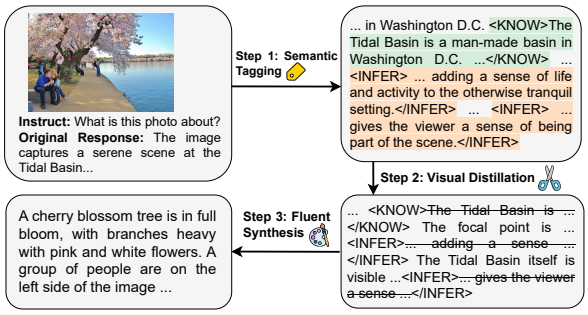

Figure 2: Three-stage Chain-of-Thought (CoT) process for response decomposition, which (1) isolates subjective inferences and factual claims via semantic tagging, (2) purifies the text through visual distillation, and (3) refines the output into a cohesive, purely visual summary.

Step 1: Semantic Tagging. The process begins by parsing the raw response R_i while strictly preserving its original wording. Subjective judgments (e.g., "the room feels cozy") are wrapped in `<INFER>` tags, and knowledge-dependent claims (e.g., "this is a Bauhaus-style lamp") are wrapped in `<KNOW>` tags. Untagged text is treated as purely visual description. This produces an annotated response $R_i^{\text{annotated}}$ that explicitly separates cognitive components without altering their content.

Step 2: Visual Distillation. Next, the annotated response is distilled into a purely visual form. Segments within `<INFER>` or `<KNOW>` tags are either rewritten into neutral, descriptive statements or deleted if unverifiable. For example, "this is likely a wedding dress" becomes "a white dress"; unverifiable claims are dropped entirely. Untagged visual statements remain unchanged. The result is a draft R_i^{draft} containing only objective, image-grounded content.

Step 3: Fluent Synthesis. Since distillation may fragment the text, a final synthesis step restores fluency and coherence. The draft response is reorganized into a single, natural paragraph while strictly forbidden from adding new content. This ensures the output R_i^{visual} is a faithful, high-quality visual summary.

Together, these steps yield two complementary artifacts: $R_i^{\text{annotated}}$, which retains the full response structure with explicit tags, and R_i^{visual}, which isolates objective descriptions. This decomposition provides the foundation for precise, component-level auditing in Phase 2.

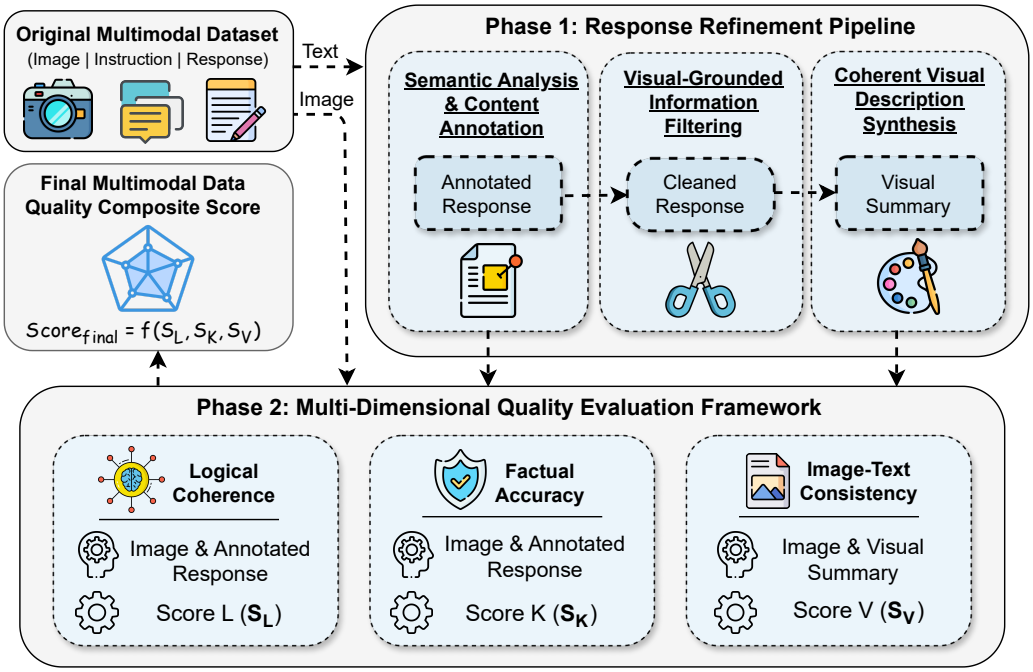

Figure 1: Overview of the two-phase EVIAN framework, which first decomposes a response into visual, inferential, and factual components and then evaluates them along the orthogonal dimensions of Image–Text Consistency, Logical Coherence, and Factual Accuracy.

3.3 Phase 2: Multi-faceted Quality Assessment

The second phase conducts a multi-faceted evaluation of each decomposed response along three orthogonal dimensions: logical coherence, factual accuracy, and image-text consistency. We employ Qwen2.5-VL-7B-Instruct-AWQ (Bai et al., 2025) as an automated auditor, which assigns interpretable 15 scores and textual rationales based on a detailed rubric. This step provides fine-grained diagnostics of different error types while producing standardized quality scores that can be aggregated for ranking and selection.

Logical Coherence (S_L). This dimension evaluates whether reasoning in the <INFER> tags follows plausibly from visual evidence. Scores increase with reasoning strength: a default of 2 when no inference is given, 3 for plausible but unsubstantiated claims, 4 for well-supported reasoning, and 5 for logically undeniable conclusions. This rubric rewards depth of reasoning while penalizing speculation.

Factual Accuracy (S_K). This dimension fact-checks knowledge claims in the <KNOW> tags against the auditors internal knowledge. Fully correct claims receive 5, minor inaccuracies lower the score to 4, and a single major error (e.g., misidentifying a capital city) caps the score at 2. In the absence of knowledge claims, the default score is 2, distinguishing informative from non-informative responses.

Image-Text Consistency (S_V). This dimension measures the alignment of the purified visual description R^{visual} with the image. The principle is consistency over completeness: omissions are acceptable, but contradictions or unverifiable assertions are heavily penalized. Perfectly faithful descriptions receive 5, minor imprecisions result in 4, and any clear contradiction drops the score to 2 or below. This ensures that only visually accurate responses achieve the highest marks.

By producing a triplet (S_L, S_K, S_V) with explicit explanations, Phase 2 delivers an interpretable and multi-dimensional quality assessment. These scores directly guide downstream data ranking and selection.

3.4 Data Ranking and Selection

To enable downstream filtering, the three-dimensional score vector $\boldsymbol{S}$ is aggregated into a single scalar:

$$S_{\text{overall}} = \frac{S_L + S_K + S_V}{3}. \quad (2)$$

This default scheme assumes equal importance, but weights can be tuned for specific applications (e.g.,

emphasizing S_K for knowledge-intensive tasks or S_L/S_V for creative captioning). To investigate the impact of these variations, we conduct a sensitivity analysis on the component weights, which is detailed in Appendix C. This flexibility ensures that data selected by EVIAN aligns with diverse modeling objectives.

4 Benchmarking Data Quality via Controlled Defect Injection

To quantitatively validate a data auditing pipelines ability to detect fine-grained flaws in logical coherence, factual accuracy, and image-text consistency, a tailored benchmark with systematically injected defects is essential, as existing datasets lack the controlled errors needed for such a targeted evaluation. To ensure consistency with prior work, we adopt the SCALE methodology (Xu et al., 2025b) as the starting point for benchmark construction. From its source pool of 500,000 multimodal samples across eight datasets (Table 1), we derive two complementary components: (i) a 50,000-sample gold standard set purified by SCALE, and (ii) a 250,000-sample challenge set obtained via random down-sampling followed by our defect injection pipeline. Together, these components yield a reproducible benchmark of 300,000 samples, designed to evaluate whether data auditing methods can distinguish clean data from semantically corrupted examples.

Table 1: Overview of the source datasets, comprising 300K samples from eight foundational datasets grouped into General Vision-Language tasks and Domain-Specific Reasoning tasks.

Dataset	Task Category
General Visual-Language Capabilities	
ShareGPT-4V (Chen et al., 2024b)	Instruction Following
LLaVA-1.5-Mix (Liu et al., 2024b)	General QA
AllSeeing-V2 (Wang et al., 2024b)	Grounding
Domain-Specific Reasoning Capabilities	
DocVQA (Mathew et al., 2021)	Document
ChartQA (Masry et al., 2022)	Chart
InfoVQA (Mathew et al., 2022)	OCR
A-OKVQA (Schwenk et al., 2022)	Knowledge
Geometry3K (Lu et al., 2021)	Mathematics

Defect Injection Pipeline. The challenge set is generated through a three-stage pipeline that leverages the Qwen3-235B-A22B-Instruct-2507-FP8 model (Team, 2025) to embed subtle, context-aware flaws. The process is guided by a principled taxonomy (Table 2) spanning three critical dimensions for auditing: *perceptual consistency*, *factual accuracy*, and *logical coherence*.

Stage 1: Content Analysis. Each source response is analyzed by an LLM to identify whether it contains external knowledge or logical reasoning. This structured analysis, output in JSON, serves as a prior to ensure that subsequent errors are coherent with the intrinsic properties of the text.

Stage 2: Contextual Error Selection. An error category is chosen via a probabilistic cascade. To counter their rarity, knowledge-related and reasoning-related errors are prioritized with probabilities of 0.8 and 0.6, respectively, while perceptual consistency serves as the default. Subtypes are selected randomly for consistency errors, whereas an additional LLM call determines the most plausible subtype for knowledge and reasoning cases.

Stage 3: Guided Rewriting. The chosen error is injected by prompting the LLM with a targeted transformation instruction. A strict system prompt constrains the model to output only the modified text, ensuring automation and reproducibility.

This injection strategy goes beyond simple noise addition: it produces realistic, semantically rich corruptions aligned with the three audit dimensions. As a result, the benchmark offers a challenging testbed for assessing whether auditing pipelines can detect not only superficial inconsistencies but also deeper factual and logical flaws.

5 Experiments

5.1 Experimental Setup

Baselines. We compare our method against a diverse set of data auditing baselines spanning visual-language pretraining filters and recent visual instruction tuning approaches. Specifically, we consider: (1) **Random Sampling**, which randomly selects 10,000 samples as a non-selective lower bound; (2) **Image-Text Similarity Filters**, including **CLIPScore** (ViT-B/32), **ALBEF**, **BLIP**, and **BLIP-2**, which rank the full data pool by holistic image-text similarity and select the top 10,000 samples; (3) **SCALE**, a multi-stage filtering method that evaluates modality quality, relevance, clarity, and task rarity using a weighted scoring scheme; and (4) **Qwen2.5-VL-7B-Instruct-AWQ**, which directly scores sample quality via model-based evaluation and selects the top 10,000 instances.

Table 2: Principled taxonomy of semantic defects used in benchmark construction, categorizing errors into three dimensions aligned with EVIANs evaluation modules: **Consistency**, **Reasoning**, and **Knowledge**.

Error Subtype	Description / Generation Strategy
Image-Text Consistency (S_V)	
Attribute	Describes an object's attribute (e.g., color, material) incorrectly.
Spatial	Details incorrect spatial relations between objects.
Action	Assigns a wrong action or state to a subject.
Fake	Introduces a plausible yet non-existent object.
Misidentification	Misidentifies an existing object.
Logical Coherence (S_L)	
Conclusion	Generalizes hastily from a single detail.
Causal	Mistakes correlation for causation between events.
Prediction	Makes a baseless prediction from scant evidence.
Procedural	Adds a flawed or superfluous step to a process.
Comparison	Forms a misleading analogy from superficial traits.
Factual Accuracy (S_K)	
Entity	Corrupts facts about a named entity.
Context	Places an object in a wrong historical/technological context.
Definition	Provides an incorrect definition of a concept.
Attribution	Misattributes a quote or work to the wrong source.

Evaluation Protocol. For all methods, we fine-tune Qwen2-VL-2B on the selected 10,000-sample subset and evaluate the resulting models using VLMEvalKit (Duan et al., 2024). All experiments share identical architectures, SFT procedures, and hyperparameters, ensuring that performance differences reflect data quality rather than training variation.

5.2 Evian Scores: Distribution and Discrimination

To evaluate EVIAN's discriminative power, we apply it to our benchmark containing 50,000 pristine and 250,000 defect-injected samples. As illustrated in Figure 4, the two groups display a clear separation: 92.3% of pristine entries score $\geq$ 3.0, while defect-injected samples form a distinct mid-range peak around 3.0. This shift indicates that EVIAN effectively penalizes semantically corrupted responses and differentiates them from high-quality data.

The separation is further quantified by a JensenShannon divergence of 0.35 and an AUC of 0.86, demonstrating the metrics strong discriminative capability. Furthermore, the concentration of defective samples in the mid-range, instead of accumulating at the lowest scores, shows that EVIAN detects subtle, context-dependent defects such as logical fallacies, not merely coarse inconsistencies. Although a small fraction of injected samples retain higher scores due to nuanced semantic ambiguities, the pronounced distributional gap overall provides strong evidence that EVIAN functions as a robust, fine-grained filter for high-quality data curation.

5.3 Downstream Task Performance

To evaluate the practical impact of EVIAN, we fine-tuned models on 10K-sample subsets curated by different methods and compared their downstream performance across multiple benchmarks. As shown in Table 3, **the model trained on the EVIAN-selected subset achieves the current best performance** (average score of 70.20), surpassing both the previous SOTA method (SCALE, 67.41) and the model trained on the full 300K unfiltered dataset (63.77). This "less is more" result highlights the diagnostic precision of EVIAN, which consistently extracts higher-quality data from a noisy pool.

These gains stem from EVIANs "Decomposition-then-Evaluation" paradigm, which addresses fine-grained defects overlooked by coarse auditing approaches. Filters such as CLIPScore and BLIP-2 provide moderate improvements but fail to capture errors like factual inaccuracies or logical fallacies. Moreover, EVIAN outperforms the **Qwen2.5-VL baseline** (70.20 vs. 66.34) using the identical auditor architecture, suggesting that the performance gains stem from our structured verification logic rather than simple knowledge distillation. By explicitly evaluating Image-Text Consistency, Logical Coherence, and Factual Accuracy, EVIAN yields targeted diagnostics that translate into stronger downstream models. For example, EVIANs leading performance on MME (1876.89) and

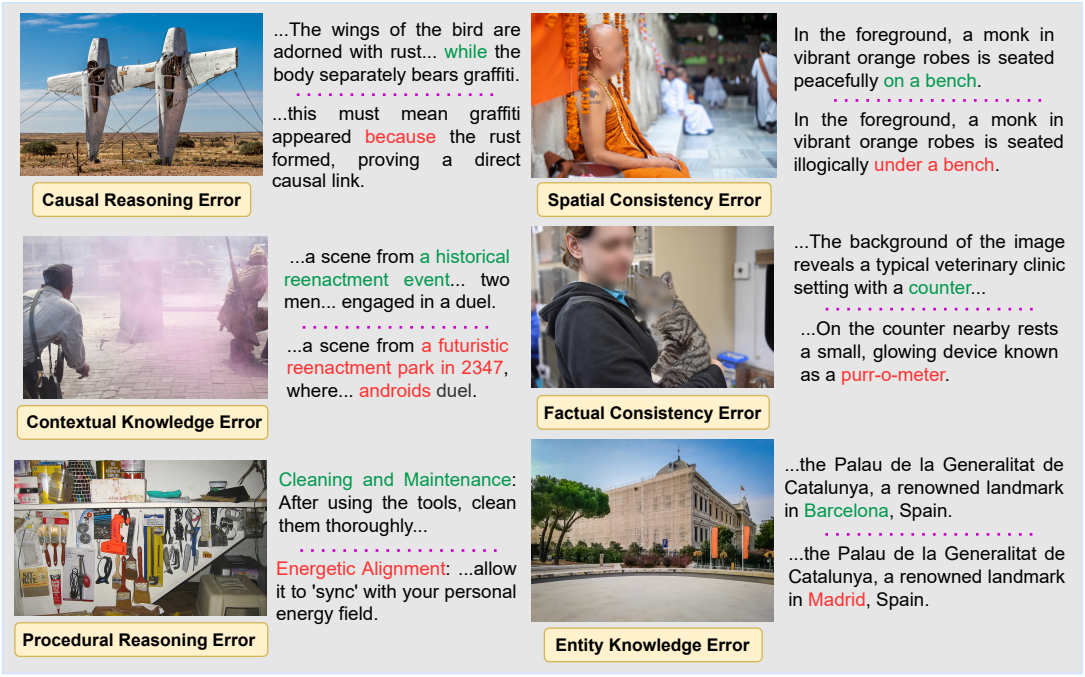

Figure 3: Examples of our controlled defect injection. For each pair, the original high-quality text (top) is rewritten to include a subtle, context-aware flaw (bottom), illustrating various error categories from our taxonomy (Table 2).

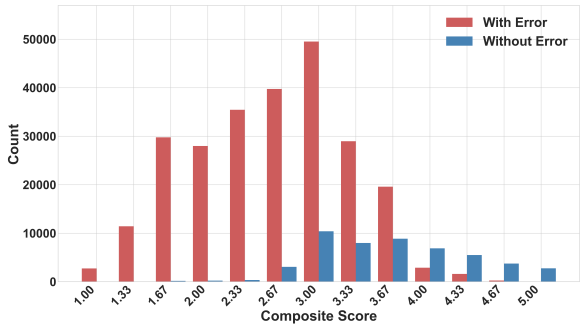

Figure 4: Score distribution comparing original and defect-injected samples, illustrating how EVIAN separates high-quality data from subtle semantic corruptions.

POPE (79.87) validates its ability to mitigate hallucinations through holistic multi-dimensional verification, while its gains on A-OKVQA (0.7493) and ScienceQA (0.7115) highlight the benefit of auditing factual and reasoning components.

Overall, these results reveal a fundamental limitation of existing curation strategies: high similarity or holistic scores do not guarantee utility and frequently obscure critical semantic defects. In contrast, EVIAN's multi-dimensional auditing yields cleaner and more reliable training data, enabling models trained on small, high-quality subsets to surpass those trained on far larger but noisier datasets. This points to a clear direction for LVLM development: progress hinges less on scaling data volume and more on fine-grained, interpretable auditing that enforces visual fidelity, factual accuracy, and logical coherence. To further verify that these gains arise from intrinsic data quality rather than a potential inductive-bias alignment between the Qwen-based auditor and target model, we conduct a cross-architecture evaluation with InternVL2-2B (Chen et al., 2024f). As shown in Appendix D, EVIAN's advantages extend to a model family with a distinct architectural lineage, reinforcing the generality of our conclusions. Finally, we validate EVIAN on the original, unmodified data distribution in Appendix E, where a 10K curated subset rivals the full 300K baseline, confirming its capability to capture intrinsic data quality signals beyond artificial defects.

5.4 Ablation Experiment

To validate the necessity of each component in EVIAN, we conducted an ablation study by selectively excluding the Logical Coherence (S_L), Factual Accuracy (S_K), and Image-Text Consistency (S_V) scores from the selection criteria. The results, summarized in Table 4, demonstrate that the full framework consistently achieves the best performance across all benchmarks. This confirms that these dimensions function as distinct and complementary factors essential for effective data auditing.

The full EVIAN framework achieves the best

Table 3: Comparisons with state-of-the-art data selection baselines on 10K subsets, where **bold** and underlined indicate the best and second-best results, respectively. "Full Data" denotes training on the entire 300K pool.

Model	MME	MMBench	SEEDBench	ScienceQA	A-OKVQA	POPE	Avg
Random	1475.76	0.5353	0.6031	0.6614	0.7092	75.50	63.18
Full Data	1553.05	0.5953	0.5743	0.6267	0.6934	78.17	63.77
CLIPScore	1565.29	0.5746	0.6170	0.6906	0.7301	74.57	65.28
ALBEF	1590.70	0.6003	0.6107	0.6748	0.7048	72.29	64.69
BLIP	1686.62	0.6183	0.6115	0.6802	0.6978	73.40	65.74
BLIP-2	1810.34	0.6317	0.6187	0.7045	0.7127	77.38	68.13
SCALE	1814.97	0.6318	0.6280	0.6916	0.7066	73.81	67.41
Qwen2.5-VL	1682.78	0.5796	0.6182	0.6797	0.7187	78.30	66.34
EVIAN (Ours)	**1876.89**	**0.6463**	**0.6359**	**0.7115**	**0.7493**	**79.87**	**70.20**

Table 4: Ablation study of the EVIAN framework, evaluating the contribution of individual components by removing them from the full pipeline. "w/o Decomposition" denotes the variant without the fine-grained decomposition stage. S_L and S_K represent the **Logical Coherence** and **Factual Accuracy** scores, respectively.

Configuration	MME	MMBench	SEEDBench	ScienceQA	A-OKVQA	POPE	Avg
w/o Decomposition	1706.70	0.6401	0.6312	0.7085	0.7170	76.93	67.93
w/o S_L	1656.62	0.3425	0.5324	0.5563	0.6288	78.45	57.27
w/o S_K	1604.91	0.6110	0.5875	0.6604	0.6629	75.77	64.21
w/o S_L, S_K (Only S_V)	1807.13	0.5605	0.6092	0.6822	0.7389	68.56	65.36
EVIAN (Full)	**1876.89**	**0.6463**	**0.6359**	**0.7115**	**0.7493**	**79.87**	**70.20**

average performance (70.20), confirming the synergistic benefit of combining all three evaluation axes. Most notably, removing Logical Coherence (S_L) leads to a significant decline in performance to 57.27. This counterintuitive outcome arises because filtering based solely on factual accuracy (S_K) and visual consistency (S_V) inadvertently favors responses that are factually correct and visually grounded but logically inconsistent. The prevalence of such inconsistent samples introduces conflicting supervision signals, which significantly impairs the model's performance on reasoning-intensive benchmarks like ScienceQA. This confirms that logical integrity is not merely an auxiliary metric but a critical factor for complex reasoning tasks.

In contrast, relying solely on Image-Text Consistency (S_V) presents an inverse trade-off. While it yields a stable average score (65.36) by avoiding logical traps, it exhibits a marked drop on POPE (68.56), falling well below the random baseline. This sharp decline reveals that visual consistency alone is an insufficient proxy for data quality; without the structural constraints of logical and factual verification, the curation process fails to filter out subtle hallucinations, leaving the model vulnerable to object fabrication. Thus, EVIANs comprehensive auditing is essential: S_V ensures visual relevance, while S_L and S_K are essential for guaranteeing logical coherence and factual precision.

6 Conclusion

In this work, our proposed visual instruction tuning data auditing method **EVIAN**, advances LVLM data quality auditing through three contributions: a 300K-sample benchmark with systematically injected defects, a "Decomposition-then-Evaluation" paradigm that separates visual, inferential, and factual components, and the EVIAN framework, which scores data along Image-Text Consistency, Logical Coherence, and Factual Accuracy. Experiments show that EVIAN-curated subsets consistently outperform models trained on much larger unfiltered datasets, and ablations confirm the necessity of each evaluation dimension. Surprisingly, our study also reveals that **dividing complex auditing into verifiable subtasks enables robust curation**, and that **Logical Coherence is the most critical factor for downstream reliability**. These results establish interpretable, fine-grained auditing as the foundation for advancing LVLMs.

7 Acknowledgments

Zimu Jia, Mingjie Xu, and Jiaheng Wei are partially supported by the CNPC Technology Project "Research on Key Technologies of Artificial Intelligence for Oil and Gas Exploration and Development" (No. 2023DJ84), and the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

8 Limitations

While EVIAN demonstrates strong effectiveness for fine-grained auditing of visual instruction data, several limitations remain. First, the framework relies on large pre-trained multimodal language models for both response decomposition and quality evaluation. Despite strong alignment between automated scores and human judgments, the auditing process may still inherit biases, blind spots, or reasoning tendencies from the underlying models, particularly in ambiguous or culturally sensitive scenarios.

Second, EVIAN assumes that model responses can be reliably decomposed into visual descriptions, subjective inferences, and factual claims. Errors introduced during this decomposition stage may propagate to subsequent evaluations and affect the final quality scores. Although our ablation results suggest that explicit decomposition is beneficial overall, improving robustness under imperfect decomposition remains an open challenge.

Third, the current pipeline incurs non-trivial computational cost due to multiple invocations of large models, which may limit its applicability in resource-constrained settings or when auditing extremely large-scale datasets. Exploring lighter-weight auditors or partially learned approximations of individual components is a promising direction for future work.

Finally, EVIAN focuses on auditing response quality with respect to visual grounding, logical coherence, and factual accuracy. Other important aspects, such as stylistic diversity, pedagogical value, or downstream task-specific preferences, are not explicitly modeled and may require complementary criteria depending on the application context.

9 Ethical Considerations

This work aims to improve the quality and reliability of visual instruction-tuning data through automated auditing. By identifying logical inconsistencies, factual errors, and visual misalignment in training data, EVIAN has the potential to reduce hallucinations and misleading outputs in downstream vision–language models, contributing to safer and more trustworthy AI systems.

Nevertheless, the use of large language and vision–language models as automated auditors raises ethical considerations. Model-based judgments may reflect biases present in their training data and influence which data distributions are preserved or suppressed during curation, potentially leading to systematic over- or under-filtering of certain types of content. Care should therefore be taken when deploying EVIAN in sensitive domains, such as medical or legal applications, where incorrect filtering decisions may have disproportionate consequences. EVIANs outputs should be viewed as decision-support signals rather than absolute ground truth, and human oversight remains important in high-stakes settings.

Finally, all experiments in this work are conducted on publicly available datasets, and no new human annotations are collected. The proposed framework is intended to support responsible data curation practices and does not introduce additional privacy or data collection concerns beyond those already present in existing vision–language datasets.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

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*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023.

Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

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

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024d. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024e. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024f. 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.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. 2025. Validating llm-as-a-judge systems in the absence of gold labels. *arXiv preprint arXiv:2503.05965*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yong-il Kim, and Kyomin Jung. 2025. Fooling the lvlm judges: Visual biases in lvlm-based evaluation. *arXiv preprint arXiv:2505.15249*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *arXiv preprint arXiv:2506.14674*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. 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. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, and 1 others. 2024c. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12):220103.

Zheng Liu, Hao Liang, Bozhou Li, Tianyi Bai, Wentao Xiong, Chong Chen, Conghui He, Wentao Zhang, and Bin Cui. 2024d. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Nazanin Mahjourian and Vinh Nguyen. 2025. Sanitizing manufacturing dataset labels using vision-language models. *arXiv preprint arXiv:2506.23465*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Zirui Pang, Haosheng Tan, Yuhan Pu, Zhijie Deng, Zhouan Shen, Keyu Hu, and Jiaheng Wei. 2025. When vlms meet image classification: Test sets renovation via missing label identification. *arXiv preprint arXiv:2505.16149*.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14247–14256.

C Schuhmann, R Vencu, R Beaumont, R Kaczmarczyk, C Mullis, A Katta, T Coombes, J Jitsev, and A Laion Komatsuzaki. 2021. 400m: Open dataset of clip-filtered 400 million image-text pairs. arxiv 2021. *arXiv preprint arXiv:2111.02114*, 2.

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.

Haitao Shi, Meng Liu, Xiaoxuan Mu, Xuemeng Song, Yupeng Hu, and Liqiang Nie. 2024a. Breaking through the noisy correspondence: A robust model for image-text matching. *ACM Transactions on Information Systems*, 42(6):1–26.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024b. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Yangfan He, Kuan Lu, Hao Feng, Yang Li, and 1 others. 2024. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Weijie Tu, Weijian Deng, and Tom Gedeon. 2025. Toward a holistic evaluation of robustness in clip models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Santosh Vasa, Aditi Ramadwar, Jnana Rama Krishna Darabattula, Md Zafar Anwar, Stanislaw Antol, Andrei Vatavu, Thomas Monninger, and Sihao Ding. 2025. Autovdc: Automated vision data cleaning using vision-language models. *arXiv preprint arXiv:2507.12414*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and 1 others. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, and 1 others. 2024b. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer.

Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. 2024c. Finetuned multimodal language models are high-quality image-text data filters. *arXiv preprint arXiv:2403.02677*.

Yiping Wang, Yifang Chen, Wendan Yan, Alex Fang, Wenjing Zhou, Kevin Jamieson, and Simon S Du. 2024d. Cliploss and norm-based data selection methods for multimodal contrastive learning. *Advances in Neural Information Processing Systems*, 37:15028–15069.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers. *arXiv preprint arXiv:2402.10412*.

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

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Jinda Xu, Yuhao Song, Daming Wang, Weiwei Zhao, Minghua Chen, Kangliang Chen, and Qinya Li. 2025a. Quality over quantity: Boosting data efficiency through ensembled multimodal data curation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21761–21769.

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. *arXiv preprint arXiv:2506.08429*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025c. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Xianghu Yue, Xueyi Zhang, Yiming Chen, Chengwei Zhang, Mingrui Lao, Huiping Zhuang, Xinyuan Qian, and Haizhou Li. 2024. Mmal: Multi-modal analytic learning for exemplar-free audio-visual class incremental tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2428–2437.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024a. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer.

Shuo Zhang, Zezhou Huang, and Eugene Wu. 2024b. Data cleaning using large language models. *arXiv preprint arXiv:2410.15547*.

Xueyi Zhang, Chengwei Zhang, Zheng Li, Xiyu Wang, Siqi Cai, Mingrui Lao, Yanming Guo, and Huiping Zhuang. 2026. Rep deep & machine learning: Exemplar-free continual video action recognition via slow-fast collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 36075–36083.

Xueyi Zhang, Peiyin Zhu, Yuan Liao, Xiyu Wang, Mingrui Lao, Siqi Cai, Yanming Guo, and Haizhou Li. 2025. Trustclip: Learning from noisy labels via semantic label verification and trust-aligned gradient projection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4388–4397.

Ruibin Zhao, Zhiwei Xie, Yipeng Zhuang, and Philip LH Yu. 2024. Automated quality evaluation of large-scale benchmark datasets for vision-language tasks. *International Journal of Neural Systems*, 34(03):2450009.

Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. 2023. Unmasking and improving data credibility: A study with datasets for training harmless language models. *arXiv preprint arXiv:2311.11202*.

A Evian Framework Implementation Details

A.1 Models and Computational Resources

All experiments were conducted on a high-performance computing node equipped with eight NVIDIA H100 (80GB) GPUs. We employed the Qwen3-235B-A22B-Instruct-2507-FP8 model for the text-heavy response decomposition and defect injection phases, and the **Qwen2.5-VL-7B-Instruct-AWQ** model for the multi-faceted quality assessment. Both models were deployed via vLLM (v0.10.0) using a greedy sampling strategy to ensure deterministic reproducibility within a unified software environment consisting of PyTorch (v2.7.1) and CUDA (v12.6). Leveraging this setup, the entire processing pipeline for the 300,000-sample benchmark was completed in nearly 28 hours.

A.2 Supervised Fine-Tuning (SFT) Details

To efficiently fine-tune the **Qwen2-VL-2B** base model, we implemented a selective update strategy, freezing the vision tower while training the projector MLP and language model. This training process was conducted on a server equipped with eight NVIDIA vGPU (48 GB) cards. It leveraged DeepSpeed ZeRO Stage 3 for memory optimization, resulting in an effective global batch size of 128. All key hyperparameters are detailed in Table 5.

Table 5: Supervised Fine-Tuning (SFT) Hyperparameters for the Base Model.

Hyperparameter	Value
Base Model	Qwen2VL-2B
Epochs	1
Learning Rate	5×10^{-6}
Batch Size (per device)	2
Gradient Accumulation Steps	8
Weight Decay	0.0
Warmup Ratio	0.1
LR Scheduler	Cosine
Max Gradient Norm	1.0
Precision	BF16
Max Sequence Length	8192
Gradient Checkpointing	Enabled
Optimization	ZeRO-3

A.3 Prompt Engineering for Phase 1: Response Decomposition

Step 1: Prompt for Semantic Tagging

Response: {response}
Your task is to precisely insert <INFER> for subjective judgments and <KNOW> for external knowledge.
Critical Guidelines for Annotation:

1. **Tag the Complete Thought:** Precisely wrap the shortest, complete phrase that conveys the entire logical idea (like a cause-and-effect statement) or the full piece of external information.
2. **Tag Interpretations of Effect/Cause:** Always tag phrases that describe the effect, purpose, or reason for a visual element.
3. **Strictly Visual is NOT Tagged:** DO NOT tag objective, verifiable descriptions of visual facts.
4. **Do Not Change Words:** Do not add, delete, or rephrase any original words, like Visible Text or Numbers.
5. **Output Format:** Your response must start with the prefix "Marked Response:".

Examples:
Input: The lighting in the room is soft, creating a cozy atmosphere. The design suggests it is from the Victorian era.
Output: Marked Response: The lighting in the room is soft, <INFER>creating a cozy atmosphere</INFER>. <INFER>The design suggests it is from the Victorian era</INFER>.
Input: This is a 1976 postage stamp from Hungary, a country in Central Europe.
Output: Marked Response: This is a 1976 postage stamp from Hungary, <KNOW>a country in Central Europe</KNOW>.
Input: The image shows a can of Coca-Cola.
Output: Marked Response: The image shows a can of Coca-Cola.

Step 2: Prompt for Visual Distillation

Instruction: {instruction}
Annotated Response: {marked_response}
Task: Process the "Annotated Response" by modifying ONLY the segments wrapped in <INFER>...</INFER> or <KNOW>...</KNOW> tags.

- Rewrite or entirely remove tagged segments to leave only what is directly and objectively visible in the image.
- **Crucially, all content NOT wrapped in tags MUST be preserved exactly as is, without any modification.**

Guidelines:

1. **Rewrite When Possible:** If a tagged idea can be rephrased as a neutral, objective, image-based description, rewrite it and remove the tags. For example, change "<INFER>creating a cozy atmosphere</INFER>" to "which illuminates the scene."
2. **Delete When Necessary:** For clearly irrelevant or purely speculative content that cannot be visually confirmed, delete the entire tagged segment (including the tags).
3. **No New Information:** DO NOT introduce any new guesses, opinions, or visual details that were not already present in the untagged parts of the original response.
4. **Output Format:** Your response must start with the prefix "Cleaned Response:".

Example:
Input Annotated Response:
A person wearing sunglasses stands under a tree. <INFER>She must be shielding her eyes from harsh sunlight.</INFER> Leaves are scattered on the ground. <KNOW>This park is famous for its autumn foliage tours.</KNOW>
Output:
Cleaned Response: A person wearing sunglasses stands under a tree. Leaves are scattered on the ground.

Step 3: Prompt for Fluent Synthesis

Instruction: {instruction}
Cleaned Response: {cleaned_response}
Task: Rephrase the "Cleaned Response" into a single, cohesive, and purely visual description.
Guidelines:

1. **Strictly Adhere to Input:** Your output MUST be a faithful reorganization of ONLY the information present in the "Cleaned Response."
2. **Preserve All Details:** Do not omit any visual information. Every object, attribute, and spatial relation from the input must be represented in your summary.
3. **No New Content or Inference:** Crucially, DO NOT add any new visual details, reasoning, assumptions, or subjective/interpretive language (e.g., "beautiful", "seems like", "creates a sense of"). Your job is to describe, not to analyze.
4. **Improve Flow:** Focus on improving sentence structure and grammatical correctness to create a natural-sounding paragraph.
5. **Output Format:** Your response must start with the prefix "Visual Summary:".

Example:
Input Cleaned Response: A white cat is on a windowsill. The background shows buildings. Light is coming through the window.
Output:
Visual Summary: A white cat sits on a windowsill where bright light is streaming in. Buildings are visible in the background.

A.4 Prompting and Rubrics for Phase 2: Multi-faceted Quality Assessment

Dimension S_L: Prompt for Logical Coherence

Input Text for Evaluation: {text_to_evaluate}
Task: You are an AI assistant designed to evaluate the correctness of logical reasoning. Your primary focus is to rigorously scrutinize the logical soundness and validity of the reasoning contained ONLY within the <INFER>...</INFER> tags, based on the visual evidence in the image.
Evaluation and Scoring Rules:

1. Isolate and Evaluate: Focus exclusively on the statements inside the <INFER> tags.
2. Assess Plausibility against Image: Judge if the inference is a logical and plausible conclusion derived from the visual information in the image.
3. Output Format:
   - Score: integer 1-5
   - Explanation: A brief evaluation of the logical rigor, noting key flaws or strengths.

Scoring Rubric:

Score 1: Grossly Illogical or Baseless. The inference is pure speculation with no connection to the image (e.g., predicting the future from a photo of a cat), or it's self-contradictory.

Score 2: Significant Logical Gaps. The inference is a major leap in logic. While loosely related to the image, it is highly unlikely or requires many unsupported assumptions. (e.g., "A person is running, <INFER>so this must be a professional athlete training for the Olympics</INFER>.")

Score 3: Plausible but Unprovable. The inference is reasonable and could be true, but it is not strongly supported by visual evidence and remains a subjective interpretation. (e.g., "The room is dim, <INFER>creating a sad atmosphere</INFER>.")

Score 4: Logically Sound. The inference is very likely correct and follows directly from strong visual evidence, with only very minor room for doubt. (e.g., "The man holds an umbrella, <INFER>suggesting it is raining or about to rain</INFER>.")

Score 5: Logically Airtight. The inference is an undeniable conclusion based on the visual facts and common-sense logic; it is virtually irrefutable. (e.g., "The wreck shows a crushed car, <INFER>indicating a high-impact collision occurred</INFER>.")

Dimension S_K: Prompt for Factual Accuracy

Input Text for Evaluation: {text_to_evaluate}
Task: You are an expert fact-checking assistant. Your task is to evaluate the factual correctness of the information contained ONLY within the <KNOW>...</KNOW> tags. Base your assessment on your internal, general knowledge.
Output Format:
Score: integer 1-5
Explanation: A brief justification for your score, specifying which facts are correct or incorrect.
Scoring Rubric:

Score 1: Entirely Incorrect or Fabricated. The information is factually wrong, nonsensical, or a complete fabrication (e.g., contains imaginary objects like the 'Luminara Scepter').

Score 2: Largely Incorrect. Contains a core factual error, even if minor details are correct. (e.g., "<KNOW>Paris, the capital of England...</KNOW>"). The presence of a single major error means the score cannot be higher than 2.

Score 3: Partially Correct but Misleading. Contains a mix of correct and incorrect information, or the information is

technically correct but presented in a highly misleading context.

Score 4: Mostly Correct. The core assertion is factually sound but contains a minor, non-critical inaccuracy (e.g., a slightly wrong year, a minor detail about a standard feature).

Score 5: Fully Correct and Accurate. Every single claim within the tags is factually sound, precise, and widely accepted.

Dimension S_V: Prompt for Image-Text Consistency

Input Text: {text_input}
Task: You are a visual consistency scoring assistant. Your task is to evaluate whether the extracted text descriptions assertions can be verified by the given image. Only assess consistency, not completeness: do NOT penalize the description for omitting image details, but DO penalize any assertions that contradict or cannot be supported by the image.
CORE SCORING GUIDELINE: Be decisive in your scoring. If the description is fully and accurately supported by the image without any errors, the score must be 5. Do not default to 4 if a 5 is warranted.
Output Format:
Score: integer 1-5
Explanation: Brief justification, indicating which assertions are verifiable and which are inconsistent or unclear.
Scoring Rubric:

Score 1: Severely inconsistent or completely unrelated. Most or all assertions contradict the image.

Score 2: Largely inconsistent. Only one or two minor assertions can be matched to the image.

Score 3: Partially consistent. Some key assertions align with the image, but others are vague, potentially incorrect, or unsupported.

Score 4: Mostly consistent. The bulk of assertions are supported by the image, but there is at least one minor imprecision or slight unsupported detail that does not mislead. Use this score for responses that are good but not perfect.

Score 5: Fully consistent and accurate. Every single assertion in the text is clearly and precisely verifiable in the image. There are no unsupported or contradictory claims. If all claims are verified, you MUST assign this score.

B Defect Injection Pipeline and Prompt Catalog

To create a challenging and diverse evaluation set, we designed and implemented a three-stage, LLM-driven pipeline for injecting controlled, contextually-relevant defects into high-quality responses. This automated pipeline ensures that the generated errors are not random but are intelligently tailored to the content of the source text.

B.1 The Three-Stage Defect Injection Pipeline

The core of our data generation process is a sequential pipeline that first analyzes the text, then selects an appropriate error type, and finally rewrites the text to introduce the defect.

Stage 1: Content Analysis First, an LLM analyzes the source text to determine if it contains logical reasoning or external knowledge. This classification serves as a prior for the subsequent error selection stage. The analysis is performed using the prompt below.

Prompt for Content Analysis

You are a text analysis expert. Analyze the following text and determine if it contains a) logical reasoning, inference, or conclusion, and b) specific external knowledge (like names of people, places, brands, historical facts).
Respond ONLY with a JSON object with two boolean keys:
{“contains_reasoning”: boolean, “contains_knowledge”: boolean}.
Text to analyze: “{text_to_analyze}”

Stage 2: Category and Subtype Selection The primary error category is selected via a probabilis-

tic cascade that prioritizes the knowledge category with a probability of 0.8 for texts flagged contains_knowledge, followed by the reasoning category with a probability of 0.6 for those with contains_reasoning, and otherwise defaults to the consistency category. This initial choice, in turn, dictates the method for subtype determination: while subtypes for the consistency category are chosen uniformly at random, a more nuanced approach is employed for the contextually-sensitive knowledge and reasoning categories, for which a second LLM call intelligently selects the most plausible subtype using the following prompt.

Prompt for Category and Subtype Selection

You are a text analysis expert. Your task is to select the single best error-injection strategy for the "Original Text" from the "Available Options".
Available Options: {error_options_text}
Original Text: "{text_to_analyze}"
Analyze the text and choose the error code from the options that is most relevant to the text's content. Respond ONLY with a JSON object containing your choice.

Stage 3: Defect Generation Finally, with a specific error subtype selected, a third LLM call rewrites the original text according to the corresponding instruction. The final prompt is constructed from a template, and a strict system prompt is used to ensure clean output.

Prompts for Defect Generation

Instruction: {prompt_instruction}
Original Text to Corrupt: "{original_text}"
- - -
You MUST provide ONLY the corrupted/rewritten text as the output. Do not include any preambles, explanations, or wrappers like 'Rewritten Text:' or the original response in your final output.

B.2 Catalog of Defect Injection Instructions

The complete set of instructions used in the defect generation stage is detailed below.

A. Consistency Errors

- consistency_attribute: Rewrite the response by changing an attribute (like color, count, or size) of one key object.
- consistency_spatial: Rewrite the response by incorrectly describing the spatial relationship between two objects (e.g., change 'on the table' to 'under the table').
- consistency_action: Rewrite the response by describing an incorrect action or state for a subject (e.g., change 'a man is sitting' to 'a man is running').
- consistency_fake: Rewrite the response to include a mention of a plausible but non-existent object.
- consistency_misidentification: Rewrite the response by misidentifying an existing object (e.g., call a 'cup' a 'bowl').

B. Reasoning Correctness Errors

- reasoning_conclusion: Your task is to rewrite the text by making a hasty generalization. The method is to grab a single detail from the text (such as one person running) and then extrapolate it into a grand conclusion that seems plausible but is actually very arbitrary (such as concluding this must be a professional marathon training session). Ensure you use reasoning words like 'so' or 'therefore' to connect this flawed logical chain.
- reasoning_causal: Your task is to confuse correlation with causation. Find two things in the text that might happen concurrently but have no direct causal link, and then forcibly establish a cause-and-effect relationship between them using words like 'because' or 'leading to'. For instance, you could take the action 'a man holding an umbrella indoors' and incorrectly present it as the cause for 'a power outage in the room', creating a deceptive misattribution.
- reasoning_prediction: Your task is to make an overly arbitrary and confident prediction based on extremely limited information. You need to take a trivial, small action (such as a child stacking blocks) and lead it directly to a very grand and distant future (such as predicting they will surely become a great architect).

This prediction needs to sound physically possible, but its logical leap must be huge and baseless.

- reasoning_procedural: Your task is to, within a normal process description, insert a step that seems plausible but is actually superfluous or based on pseudoscience. This step must not cause the entire process to fail but will make it logically flawed. For instance, when describing the process of brewing tea, you could add a step claiming that 'before adding water, you need to let the tea leaves sit for a minute to absorb the room's energy,' thereby making the process imprecise.

- reasoning_comparison: Your task is to construct a faulty analogy. You need to find two things that have only minor superficial similarities but are completely different in their core essence to make a comparison, and then draw a misleading conclusion from it. A classic example is to compare 'company strategy' to a 'car engine' and then argue that 'as long as there's enough fuel (funding), success is guaranteed,' an analogy that deliberately ignores more critical factors like the 'steering wheel (strategic direction)'

C. External Knowledge Errors

- knowledge_entity: If the response mentions a real-world named entity, rewrite it by corrupting that entity (e.g., 'Eiffel Tower in London').

- knowledge_context: Rewrite the response to place an object or scene in a wrong historical or technological context.

- knowledge_definition: If the response defines a concept, rewrite it to provide an incorrect definition.

- knowledge_attribution: If the response mentions a creation or quote, misattribute it to the wrong source.

C Sensitivity Analysis on Component Weights

To further validate the robustness of EVIAN and explore the impact of different quality dimensions on downstream performance, we conducted a sensitivity analysis by adjusting the aggregation weights refer to Section 3.4.

We designed three biased weighting schemes, where one dimension is assigned a dominant weight of 60% ($w = 0.6$), while the remaining two dimensions are assigned 20% ($w = 0.2$) each. The configurations are:

- **Vision-Centric:** $w_V = 0.6, w_L = 0.2, w_K = 0.2$. Prioritizes visual fidelity.

- **Reason-Centric:** $w_L = 0.6, w_V = 0.2, w_K = 0.2$. Prioritizes logical soundness.

- **Knowledge-Centric:** $w_K = 0.6, w_V = 0.2, w_L = 0.2$. Prioritizes factual correctness.

We selected the top-10k samples using each scheme and fine-tuned the model under the same protocol as our main experiments. The results are reported in Table 6.

The results suggest that emphasizing logical coherence is particularly beneficial under the evaluated setting, as the *Reason-Centric* configuration attains the highest average score (68.28) and shows consistent advantages on reasoning-intensive benchmarks. In contrast, the *Vision-Centric* scheme demonstrates strong effectiveness in reducing hallucinations, as reflected by its POPE performance, but does not consistently match the balanced EVIAN setting, indicating that visual accuracy alone is insufficient for broader capability gains. The *Know-Centric* configuration yields comparatively lower performance (65.99), suggesting potential trade-offs between strict factual filtering and the retention of samples that support multi-step reasoning. Overall, these findings indicate that while balanced weighting provides a robust default, the EVIAN framework allows practitioners to adjust score weights to prioritize task-specific objectives, such as hallucination mitigation via S_V or enhanced reasoning via S_L.

D Generalizability Across Model Architectures

Motivation: Addressing Auditor–Target Architectural Bias. A potential concern regarding our main results is the possibility that EVIAN overfits to the inductive biases shared by the auditor (Qwen2.5-VL-7B) and the fine-tuned target model (Qwen2-VL-2B). Since these two models belong to the same family, one could argue that EVIAN does not truly identify intrinsically high-quality samples, but instead selects data that align with Qwen-specific reasoning heuristics. This concern

Table 6: Sensitivity analysis of EVIAN under different weighting schemes. We report performance on 10K-sample subsets selected by prioritizing different quality dimensions (60% weight). **Reason-Centric** selection achieves the best average performance, highlighting the importance of logical coherence. **Bold** denotes the best result, and underlined denotes the second best.

Configuration	MME	MMBench	SEEDBench	ScienceQA	A-OKVQA	POPE	Avg
Vision-Centric ($S_V \uparrow$)	1779.15	0.6109	0.6163	0.6905	0.7160	**77.76**	67.45
Reason-Centric ($S_L \uparrow$)	**1803.53**	**0.6373**	**0.6312**	**0.7030**	**0.7228**	75.81	**68.28**
Know-Centric ($S_K \uparrow$)	1748.02	0.5995	0.6034	0.6731	0.7021	75.74	65.99

is particularly acute for the **Logical Coherence** dimension, which depends on the auditors internal reasoning pathways and, unlike Visual Consistency, could in principle be architecture-specific. To rigorously stress-test this hypothesis, we adopt **InternVL2-2B** (Chen et al., 2024f) as an alternative target model. InternVL2 differs substantially from Qwen in both its vision encoder and language backbone, while maintaining a comparable 2B scale, allowing us to isolate architectural effects from capacity effects.

Experimental Setup. We follow the exact protocol of Section 5.1 and fine-tune InternVL2-2B using 10k-sample subsets selected by three representative methods: Random Sampling, BLIP-2, and EVIAN. This setup ensures that the only change relative to the main experiment is the replacement of the downstream architecture. Any shift or preservation of performance patterns therefore directly reflects whether EVIAN captures genuinely architecture-agnostic indicators of data quality, or merely capitalizes on Qwen-specific inductive biases.

Results and Interpretation. The results are presented in Table 7. Across all benchmarks, the performance ordering observed in our main experiments, **EVIAN** $>$ **BLIP-2** $>$ **Random**, remains fully preserved when InternVL2-2B is used as the downstream model. The consistency of this ranking across architectures strongly contradicts the architectural-bias hypothesis and indicates that EVIAN captures transferable signals of data quality rather than family-specific preferences. Notably, the largest gains again appear on reasoning-centric tasks such as **ScienceQA** (92.46 vs. 91.32) and **A-OKVQA** (76.24 vs. 74.67), demonstrating that the logical flaws identified by the Qwen-based auditor are also detrimental for a model with a completely different architecture. This shows that **logical coherence is a genuinely universal data-quality factor**, and that enforcing it yields robust improvements regardless of the model family used for fine-tuning.

E Effectiveness on Real-World Data

To definitively address potential concerns regarding the generalizability of our method beyond the defect-injected benchmark, we conducted an additional evaluation on the original, unmodified 300K source dataset. This experiment aims to verify whether the data quality improvements observed with EVIAN stem solely from filtering out artificial noise or if the framework captures intrinsic data quality signals applicable to real-world scenarios. Our findings confirm the latter: the model fine-tuned on the small, EVIAN-curated subset outperforms the model trained on the entire original dataset, validating the efficacy of our pipeline in extracting high-value samples from standard distribution data.

Experimental Setup. We established a rigorous baseline using the "Origin Full" configuration, where the Qwen2-VL-2B model was fine-tuned on the complete pool of 300,000 raw samples derived from the eight foundational datasets prior to any defect injection. This represents the upper bound of data quantity without our intervention. This baseline was compared against the "EVIAN (10K)" model, which was fine-tuned on the top-10,000 samples selected by our framework. Both models utilized identical model architectures and training hyperparameters and were evaluated using the standard VLMEvalKit suite to ensure a fair comparison focused strictly on data efficiency and quality.

Results. As shown in Table 8, the EVIAN-selected subset achieves a higher average score of 70.20 compared to 70.07 for the full original dataset, despite utilizing only approximately 3.3% of the total training volume. While the full dataset exhibits a slight advantage on knowledge-heavy benchmarks such as MMBench, SEEDBench, and

Table 7: Cross-architecture validation using **InternVL2-2B**. We report the performance on 10K subsets. **Bold** denotes the best result, and underlined denotes the second best. Note that the MME and MMBench columns have been mapped to their standard metric scales based on the raw logs. Our **EVIAN** consistently outperforms baselines even on a different model architecture.

Model	MME	MMBench	SEEDBench	ScienceQA	A-OKVQA	POPE	Avg
Random	1776.87	0.3857	0.6064	0.7313	0.5467	87.18	62.94
BLIP-2	1754.31	0.6945	0.6888	0.9132	0.7467	87.10	75.68
EVIAN (Ours)	**1796.41**	**0.7096**	**0.6971**	**0.9246**	**0.7624**	**87.37**	**76.82**

Table 8: Comparison between the full original dataset and the EVIAN-selected subset. The "Origin Full" model is trained on the unmodified 300K source data. EVIAN achieves a higher average score using only 1/30th of the data, demonstrating superior data efficiency and quality verification.

Model	MME	MMBench	SEEDBench	ScienceQA	A-OKVQA	POPE	Avg
Origin Full (300K)	1715.03	**0.6734**	**0.6416**	**0.7287**	**0.7511**	79.68	70.07
EVIAN (10K)	**1876.89**	0.6463	0.6359	0.7115	0.7493	**79.87**	**70.20**

ScienceQA, likely attributable to the broader exposure to diverse facts inherent in larger data scales, EVIAN secures a substantial lead on MME (+161.86) and maintains highly competitive performance on hallucination and reasoning tasks. This empirical evidence reinforces the "less is more" principle, suggesting that fine-grained auditing can effectively identify and prioritize high-utility samples even within generally high-quality data distributions.