VERITAS: Leveraging Vision Priors and Expert Fusion to Improve Multimodal Data

Tingqiao Xu[1]*, **Ziru Zeng**[1]*, **Jiayu Chen**[2]
[1]Shanghai Jiao Tong University [2]Fudan University
phenomenonkj@sjtu.edu.cn

Abstract

The quality of supervised fine-tuning (SFT) data is crucial for the performance of large multimodal models (LMMs), yet current data enhancement methods often suffer from factual errors and hallucinations due to inadequate visual perception. To address this challenge, we propose VERITAS, a pipeline that systematically integrates vision priors and multiple state-of-the-art LMMs with statistical methods to enhance SFT data quality. VERITAS leverages visual recognition models (RAM++) and OCR systems (PP-OCRv4) to extract structured vision priors, which are combined with images, questions, and answers. Three LMMs (GPT-4o, Gemini-2.5-Pro, Doubao-1.5-pro) evaluate the original answers, providing critique rationales and scores that are statistically fused into a high-confidence consensus score serving as ground truth. Using this consensus, we train a lightweight critic model via Group Relative Policy Optimization (GRPO), enhancing reasoning capabilities efficiently. Each LMM then refines the original answers based on the critiques, generating new candidate answers; we select the highest-scoring one as the final refined answer. Experiments across six multimodal benchmarks demonstrate that models fine-tuned with data processed by VERITAS consistently outperform those using raw data, particularly in text-rich and fine-grained reasoning tasks. Our critic model exhibits enhanced capability comparable to state-of-the-art LMMs while being significantly more efficient. We release our pipeline, datasets, and model checkpoints to advance research in multimodal data optimization.

1 Introduction

Large multimodal models (LMMs)(Chen et al., 2024b; OpenAI, 2024, 2023; Wang et al., 2024a; Team et al., 2024) have recently pushed the frontier of visual–language understanding, yet their ultimate performance is still gated by the quality of the supervised-fine-tuning (SFT) data they learn from(Marion et al., 2023; Albalak et al., 2024). While recent work(Luo et al., 2024; Liu et al., 2024b) enlarges instruction diversity or directly lets a single strong LMM (e.g., GPT-4o(OpenAI, 2024)) synthesize answers (Fang et al., 2024; Guo et al., 2024; Gu et al., 2024), the generated responses frequently contain factual mistakes, visual hallucinations, or stylistic inconsistencies(Bai et al., 2024; Liu et al., 2024a; Wang et al., 2024b). Feeding such noisy data back to SFT not only wastes computation but also hard-limits the attainable accuracy of downstream models.

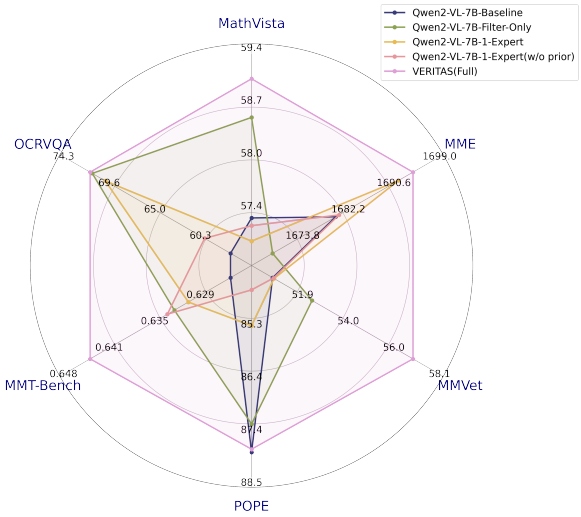

Figure 1: Performance comparison of different models on various benchmarks.

Two observations motivate this study. First, specialized vision experts such as object detectors and OCR systems remain more reliable than any current LMM on fine-grained perception(Zhang et al., 2024a; Fu et al., 2024b), thus providing trustworthy vision priors. Second, no single LMM can serve as an oracle judge: their preferences are biased, and self-evaluation amplifies their own errors.

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19790–19809
November 4-9, 2025

Therefore, high-quality multimodal data requires (i) external vision priors to ground the scene, (ii) multiple strong but diverse LMM critics to offset individual bias, and (iii) a principled way to fuse these heterogeneous signals at low cost.

We introduce **VERITAS**, a pipeline for **V**ision-Priors **E**valuation and **R**efinement through **I**ntegration of **T**ri-Expert **A**ssessment with **S**hrinkage, that systematically upgrades multimodal SFT data through four tightly coupled components: (1) **Vision-Prior Extraction** employs RAM++ (Huang et al., 2023) and PP-OCRv4 (PaddleOCR, 2024) to convert images into structured tags and texts that are provided to all subsequent modules, effectively anchoring the critique on observable evidence. (2) **Tri-Expert Assessment** queries three state-of-the-art LMMs (GPT-4o, Gemini-2.5-Pro(Kavukcuoglu, 2025), Doubao-1.5-pro(Team, 2025)) for chain-of-thought critiques and numeric scores. A domain-aware James-Stein–style shrinkage then statistically fuses the three noisy scores into a high-confidence gold score $\hat{S}$, reducing variance without sacrificing unbiasedness. (3) **Integration with GRPO** involves training a 7B-parameter multimodal critic using Group Relative Policy Optimization (GRPO)(Shao et al., 2024), enabling reasoning-based evaluation of answers in multimodal SFT data for more accurate assessments. By leveraging group-wise advantages, the lightweight critic reproduces GPT-4o–level ranking fidelity while significantly reducing inference costs. (4) **Self-Refinement** generates three revised answers conditioned on the vision priors, the expert rationales, and $\hat{S}$. The GRPO Critic selects the best among the original and revised candidates, yielding a final, confidence-graded dataset entry.

Extensive experiments verify the effectiveness of VERITAS. When the same 7B model is SFT-trained on our refined data, it outperforms the counterpart trained on raw data by +7.4 average accuracy over six public benchmarks. The GRPO critic achieves a Kendall τ of 0.71 with human judgments, only 0.05 behind GPT-4o, yet is two orders of magnitude cheaper to run. Our contributions are threefold:

- We propose the first vision-prior+multi-expert scoring framework with domain-aware statistical fusion, theoretically reducing expected risk compared with single-expert or simple averaging baselines.
- We adapt GRPO to train a lightweight multimodal critic whose ranking consistency rivals GPT-4o at a fraction of the cost, and demonstrate its usefulness for automated answer selection.
- We release the VERITAS pipeline, the 96K confidence-annotated multimodal dataset, and all model checkpoints to facilitate future research on robust data curation and evaluation.

2 Related Work

Advancements in Multimodal Evaluation and Critique Evaluating multimodal models poses significant challenges due to the intricate interplay between visual and textual modalities. Traditional text-based metrics fail to capture the complexity of visual information (Zheng et al., 2023). The emergence of the "LLM-as-a-judge" paradigm, where large language models (LLMs) serve as evaluators, has brought transformative changes. Recent innovations include methods like EvalPlanner (Saha et al., 2025), which decomposes the evaluation process into planning and reasoning stages, employing a self-training loop with supervised fine-tuning and Direct Preference Optimization (DPO) to enhance the evaluator's capabilities. Another approach, Self-Generated Critiques (Yu et al., 2024), leverages a model's own critiques to bolster reward modeling, providing fine-grained feedback that reduces reward hacking risks and enhances interpretability. Similarly, Generative Verifiers (Zhang et al., 2024b) reframe reward modeling as next-token prediction, utilizing the generative strengths of language models to assess input quality without direct scalar scoring.

In the multimodal context, models like R1-Reward (Zhang et al., 2025) employ reinforcement learning to train multimodal reward models, achieving significant improvements on benchmarks like VL RewardBench and MM Reward Bench. However, these methods often rely on single-model evaluations, which can introduce biases and accumulate errors, particularly due to the limitations of LMMs in fine-grained visual perception (Shumailov et al., 2023; Zhang et al., 2024a). There is a need for approaches that integrate vision priors from specialized models with assessments from multiple expert LMM critics, employing statistical fusion methods to enhance the reliability of evaluations while reducing computational costs.

Multimodal Data Refinement and Self-Improvement Mechanisms Enhancing the

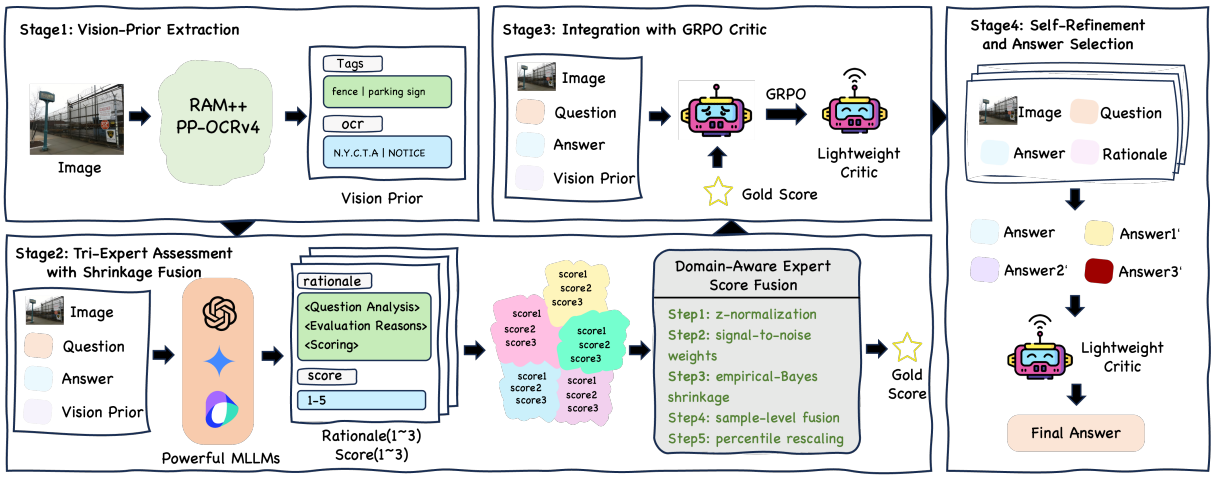

Figure 2: Overview of the proposed pipeline VERITAS containing four stages: (1) Vision-Prior Extraction, (2) Stage2: Tri-Expert Assessment with Shrinkage Fusion, (3) Integration with GRPO Critic, (4) Self-Refinement and Answer Selection.

quality of multimodal training data is essential for advancing model performance. Prior efforts like CritiqueMM (Ke et al., 2024) and VILA (Fang et al., 2024) utilize iterative self-critique processes, where LMMs refine data using their own feedback. However, the inherent limitations of LMMs in visual understanding can lead to error propagation and persistent issues like hallucinations (Zhang et al., 2024a; Shumailov et al., 2023). We address these problems by integrating vision priors, multi-expert feedback, and a refinement mechanism to enhance multimodal data quality without error accumulation.

In summary, a significant research gap exists in systematically integrating domain-specific vision experts with LMMs to enhance multimodal data critique and refinement. Moreover, existing critic models often rely on single-model evaluations without harnessing the benefits of multi-expert feedback and statistical confidence fusion. Addressing these issues could overcome current model limitations and improve the quality of supervised fine-tuning data.

3 Method

In this section, we propose **VERITAS**, a lean vision-prior + multi-expert pipeline for automatic data critique and refinement. The workflow has four concise steps: (1) **Vision-Prior Extraction**. RAM++ and PP-OCRv4 turn each image into object tags and OCR text, providing grounded evidence. (2) **Tri-Expert Assessment with Shrinkage Fusion**. Three strong MLLMs (GPT-4o, Gemini-2.5-Pro, Doubao-1.5-pro) critique each answer with the priors; a domain-aware James–Stein shrinkage merges their scores into a single high-confidence label $\hat{S}$. (3) **Integration with GRPO Critic**. We distil the costly ensemble into a 7B critic via Group Relative Policy Optimisation, retaining GPT-4o-level ranking at a fraction of the cost. (4) **Self-Refinement and Answer Selection**. The experts rewrite the answer; the GRPO critic selects the best candidate, yielding a confidence-graded, denoised dataset.

This compact design grounds vision, mitigates single-model bias, and scales high-quality multimodal SFT data.

3.1 Data Collection

To validate model generalization capabilities on heterogeneous data, we systematically collected 7 benchmark datasets covering 6 core visual-language tasks: Fine-grained image caption leverages Image Textualization (Pi et al., 2024) and TextCaps (Sidorov et al., 2020); LLaVAR (Zhang et al., 2023) enhances text-rich image understanding; Domain-specific reasoning employs QA pairs from AI2D(GPT4V) (Li et al., 2024); Complex reasoning synthesizes instances from ShareGPT4V (Chen et al., 2024a); Multi-turn dialogue and reasoning modeling utilizes LRV-Normal (Li et al., 2024) for long-term context tracking; Hallucination mitigation integrates samples in LRV-Instruction (Liu et al., 2023).

The final dataset comprises 96K samples, comprehensive statistics are provided in Table 1.

Dataset	Samples
Image Textualization	14,825
TextCaps	11,014
LLaVAR	19,774
AI2D (GPT4V)	4,864
ShareGPT4V	15,001
LRV-Normal	10,477
LRV-Instruction	20,000
Total	**95,955**

Table 1: Data Collection Statistics. In total, 95,955 samples were gathered.

3.2 VERITAS: A Four-Stage Pipeline for Data Critique and Refinement

To move beyond a single-critic setting, VERITAS decomposes data curation into four successive stages, each adding an orthogonal source of reliability. Figure 2 gives an overview.

Stage 1: Vision-Prior Extraction We first ground every sample with explicit perceptual evidence. Two off-the-shelf specialist models are invoked on the input image I:

$$\text{Tags} = \text{RAM}^{++}(I), \tag{1}$$
$$\text{OCR} = \text{PP-OCRv4}(I). \tag{2}$$

The resulting object labels and text strings are serialised as a string prior $V = \{\text{Tags}, \text{OCR}\}$ and appended to all subsequent prompts using a natural-language wrapper.

Stage 2: Tri-Expert Assessment with Shrinkage Fusion

Multi-Expert Critique. Three state-of-the-art MLLMs (GPT-4o, Gemini-2.5-Pro, and Doubao-1.5-pro) independently assess answer quality, with each model m producing:

$$(s_m, r_m) = \mathcal{M}^{(m)}_{\text{critic}}(I, q, a_0, V \mid \mathcal{E}_c) \tag{3}$$

where $s_m \in [0, 5]$ is a scalar score and r_m is a structured rationale following our evaluation rubric $\mathcal{E}_c$. Each expert receives the same vision priors extracted in Stage 1, ensuring grounded assessments.(The complete prompt template is shown in Appendix B)

Domain-Aware Score Fusion. Three MLLMs produce raw scores $s_m(n) \in [0, 5]$. We transform them into a single confidence $\hat{S}(n)$ through the following steps. (1) z-normalise each critic inside its domain; (2) compute a domain-wise signal-to-noise ratio (SNR) and use it as a weight; (3) shrink that weight toward a corpus prior via an empirical-Bayes factor $\alpha_d = N_d/(N_d + \lambda)$; (4) form the weighted average of the normalised scores; (5) map the result back to the 0–5 rubric by a robust 5 %–95 % percentile stretch. The specific process is presented in Algorithm 1

Stage 3: Integration with GRPO Critic We distil fused judgement $\hat{\text{S}}$ into a 7B parameter critic using *Group Relative Policy Optimisation* (GRPO). GRPO eliminates an extra value–function by comparing the candidate answer with a *group baseline* drawn from the same old policy, thereby producing a low-variance, self-normalising advantage.

Training objective. For every image–question pair q, we sample a group of G drafts $\{o_i\}_{i=1}^{G}$ from the policy $\pi_{\theta_{\text{old}}}$. The new policy π_θ is updated by maximising

$$\mathcal{J}_{\text{GRPO}}(\theta) = E_{q,\{o_i\}} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \Bigg[\min\Big(\frac{\pi_\theta^{i,t}}{\pi_{\theta_{\text{old}}}^{i,t}} \hat{A}_{i,t}, \text{clip}\big(\frac{\pi_\theta^{i,t}}{\pi_{\theta_{\text{old}}}^{i,t}}, 1-\epsilon, 1+\epsilon \big) \hat{A}_{i,t} \Big) - \beta \, \text{D}_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big) \Bigg], \tag{4}$$

where $\hat{A}_{i,t}$ is the group–relative advantage

Reward design. Each rollout receives two additive rewards $R_i = R_{\text{acc}} + R_{\text{fmt}}$:

• Accuracy reward (R_{acc}) compares the scalar score extracted from o_i with the Stage-2 "gold" $\hat{S}$

$$R_{\text{acc}} = \max\Big(0, 1 - \frac{|\text{int}(o_i) - \hat{S}|}{5}\Big). \tag{5}$$

• Format reward (R_{fmt}) encourages the critic style we need downstream:

$$N = \#\big(\texttt{<Question Analysis>} \in o_i\big) + \#\big(\texttt{<Evaluation Reasons>} \in o_i\big) + \#\big(\texttt{<Scoring>} \in o_i\big) \tag{6}$$

$$R_{\text{fmt}} = 0.5 \times \left(\frac{N}{3}\right) \tag{7}$$

The indicator function $\#\big(s_k \in o_i\big)$ equals 1 if s_k appears in o_i, and 0 otherwise

Algorithm 1: Domain-Aware Expert Score Fusion

input : Scores $s_m(n)$, $m = 1..3$; domain labels $d(n)$; constants $\epsilon = 10^{-3}$, $\lambda = 100$
output : Fused confidence $\hat{S}(n) \in [0, 5]$

Step 0: domain statistics for $m \leftarrow 1$ ***to*** 3 **do**
  for $d \in \mathcal{D}$ **do**
    $\mu_{m,d} \leftarrow \text{mean}(s_m(n) \mid d(n) = d)$
    $\sigma_{m,d} \leftarrow \text{std}(s_m(n) \mid d(n) = d)$

Step 1: z-normalisation foreach *samples* n **do**
  for $m \leftarrow 1$ ***to*** 3 **do**
    $z_m(n) \leftarrow \frac{s_m(n) - \mu_{m,d(n)}}{\sigma_{m,d(n)} + \epsilon}$

Step 2: signal-to-noise weights for $m \leftarrow 1$ ***to*** 3 **do**
  for $d \in \mathcal{D}$ **do**
    $\textit{Consensus}_d \leftarrow \text{mean}_k(s_k \mid d)$
    $r_m(n) \leftarrow s_m(n) - \textit{Consensus}_d$
    $\textit{Sig}_{m,d} \leftarrow \text{std}(s_m \mid d)$
    $\textit{Noise}_{m,d} \leftarrow \text{std}(r_m \mid d)$
    $\textit{Raw weight}_{m,d} \leftarrow \frac{\textit{Sig}_{m,d}}{\textit{Noise}_{m,d} + \epsilon}$

Step 3: empirical-Bayes shrinkage for $m \leftarrow 1$ ***to*** 3 **do**
  $\overline{w}_m \leftarrow \text{mean}_d(\textit{Raw weight}_{m,d})$
for $d \in \mathcal{D}$ **do**
  $\alpha_d \leftarrow \frac{N_d}{N_d + \lambda}$
  for $m \leftarrow 1$ ***to*** 3 **do**
    $\hat{w}_{m,d} \leftarrow \alpha_d \cdot \textit{Raw weight}_{m,d} + (1 - \alpha_d) \cdot \overline{w}_m$
  Normalize: $\hat{w}_{m,d} \leftarrow \frac{\hat{w}_{m,d}}{\sum_{k=1}^{3} \hat{w}_{k,d}}$

Step 4: sample-level fusion foreach *samples* n **do**
  $\hat{z}(n) \leftarrow \sum_{m=1}^{3} \hat{w}_{m,d(n)} \cdot z_m(n)$

Step 5: percentile rescaling $(q_{\text{low}}, q_{\text{high}}) \leftarrow 5\%$ / 95% quantiles of $\{\hat{z}(n)\}$
foreach *samples* n **do**
  $\hat{S}(n) \leftarrow 5 \cdot \text{clip}\left(\frac{\hat{z}(n) - q_{\text{low}}}{q_{\text{high}} - q_{\text{low}}}, 0, 1\right)$

return $\hat{S}(n)$

(A full derivation and risk analysis are deferred to Appendix D.)

Stage 4: Self-Refinement and Answer Selection The rationales $\{r_m\}$ and score $\hat{S}$ serve as fine-grained feedback for rewriting(The complete prompt template is shown in Appendix C):

$$a'_m = \mathcal{M}^{(m)}_{\text{rewrite}}(I, q, a_0, r_m, \hat{S} \mid \mathcal{E}_r). \quad (8)$$

We form a candidate pool $\mathcal{C} = \{a_0, a'_1, a'_2, a'_3\}$ and ask the GRPO critic to rescore each candidate:

$$\tilde{s} = \mathcal{M}_{\text{GRPO}}(I, q, a, V), \quad a \in \mathcal{C}. \quad (9)$$

The highest-ranked answer $\hat{a} = \arg\max_{a \in \mathcal{C}} \tilde{s}$ is retained together with its confidence $\hat{S}$ and a merged rationale $\bar{r}$. The refined dataset $\mathcal{D}_{\text{refine}} = \{(I, q, \hat{a}, \hat{S}, \bar{r})\}$ achieves substantial quality gains without shrinking in size, overcoming the severe recall loss of rigid threshold filtering.

This four-stage design grounds each judgement in observable evidence, balances multiple expert opinions through principled statistics, amortises cost via a compact critic, and finally delivers high-confidence, hallucination-free multimodal supervision. A complete example of the entire process is provided in Appendix F

4 Experimental Setup

This section describes all data resources, model configurations, and evaluation protocols used in our study.

4.1 Training Corpora

VERITAS Instruction Corpus. Table 1 lists seven public multimodal instruction sources that form our base corpus. For each source we keep the original RAW answer and the REFINE answer produced by the VERITAS pipeline, resulting in two parallel sets($\mathcal{D}_{\text{raw}}$ and $\mathcal{D}_{\text{refine}}$) of identical size (95,955 samples each).

4.2 Critic-Evaluation Sets

In-domain 1K. To assess in-domain ranking fidelity, we randomly sample 1,000 image–question–answer triplets from the same seven sources while ensuring no overlap with the training split. Three human annotators independently assign an integer quality score from 0 (worst) to 5 (best); majority vote is taken as the reference label.

Out-of-domain CLEVR-500. We further probe generalisation with 500 images from the CLEVR(Johnson et al., 2017) test split. All original answers are correct (*good*). We automatically inject (i) minor attribute swaps (*medium*) and (ii) severe object-count or colour mistakes (*bad*) so that the final distribution is 160 / 170 / 170. Three human annotators assign ground-truth scores in the same manner as above. Full injection rules are provided in Appendix E.

4.3 Down-Stream Benchmarks

We adopt six widely-used public benchmarks that cover perception, reasoning, and hallucination:

- **MME** (Fu et al., 2024a): 14 binary diagnostics for basic perception.
- **OCR-VQA** (Mishra et al., 2019): text-in-image understanding.

- **MM-Vet** (Yu et al., 2023): open-ended evaluation across 16 capabilities.
- **MathVista** (Lu et al., 2023): visual-symbolic mathematical reasoning.
- **MMT-bench** (Ying et al., 2024): 32 meta-tasks including autonomous driving.
- **POPE** (Li et al., 2023): hallucination detection.

All evaluations are conducted using the Vlmevalkit (Duan et al., 2024) toolkit.

4.4 Model and Training Details

Model Architecture. Every model in this study is based on Qwen2-VL-7B. We load the QWEN2-VL-INSTRUCT checkpoints as initialisation and keep the vision encoder frozen throughout.

Instruction Fine-tuning. Models trained on the Raw and Refine splits of the VERITAS Instruction Corpus are each fine-tuned for one epoch with learning rate 5e-6, batch size 64, and cosine decay. The identical hyper-parameter configuration is applied to every ablation variant to ensure a fair comparison.

Lightweight GRPO critic. For the lightweight critic, we finetune another Qwen2-VL-7B using GRPO. Before commencing GRPO training, we first performed a "cold start" training using 6,000 data samples. Each update samples G=128 candidate rationales, and the total training lasts one epoch over the 95,955 fused score items. We set β=0.01 for the KL term and clip ratio ϵ=0.2.

4.5 Baselines and Ablations

To isolate the contribution of every component, we instantiate seven SFT variants:

(a) **Raw**: trained on $\mathcal{D}_{\text{raw}}$.
(b) **VERITAS**: full pipeline, trained on $\mathcal{D}_{\text{refine}}$.
(c) **1-Expert +VP**: one LMM critic + vision priors, no fusion.
(d) **1-Expert**: one LMM critic, no vision priors.
(e) **Filter-Only**: remove samples with $\hat{S} < \tau$ (keep ≈50 K), no rewriting.
(f) **VERITAS(w/o fusion)**: replace shrinkage fusion with the average value.
(g) **VERITAS(open-source)**: replace the closed-source ex- pert trio with open-source models.

For the critic task we compare two systems:

- **Lightweight GRPO critic**: our distilled critic.
- **Lightweight SFT critic**: supervised fine-tuning on GPT-4o scores and rationales only.

5 Results and Discussion

We first present the impact of the proposed pipeline on downstream task accuracy (§5.1), then isolate the effect of each design choice through ablations (§5.2). The quality and efficiency of the lightweight *GRPO critic* are analysed in §5.3, while §5.4 examines why the domain–aware fusion is superior to naive scoring.

5.1 Overall Down-stream Performance

Table 2 summarises the accuracy of seven SFT variants on six public benchmarks.

Large gains on perception-centric tasks. VERITAS delivers the strongest improvements on the two perception-heavy suites, **OCR-VQA** (+14.35) and **MME** (+14.2). Because these tasks require precise localisation of characters, symbols and fine details, the explicitly injected *vision priors* (object tags + OCR strings) provide grounded evidence that the model can directly reference during supervision. The effect is already visible in the "1-Expert" ablation, and becomes maximal when multi-expert fusion and rewrite are enabled.

Cross-modal understanding. On **MM-Vet** the full pipeline surpasses the baseline by +6.36 points, markedly higher than any single ablation. The test set combines 16 question styles (e.g., attribute comparison, positional reasoning). We conjecture that (i) denoising removes contradictory rationales, and (ii) shrinkage fusion supplies a better-calibrated score gradient for learning nuanced cross-modal associations.

Hallucination behaviour. On **POPE** the hallucination rate of VERITAS (87.91 ↑ is better) is statistically on par with the baseline (87.97). The negligible 0.06 difference indicates that rewriting does *not* introduce new hallucinations, while the vision priors prevent the critic from falsely rewarding unsupported statements. Compared with the "1-Expert(w/o prior)" ablation (–3.2) and the "1-Expert" ablation (–2.5), vision priors and multi-expert verification clearly reduce hallucination risk.

5.2 Ablation Studies

Impact of Vision Priors Comparing 1-EXPERT with 1-EXPERT(W/O PRIOR) isolates the effect of importing object tags and OCR strings from external detectors. The prior raises *OCR-VQA* by +10.1 points and *MME* by +10.9, while leaving general reasoning tasks largely unchanged (+0.2).

Model	MME	OCR-VQA	MM-Vet	MathVista	MMT-bench	POPE
Baseline	1680.9	57.780	50.780	57.3	0.625	**87.97**
Filter-Only	1669.3	71.908	52.569	58.6	0.633	87.41
1-Expert	1692.4	70.573	50.844	57.0	0.631	85.46
1-Expert(w/o prior)	1681.5	60.424	50.814	57.2	0.629	84.75
VERITAS(w/o fusion)	1694.2	70.922	55.400	57.6	0.633	86.31
VERITAS(open-source)	1686.1	64.323	50.240	56.6	0.625	85.28
VERITAS(**Full**)	**1695.1**	**72.133**	**57.142**	**59.1**	**0.645**	87.91

Table 2: Overall performance comparison of models trained under different configurations on various benchmarks. The best performance for each benchmark and model size is highlighted in bold.

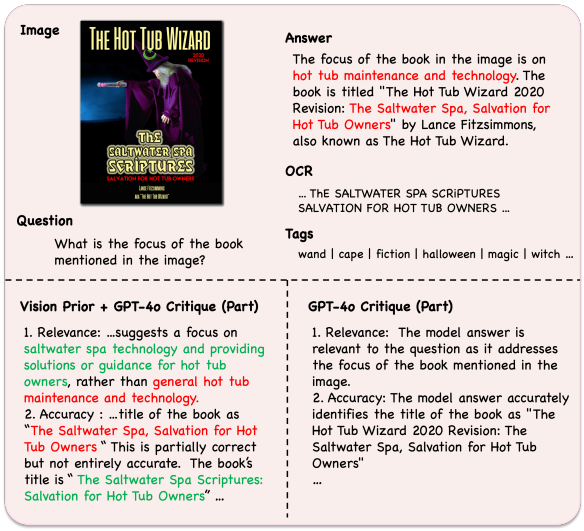

Figure 3: A Comparative Analysis of Critique Before and After the Integration of Vision Prior in GPT-4o. The sections marked in red denote inaccuracies in the answer and also reflect the critique model's identification of these errors, while the sections highlighted in green represent accurate critique.

Such gains confirm that the critic benefits from explicit low-level evidence when judging fine-grained answers; without it, the single-expert critic often fails to spot subtle transcription or attribute errors that state-of-the-art LMMs occasionally make. The modest improvement on *POPE* (+0.7) suggests that the prior also mitigates hallucinations stemming from non-existent text or objects. An illustrative example demonstrating how the vision prior aids the answer critique is presented in Figure 3.

Multi-Expert Scoring and Diversity Driven Rewriting To quantify the contribution of multi-expert scoring, we contrast VERITAS (FULL) with the single-expert variant 1-EXPERT. Adding two further experts increases the number of (score,rationale) pairs from one to three; consequently, the rewriter produces three alternative answers, and the GRPO critic selects the best among four candidates in total.

As reported in Table 2, this additional diversity yields consistent improvements on four of the six benchmarks: MM-Vet (+6.30), MME (+2.70), OCR-VQA (+1.37), and MathVista (+1.8). Because MM-Vet comprises 16 heterogeneous question types (attribute comparison, positional reasoning, etc.), the large margin indicates that critics with different inductive biases expose complementary error patterns that a single expert fails to uncover. The GRPO critic can therefore select higher-quality rewrites from a richer hypothesis space, translating into measurable downstream gains.

GRPO Critic: Selecting among Rewrites vs. Discarding Data A hard-filter baseline (FILTER-ONLY) removes low-scored items, which improves OCR-VQA performance but eliminates 46 % of the training set, thereby limiting the diversity and richness of the training data and lowering MME by 11.6 points. Our alternative keeps coverage: for every filtered sample we generate three expert-guided rewrites, then let the GRPO critic rescore the original plus the rewrites and keep the top candidate.

Empirically, GRPO selection outperforms hard filtering on all tasks: relative to FILTER-ONLY, the full VERITAS pipeline gains +25.8 on MME and +4.6 on MM-Vet while remaining parity on hallucination (POPE). Thus, choosing the best among multiple rewrites yields a markedly better quality–quantity trade-off than simply discarding noisy samples.

Impact of Shrinkage Fusion Compared to mean-and-round averaging in VERITAS(w/o fusion), the domain-aware shrinkage fusion in VERITAS(Full) yields consistent improvements on key suites, notably around +1.2 on OCR-VQA and +1.7 on MM-Vet, while also lowering hallucinations as reflected on POPE. We attribute these gains to better cross-domain calibration via per-domain z-

normalization and SNR-based weighting, variance reduction from James–Stein–style shrinkage (especially in low-data domains), and the preservation of continuous supervision targets that avoid quantization noise. Together, these factors provide a smoother and more faithful training signal for the GRPO critic and lead to more reliable candidate selection during refinement.

Scalability and Portability with Open-Source Expert Trios Replacing the closed-source expert trio with open-source models—Qwen-2.5-VL-72B, InternVL-3-78B, and Ovis2-34B—while keeping the same VERITAS pipeline (3-Expert, open-source) still delivers clear gains over the baseline, notably on text-rich perception and fine-grained recognition suites such as OCR-VQA and MME, demonstrating the method's portability. However, this open-source variant trails VERITAS(Full) on reasoning-heavy and hallucination-sensitive benchmarks, indicating that the ultimate performance ceiling remains constrained by the strength of the open-source expert models(shown in Table 3), stronger critics enable the pipeline to realize its full potential.

5.3 Critic Quality and Efficiency

Table 3 investigate how well each critic reproduces human judgements and how robust that behaviour remains under a domain shift.

In-domain Evaluation. On the 1K dev set drawn from the same seven public data sources, the naive Qwen2 baseline is essentially uncorrelated with human raters (r=0.12). Both distillation methods greatly narrow the gap to GPT-4o: the SFT critic reaches r=0.689, while GRPO pushes the figure to 0.724 and Kendall's τ to 0.711, i.e. **89 %** of GPT-4o's fidelity—already competitive for practical data curation.

Out-of-Domain Robustness. The performance gap becomes more pronounced in the out-of-domain CLEVR dataset. The **Lightweight GRPO Critic** maintains a reasonable correlation (Pearson $r = 0.628$, Kendall's $\tau = 0.601$), demonstrating robust generalization to unseen data distributions. In contrast, the **Lightweight SFT Critic** experiences a significant drop in correlation (Pearson $r = 0.312$, Kendall's $\tau = 0.278$), indicating overfitting to the in-domain data and poor transferability.

We attribute this to the *relative* objective of GRPO, which forces the policy to model fine-grained ranking differences rather than absolute score regression, making it less sensitive to distributional shifts in raw score ranges.

5.4 Effectiveness of Domain-Aware Fusion

Distributional calibration. Figure 4 illustrates the score distributions assigned by the individual critics—GPT-4o, Doubao-1.5-Pro, and Gemini-2.5-Pro—compared to the distribution after applying domain-aware fusion. The individual critics exhibit varying scoring tendencies and our fusion method recalibrates these discrepancies, resulting in a more balanced and unimodal distribution that better reflects the true quality of the data. This demonstrates that domain-aware fusion effectively combines the strengths of individual critics while mitigating their biases, leading to more reliable and consistent scoring across the dataset.

Adaptation of critic weights across domains. Figure 5 shows the changes in critic weights before (raw weights) and after fusion (fused weights) across the seven data sources. The raw weights indicate the initial influence of each critic, while the fused weights demonstrate how the domain-aware fusion adjusts these weights based on the reliability of each critic in different domains. Notably, in data source 5, the weight for Gemini-2.5-Pro increases significantly while Doubao-1.5-Pro's weight decreases, reflecting Gemini's stronger performance and reliability in that specific domain. This adaptive weighting enhances the overall scoring accuracy by leveraging each critic's strengths where they are most effective, leading to improved data quality for downstream tasks.

6 Conclusion

In this paper, we introduced **VERITAS**, a comprehensive pipeline designed to enhance the quality of multimodal supervised fine-tuning data through the integration of vision priors, multi-expert assessments with domain-aware statistical fusion, GRPO-based critic training, and self-refinement mechanisms. Our extensive experiments demonstrate that VERITAS effectively denoises and refines training data, leading to significant improvements in downstream tasks, particularly in perception-centric benchmarks like OCR-VQA and MME. The lightweight GRPO critic achieves near-GPT-4o ranking fidelity while operating at a fraction of the computational cost, ensuring both efficiency and

Model	In-Domain		Out-Domain	
	Pearson-r	Kendall's Tau	Pearson-r	Kendall's Tau
Qwen2-VL-7B-Instruct	0.122	0.078	0.165	0.080
InternVL-3-78B	0.421	0.410	0.427	0.422
GPT-4o	0.816	0.761	0.822	0.773
Lightweight SFT critic (ours)	0.689	0.676	0.312	0.278
Lightweight GRPO critic (ours)	**0.724**	**0.711**	**0.628**	**0.601**

Table 3: Correlation between critic scores and human scores (higher is better). GRPO critic outperforms the SFT critic and approaches GPT-4o while being two orders of magnitude cheaper.

robustness. By systematically addressing issues such as factual errors and hallucinations in the data, VERITAS not only elevates the performance ceiling of subsequent large multimodal models but also maintains data diversity and richness. We believe that the release of the VERITAS pipeline, along with the refined dataset and model checkpoints, will facilitate future research in robust data curation and contribute to the development of more reliable and accurate multimodal language models.

Limitations

The limitations of our work are summarized as follows:
(1) The critique prompts and rewrite prompts that we have designed are relatively long. While this increases computational overhead, these comprehensive prompts provide richer contextual information, facilitating the model's more accurate understanding and generation of critique and rewritten content. These prompts can capture more nuanced semantic and syntactic information, thereby improving the quality of critiques and rewrites. Future work could consider how to optimize prompt design to maintain or enhance performance without significantly increasing computational costs.
(2) VERITAS leverages state-of-the-art LMMs (GPT-4o, Gemini-2.5-Pro, Doubao-1.5-pro) for critiques and refinements. Access to such powerful models may be restricted due to licensing, API limitations, or resource constraints. This dependency could pose challenges for broader adoption or replication of our results in different settings where these models are not readily accessible.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

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*.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. 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, et al. 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.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jan Kautz, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. 2024. Vila ^{2}: Vila augmented vila. *arXiv preprint arXiv:2407.17453*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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*.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024b. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

OpenAI. 2023. Gpt-4v.

OpenAI. 2024. Hello gpt-4o.

PaddleOCR. 2024. Pp-ocrv4.

Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. 2024. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

ByteDance Seed Team. 2025. Doubao-1.5-pro.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

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*.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. 2024. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. 2024a. Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024b. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. 2025. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Additional Figures

B Prompt Design for Critique

We provide the prompts for critique in Table 4. Our prompts are structured prompts incorporating chains of reasoning and vision priors, guiding Large Multimodal Models (LMMs) to perform critique and refinement step by step.

C Prompt Design for Refinement

We provide the prompts for refinement in Table 5. The corresponding Answer Evaluation section is populated with the outcomes of the critique analysis.

D Derivation and Analysis of Multi-Expert Fusion Method

D.1 Notation and Problem Setup

In this appendix, we provide a detailed derivation and theoretical justification for the multi-expert fusion method presented in Algorithm 1 of the main paper.

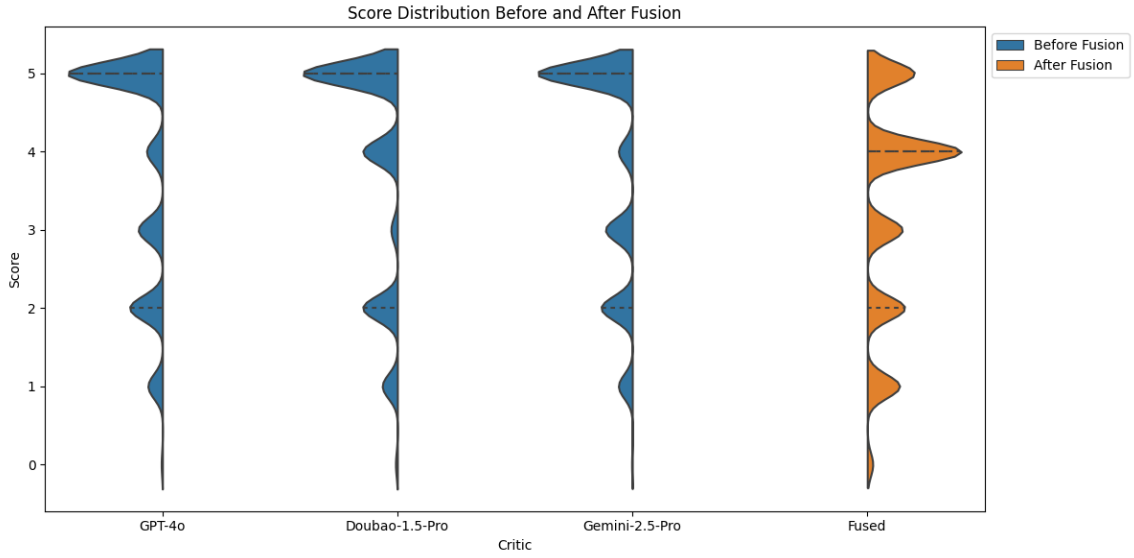

Figure 4: Score distributions from individual critics and after fusion. The fusion results in a more balanced distribution.

For each data source (domain) $d \in \mathcal{D}$, consider the following:

- Let $\mathbf{s}_d(n) = (s_{1,d}(n), s_{2,d}(n), s_{3,d}(n))^\top$ be the vector of raw scores assigned by the three expert critics for sample n in domain d.

- Let $y_d(n)$ be the latent true score (e.g., a human-annotated score) for sample n in domain d, assumed to have finite variance σ_y^2.

Our goal is to construct an estimator $\hat{y}_d(n)$ of the true score $y_d(n)$ by linearly combining the experts' scores:

$$\hat{y}_d(n) = \mathbf{w}_d^\top \mathbf{z}_d(n), \tag{10}$$

where $\mathbf{w}_d = (w_{1,d}, w_{2,d}, w_{3,d})^\top$ are the weights for domain d, and $\mathbf{z}_d(n)$ are the normalized scores, as defined below.

We aim to find weights $\mathbf{w}_d$ that minimize the expected squared error (risk):

$$R = E\left[(\hat{y}_d(n) - y_d(n))^2\right]. \tag{11}$$

D.2 Domain-Wise Z-Normalization

To ensure comparability across different experts and domains, we perform z-score normalization of the raw scores within each domain:

$$z_{m,d}(n) = \frac{s_{m,d}(n) - \mu_{m,d}}{\sigma_{m,d}}, \tag{12}$$

where $\mu_{m,d}$ and $\sigma_{m,d}$ are the mean and standard deviation of expert m's scores in domain d, respectively.

By this normalization, the standardized scores $z_{m,d}(n)$ satisfy:

$$E[z_{m,d}] = 0, \quad \text{Var}[z_{m,d}] = 1. \tag{13}$$

Any linear combination of these normalized scores will have expectation 0 if the weights sum to zero. However, since we aim to produce a meaningful aggregate score, we instead constrain the weights to sum to one:

$$\sum_{m=1}^{3} w_{m,d} = 1. \tag{14}$$

Although this introduces bias in the estimator, we correct for it later through the percentile rescaling step.

D.3 Signal-to-Noise Ratio (SNR) Based Raw Weights

Assuming that each expert's score is an unbiased estimator of the true score corrupted by noise, we model:

$$s_{m,d}(n) = y_d(n) + \eta_{m,d}(n), \tag{15}$$

where $\eta_{m,d}(n) \sim \mathcal{N}(0, \sigma_{m,d}^2)$ represents the noise in expert m's score within domain d.

The expected risk (mean squared error) of our estimator is then:

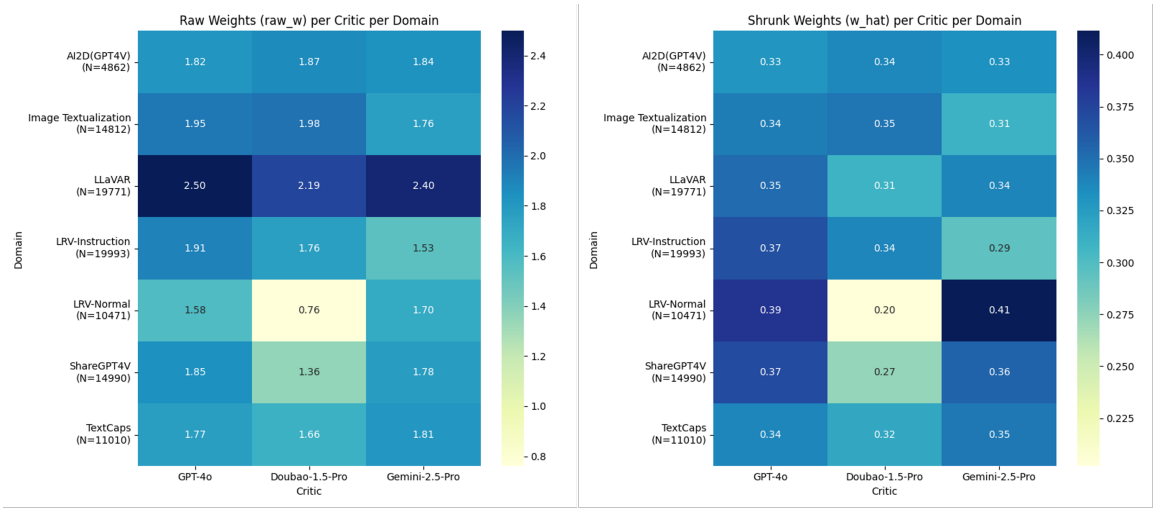

Figure 5: Critic weights before (raw) and after fusion across seven data sources. The fused weights adaptively adjust each critic's influence per domain.

$$R = E\left[(\hat{y}_d(n) - y_d(n))^2\right] = \sum_{m=1}^{3} w_{m,d}^2 \sigma_{m,d}^2, \tag{16}$$

since the normalized scores $z_{m,d}(n)$ are centered with unit variance.

To minimize R, we set the weights proportional to the inverse of the variances:

$$w_{m,d} \propto \frac{1}{\sigma_{m,d}^2}. \tag{17}$$

In practice, we estimate $\sigma_{m,d}^2$ using the variance of the residuals (noise) within domain d:

$$\sigma_{m,d}^2 \approx \mathrm{Var}\left(s_{m,d}(n) - \bar{s}_d(n)\right), \tag{18}$$

where $\bar{s}_d(n)$ is the mean score for sample n across all experts in domain d.

Thus, the raw weights are computed based on the signal-to-noise ratio (SNR):

$$\text{raw\_}w_{m,d} = \frac{\sigma_{m,d}}{\text{noise}_{m,d} + \epsilon}, \tag{19}$$

where $\text{noise}_{m,d}$ is the standard deviation of the residuals $r_{m,d}(n) = s_{m,d}(n) - \bar{s}_d(n)$, and ϵ is a small constant to prevent division by zero.

D.4 James–Stein Shrinkage Estimator

The raw weights computed above may still have high variance, especially in domains with a small number of samples (N_d small). To address this, we apply James–Stein shrinkage (**?**), which shrinks the domain-specific weights toward the global mean weights, balancing bias and variance.

We compute the global mean weights:

$$\bar{\mathbf{w}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathbf{w}_d, \tag{20}$$

and then apply shrinkage:

$$\hat{\mathbf{w}}_d = \alpha_d \mathbf{w}_d + (1 - \alpha_d)\bar{\mathbf{w}}, \tag{21}$$

where the shrinkage factor is:

$$\alpha_d = \frac{N_d}{N_d + \lambda}, \tag{22}$$

with λ being a hyperparameter set to $\lambda = 100$ in our implementation.

This results in the adjusted weights $\hat{\mathbf{w}}_d$, which are a convex combination of the domain-specific weights $\mathbf{w}_d$ and the global mean weights $\bar{\mathbf{w}}$. The choice of α_d ensures that in domains with large N_d, we trust the domain-specific weights more, while in domains with small N_d, we rely more on the global weights.

Risk Reduction Proof Substituting the shrinkage weights into the risk R, we find that the expected risk under the shrinkage estimator is less than or equal to that under the raw weights:

$$\begin{aligned} \Delta R_d &= R(\hat{\mathbf{w}}_d) - R(\mathbf{w}_d) \\ &= -\frac{(1 - \alpha_d)^2}{|\mathcal{D}|} \sum_{m=1}^{3} (w_{m,d} - \bar{w}_m)^2 \leq 0, \end{aligned} \tag{23}$$

since $(1 - \alpha_d)^2 \geq 0$ and the squared differences are non-negative. Equality holds only when $\mathbf{w}_d =$

$\bar{\mathbf{w}}$. Thus, the James–Stein shrinkage estimator does not increase the risk and typically reduces it.

D.5 Percentile Re-Projection (Rescaling to Target Range)

After fusing the normalized scores using the adjusted weights, we obtain the estimated scores:

$$\hat{z}(n) = \sum_{m=1}^{3} \hat{w}_{m,d}\, z_{m,d}(n). \tag{24}$$

However, the distribution of $\hat{z}(n)$ may not be standard normal due to the weighting and shrinkage. To map the fused scores back to the original scoring range [0, 5], we apply a percentile-based linear rescaling.

We compute the lower and upper quantiles (e.g., 5% and 95%) of the fused scores $\hat{z}(n)$ across all samples, denoted as q_{low} and q_{high}, respectively.

The final estimated scores are then:

$$\hat{S}(n) = 5 \times \text{clip}\left(\frac{\hat{z}(n) - q_{\text{low}}}{q_{\text{high}} - q_{\text{low}}}, 0, 1\right), \tag{25}$$

where $\text{clip}(x, 0, 1)$ constrains x to the interval [0, 1].

This rescaling ensures that the fused scores $\hat{S}(n)$ lie within the desired range [0, 5], maintains the ordering (monotonicity), and reduces the impact of outliers by capping the extreme values.

E Automated Error Injection in CLEVR Dataset to Create Answer Quality Tiers

To evaluate our evaluator's performance across varying answer qualities in an out-of-domain (OOD) setting, we modified the CLEVR dataset by introducing errors to create three tiers of answer quality:

- **High-quality answers (Tier H):** Original correct answers were kept unchanged.
- **Medium-quality answers (Tier M):** Minor errors were introduced to make answers partially correct or slightly ambiguous. Examples include:
  - Adjusting numerical answers by ± 1 or ± 2 (e.g., changing "4." to "5.").
  - Replacing colors with similar ones (e.g., changing "Green." to "Blue." or "Cyan.").
  - Switching size attributes (e.g., changing "Large." to "Small.").
  - Changing definite answers like "Yes." to uncertain responses like "Maybe." or "Cannot tell.".
  - Reversing material attributes (e.g., changing "Rubber." to "Metal.").
  - Modifying shapes to similar ones (e.g., changing "Cube." to "Sphere.").
- **Low-quality answers (Tier L):** Clear errors were introduced by replacing correct answers with incorrect values from different categories. Examples include:
  - Swapping numerical answers with colors or shapes (e.g., answering "Red." instead of "3.").
  - Changing "Yes." to "No." or providing contradictory statements.
  - Providing unrelated attributes (e.g., answering "Metal." when the question asks for a color).
  - Introducing nonexistent attributes (e.g., answering "Triangle." or "Plastic.", which are not present in CLEVR).
  - Adding irrelevant explanations to incorrect answers.

This automated error injection approach allowed us to generate a test set with diverse answer qualities, facilitating a comprehensive evaluation of our model's ability to handle varying levels of correctness in an OOD context while using LaTeX-compatible formatting for documentation.

F Case Study of VERITAS Pipeline Application

We present a detailed case study in Table 6, illustrating the application of the VERITAS pipeline on a sample multimodal input. This example demonstrates each stage of our methodology—vision priors extraction, multi-expert scoring and fusion, GRPO critic assessment, and answer refinement—showcasing how our approach systematically enhances the quality of the answer. Each section is annotated to correspond with the steps outlined in our Method, providing a clear and practical illustration of the VERITAS pipeline in action.

Table 4: Critique prompt

Critique Prompt

You are an expert in assessing the quality of answers in image-based question-and-answer pairs, skilled at deeply understanding user queries and using this understanding to comprehensively and thoroughly evaluate the quality of model responses. I will provide you with a user's **[Question]** and **[Image]**, along with a **[Model Answer]**, and also provide **[Image Tags]** and **[OCR Results]**.

You need to refer to the [Inspection Points], follow the [Analysis Process] to deeply analyze the [Question], step by step evaluate the [Model Answer] quality, and finally according to the [Scoring Standards] and [Precautions], provide the model answer quality analysis results and scoring.

[Inspection Points]
The inspection points list 5 common elements for model quality evaluation, but do not cover all points. You must complement the inspection points based on the characteristics of the question:

Relevance
Whether the model answer is highly relevant to the question and meets every user requirement;
Accuracy
Whether the model answer contains any inaccuracies that contradict established knowledge/statements/theorems of the real world;
Factual Correctness
Whether the information provided in the answers is accurate, based on the image, and contains no unreasonable content;
Logical Coherence
Whether the language expression is fluent, the logic is clear, whether the various parts of the answer are organically combined, and whether the structure is clear;
Readability
Note whether the model answer is clearly organized and highly readable; having subheadings and key summaries is better than plain narration, but pay attention to the match between subheadings and subsequent explanations; check for any truncation or garbled text.

[Analysis Process]
Step 1, deeply analyze the [Question], identify all user needs, ensure your understanding of the question is accurate, and record the analysis process in detail at [Question Analysis];
Step 2, break down the [Model Answer] into multiple independent sentences. After dividing the answer into several parts, check each sentence against the provided [Question];
Step 3, Element Identification Verification: Analyze each sentence, check whether every element mentioned in the answer is included in the [Image Tags]. If any element is not in the [Image Tags], carefully verify the image to check for errors in the answer;
Step 4, Detail Matching: Analyze each sentence, check whether the specific properties (such as color, size, shape, etc.) of some elements described in the answer match the [Image Tags]. If inconsistent, carefully verify the image to check for errors in the answer;
Step 5, Scenario and Relationship Verification: Analyze each sentence, check whether the description of the image scene in the answer is consistent with the scene reflected by the [Image Tags], and whether each description of the relationship between objects is correct. If inconsistent, carefully verify the image to check for errors;
Step 6, OCR Verification: Analyze each sentence, check whether the description or paraphrasing of text in the image matches the [OCR Results]. If inconsistent, carefully verify the image. If there are errors in the [OCR Results], ignore the erroneous parts;
Step 7, read the entire [Model Answer], analyze whether the logic of the model answer is coherent and whether it is readable;
Step 8, output the analysis process and conclusions of each information point in the model answer in an ordered list, and finally give an overall evaluation of the [Model Answer]. Note that in your analysis, you must provide a reasonable explanation for the process and conclusions presented in your [Quality Evaluation];
Finally, score according to the evaluation reasons, and output the score in [Output Results].Note that you only need to output the score.

[Scoring Standards]
Scores range from 0-5 points, where 0-1 is very poor, 2-3 is average, and 4-5 is excellent.
0-1 points typically indicate: the appearance of garbled text, meaningless mix of Chinese and English, answers cut off mid-way, the model completely failing to respond to user needs, or hallucinatory issues;
2-3 points typically indicate: the answer lacks key information, has low utility, partially fails to respond to user needs, has low explanatory value, poor logical progression, and unrefined language;
4-5 points typically indicate: most of the answer is correct, with minor errors scoring 4 points, and a completely excellent and error-free response scoring 5 points.

[Precautions]
First, please be strict; if any element in the [Inspection Points] is incorrect, regardless of the size of the error, the score must be reduced.

Continued on next page

Continued from previous page

Critique Prompt
Second, when evaluating the quality of the model answer, you must remain objective and not let the length of the model answer affect your evaluation. Third, there may be some errors in the [Image Tags], including missing details or incorrect information. If there are errors, you can ignore them. Fourth, there may be some errors in the [OCR Results], including missing details or incorrect information. If there are errors, you can ignore them. Fifth, if the model's answer is irrelevant to the question, contains factual errors, or generates harmful content, the score must be 0-1 points. Sixth, if the model's response is inconsistent with the [Image Tags] or [OCR Results], please verify carefully. If the answer is confirmed to be incorrect, the score must not exceed 3 points.
Strictly follow the format below for output content: **[Question Analysis]** <Question Analysis>: xxx **[Quality Evaluation]** <Evaluation Reasons>: xxx **[Output Results]** <Scoring> xxx
The image-based question-and-answer data is as follows: **[Image]** {} **[Question]** {} **[Model Answer]** {} The image tags are as follows, when there are multiple images, the tags will be provided in the order of the image numbers: **[Image Tags]** {} The OCR results are as follows, when there are multiple images, the OCR results will be provided in the order of the image numbers: **[OCR Results]** {} Please start the assessment:

Table 5: Rewrite prompt

Rewrite Prompt

You are an expert in correcting and rewriting answers for image-based question-and-answer pairs, adept at deeply understanding user queries and using this as a basis to comprehensively and deeply inspect the quality of model answers. I will provide you with a [Question] and [Image] from a user, as well as a [Model Answer], along with an [Answer Evaluation]. You need to deeply analyze the [Question] and [Image], refer to the [Answer Evaluation] and [Rewriting Dimensions], and follow the [Rewriting Process] to correct and rewrite the [Model Answer], amending any errors in the [Model Answer], and output the new answer under [New Answer].

[Rewriting Dimensions]
1. Detail Richness: Add relevant details to the answer without deviating from the core of the question, making the answer more informative and in-depth.
2. Language Fluency: Ensure the answer uses accurate vocabulary, the sentence structure is reasonable, and it reads smoothly and naturally.
3. Vocabulary Accuracy: Use more precise vocabulary to ensure the accuracy and professionalism of the expression.
4. Content Completeness: Ensure the answer covers all necessary information and avoids missing key points.

[Rewriting Process]
Step 1, deeply analyze the [Question], identify all user needs, ensure your understanding of the question is accurate, and record the analysis process in detail at [Question Analysis];
Step 2, deeply understand the [Model Answer], based on the [Answer Evaluation], analyze whether there are any errors in the [Model Answer], and record the analysis process in [Correction Suggestions];
Step 3, deeply understand the [Model Answer], analyze whether there are areas for improvement in the [Model Answer] across the four dimensions of [Rewriting Dimensions], and record the analysis process in [Rewriting Suggestions];
Step 4, modify the answer based on [Correction Suggestions] and [Rewriting Suggestions], and output the revised answer under [New Answer].

[Precautions]
There may be some errors in the [Answer Evaluation], including missing details or incorrect information. If there are errors, you can ignore them.

Strictly follow the format below for output content:
[Question Analysis]
<Question Analysis>:
xxx
[Correction Suggestions]
<Correction Suggestions>:
xxx
[Rewriting Suggestions]
<Rewriting Suggestions>:
xxx
[New Answer]
<New Answer>
xxx

The image-based question-and-answer data is as follows:
[Image]
{}
[Question]
{}
[Model Answer]
{}
The answer evaluation is as follows:
[Answer Evaluation]
{}
Please start the assessment:

Table 6: Case Study: Applying VERITAS Pipeline

Case Study: Applying VERITAS Pipeline

[Image]

[Question]
<image>What is the name of the magazine mentioned in the image?
[Answer]
The name of the magazine is HoustoniaMag.com.

Stage 1: Vision Priors Extraction
[OCR]
DIGITAL | FEATURES | WULOME10,HoustoniaMag.com | 40 | 60 | 64 | On the | cover | IVEWITHS | We Go | .HLUM.SS | Our Type | ofIssue | THECLASSROOM
[Tags]
article | blanket | gymnast | magazine | man | news | paper

Stage 2: Multi-Expert Scoring and Fusion
[GPT-4o Critique]
[Question Analysis]
<Question Analysis>:
The user is asking for the name of the magazine mentioned in the image. The primary requirement is to identify the magazine's name accurately from the provided image.
[Quality Evaluation]
<Evaluation Reasons>:
1. The model answer states, "The name of the magazine is HoustoniaMag.com." This is close but not exact; the correct magazine name should be "HoustoniaMag" without ".com".
2. The OCR does include "HoustoniaMag.com," but the ".com" denotes a website domain rather than the magazine's title.
3. The image tags include "magazine," indicating that the image contains a magazine.
4. The answer is relevant and mostly accurate in identifying the publication, but it conflates the website URL with the magazine name, resulting in a minor factual inaccuracy.
5. The answer remains clear and logically coherent.
[Output Results]
<Scoring>
4

[Doubao-1.5-pro Critique]
[Question Analysis]
<Question Analysis>:
The user's need is to determine the name of the magazine shown in the provided image. The question - asks for a specific name and is straightforward.
[Quality Evaluation]
<Evaluation Reasons>:
1. Relevance: The answer directly addresses the question, providing a name for the magazine, so it is relevant.
2. Accuracy: HoustoniaMag.com is not the name of the magazine. The name of the magazine seems to be "FEATURES" as seen at the top - left of the image. So, there is an inaccuracy.
3. Factual Correctness: The answer is factually incorrect as it misidentifies the magazine name. HoustoniaMag.com appears to be a website associated with the magazine rather than the magazine's name itself.
4. Logical Coherence: The answer is a single - sentence response, and the logic is clear in terms of what it is trying to convey, but it is wrong.

Continued on next page

Continued from previous page

Case Study: Applying VERITAS Pipeline

5. Readability: The answer is short and easy to read, but the content is wrong. Overall, due to the factual error in identifying the magazine name, the answer does not meet the requirements well.
[Output Results]
<Scoring>
1

[Gemini-2.5-Pro Critique]
[Question Analysis]
<Question Analysis>:
The user wants to know the name of the magazine shown or mentioned in the provided image. The image displays what appears to be pages from a magazine, featuring articles, advertisements, and branding elements. The goal is to locate the text identifying the magazine's title.
[Quality Evaluation]
<Evaluation Reasons>:
1. Relevance: The model answer directly addresses the question by providing a name.
2. Factual Correctness: The model identifies "HoustoniaMag.com". This text is indeed present in the image, specifically under the "DIGITAL" section header ("WELCOME TO HoustoniaMag.com") and also referenced near the bottom right ("houstoniamag.comipad"). However, this is the website address associated with the magazine. The actual magazine title, "Houstonia", is also visible on a small representation of the magazine cover in the top right corner and implied by the ".com" domain. While "HoustoniaMag.com" is *mentioned* and relates directly to the magazine, it's technically the website URL, not the primary title of the publication itself. The most accurate answer would be "Houstonia". Therefore, the model's answer is slightly inaccurate because it provides the website domain instead of the magazine's title, although the domain name is explicitly mentioned in the image.
3. OCR Verification: The OCR results confirm the presence of "HoustoniaMag.com". Visual inspection confirms this text is prominently displayed in the digital section. The OCR also picks up "Houstonia" under the "FIND IT:" section and "Houstonia Digital", further supporting that "Houstonia" is the core name.
4. Logical Coherence
Readability: The answer is clear, concise, and easy to read.
5. Overall, the model successfully located text related to the magazine's identity mentioned in the image. However, it chose the website address instead of the magazine's actual title, which is also present. This represents a minor factual inaccuracy in identifying the specific *type* of name (website vs. title).
[Output Results]
<Scoring>
4

[Fusion Score]
Final Confidence Score: **4** (on a scale of 0 to 5)

Stage 3: Answer Refinement and Selection
[Answer Refinement Using GPT-4o's Critique]
The name of the magazine mentioned in the image is Houstonia.

[Answer Refinement Using Doubao-1.5-pro's Critique]
The name of the magazine shown in the image is FEATURES.

[Answer Refinement Using Gemini-2.5-Pro's Critique]
The name of the magazine mentioned in the image is Houstonia. The text "HoustoniaMag.com" refers to the magazine's website."

[Lightweight GRPO Critic Score for Raw Answer]
[Question Analysis]
<Question Analysis>:
The user is asking for the name of the magazine mentioned in the image. The key requirement is to identify the name of the magazine from the image.
[Quality Evaluation]
<Evaluation Reasons>:
1. The model answer states, "The name of the magazine is HoustoniaMag.com." which is partially incorrect because it identifies the website rather than the magazine's actual name.
2. The image shows the text "HoustoniaMag.com" in the top right corner, which is not the name of the magazine but rather a website of the magazine.
3. The model answer is somewhat relevant (it references a related entity), it fails to provide the actual magazine name
4. The model answer is logically readable.
[Output Results]
<Scoring>
4

[Lightweight GRPO Critic Score for GPT-4o Rewritten Answer]

Continued on next page

Continued from previous page

Case Study: Applying VERITAS Pipeline

[Question Analysis]
<Question Analysis>:
The user is asking for the name of the magazine mentioned in the image. The key requirement is to identify the name of the magazine from the image.
[Quality Evaluation]
<Evaluation Reasons>:
1. The model answer states, "The name of the magazine mentioned in the image is Houstonia."
2. The image shows “HoustoniaMag.com,” which aligns with the official domain of the magazine “Houstonia,” providing strong evidence for the magazine’s name.
3. The model answer is factually correct based on the image.
4. The model answer is logically coherent and readable.
[Output Results]
<Scoring>
5

[Lightweight GRPO Critic Score for Doubao-1.5-pro Rewritten Answer]
[Question Analysis]
<Question Analysis>:
The user is asking for the name of the magazine shown in the image. The key requirement is to identify the name of the magazine from the image.
[Quality Evaluation]
<Evaluation Reasons>:
1. The model answer states, "The name of the magazine shown in the image is FEATURES," which is incorrect.
2. The image indicates the magazine is Houstonia (e.g., “HoustoniaMag.com”), while “FEATURES” appears to be a section header, not the magazine title.
3. The model answer is relevant to the question but does not fully meet the user’s requirement for the name of the magazine.
4. The model answer is factually incorrect based on the image.
5. The model answer is logically coherent and readable but does not fully meet the user’s requirement for the name of the magazine.
[Output Results]
<Scoring>
1

[Lightweight GRPO Critic Score for Gemini-2.5-Pro Rewritten Answer]
[Question Analysis]
<Question Analysis>:
The user is asking for the name of the magazine mentioned in the image. The key requirement is to identify the name of the magazine from the image.
[Quality Evaluation]
<Evaluation Reasons>:
1. The model answer states, "The name of the magazine mentioned in the image is Houstonia." This is correct and matches the text in the image.
2. "The text ’HoustoniaMag.com’ refers to the magazine’s website," is irrelevant to the user’s question and introduces unnecessary information.
3. The model answer is factually correct about the magazine name based on the image.
4. The response is generally relevant but could be more concise by omitting unrelated website details.
5. The answer is clear and readable.
[Output Results]
<Scoring>
4

[Final Selected Answer]
The name of the magazine mentioned in the image is **Houstonia**.