

When Vision-Language Models Judge Without Seeing: Exposing Informativeness Bias

Xiaohan Zou^{1*} Roshan Sridhar² Mohammadtaher Safarzadeh² Dan Roth²

¹The Pennsylvania State University ²Oracle AI

xfz5266@psu.edu, {roshan.sridhar, mohammadtaher.safarzadeh, dan.roth}@oracle.com

Abstract

The reliability of VLM-as-a-Judge is critical for the automatic evaluation of vision-language models (VLMs). Despite recent progress, our analysis reveals that VLM-as-a-Judge often pays limited attention to the image when making decisions. Instead, they often blindly favor the more informative answer, even when they can recognize it conflicts with the image content. We call this problem *informativeness bias*, which significantly undermines judge reliability. To address it, we propose BIRCH (Balanced Informativeness and CoRrectness with a Truthful AnCHor), a judging paradigm that first corrects inconsistencies with the image content in candidate answers, and then compares the answers against this corrected version. This shifts the judge’s focus from informativeness to image-grounded correctness. Experiments on multiple models and benchmarks show that BIRCH reduces informativeness bias by up to 17%, resulting in performance gains of up to 9.8%. Our work reveals an overlooked but fundamental flaw in current VLM-as-a-Judge systems and highlights the need for more principled designs.

1 Introduction

Vision-Language Models (VLMs) have demonstrated impressive performance across a wide range of multimodal perception and reasoning tasks (OpenAI et al., 2024b; Fu et al., 2024; Yue et al., 2024). Building on this, they are now widely used as automatic judges to evaluate responses and guide reward generation (Chen et al., 2024a; Yasunaga et al., 2025). VLM-as-a-Judge is typically applied in pairwise comparison tasks, where two candidate answers to a question about an image are compared to determine which is better. The judgment is then evaluated against human annotations (Jing and Du,

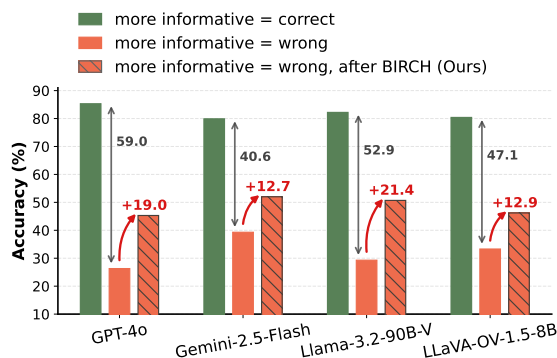


Figure 1: VLM-as-a-Judge performs well when the more informative answer happens to be correct (■ more informative = correct). But when the more informative answer is actually wrong (■ more informative = wrong), accuracy drops sharply by 40.6-59.0% across models. This shows judges over-prioritize informativeness over truth. Our method BIRCH greatly improves performance in these cases, thereby boosting overall results.

2025; Yu et al., 2025; Zhou et al., 2024), as shown in Figure 2 (A), (B) and (C).

Despite the increasing use of VLM-as-a-Judge, its reliability is still underexplored. In LLM-as-a-Judge, several biases have been reported, such as positional bias, where judges favor certain answer positions (Zheng et al., 2023), length bias, where longer responses are preferred (Wang et al., 2024a), and authority bias, where fabricated citations reduce judgment accuracy (Chen et al., 2024b). These issues raise concerns about fairness and trust in evaluation tasks (Gu et al., 2025; Shi et al., 2025; Ye et al., 2025). However, the vision-language setting remains less explored. Existing work on VLM-as-a-Judge mainly focuses on increasing task diversity and difficulty (Chen et al., 2024a; Li et al., 2025b; Yasunaga et al., 2025), but often overlooks whether vision-language judgment is being made and evaluated as intended.

To investigate this, we analyzed popular models and found that they pay little attention to images

*Work done during internship at Oracle AI.

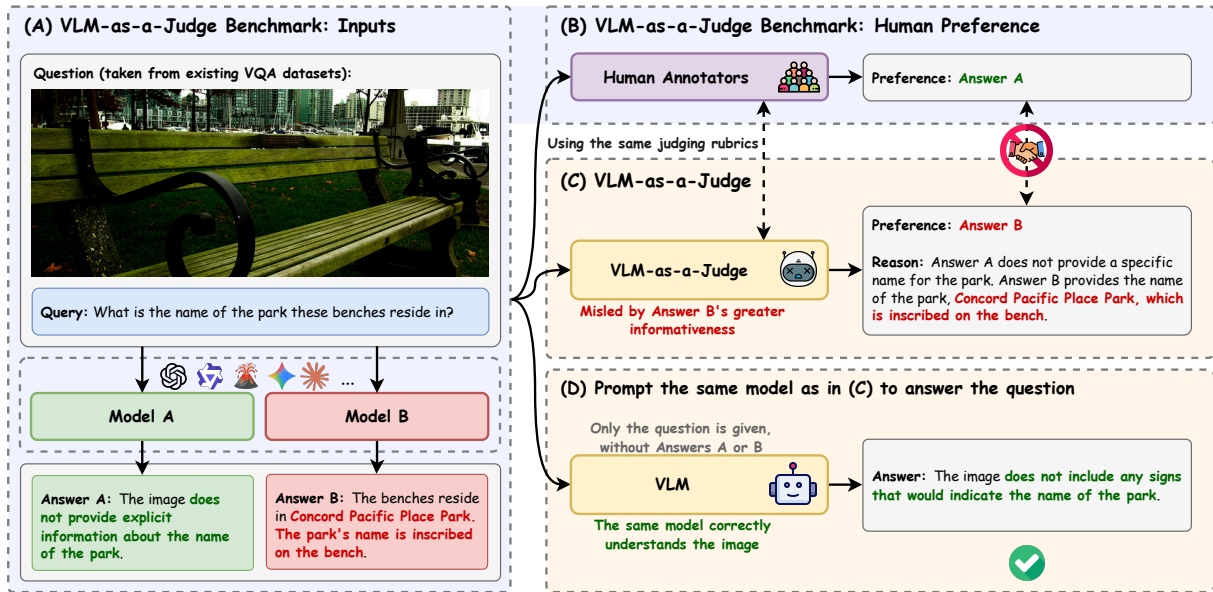


Figure 2: (A) **VLM-as-a-Judge input**: A VQA question (query + image) is given to two different response models, producing candidate answers A and B. The question asks for the park’s name. Answer B looks more informative by giving a name and claiming it is inscribed on the bench, while Answer A is correct since the name is not visible in the image. (B) **Human preference**: Human annotators, given the query, image, answers, and judging rubrics, choose Answer A as the preference label. (C) **VLM-as-a-Judge decision**: The judge receives the same input and is prompted with the same judging rubrics as the humans. It is misled by Answer B’s greater informativeness, mistakenly selecting it because it provides a specific (but wrong) name, while Answer A does not—misaligning with human judgment. (D) **Direct model response**: When asked to answer the question directly, the same VLM correctly sees that the name cannot be determined from the image. Here, (A) and (B) are covered by existing benchmarks. (C) and (D) highlight a gap between VLM-as-a-Judge’s actual recognition ability and its judging behavior, caused by being misled by informativeness. This paper aims to close that gap for more reliable judgment.

when making judgments. Removing the image from their inputs causes only a minimal drop in performance (Section 4.1). A deeper look reveals a striking behavior: VLM-as-a-Judge is heavily deceived by the answer *informativeness*, following the definition in Lin et al. (2022); Wu et al. (2025):

Informativeness: How richly an answer goes beyond a minimal reply, by adding explanations, facts, context, or examples.

They often favor the more informative answer despite clear inconsistencies with the image, even when they can recognize those inconsistencies (Section 4.2). For example, in Figure 2 (A), Answer A is more truthful since it avoids a false claim about the park’s name, but is less informative. Answer B is more informative by providing a name, though it is wrong. This shows that greater informativeness can increase the risk of speculation or error (Lin et al., 2022; Wu et al., 2025). However, VLM-as-a-Judge still incorrectly selects Answer B (Figure 2 (C)), despite knowing the image gives no clue about the park’s name (Figure 2 (D)). As a result,

VLM-as-a-Judge performs much worse when the more informative answer is wrong than when it happens to be truthful, as shown in Figure 1.

In Section 4.4, we further show that informativeness bias is much stronger than the commonly discussed length bias (Zheng et al., 2023; Gu et al., 2025) and remains significant even after controlling for length. This suggests that, although more informative answers tend to be longer, the two biases are not equivalent. Consequently, prior strategies for reducing length bias, such as instructing models to ignore length (Zheng et al., 2023) or equalizing response lengths (Hu et al., 2025), are insufficient, since informativeness bias cannot be resolved by addressing length effects alone.

In summary, informativeness bias is a key factor that pulls VLM-as-a-Judge away from image content, undermining reliability. As a result, their judging ability lags behind their vision-language understanding, leaving much of that capability underused. To bridge this gap and make VLMs more reliable for evaluation, we propose a novel judging paradigm **BIRCH** (Balanced Informativeness

and CoRrectness with a Truthful AnCHor). The core idea is for the judge to first create a truthful "anchor" that encodes the necessary image content into text. Candidate answers are then compared against this anchor, bringing the judge's attention back to the image and reducing informativeness bias. To avoid over-prioritizing correctness and unfairly rejecting valid details, the anchor is made as informative as the candidates. We achieve this by asking the judge to revise the candidate answers using the query and image, removing unsupported claims with explanations and fixing inconsistencies. The resulting anchor is both accurate and informative, allowing every detail in the candidates to be verified against it, as described in Section 5.

We conduct a comprehensive evaluation of 8 state-of-the-art judging models, covering both open-source and commercial systems, on two benchmarks. BIRCH consistently achieves better performance than all baselines. We also confirm that BIRCH substantially reduces informativeness bias. As shown in Figure 1, it greatly improves accuracy when the more informative answer is untruthful, demonstrating a stronger focus on image consistency, which drives the performance gains.

Our contributions can be summarized as follows:

1. We identify *informativeness bias*, where VLM-as-a-Judge is distracted by the informativeness of answers and pays little attention to image content. This leaves much of its vision-language capability underused for judgment tasks, thereby undermining reliability (Section 4.1, 4.2, and 4.3).
2. We show that informativeness bias is distinct from and cannot be reduced to the commonly discussed length bias (Section 4.4), and thus calls for a solution that goes beyond simply removing length effects.
3. We propose BIRCH, a novel judging paradigm that mitigates informativeness bias and refocuses the judge's attention on the image, enabling better use of VLM capabilities and improving overall performance (Section 5).

2 Related Work

2.1 LLM/VLM-as-a-Judge Benchmarks

Developing effective LLM/VLM-as-a-Judge systems is essential for evaluating and aligning LLMs and VLMs (Ouyang et al., 2022; Grattafiori et al., 2024). Existing benchmarks such as Anthropic Helpful and Harmless (Bai et al., 2022), OpenAI

Summarization (Stiennon et al., 2020), RewardBench (Lambert et al., 2025), and JudgeBench (Tan et al., 2025) have significantly advanced evaluation in the text modality (LLMs).

To extend this progress to the multimodal domain, MLLM-as-a-Judge (Chen et al., 2024a) takes the first step in quantifying VLMs' performance as judges. VL-RewardBench (Li et al., 2025b) targets scenarios where even state-of-the-art models struggle. Multimodal RewardBench (Yasunaga et al., 2025) incorporates expert knowledge domains and safety-related concerns. Judge Anything (Pu et al., 2025) expands this scope to multimodal generation tasks, incorporating both video and audio inputs. MM-RLHF (Zhang et al., 2025b) improves size, diversity, annotation granularity, and includes safety-related tasks. However, despite this progress, existing works still lack a critical examination of the reliability of VLM-as-a-Judge behavior.

2.2 Biases in LLM/VLM-as-a-Judge

Recent work has found several cognitive biases that affect how LLMs evaluate (Gu et al., 2025). Many studies (Zheng et al., 2023; Shi et al., 2025) have discussed biases such as position, verbosity, and self-enhancement. Others (Koo et al., 2024) highlight compassion fade, bandwagon effect, and attentional bias. Additional work (Chen et al., 2024b; Stureborg et al., 2024) has noted fallacy-oversight, authority, and beauty bias. Ye et al. (2025) propose an automated pipeline to measure diverse biases, while Wei et al. (2025) analyze biases across different prompt templates.

However, this topic remains underexplored in the multimodal domain. While modality bias, where multimodal LLMs over-rely on language (Zheng et al., 2025; Li et al., 2024; Liu et al., 2025), is well known, its impact on judging behavior remains unclear. The most related work, Li et al. (2025c), shows that over-relying on text-only patterns during training hurts the generalization of multimodal reward models. Our work goes further by linking this issue to over-prioritizing informativeness.

2.3 Debiasing in LLM/VLM-as-a-Judge

To reduce bias in LLM-as-a-Judge, studies have asked judges to follow predefined rules (Zeng et al., 2024), provided reference answers to avoid context confusion (Zheng et al., 2023; Zhang et al., 2025a), swapped response positions to reduce positional bias (Wang et al., 2024a), or equalized answer length to reduce length bias (Hu et al., 2025).

However, these methods are not designed to handle informativeness bias and are therefore ineffective for this issue.

On the multimodal side, Li et al. (2025c) reduces bias from textual spurious correlations by shifting the training distribution toward better multimodal understanding. Our work instead identifies informativeness bias as a cause and addresses it by enforcing attention to image consistency without additional training.

3 Analysis Setup

3.1 Pairwise Comparison in VLM-as-a-Judge

We formalize VLM-as-a-Judge using a pairwise preference comparison paradigm (Chen et al., 2024a). Let $\mathcal{D} := \{d_i\}_{i=1}^N$ be a dataset with N preference examples. Each example $d_i = (Q_i, I_i, y_i^c, y_i^r)$ contains a question about an image, the image itself, and two responses labeled by human preference. Here, Q_i is the textual query, I_i is the associated image, y_i^c is the preferred response, and y_i^r is the less preferred one. For simplicity, we drop the subscript i when the context is clear.

The judge model $f(Q, I, y^c, y^r)$ takes a sample as input and decides which answer better addresses the question. Its performance is measured by accuracy against human preferences:

$$\text{Acc} = \mathbb{E}_{(Q, I, y^c, y^r) \sim \mathcal{D}} \left[\mathbb{I}(f(Q, I, y^c, y^r) = y^c) \right].$$

3.2 Quantifying Image Reliance

VLMs are known to underuse visual content and over-rely on text (Li et al., 2024; Fu et al., 2025; Liu et al., 2025; Leng et al., 2024). We ask whether they show a similar tendency when used as judges and measure how much they rely on visual input. Specifically, we define the *Image Reliance Score (IRS)* as the accuracy gap of judge model f when evaluated with and without the image I as input:

$$\text{IRS} = \mathbb{E}_{(Q, I, y^c, y^r) \sim \mathcal{D}} \left[\mathbb{I}(f(Q, I, y^c, y^r) = y^c) \right] - \mathbb{E}_{(Q, y^c, y^r) \sim \mathcal{D}} \left[\mathbb{I}(f(Q, y^c, y^r) = y^c) \right].$$

A VLM judge that depends more on image content will have a higher IRS, since its accuracy improves more when the image is included.

3.3 Judging Informativeness

General-purpose alignment aims to produce responses that balance correctness and informativeness (Zheng et al., 2023). This is reflected in common judging prompts, which often include rubrics

such as: "Your evaluation should consider factors such as accuracy, helpfulness, relevance, depth, and level of detail" (Zheng et al., 2023; Chen et al., 2024a), encouraging responses to go beyond a minimal correct answer and provide additional context (Lin et al., 2022). Here, *accuracy* corresponds to correctness, which requires using the image to detect errors or inconsistencies, while *helpfulness, relevance, depth, and level of detail* correspond to informativeness, which can be judged from the answers alone. We focus on how informativeness affects VLM-as-a-Judge behavior, so we separate it from the other judging criteria.

Based on this, we define the rubrics for informativeness as *helpfulness, relevance, depth, and level of detail*, with *accuracy* removed from the original set. We then ask GPT-4o to decide which answer, y^c or y^r , is more informative using these rubrics. The full prompt is in Appendix A. Only the textual query Q and the two candidate answers (y^c, y^r) are given, while the image I is withheld, since informativeness does not depend on it. GPT-4o is explicitly instructed to judge only on informativeness, without considering correctness or consistency with the image. The output is y^i , the more informative answer, where y^i is either y^c or y^r . See Appendix C.1 for an evaluation of the reliability of GPT-4o in measuring informativeness

3.4 Measuring Informativeness Bias

Several studies have shown that prioritizing informativeness can come at the cost of correctness (Zhou et al., 2023; Wu et al., 2025). We define *Informativeness Bias (IB)* as the degree to which a judge favors informativeness over correctness.

We split the preference dataset \mathcal{D} into two subsets. The *informativeness-driven subset* \mathcal{D}_{IDS} contains cases where the human-preferred answer y^c is more informative than y^r (i.e., $y^c = y^i$), meaning informativeness is a valid factor. The *correctness-driven subset* \mathcal{D}_{CDS} includes cases where the preferred answer y^c is less informative (i.e., $y^r = y^i$), meaning correctness is the main factor and informativeness is misleading. We compute IB as:

$$\text{IB} = \mathbb{E}_{d \sim \mathcal{D}_{\text{IDS}}} \left[\mathbb{I}(f(d) = y^c) \right] \quad (\text{Acc}_{\text{IDS}}) \\ - \mathbb{E}_{d \sim \mathcal{D}_{\text{CDS}}} \left[\mathbb{I}(f(d) = y^c) \right] \quad (\text{Acc}_{\text{CDS}}).$$

In other words, we treat human annotators as unbiased, and obviously $\text{IB} = 0$ for humans since their Acc_{IDS} and Acc_{CDS} are both 100%. For an unbiased judge, IB should also be close to zero,

Models	MLLM-as-a-Judge			VL-RewardBench		
	With Image	No Image	IRS	With Image	No Image	IRS
GPT-4o	66.45	65.46	0.99	61.12	57.92	3.20
Gemini-2.5-Flash	66.92	64.94	1.98	56.25	53.71	2.54
Llama-3.2-V-90B	65.28	62.46	2.82	59.46	55.44	4.02
LLaVA-1.5-13B	56.21	53.32	2.89	44.76	42.63	2.13
LLaVA-1.6-34B	58.95	59.96	-1.01	48.83	46.31	2.52
LLaVA-OV-1.5-8B	65.35	65.48	-0.13	56.01	54.13	1.88

Table 1: The image reliance score (IRS, the accuracy gain from including the image in the judge’s input, higher values mean stronger image dependence) remains minimal across all models and benchmarks.

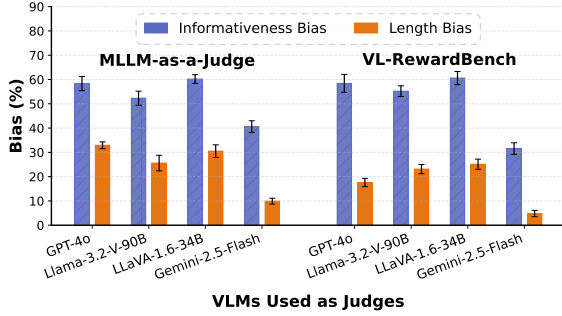


Figure 3: Informativeness bias is consistently stronger than length bias across all models and benchmarks, with both showing low variance, indicating informativeness has a greater impact on the judge’s behavior than length.

meaning it performs equally well whether or not informativeness is misleading. A positive IB indicates the judge performs better on \mathcal{D}_{IDS} , showing a preference for informativeness. An ideal judge should achieve high Acc_{IDS} and Acc_{CDS} for strong overall performance, while keeping IB low to balance informativeness and correctness.

3.5 Measuring Length Bias

Prior studies show that LLM-as-a-Judge often favors longer responses, known as *length bias* (Zheng et al., 2023; Ye et al., 2025). Since informativeness is often tied to length, with more informative answers usually being longer, we also measure *Length Bias* (LB) for comparison to investigate how the two biases relate.

Following the definition of informativeness bias, we split the dataset \mathcal{D} into two subsets, one where the longer answer is preferred and one where the shorter answer is preferred. We then calculate LB as the difference in a judge’s accuracy between them: $LB = Acc_{y^c \text{ is longer}} - Acc_{y^c \text{ is shorter}}$. The magnitude $|LB|$ shows how strongly the judge is influenced by length, while the sign shows whether it favors longer or shorter responses.

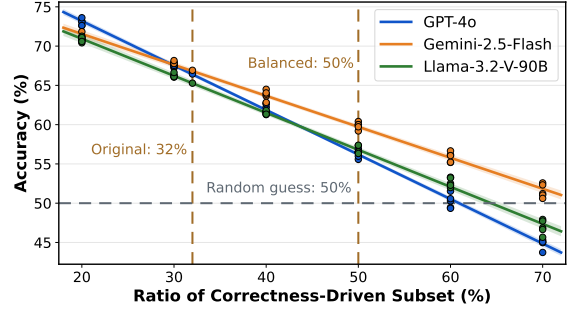


Figure 4: Judge performance is strongly affected by how often informativeness is misleading, i.e., the ratio of correctness-driven subset (\mathcal{D}_{CDS}). We control this ratio in MLLM-as-a-Judge by downsampling. Judges achieve good accuracy ($>65\%$) at the original low ratio (32%), but when the data is balanced (50%), their accuracy drops to only slightly above random guessing.

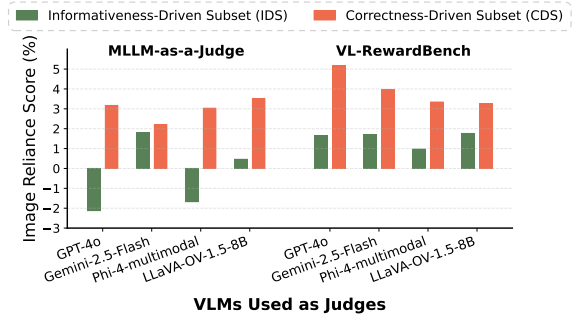


Figure 5: Image reliance score is consistently higher on correctness-driven subsets than on informativeness-driven subsets across all models and benchmarks, showing that images matter more when informativeness is misleading and correctness drives the decision.

4 Exposing Informativeness Bias

4.1 Judges Are Nearly Blind to Images

Table 1 shows that the Image Reliance Score (IRS) is minimal across all models and benchmarks, meaning image content provides only negligible gains (less than 3% on MLLM-as-a-Judge (Chen et al., 2024a) and under 5% on VL-RewardBench (Li et al., 2025b)). We further divide MLLM-as-a-Judge into different source domains and show in Figure 11 that adding the image yields only slight accuracy improvements across all domains, and in some cases even reduces performance. This suggests that VLM-as-a-Judge often ignores the image when evaluating answers to VQA questions, raising concerns about their reliability and highlighting the need for deeper investigation of this issue.

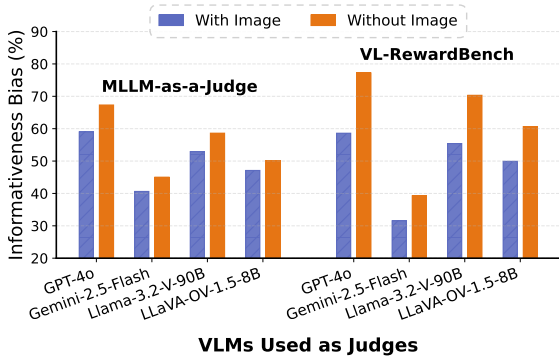


Figure 6: Informativeness bias is higher when the image is removed than when it is included in the input.

4.2 Judges Are Deceived by Informativeness

Judges over-prioritize informativeness. As shown in Figure 3, we find significant informativeness bias (IB): over 30% for Gemini-2.5-Flash and over 50% for other models. We also report IB variance by downsampling \mathcal{D}_{IDS} and \mathcal{D}_{CDS} at different ratios (see Appendix C.3), showing reasonably low variance. The high IB is also clear in Figure 9, where Acc_{IDS} is always much higher than Acc_{CDS} . These indicate that VLM judges perform far worse when informativeness is misleading, revealing their strong tendency to overvalue informativeness.

Judge performance heavily depends on how often informativeness is misleading. We downsample to control the ratio of the correctness-driven subset \mathcal{D}_{CDS} and examine how performance changes with it. A higher ratio means informativeness is misleading more often. As shown in Figure 4, judge accuracies decrease as the proportion of \mathcal{D}_{CDS} grows. Since the proportion of \mathcal{D}_{CDS} is low in the original benchmark (32%), judges can appear strong overall (>65%) even with poor Acc_{CDS} (Figure 9). But with a balanced ratio (50%), which better reflects their real performance, overall accuracy drops to around 55%, only slightly above random guessing. This reveals a serious and previously overlooked unreliability in VLM-as-a-Judge.

Judges are misled by informativeness, even when they know the correct answer. We show that poor Acc_{CDS} is not due to a lack of vision-language ability in the judge. As illustrated in Figure 2 (D) and additional examples in Appendix D, when directly prompted to answer the question Q , the judge often produces a correct response that aligns with the human-chosen answer and points out inconsistencies or errors in the rejected one.

Models	MLLM-as-a-Judge	VL-RewardBench
GPT-4o	59.0 → 44.6	58.6 → 45.5
Gemini-2.5-Flash	40.6 → 33.66	31.6 → 26.0
Llama-3.2-Vision-90B	52.9 → 40.2	55.4 → 39.1
LLaVA-1.6-34B	60.2 → 36.7	60.6 → 42.3

Table 2: Informativeness bias before → after equalizing answer length to remove length bias. Although they decrease after eliminating length bias, the values remain high, indicating that it is the root cause of length bias.

Yet, when acting as a judge, it still selects the human-rejected answer as the better choice, contradicting its own response. A common pattern is that the chosen answer appears more informative, offering detailed explanations or direct answers, but without considering correctness.

These findings show that judges can understand questions and spot conflicts between answers and the image. However, they are deceived by informativeness, giving it far more weight than correctness.

4.3 Informativeness Bias Pulls Judge’s Attention Away from the Image

From Section 4.1 and 4.2, we hypothesize that informativeness bias (IB) distracts the judge’s attention from the image, leading to a low image reliance score (IRS). We now verify this connection.

Image reliance is higher on correctness-driven cases \mathcal{D}_{CDS} . As shown in Figure 5, IRS is consistently higher on correctness-driven cases \mathcal{D}_{CDS} than on informativeness-driven cases \mathcal{D}_{IDS} across all models and benchmarks, indicating that judges rely more on images when correctness drives the decision. This supports our hypothesis: when informativeness is misleading, correctness becomes crucial, and as correctness cannot be verified without the image, visual input is far more valuable.

Informativeness bias worsens when the judge lacks image input. We test how removing the image from the judge’s input affects IB. As shown in Figure 6, judges without images consistently show stronger IB than those with images, with gaps of up to 18.6%. This also supports our claim: without the image, the judge cannot verify correctness and is forced to rely only on informativeness.

4.4 Informativeness Bias Is Distinct from Length Bias

Figure 3 compares informativeness bias (IB) with length bias (LB). LB is consistently much weaker, trailing IB by about 30% or more across all models

and benchmarks. Figure 8 further breaks this down and shows IB is higher than LB in nearly all domains. This suggests that IB is not merely a proxy for LB; rather, it might be a more substantial issue.

To investigate further, we remove the effect of LB by creating a version of the MLLM-as-a-Judge benchmark where the length difference between candidate answers is below a small threshold (see Appendix C.5). Table 2 shows IB after this length equalization. While IB decreases, it remains substantial (26%-45.5%). This suggests that, although IB and LB are correlated, IB cannot be resolved by addressing LB alone, as a substantial portion of IB remains after controlling for length. The correlation is expected, since more informative answers are often longer, making a preference for informativeness appear as a preference for length. However, mitigating IB requires directly addressing informativeness bias rather than focusing only on length effects (Zheng et al., 2023; Hu et al., 2025).

5 Method

Based on Section 4, VLM-as-a-Judge can answer questions correctly but gets distracted by informativeness and fails to focus on the image. Therefore, we explore whether using its own answer as a reference can refocus attention on image consistency. Unlike prior work that improves LLM-as-a-Judge by providing ground-truth references (Zheng et al., 2023; Hagos et al., 2024; Zhou et al., 2025), we instead use a self-generated reference.

5.1 Reducing Informativeness Bias

Figure 12 illustrates the Standard Ref pipeline. The model first answers the question directly, encoding key image details relevant to the question into a textual reference. It then judges by comparing the two candidate answers to this reference. This naturally shifts the focus from informativeness alone to consistency with the image, since the reference already includes image details.

Figure 12 (A) shows a successful case: with its own answer as reference, the judge correctly detects that Answer B gives a park name not visible in the image and selects Answer A instead. Table 4 confirms that this greatly reduces informativeness bias (IB) and improves Acc_{CDS}, boosting performance when correctness should drive the decision.

However, this approach can also cause judges to over-prioritize correctness and unfairly reject valid details in candidates simply because they are

missing from the reference. In Figure 12 (B), Answer A is rejected just for mentioning the correct position of a sign omitted in the reference. Table 4 shows a clear drop in Acc_{IDS}, meaning the judge performs worse when informativeness should guide the decision. This matches prior findings (Wu et al., 2025; Zhou et al., 2023) that greater emphasis on correctness often leads to over-caution about extra information. As a result, while this method improves Acc_{CDS} and lowers IB, it still reduces overall accuracy compared to the base pipeline.

5.2 Our Method: BIRCH

We propose BIRCH, which balances informativeness and correctness using an "anchor" that is both truthful and as informative as the candidates. As shown in Figure 7, the judge first modifies each candidate answer by correcting inconsistencies with the image or removing unsupported claims with explanations. The corrected answers are then merged into a single anchor. This ensures the generated anchor encodes the truthful version of all relevant image details included in the candidates. As a result, when candidates are compared to the anchor, every detail can be verified against it, preserving informativeness while enforcing correctness.

6 Experiments

6.1 Evaluation Settings

We evaluate our approach on MLLM-as-a-Judge (Chen et al., 2024a) and VL-RewardBench (Li et al., 2025b) benchmarks. We use a pairwise comparison setup (Chen et al., 2024a; Li et al., 2025b) to measure agreement between model and human judgments. We manually corrected human labeling errors that reduced benchmark reliability and evaluated on the revised data (see Appendix B).

Our experiments include open-source models such as Phi-4-multimodal (Abdin et al., 2024), LLaVA-OneVision-1.5-8B (Li et al., 2025a), LLaVA-1.5-13B, LLaVA-1.6-34B (Liu et al., 2024), and Llama-3.2-Vision-90B (Grattafiori et al., 2024), as well as commercial models like GPT-4o (OpenAI et al., 2024b), Gemini-2.5-Flash, and Gemini-2.5-Pro (Comanici et al., 2025).

6.2 Evaluation of BIRCH

Main Results. As shown in Table 3, BIRCH achieves the best performance across all tested models and benchmarks, with improvements over Base (a standard judge pipeline following (Chen

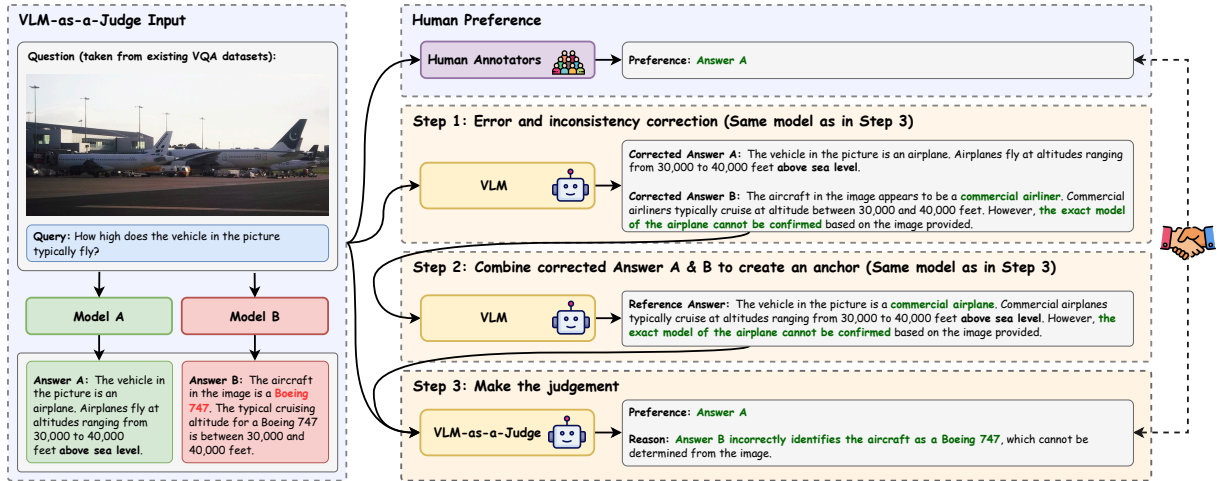


Figure 7: Illustration of the proposed BIRCH. **Step 1:** Each candidate answer is checked and corrected based on the image. In Answer B, the unsupported detail "Boeing 747" is replaced with "a commercial airliner" and clarified as "the exact model cannot be confirmed". **Step 2:** The corrected answers are merged into an anchor that is both truthful and informative (all details from the candidates are either corrected or preserved). **Step 3:** Candidates are compared with the anchor. This forces the judge to focus on the image content encoded in the anchor, balance informativeness with correctness, and avoid both over-prioritizing and over-penalizing informativeness.

Models	MLLM-as-a-Judge (Chen et al., 2024a)				VL-RewardBench (Li et al., 2025b)			
	Base	Image Caption	Standard Ref	BIRCH	Base	Image Caption	Standard Ref	BIRCH
Proprietary Models								
GPT-4o	66.45	65.79	65.14	75.78	61.12	59.36	60.42	66.95
Gemini-2.5-Flash	66.92	66.90	68.34	73.37	56.25	56.65	57.84	61.08
Gemini-2.5-Pro	67.63	68.04	69.21	73.37	57.55	57.33	59.21	61.49
Open-Source Models								
Llama-3.2-Vision-90B	65.28	66.73	63.18	75.12	59.46	59.94	59.21	64.52
LLaVA-OneVision-1.5-8B	65.35	65.67	66.21	71.17	56.01	57.07	57.95	60.01
LLaVA-1.5-13B	56.21	54.69	53.28	59.48	44.76	46.24	46.59	47.32
LLaVA-1.6-34B	58.95	58.88	58.02	61.54	48.83	49.29	50.19	52.03
Phi-4-multimodal	59.19	59.07	59.11	63.11	54.72	54.49	55.45	57.45

Table 3: Accuracy of VLM-as-a-Judge. Our proposed BIRCH significantly improves the performance of both open-source and proprietary models across different benchmarks.

Methods	Acc	Acc _{IDS}	Acc _{CDS}	IB	LB
GPT-4o (Base)	66.5	85.3	26.3	59.0	32.9
+ Standard Ref	65.1 (-1.4)	80.7 (-4.6)	31.8 (+5.5)	48.9 (-10.1)	16.9 (-16.0)
+ BIRCH (Ours)	75.8 (+9.3)	90.1 (+4.8)	45.3 (+19)	44.8 (-14.2)	20.3 (-12.6)

Table 4: Changes across different metrics on MLLM-as-a-Judge (Chen et al., 2024a). Our BIRCH reduces informativeness bias (IB) by significantly improving Acc_{CDS} while keeping strong Acc_{IDS}, explaining its strong overall performance.

Text	Average Score
Answer 1	3.1034
Answer 2	3.2126
Anchor (Gemini-2.5-Flash)	4.3793
Anchor (LLaVA-1.5-13B)	3.7364

Table 5: Comparison of GPT-5 quality scores for generated anchors and candidate answers, using inputs from MLLM-as-a-Judge (Chen et al., 2024a).

et al., 2024a) with no extra input) ranging from 2.73% to 9.84%. Improvements are larger on stronger models like GPT-4o, Gemini, and Llama-3.2. Although these models are affected by informativeness bias, their stronger vision ability helps them detect and correct errors in candidate answers and generate more reliable anchors. Gains are

smaller for weaker models such as LLaVA-1.5/1.6, which lack strong vision understanding, and for Phi-4-Multimodal, which shows less informativeness bias to begin with (see Appendix C.6), leaving less room for improvement.

Ablation Studies. BIRCH introduces an anchor gen-

Method	Base	BIRCH
Inference Time (s)	3.87	7.47 (+3.6)

Table 6: Inference time comparison per data point.

eration module that corrects and merges candidate answers to assist the judge. To evaluate this design, Table 3 compares it with two alternatives: (1) Image Caption, which generates an image caption and uses it with the query, image, and candidate answers for judging, as illustrated in Figure 13, and (2) Standard Ref, which produces a reference answer without using the candidates and then compares the candidates against it, as described in Section 5.1. BIRCH outperforms all baselines. Image Caption and Standard Ref sometimes slightly outperform Base, but in some cases they perform worse than Base.

6.3 Why BIRCH Works Well?

Quantitative Results. We analyze changes across multiple metrics with and without BIRCH in Table 2 and 8. BIRCH shows a much lower IB than both Base and Standard Ref across all models. For GPT-4o, BIRCH boosts Acc_{CDS} by 19% over Base, much higher than the gain from Standard Ref, and also raises Acc_{IDS} by 4.8% at the same time, resulting in a 9.3% overall accuracy gain. In contrast, Standard Ref causes a 4.6% drop in Acc_{IDS} , which finally lowers overall accuracy. These results show that BIRCH maintains a good balance without over-prioritizing or over-penalizing informativeness. Moreover, BIRCH also clearly reduces length bias (LB), consistent with our finding that IB and LB are not independent and that mitigating IB also alleviates LB.

Case study. We provide qualitative examples in Figure 7, 12, 13 and Appendix D to illustrate how BIRCH balances informativeness and correctness, and why Standard Ref and Image Caption fail. Base is easily misled by a more informative but incorrect answer, even when it clearly conflicts with the judge’s own response. BIRCH creates an anchor by correcting or removing inconsistencies while preserving valid details. This allows the judge to verify every detail in the candidates and make correct judgments. In contrast, Standard Ref and Image Caption generate references without using the candidate answers and often miss details mentioned in them. As a result, Standard Ref makes

the judge overly cautious and reject correct details that are not in the reference, while Image Caption accepts unsupported details when they cannot be verified against the caption. BIRCH reduces both incorrect rejection and acceptance and achieves a better balance.

6.4 Anchor Quality

The effectiveness of BIRCH depends on the quality of the generated anchors. These anchors should contain fewer factual errors than the original candidate answers to help the judge verify image details. To evaluate this, we use GPT-5 to score the correctness of the anchors and the two original candidate answers for Gemini-2.5-Flash and LLaVA-1.5-13B. As shown in Table 5, the anchors achieve higher average scores than both original answers, indicating that VLMs can correct many image inconsistencies and produce more accurate anchors.

6.5 Inference Efficiency

Although anchor generation in BIRCH involves two steps, answer correction and merging, they are implemented in a single pass, as shown in Table 20. Therefore, BIRCH requires only one additional pass compared to the plain judge pipeline Base.

We further illustrate this by reporting the average inference time per data point in seconds for LLaVA-One-Vision-1.5-8B on a single A100 GPU for BIRCH and Base in Table 6. The inference time of BIRCH is about twice that of Base, which is expected since it performs one extra pass. This overhead is comparable to or lower than many LLM-as-a-judge methods (Tang et al., 2024; Khan et al., 2024) that require multiple extra calls.

7 Conclusion

We reveal and address a fundamental flaw in current VLM-as-a-Judge systems, namely *informativeness bias*, the tendency to blindly favor more informative answers while giving little attention to the image. To improve reliability, we propose BIRCH, which compares candidate answers against a self-generated anchor that corrects or verifies all image details mentioned. This shifts the judge’s focus back to image content, ensuring every detail is properly verified and thus balancing informativeness with correctness. BIRCH substantially reduces informativeness bias and delivers consistent gains across multiple VLMs and benchmarks.

8 Limitations

Our work analyzes the issues that undermine the reliability of VLM-as-a-Judge and proposes BIRCH to reduce informativeness bias and refocus on image content for more truthful judgments. BIRCH is simple to implement and requires no extra training.

However, informativeness bias still remains after applying BIRCH (Table 3), which shows that it does not fully resolve the problem. Future work could address the following issues. BIRCH builds an anchor by correcting candidate answers to encode key image details and guide the judge toward correctness. However, the model can still be influenced by informativeness during this process and fail to remove or fix incorrect claims, which reduces the quality of the anchor. Besides, even when BIRCH generates a reliable and accurate anchor, the judge may still favor a more informative but incorrect answer (Appendix D.3).

In addition, BIRCH is designed to boost performance by reducing informativeness bias. However, for models like Phi-4-Multimodal, where the bias is already small (Section C.6), the gains are limited (Table 3). Future work should explore ways to further strengthen judging ability in such models.

9 Ethical Considerations

In this paper, we introduce BIRCH, a new judging framework designed to make VLM-as-a-Judge more reliable and better aligned with human preferences. However, human annotators themselves are subject to bias, which in turn makes aligning VLM-as-a-Judge with their preferences also biased. We leave mitigating such biases in both benchmarks and VLM-as-a-Judge to future work.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. *Phi-4 technical report*. *Preprint*, arXiv:2412.08905.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. *Training a helpful and harmless assistant with reinforcement learning from human feedback*. *Preprint*, arXiv:2204.05862.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinyu Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. *MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark*. In *Forty-first International Conference on Machine Learning*.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. *Humans or llms as the judge? a study on judgement biases*. *Preprint*, arXiv:2402.10669.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. *Preprint*, arXiv:2507.06261.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. *Mme: A comprehensive evaluation benchmark for multimodal large language models*. *Preprint*, arXiv:2306.13394.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. *Blink: Multimodal large language models can see but not perceive*. In *Computer Vision – ECCV 2024*, pages 148–166. Cham. Springer Nature Switzerland.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. *A survey on llm-as-a-judge*. *Preprint*, arXiv:2411.15594.
- Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. *Recent advances in generative ai and large language models: Current status, challenges, and perspectives*. *IEEE Transactions on Artificial Intelligence*, 5(12):5873–5893.
- Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Zhengyu Chen, and Hui Xiong. 2025. *Explaining length bias in LLM-based preference evaluations*. In

- ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models.*
- Liqiang Jing and Xinya Du. 2025. [FGAIF: Aligning large vision-language models with fine-grained AI feedback](#). *Transactions on Machine Learning Research*.
- Akbar Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. [Debating with more persuasive LLMs leads to more truthful answers](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23662–23733. PMLR.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. [Benchmarking cognitive biases in large language models as evaluators](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [RewardBench: Evaluating reward models for language modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. [Mitigating object hallucinations in large vision-language models through visual contrastive decoding](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. [Naturalbench: Evaluating vision-language models on natural adversarial samples](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. [LLaVA-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2025b. [V1-rewardbench: A challenging benchmark for vision-language generative reward models](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24657–24668.
- Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2025c. [The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models](#). In *Forty-second International Conference on Machine Learning*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2025. [Paying more attention to image: A training-free method for alleviating hallucination in LLMs](#). In *Computer Vision – ECCV 2024*, pages 125–140, Cham. Springer Nature Switzerland.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark and others. 2024a. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024b. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. 2025. [Judge anything: Mllm as a judge across any modality](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25*, page 5742–5753, New York, NY, USA. Association for Computing Machinery.

- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in llm-as-a-judge](#). *Preprint*, arXiv:2406.07791.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. [Large language models are inconsistent and biased evaluators](#). *Preprint*, arXiv:2405.01724.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. [Judgebench: A benchmark for evaluating LLM-based judges](#). In *The Thirteenth International Conference on Learning Representations*.
- Tianyi Tang, Hongyuan Lu, Yuchen Jiang, Haoyang Huang, Dongdong Zhang, Xin Zhao, Tom Kocmi, and Furu Wei. 2024. [Not all metrics are guilty: Improving NLG evaluation by diversifying references](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6596–6610, Mexico City, Mexico. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, and 1 others. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024c. [Cogvlm: Visual expert for pre-trained language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499. Curran Associates, Inc.
- Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. 2025. [Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates](#). In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Tianyi Wu, Jingwei Ni, Bryan Hooi, Jiaheng Zhang, Elliott Ash, See-Kiong Ng, Mrinmaya Sachan, and Markus Leippold. 2025. [Balancing truthfulness and informativeness with uncertainty-aware instruction fine-tuning](#). *Preprint*, arXiv:2502.11962.
- Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. [Multimodal rewardbench: Holistic evaluation of reward models for vision language models](#). *Preprint*, arXiv:2502.14191.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. [Justice or prejudice? quantifying biases in LLM-as-a-judge](#). In *The Thirteenth International Conference on Learning Representations*.
- Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiao Cheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2025. [Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19985–19995.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. [Evaluating large language models at evaluating instruction following](#). In *The Twelfth International Conference on Learning Representations*.
- Qiyuan Zhang, Yufei Wang, Tiezheng YU, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025a. [Reviseval: Improving LLM-as-a-judge via response-adapted references](#). In *The Thirteenth International Conference on Learning Representations*.
- YiFan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu,

Bin Wen, Tingting Gao, Zhang Zhang, Fan Yang, Di ZHANG, Liang Wang, and Rong Jin. 2025b. [MM-RLHF: The next step forward in multimodal LLM alignment](#). In *Forty-second International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. 2025. [MLLMs are deeply affected by modality bias](#). *Preprint*, arXiv:2505.18657.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [LIMA: Less is more for alignment](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jin Peng Zhou, Sébastien M. R. Arnold, Nan Ding, Kilian Q. Weinberger, Nan Hua, and Fei Sha. 2025. [Graders should cheat: privileged information enables expert-level automated evaluations](#). *Preprint*, arXiv:2502.10961.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. [Calibrated self-rewarding vision language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

A Prompt Templates

A.1 Prompt for Comparing Informativeness

Table 18 presents the prompt we used with GPT-4o to select the more informative candidate answer y^i .

A.2 Prompt used in Standard Ref

Table 19 presents the prompt we used to generate the reference answer in Standard Ref baseline.

A.3 Prompts used in BIRCH

Prompt for generating anchor Table 20 shows the prompt we used to generate the anchor in BIRCH.

Prompt for judging Table 21 shows the prompt we use in BIRCH to judge with the anchor generated using the prompt in Table 20.

B Benchmark Details

B.1 Introduction

MLLM-as-a-Judge (Chen et al., 2024a) is the first benchmark to evaluate multimodal LLMs as judges. It contains 4,414 image-instruction pairs from 14 datasets covering tasks such as image captioning, math reasoning, text reading, and infographic understanding. Candidate answers are generated using six MLLMs: GPT-4V (OpenAI et al., 2024a), Gemini-Pro-Vision (Team et al., 2025), LLaVA-1.5-13b, LLaVA-1.6-34b (Liu et al., 2024), CogVLM (Wang et al., 2024c), and Qwen-VL-Max (Wang et al., 2024b). The benchmark provides three settings: (1) scoring evaluation, where each answer is rated from 1 to 5, (2) pairwise comparison, where two answers are compared to select the better one, and (3) batch ranking, where answers are ordered by quality. In our work, we use only the pairwise comparison setting, which includes 7,756 samples.

VL-RewardBench (Li et al., 2025b) focuses on the pairwise comparison setting, with 1,247 samples from 7 datasets. It includes hallucination-oriented queries in addition to general instruction-following and reasoning tasks, and provides error-type labels. To ensure the samples are challenging enough to expose model weaknesses, it only keeps cases where multiple smaller models (such as LLaVA-1.5/1.6-7B (Liu et al., 2024) and Qwen2-VL-7B (Wang et al., 2024b)) make errors.

B.2 Preprocessing

We found that human annotators made many errors when labeling the better candidate answer in the benchmarks we used. Figures 14, 15, 16 and 17 show examples of such errors. These mistakes undermine the reliability of evaluating VLM-as-a-Judge. To address this, we manually reviewed the data samples and corrected the identified errors before evaluation.

C More Experiment Details

C.1 Reliability of Informativeness Measuring

As described in Section 3.3, we use GPT-4o to measure informativeness, and our analysis depends on its reliability. To validate this, we randomly sample 100 examples from each benchmark, MLLM-as-a-Judge (Chen et al., 2024a) and VL-RewardBench (Li et al., 2025b), and ask human annotators to decide which candidate is more informative based on text alone, using the same rubric as GPT-4o (Figure 18). We then measure agreement between human and GPT-4o using match rate, which is the percentage of matching cases in $[0, 1]$, and Cohen’s kappa (Cohen, 1960) in $[-1, 1]$, where higher values indicate better agreement.

Table 7 shows that the match rate exceeds 0.9 and Cohen’s kappa exceeds 0.8, indicating strong agreement between GPT-4o and human judgments and confirming its reliability for comparing informativeness.

	MLLM-as-a-Judge	VL-RewardBench
Match rate	0.91	0.94
Cohen’s kappa	0.82	0.88

Table 7: Agreement between human annotators and GPT-4o on informativeness measurement.

C.2 Image Reliance Across Domains

Figure 11 shows a per-domain comparison of GPT-4o (OpenAI et al., 2024a), Llama-3.2-Vision-90B (Grattafiori et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025), with and without image input on the MLLM-as-a-Judge (Chen et al., 2024a) benchmark. Most domains lie close to the diagonal, showing that images add little benefit and sometimes even reduce performance.

C.3 Downsampling Details for Section 4.2

We compute informativeness bias (IB) and length bias (LB) by subtracting accuracy on different sub-

sets (Section 3.4 and 3.5). Since this can be affected by randomness, variance, and differences in sample selection or subset size, we mitigate it by down-sampling each subset at different ratios multiple times. We report the mean across all samples, and show variance bars in Figures 3 and 10.

Specifically, for IB the two subsets are \mathcal{D}_{IDS} and \mathcal{D}_{CDS} , and for LB they are $\mathcal{D}_{y^c \text{ is longer}}$ and $\mathcal{D}_{y^c \text{ is shorter}}$. We downsample each at 40%, 60%, 80%, and 100% (full set), with 10 samples per ratio. This results in $4 \times 4 \times 10 = 160$ downsamplings for each IB or LB computation.

The variance from this process is reasonably low (see Figures 3 and 10), confirming the validity of our results.

C.4 Informativeness Bias Across Domains

Figure 8 compares informativeness bias (IB) and length bias (LB) across domains in MLLM-as-a-Judge, showing that IB is higher than LB in almost all cases. Figure 9 presents Acc_{IDS} and Acc_{CDS} , the accuracies for the informativeness-driven and correctness-driven subsets. A large gap appears, with Acc_{IDS} much higher. Together, these results indicate that IB is far more significant than LB.

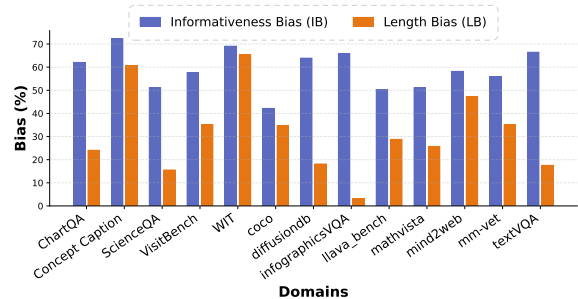


Figure 8: Informativeness bias (IB) and length bias (LB) comparison for different domains on MLLM-as-a-judge benchmark. IB is stronger than LB across all domains, confirming that informativeness is a greater source of bias than length.

C.5 Ablation of Length Bias

Inspired by (Ye et al., 2025; Hu et al., 2025), we separate length bias from informativeness bias by equalizing the lengths (measured in words) of the two answers. Specifically, before computing informativeness bias, we extend the shorter answer by asking GPT-4o to lengthen its existing content until it matches the longer one. We strictly instruct GPT-4o not to add any new information beyond

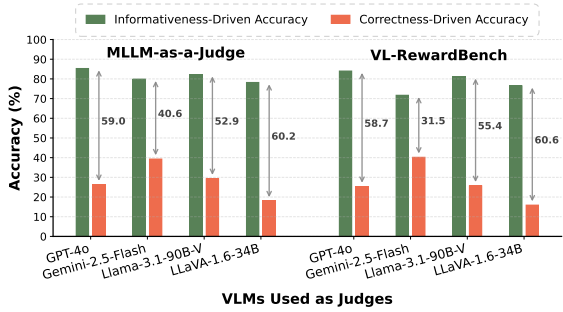


Figure 9: Accuracy on informativeness-driven (Acc_{IDS}) and correctness-driven (Acc_{CDS}) subsets of GPT-4o. Acc_{CDS} is much lower than Acc_{IDS} , revealing weak performance in cases where correctness should drive the decision.

the original answer, and to increase length only by repeating existing content or adding verbosity.

After this, we recheck the dataset and keep only samples where the length difference between the two answers is within a small threshold (5 words). This process removes the effect of length bias, since both answers are made nearly equal in length.

C.6 Lower Informativeness Bias for Phi-4-Multimodal

Figure 10 compares informativeness bias (IB) and length bias (LB) for Phi-4-Multimodal (Abdin et al., 2024). To reduce randomness and variance, we report mean and variance computed by downsampling \mathcal{D}_{IDS} and \mathcal{D}_{CDS} at different ratios multiple times, as described in Appendix C.3. Phi-4-Multimodal shows IB below 30% on both benchmarks, significantly lower than other models (see Figure 3, where IB is around 60%).

This explains why BIRCH yields smaller improvements on Phi-4-Multimodal, as it is designed to reduce IB, and Phi-4-Multimodal already has low IB. As shown in Table 8, BIRCH improves Acc_{CDS} by only 5.3% and reduces IB by 2.0% for Phi-4-Multimodal, much lower than the gains observed on other models.

C.7 Why BIRCH Works Well

Table 8 extends Table 4 and reports changes across multiple metrics for other models before and after applying BIRCH on the MLLM-as-a-Judge (Chen et al., 2024a) benchmark.

BIRCH consistently reduces informativeness bias (IB) by substantially improving Acc_{CDS} while keeping Acc_{IDS} strong across all models, leading to higher overall performance. Although designed

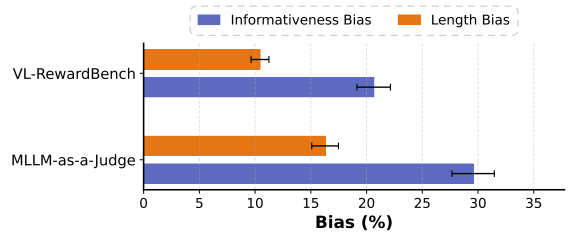


Figure 10: Informativeness bias (IB) and length bias (LB) for Phi-4-Multimodal (Abdin et al., 2024). The mean and variance are computed by downsampling \mathcal{D}_{IDS} and \mathcal{D}_{CDS} at different ratios, each sampled multiple times. Both IB and LB of Phi-4-Multimodal are significantly lower than in other models (see Figure 3). However, a clear gap between IB and LB remains, supporting our claim that IB drives LB.

to mitigate IB, BIRCH also greatly reduces length bias (LB), further supporting our finding that IB and LB are intertwined and that addressing IB also reduces LB.

D More Case Studies

D.1 Judges Misled by Informativeness, Mitigated by a Reference Answer

Figures 22, 23 (MLLM-as-a-Judge), and 24 (VL-RewardBench) provide additional qualitative examples showing how VLM-as-a-Judge is misled by a more informative but wrong answer, even when it clearly conflicts with the judge’s own response. However, when the judge uses its own answer as a reference, it can easily detect inconsistencies with the image and select the correct choice.

D.2 Why BIRCH outperforms Standard Ref?

Figures 25, 26, 27 (MLLM-as-a-Judge), and 28 (VL-RewardBench) present qualitative examples showing why Standard Ref, which generates a reference without considering the candidate answers, performs poorly in some cases, and why our BIRCH works better.

In Standard Ref, many details from the candidate answers may be missing in the reference. As a result, the judge often treats these extra but unverifiable details as wrong, even if they are actually correct. This leads to over-caution, an excessive focus on correctness, and weaker performance on informativeness-driven cases.

In contrast, BIRCH ensures that every detail in the candidates is reflected in the reference, either corrected, removed with explanation, or preserved

Methods	Acc	Acc _{IDS}	Acc _{CDS}	IB	LB
GPT-4o	66.5	85.3	26.3	59.0	32.9
+ BIRCH (Ours)	75.8 (+9.3)	90.1 (+4.8)	45.3 (+19)	44.8 (-14.2)	20.3 (-12.6)
Gemini-2.5-Flash	66.9	79.9	39.3	40.6	9.9
+ BIRCH (Ours)	73.4 (+6.5)	83.4 (+3.5)	52.0 (+12.7)	31.4 (-9.2)	9.2 (-0.7)
Llama-3.2-90B	65.3	82.2	29.3	52.9	26.7
+ BIRCH (Ours)	75.1 (+9.8)	86.6 (+4.4)	50.7 (+21.4)	35.9 (-17.0)	20.4 (-6.3)
LLaVA-OV-1.5-8B	65.4	80.4	33.3	47.1	25.8
+ BIRCH (Ours)	71.2 (+5.8)	82.9 (+2.5)	46.2 (+12.9)	36.7 (-10.4)	20.7 (-5.1)
Phi-4-multimodal	59.2	68.6	39.1	29.5	19.3
+ BIRCH (Ours)	63.1 (+3.9)	71.9 (+3.3)	44.4 (+5.3)	27.5 (-2.0)	14.6 (-5.3)

Table 8: Overall accuracy, accuracy on informativeness-driven cases Acc_{IDS} and correctness-driven cases Acc_{CDS}, informativeness bias (IB), and length bias (LB) before and after applying BIRCH on the MLLM-as-a-Judge (Chen et al., 2024a) benchmark. BIRCH reduces IB by greatly improving Acc_{CDS} while keeping Acc_{IDS} strong, leading to higher overall performance. Moreover, although designed to mitigate IB, BIRCH also greatly reduces LB.

if valid. This way, the judge can always verify each detail against the reference, maintaining attention to correctness with image content while also preserving informativeness.

D.3 Failure Case Analysis

We analyze failure cases where informativeness is misleading and explain why informativeness bias can persist even after applying BIRCH.

In Figure 29, the correct answer is 5,200 million. Answer A is correct, while Answer B incorrectly states 5.2 million. BIRCH fails to correct this inconsistency and includes it in the anchor. This may be due to limited visual or reasoning ability of the VLM, or because the model is still influenced by informativeness during anchor generation.

In Figure 30, Answer 1 correctly states that the balloon is filled with a light gas. Answer 2 is more informative but does not specify what the balloon is filled with, so Answer 1 is better. In this case, the anchor is reliable, but BIRCH is still misled by the more informative Answer 2. This shows that anchors alone are not sufficient to eliminate informativeness bias, and we leave this limitation to future work.

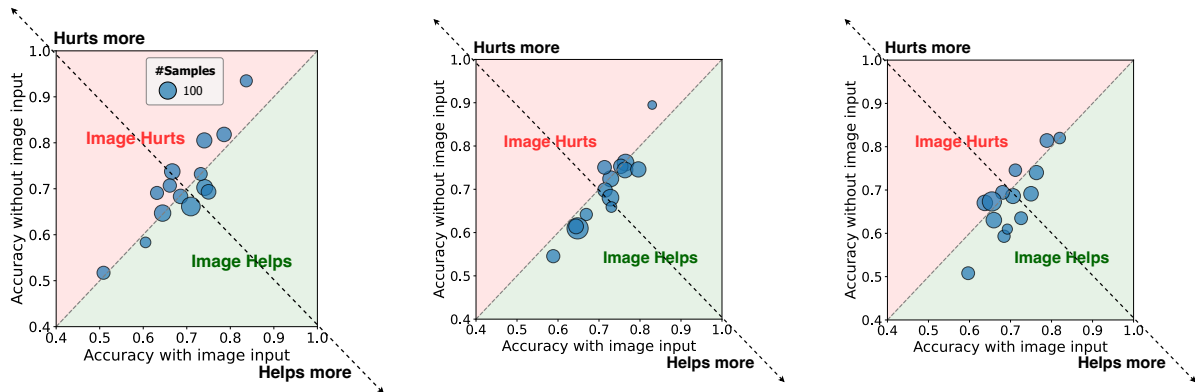


Figure 11: Per-domain comparison of GPT-4o (left), Llama-3.2-Vision-90B (middle), and Gemini-2.5-Flash (right), with and without image input, on the MLLM-as-a-Judge benchmark. Each scatter point represents a domain, with its size indicating the number of samples. The x-axis shows accuracy with image input; the y-axis shows accuracy without. Points in the green area indicate that the image improves accuracy, while points in the red area indicate that the image reduces accuracy. The farther a point is from the diagonal, the greater the impact (positive or negative) of the image. For all models, most domains cluster near the diagonal, showing that images provide little additional benefit.

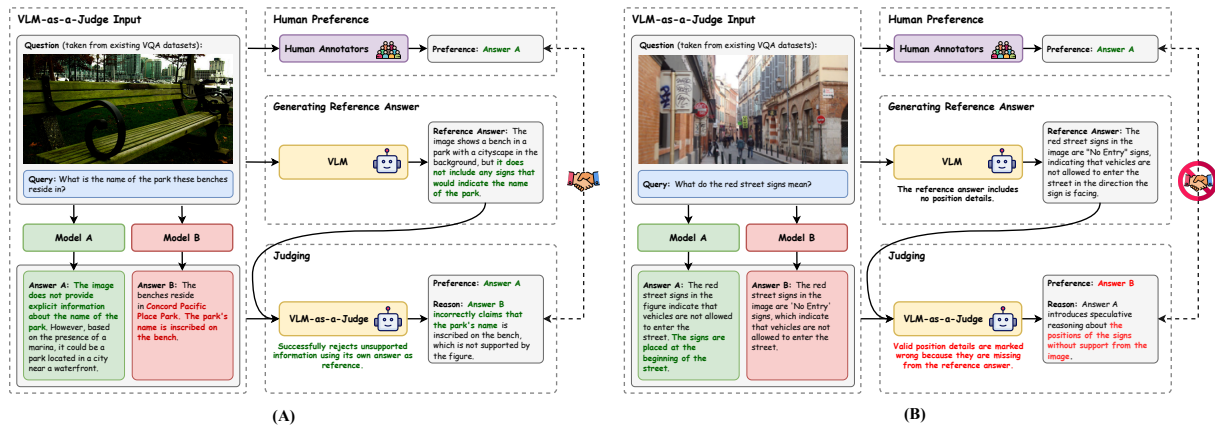


Figure 12: Successful and failed cases with the Standard Ref pipeline, where the judge first generates its own answer as a reference and then compares Answer A and B against it to decide which is better. **(A) Successful case:** Using the judge's answer "it does not include any signs that would indicate the name of the park" as reference, the judge correctly identifies an error in Answer B, which claims the park name is inscribed on the bench when it is not. This helps the judge overcome informativeness bias and make the right decision. **(B) Failed case:** Standard Ref can over-prioritize correctness and unfairly penalize valid details missing from the reference. Here, Answer A is rejected simply for mentioning the position (marked in green) of the "No Stop" sign, information that is valid but absent from the reference, leading it to be mistakenly judged as "unsupported by the image".

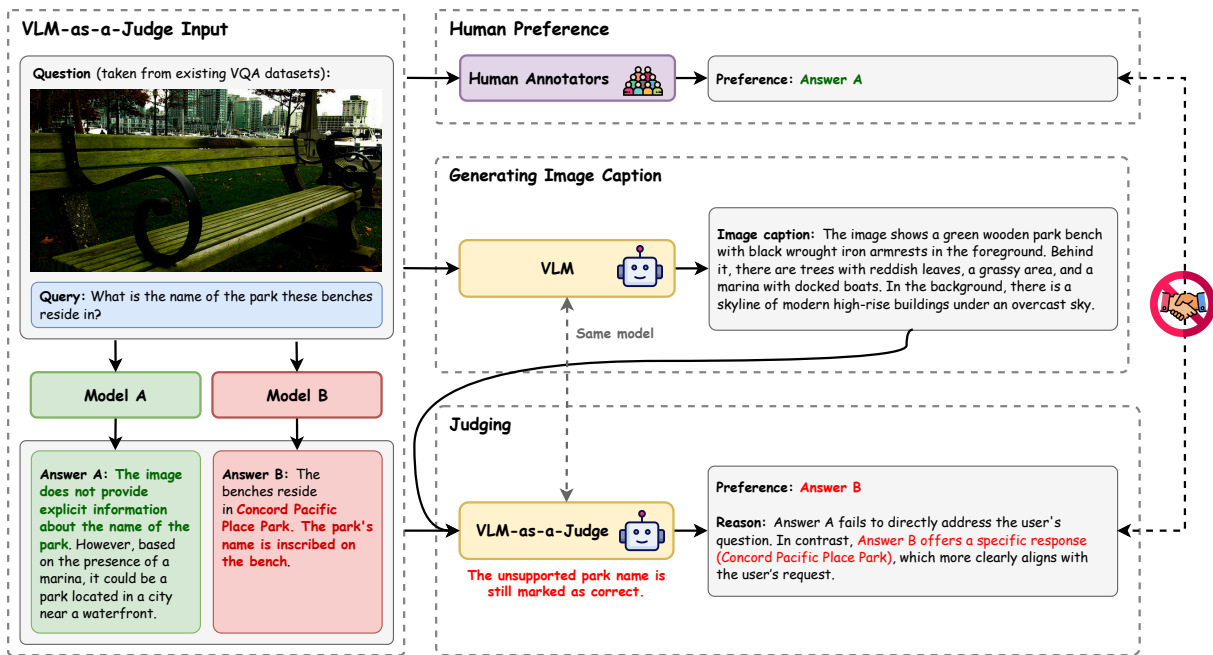



Figure 13: Illustration of the Image Caption baseline and why it performs poorly. First, the same VLM later used as the judge is asked to generate a caption for the image. Then, the judge takes the query, image, two candidate answers, and the caption as input, and decides which answer better addresses the query. However, the caption does not mention whether the park name is visible, so the judge still accepts the unsupported park name in Answer B as correct. This highlights the need for BIRCH, which ensures that every detail in the candidate answers is also represented in the reference.

Benchmark: VL-RewardBench



Query: Is the white food to the right of the meat on the right side?

Answer A: No, the white food is to the left of the meat on the right side. It is rice.

Answer B: Yes, the white food is rice, which is located to the right of the meat on the right side of the plate.

Human preference: Answer B

Figure 14: Example showing a human labeling error in VL-RewardBench. The white food (rice) is clearly to the left of the meat on the right side, but humans incorrectly chose B.

Benchmark: VL-RewardBench



Query: In which direction should I push the red car to bring it closer to the american flag? Up, down, right or left?

Answer A: To bring the red car closer to the American flag, you should push it to the left.

Answer B: To determine the direction in which to push the red car to bring it closer to the American flag, we can analyze the position of the red car relative to the American flag. Observing the flags' arrangement, **the American flag is positioned to the right of the red car**. Therefore, to move the car closer to the American flag, we would need to push the car towards the left. Based on this spatial analysis, the correct direction to push the red car is "left".

Human preference: Answer B

Figure 15: Example showing a human labeling error in VL-RewardBench. Answer B incorrectly states that the American flag is to the left of the red car, yet it was still chosen by humans.



Query: How many flags can be seen?

Answer A: There are many flags present in the scene, too many to count. They appear to be hanging on a fence and arranged in a long line along the roadside.

Answer B: There are 27 flags visible in this image. The flags appear to be hanging from vertical poles and arranged in a line along the street leading up to the clock tower.

Human preference: Answer B

Figure 16: Example showing a human labeling error in VL-RewardBench. Although precisely counting the flags here is very difficult, the number is clearly much higher than the 27 claimed in Answer B. Answer A, while not giving an exact count, is better since it avoids misleading information. However, the human still chose Answer A.



Query: How many tracks can be seen in the snow?

Answer A: In the snowy area, many tracks can be seen. This implies that many skiers may have skied in this ski resort, but the picture only captures the boy skiing alone in the snow.

Answer B: In the snowy ground, only a single trace can be seen. This suggests that the skier might be skiing alone, or they are the first one to ski on this particular path.

Human preference: Answer B

Figure 17: Example showing a human labeling error in VL-RewardBench. The image clearly shows multiple tracks on the snow ground, matching Answer A, but humans incorrectly chose B.

Instruction

Please act as an impartial judge and evaluate the informativeness of two candidate responses from different assistants for the same user instruction, which originally includes an image and a textual instruction. Your only job is to decide which response is more informative. You will not be given the image. Do NOT consider factual correctness or image relevance, just evaluate informativeness.

Definition of informativeness:

1. Relevance: Does the response address the user's intent clearly?
2. Completeness: Does it provide necessary information or explanation the user would need?
3. Conciseness: Does it avoid unnecessary details or digressions?

Your evaluation should focus on these dimensions: helpfulness, relevance, depth, and the level of detail. There is no tie. Choose the more informative one. Output your reasoning first, then the decision.

Output format

Analysis: <1-2 sentence explanation>

Judgement: [[A]] or [[B]]

Inputs

[The Start of User Instruction] { user query } [The End of User Instruction]

[The Start of Assistant A's Answer] { Answer A } [The End of Assistant A's Answer]

[The Start of Assistant B's Answer] { Answer B } [The End of Assistant B's Answer]

Figure 18: Prompt for comparing informativeness. We provide GPT-4o with the original query and two candidate answers (without the image) and ask it to focus only on helpfulness, relevance, depth, and level of detail when determining which answer is more informative, ignoring image alignment correctness.

Instruction

You are a helpful and knowledgeable assistant. The user has provided an image and a question or instruction related to it. Carefully analyze the image and follow the instruction to generate a clear, concise answer.

Inputs

{ image }

[The Start of User Instruction] { user query } [The End of User Instruction]

Figure 19: Prompt for generating reference answers in Standard Ref baseline.

Instruction

Please act as a skilled reviser and editor to refine AI assistant responses based on a user instruction referring to a figure.

You will be given two assistant-generated responses. These responses are NOT guaranteed to be correct or truthful. Your tasks are:

1. Revise each response individually: Correct any errors or inconsistencies with the image content. Ensure the response is accurate, truthful, and fully addresses the user's instruction. Carefully search for mistakes, since the response may not be fully correct or aligned with the image. Only make revisions when necessary, and DO NOT alter the original style or tone. Remove irrelevant or speculative parts that cannot reasonably be inferred from the image, and leave comments explaining what was removed and why. If the original response is good enough, leave it unchanged.
2. Merge the two revised responses into one unified response: Eliminate redundancy and overlap between the two responses. When both responses mention the same idea, keep the version that is more accurate, helpful, relevant, comprehensive, creative, and detailed. Maintain as much of the original style and tone as possible while merging.

Be precise and objective in your edits.

Output format

Revised Response A: <revised response A>

Revised Response B: <revised response B>

Merged Response: <merged response>

Inputs

{ image }

[The Start of User Instruction] { user query } [The End of User Instruction]

[The Start of Assistant A's Answer] { Answer A } [The End of Assistant A's Answer]

[The Start of Assistant B's Answer] { Answer B } [The End of Assistant B's Answer]

Figure 20: Prompt for generating an anchor in BIRCH.

Instruction

Please act as an impartial judge and evaluate the quality of two responses generated by different AI assistants for the same user question and figure.

You will be provided with:

1. A user instruction
2. A figure (relevant to the instruction)
3. A reliable and accurate reference answer
4. Two candidate responses (Assistant A and Assistant B)

You should carefully examine the figure and compare the two responses with the reference answer.

You should then choose the assistant that better follows the user's instructions and more effectively answers the user's question.

Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses.

Do not let the position or order of responses influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.

Output format

Analysis: <1-2 sentence explanation comparing Assistant A and Assistant B.>

Judgement: [[A]] or [[B]] (Choose [[A]] if Assistant A is better, or [[B]] if Assistant B is better)

Inputs

{ image }

[The Start of User Instruction] { user query } [The End of User Instruction]

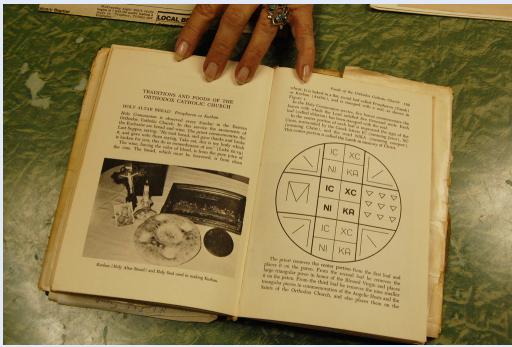
[The Start of Assistant A's Answer] { Answer A } [The End of Assistant A's Answer]

[The Start of Assistant B's Answer] { Answer B } [The End of Assistant B's Answer]

[The Start of Reference Answer] { Reference Answer } [The End of Reference Answer]

Figure 21: Prompt for judging with reference answer in BIRCH.

Benchmark: MLLM-as-a-Judge



Query: This is a task of text reading on natural image. Please analyze this figure in detail and answer the following question with reason based on this figure: what is written on the top of the left page?

Zoom in to see the image details more clearly.

Candidate Answers

Answer A (chosen): TRADITIONS AND FOODS OF THE ORTHODOX CATHOLIC CHURCH

Answer B (rejected): This is a picture of two open books. The left page has a black and white photo of an altar with a loaf of bread on it. The caption below the photo reads "Kurban (Holy Altar) and Holy Seal used in making Kurban." **The text on the right page is titled "Traditions and Foods of the Orthodox Catholic Church."** The first paragraph on this page begins with "Holy Communion is observed every Sunday in the Orthodox Catholic Church. In the Eastern Orthodox Church, wine, having the color of blood, and pure unleavened bread are used." Based on the information provided, we can infer that **the text on the top of the left page is "Kurban (Holy Altar) and Holy Seal used in making Kurban."**

VLM-as-a-Judge (GPT-4o)

Without Reference:

- **Preference:** Answer B
- **Reason:** Assistant A's response directly provides the text 'TRADITIONS AND FOODS OF THE ORTHODOX CATHOLIC CHURCH' as the answer to the user's question. **However, this is incorrect because the provided image shows that this text is on the top of the right page, not the left.** Assistant A does not analyze the image in detail or provide reasoning to support the answer. Assistant B's response provides a detailed description of the image, including the content of both pages. **The response of Assistant B demonstrates a more comprehensive analysis of the image and reasoning process compared to Assistant A.**

With Reference:

- **Preference:** Answer A
- **Reference answer generated by the judge itself:** The text written at the top of the left page is: "TRADITIONS AND FOODS OF THE ORTHODOX CATHOLIC CHURCH. This title introduces the content on the page, which discusses the preparation and significance of "Holy Altar Bread" in the Orthodox Catholic Church.
- **Reason:** Assistant A correctly identifies the text at the top of the left page as "TRADITIONS AND FOODS OF THE ORTHODOX CATHOLIC CHURCH," which matches the reference answer. Assistant B, on the other hand, provides a detailed description of the image **but incorrectly infers that the text at the top of the left page is "Kurban (Holy Altar) and Holy Seal used in making Kurban," which is actually the caption below the image.** This misidentification makes Assistant B's response less accurate despite its detailed explanation.

Figure 22: In this example, VLM-as-a-Judge prefers Answer B, which gives more details about the book but makes two mistakes: it treats the figure caption as text on the top of the left page, and it misplaces the actual text from the top of the left page onto the right page. The model does recognize the correct text on the top of the left page, yet still judges Answer B as correct. Using the model's own generated answer as a reference helps mitigate this issue.

Benchmark: MLLM-as-a-Judge



Query: Please analyse this figure in detail and follow the following instruction based on this figure: Generate a question based on the given image such that it receives the following answer: cigarette

Zoom in to see the image details more clearly.

Candidate Answers

Answer A (rejected): The image depicts a person riding a motorcycle with a helmet on. The person has tattoos on their left arm and is wearing a sleeveless shirt, which allows for the visibility of the tattoos. **On the person's right hand**, there is an object that appears to be a cigarette between the fingers, suggested by its slim, cylindrical shape and the way it's being held, typical of how one would hold a cigarette. The setting appears to be outdoors during daytime, as indicated by the sunlight and shadows. Based on these observations, a suitable question to elicit the response 'cigarette' could relate to **what the person is holding in their right hand** while operating the motorcycle.

Answer B (chosen): What is the man on the motorcycle **holding in his mouth**? The answer is "cigarette". The man's mouth is slightly open and there is a small, cylindrical object between his lips. The object is likely a cigarette because it is thin and has a filter on the end. The man is also holding the object in a way that suggests that he is smoking it.

VLM-as-a-Judge (LLaMA-3.2-Vision-Instruct-90B)

Without Reference Answer:

- **Preference:** Answer B
- **Reason:** Assistant A provides a detailed description of the image, focusing on the person riding the motorcycle and the object in their right hand, which is identified as a cigarette. The assistant suggests a question that aligns with the user's instruction by asking about what the person is holding in their right hand. The response is comprehensive and adheres to the instruction effectively. Assistant B proposes a question about what the man is holding in his mouth. While the response is relevant, **it assumes the cigarette is in the mouth without clear evidence from the image, which could lead to inaccuracies**. Additionally, **the explanation is less detailed compared to Assistant A's response**.

With Reference Answer:

- **Preference:** Answer B
- **Reference answer generated by the judge itself:** What object is the person riding the motorcycle **holding in their mouth**?
- **Reason:** Assistant A provides a detailed description of the image but **incorrectly identifies the cigarette as being held in the person's right hand**, which is inconsistent with the reference answer and the image. **Assistant B correctly identifies the cigarette as being held in the person's mouth**, aligning with the reference answer and the image. Assistant B's response is more accurate and helpful.

Figure 23: In this example, VLM-as-a-Judge prefers Answer A, which provides more details about the environment but incorrectly states that the cigarette is in the person's right hand when it is actually in their mouth. The model does recognize the correct location of the cigarette, yet still judges Answer A as correct. Using its own generated answer as a reference helps mitigate this issue.

Benchmark: VL-RewardBench



Query: Convey the specifics of the image with meticulous attention to detail.

Candidate Answers

Answer A (chosen): A black cat with white fur around its belly is sprawled out on a laptop keyboard, which has an open screen.

Answer B (rejected): A black cat lounges on a laptop, its fur in stark contrast to the silver computer. A white patch adorns the animal's chest, while its ears perk up attentively.

VLM-as-a-Judge (Phi-4-Multimodal)

Without Reference Answer:

- **Preference:** Answer B
- **Reason:** Assistant A misses some details, such as the cat's posture, the silver color of the laptop, and the cat's attentive ears. Assistant B offers a more detailed response, highlighting the black cat's contrast with the silver laptop.

With Reference Answer:

- **Preference:** Answer A
- **Reference answer generated by the judge itself:** A black cat with a white patch of fur on its lower belly is sprawled on a black laptop. The laptop screen is open, displaying a blue background.
- **Reason:** Assistant B, while descriptive, incorrectly states that the laptop is silver, deviating from the image and reference answer.

Figure 24: In this example, VLM-as-a-Judge prefers Answer B, which provides more details about the cat but incorrectly claims the laptop is silver when it is actually black. The model does recognize the correct color of the laptop, yet still judges Answer B as correct. Using its own generated answer as a reference helps mitigate this issue.

Benchmark: MLLM-as-a-Judge



Query: Please analyse this figure in detail and answer the following question based on this figure: What do the red street signs mean?

Candidate Answers

Answer A (rejected): The red street signs in the image are 'No Entry' signs, which indicate that vehicles are not allowed to enter the street.

Answer B (chosen): The red street signs in the figure indicate that vehicles are not allowed to enter the street. The signs are placed at the beginning of the street and are visible to drivers approaching the street. The red street signs are a traffic control device used to regulate traffic flow and keep pedestrians safe.

VLM-as-a-Judge (GPT-4o)

Standard Ref:

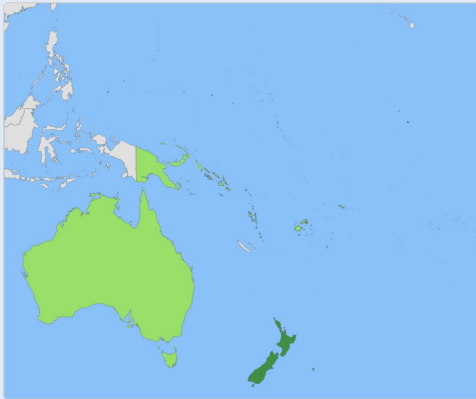
- **Preference:** Answer A
- **Reference answer:** The red street signs in the image are circular with a white horizontal bar in the center. These are "No Entry" signs, indicating that vehicles are not allowed to enter the street in the direction the sign is facing.
- **Reason:** Assistant A correctly identifies the red street signs as "No Entry" signs. Assistant B also identifies the signs as indicating "No Entry" and elaborates on their function as traffic control devices, emphasizing their role in regulating traffic and ensuring safety. However, Assistant B's response includes speculative details (e.g., placement at the beginning of the street) that are not supported by the image. Assistant A is more aligned with the reference answer.

BIRCH:

- **Preference:** Answer B
- **Anchor:** The red street signs in the image are 'No Entry' signs, clearly indicating that vehicles are prohibited from entering the street. Positioned at the entrance and visible to approaching drivers, these signs serve as a traffic control measure to regulate access and ensure pedestrian safety.
- **Reason:** Assistant A provides a concise and accurate explanation of the red street signs, aligning well with the reference answer but lacks depth and detail. Assistant B offers a more comprehensive response, including additional context about the placement and purpose of the signs, which enhances the explanation and aligns more closely with the reference answer's depth and detail.

Figure 25: In this example, Standard Ref rejects the better Answer B, judging its placement of the sign as speculative since this detail is not in the reference. This shows how it over-penalizes informativeness. In contrast, BIRCH ensures such placement details are included in the generated anchor, allowing it to correctly verify Answer B and achieve a better balance between informativeness and correctness.

Benchmark: MLLM-as-a-Judge



Query: Which country is highlighted?

Candidate Answers

Answer A (chosen): The image you've provided is a map showing a large portion of the Oceania region. **The highlighted country, in green, is Australia.** Australia is the only country in the world that is also a continent by itself and is characterized by its considerable size covering the majority of the landmass shown at the center of the image. To the southeast of Australia, the two large islands highlighted in green represent New Zealand, which is another country.

Answer B (rejected): New Zealand is highlighted.

VLM-as-a-Judge (LLaMA-3.2-Vision-Instruct-90B)

Standard Ref:

- **Preference:** Answer B
- **Reference answer:** The country highlighted on the map is New Zealand, which is located to the southeast of Australia and consists of two main islands, North Island and South Island, as well as numerous smaller islands.
- **Reason:** Assistant A incorrectly identifies the highlighted country as Australia, despite the image clearly showing New Zealand in dark green. Assistant B correctly identifies the highlighted country as New Zealand, aligning perfectly with the reference answer and the image. However, Assistant B's response is brief and lacks additional context, though this does not affect its correctness.

BIRCH:

- **Preference:** Answer A
- **Anchor:** The image depicts a map of the Oceania region, with two countries highlighted in green: Australia and New Zealand. **Australia, the larger landmass located at the center-left of the image,** is both a country and a continent. To the southeast of Australia, New Zealand is highlighted, consisting of two smaller islands. Both countries are distinctly marked in green.
- **Reason:** Assistant A provides a detailed and accurate response, **identifying both Australia and New Zealand as the highlighted countries** in the image. The explanation includes relevant context about their locations and characteristics, aligning closely with the reference answer. In contrast, Assistant B's response is incomplete and incorrect, as it **only mentions New Zealand and ignores Australia, which is also highlighted** in the image.

Figure 26: In this example, Standard Ref rejects the better Answer A, assuming Australia is not the highlighted country because the reference answer does not mention it. This shows how it over-penalizes informativeness. In contrast, BIRCH includes the valid mention of Australia in the generated anchor, allowing it to correctly verify Answer A and achieve a better balance between informativeness and correctness.

Benchmark: MLLM-as-a-Judge



Query: What is the make of the car on the left?

Candidate Answers

Answer A (rejected): The make of the car on the left appears to be Volkswagen, as indicated by the logo visible on its front.

Answer B (chosen): The car on the left appears to be a Volkswagen (VW), as indicated by the distinctive front grille design and the logo on the hood. Volkswagen is a German multinational automotive manufacturing company known for its wide range of vehicles, including sedans, hatchbacks, and SUVs. The specific model of this car is not clearly identifiable from this angle and without more context, but **it's part of the brand's lineup of electric or hybrid vehicles, as evidenced by the presence of a charging cable and plug connected to the car.**

VLM-as-a-Judge (Gemini-2.5-Flash)

Standard Ref:

- **Preference:** Answer A
- **Reference answer:** The car on the left is a Volkswagen, as indicated by the logo on the front grille.
- **Reason:** Assistant A provides a concise and accurate response, correctly identifying the make of the car on the left as Volkswagen based on the visible logo. Assistant B also identifies the car as a Volkswagen **but speculates about the car's model and type (electric or hybrid), which is not supported by the image.** This extra information does not enhance the response and could potentially mislead.

BIRCH:

- **Preference:** Answer B
- **Anchor:** The car on the left appears to be a Volkswagen (VW), as identified by the logo on its front grille. **It is connected to a charging station via a cable, indicating it is likely part of Volkswagen's lineup of electric or hybrid vehicles.** The specific model, however, cannot be determined from this angle.
- **Reason:** Assistant A provides a concise and accurate answer, identifying the make of the car as Volkswagen based on the visible logo. However, it lacks additional context or detail. Assistant B also correctly identifies the make as Volkswagen and provides more depth by mentioning the grille design, the charging cable, and **the likelihood of the car being part of Volkswagen's electric or hybrid lineup.** This additional context aligns well with the reference answer, making Assistant B's response more comprehensive and informative.

Figure 27: In this example, Standard Ref rejects the better Answer B, reasoning that electric or hybrid vehicles cannot be inferred from the image (even though the charging cable and plug suggest this) because the reference answer does not mention it. This shows how it over-penalizes informativeness. In contrast, BIRCH includes the likelihood of the car being electric or hybrid in the generated anchor, allowing it to correctly verify Answer B and achieve a better balance between informativeness and correctness.

Benchmark: VL-RewardBench



Query: Look at this image and determine which of the following animals is present: Boar, Badger Bird, Unknown

Candidate Answers

Answer A (rejected): Boar

Answer B (chosen): Based on the image, which shows a wild boar or pig foraging at a feeding station in a dark forest at night, the animal present is a Boar.

VLM-as-a-Judge (LLaVA-OneVision-1.5-8B)

Standard Ref:

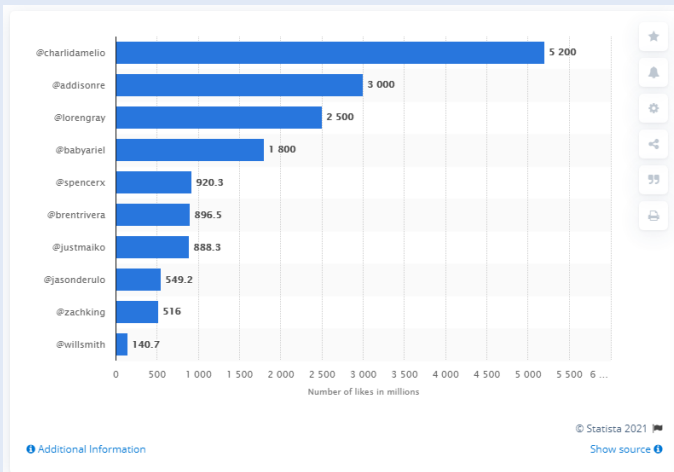
- **Preference:** Answer B
- **Reference answer:** The animals present in the image are boars.
- **Reason:** Assistant B adds **extra context (a boar foraging at a feeding station) which is not supported** by the image or the reference answer.

BIRCH:

- **Preference:** Answer A
- **Anchor:** Based on the image, which shows a wild boar foraging at a feeding station in a dark forest at night, the animal present is a Boar.
- **Reason:** Assistant B's response goes further by including details about the setting (a wild boar foraging at a **feeding station** in a dark forest at night), showing a clear understanding of the image and supports the conclusion.

Figure 28: In this example, Standard Ref rejects the better Answer B, judging “boar foraging at a feeding station” as unsupported because the reference answer does not mention it. This shows how it over-penalizes informativeness. In contrast, BIRCH includes this context in the generated anchor, allowing it to correctly verify Answer B and achieve a better balance between informativeness and correctness.

Benchmark: MLLM-as-a-Judge



Query: How many likes did Charli D'Amelio have on her content?

Candidate Answers

Answer A (chosen): Charli D'Amelio had 5,200 million likes on her content.

Answer B (rejected): Charli D'Amelio had 5.2 million likes on her content. The figure shows the number of likes on the content of several TikTok users. Charli D'Amelio is the first user on the list, and the number of likes on her content is shown as 5.2 million.

VLM-as-a-Judge (GPT-4o)

BIRCH:

- **Preference:** Answer B
- **Anchor:** Charli D'Amelio had 5.2 million likes on her content. This is evident from the bar chart, where she is listed as the first user, indicating the highest number of likes among all the individuals shown.
- **Reason:** Answer B identifies her like count as 5.2 million, which matches the reference answer.

Figure 29: Failure case study. BIRCH generates an anchor with incorrect visual reasoning or is misled by the wrong Answer B, using 5.2 million as the like count instead of the correct 5,200 million. Based on this anchor, the judge incorrectly prefers Answer A.

Benchmark: MLLM-as-a-Judge



Query: What is filled in that oval object the boy is holding?

Candidate Answers

Answer A (chosen): The oval object the boy is holding appears to be a balloon. It is filled with a light gas, which is commonly used in balloons to make them float.

Answer B (rejected): That is a balloon. The object is oval-shaped, and the boy is holding its string. The object is also reflective, as light is reflecting off of its surface. This indicates that the object is not porous and is likely made of a smooth material. The object is also lightweight, as it is floating in the air. All of these properties indicate that the object is a balloon.

VLM-as-a-Judge (GPT-4o)

BIRCH:

- **Preference:** Answer B
- **Anchor:** The oval object the boy is holding is a white balloon, and he is holding its string. Given that it appears lightweight and to be floating in the air, it is filled with gas, likely helium.
- **Reason:** Answer B offers a more detailed description.

Figure 30: Failure case study. BIRCH generates a reliable anchor stating that the balloon held by the boy is filled with a light gas, likely helium. However, the judge still incorrectly selects the more informative Answer B, which does not specify what the balloon is filled with.