SDGBiasBench: Benchmarking and Mitigating Vision–Language Models' Biases in Sustainable Development Goals

Zihang Lin [1,2,*] **Huaiyuan Qin** [2,†] **Muli Yang** [2] **Hongyuan Zhu** [2,✉]

[1]Nanyang Technological University
[2]Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore
[†]Project Lead. [✉]Corresponding Author.
linz0071@e.ntu.edu.sg, {qinhy, yangml, zhuh}@a-star.edu.sg

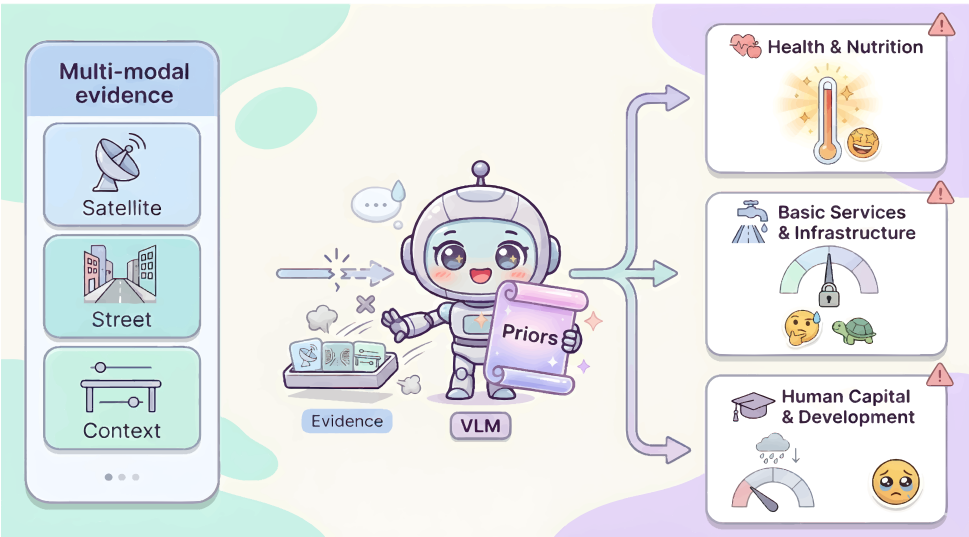

Figure 1: Sustainable Development Goals (SDGs) monitoring demands multi-step reasoning over satellite imagery, street-level cues, and structured context. However, VLMs always tend to under-use or mis-integrate this evidence and instead rely on SDG-specific priors, yielding directionally skewed predictions across SDG domains. We therefore introduce ***SDGBiasBench*** to expose decision- and estimation-level SDG bias under controlled evidence availability.

Abstract

Assessing progress toward the Sustainable Development Goals (SDGs) requires multi-step reasoning over visual cues, contextual knowledge, and development indicators, where incomplete evidence use and imperfect evidence integration can introduce hidden prediction biases. Real-world SDG monitoring further spans both qualitative judgments and quantitative estimation. However, existing benchmarks typically evaluate these aspects in isolation, obscuring systematic biases that emerge when models substitute priors for evidence. To address this gap, we propose **SDGBiasBench**, a large-scale benchmark suite for SDG-oriented vision–language reasoning. Spanning **500k** expert-involved multiple-choice questions and **50k** regression tasks, the benchmark enables comprehensive assessment of both decision-level and estimation-level bias in Vision–Language Models (VLMs). Evaluations on SDGBiasBench reveal an intrinsic **SDG bias** in current VLMs, where predictions are frequently driven by SDG-specific priors rather than reliable multi-modal cues. To mitigate such bias, we propose **CADE** (**C**ontrastive **A**daptive **D**ebias **E**nsemble), a training-free, plug-and-play method that leverages modality-specific answer priors. CADE yields significant gains on the proposed benchmark, improving multiple-choice accuracy by up to 25% and reducing regression MAE by up to 12 points across multiple VLMs. We hope our work can foster the development of more fair and reliable AI systems for sustainable development. Code is available at: SDGBiasBench.

1 Introduction

Established by the United Nations, Sustainable Development Goals (SDGs) define a shared global agenda for sustainable and inclusive development across social, economic, and environmental dimensions (United Nations General Assembly, 2015). Vision–language foundation models (Bommasani et al., 2021; Radford et al., 2021; Alayrac et al.,

*This paper was completely accomplished when Zihang Lin interned under Dr. Hongyuan Zhu's supervision at A*STAR.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 1374–1394
October 24-29, 2026 ©2026 Association for Computational Linguistics

2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023, 2024; Chen et al., 2023; Bai et al., 2025; Chen et al., 2024; Li et al., 2024) provide a promising paradigm for SDG-related monitoring by jointly leveraging visual and textual signals, complementing prior sustainability work that uses satellite imagery to measure development outcomes (Jean et al., 2016; Burke et al., 2021; Yeh et al., 2021).

Prior work has documented geographic and socio-economic biases in foundation models used in sustainability and geospatial settings (Manvi et al., 2024; Huang et al., 2025b; Pouget et al., 2024). Since SDG applications are closely related to global development, resource allocation, and policy-making, biased behavior in these models risks amplifying inequities and producing unreliable evidence (Bender et al., 2021; Mitchell et al., 2019; Hardt et al., 2016). Ensuring fair and trustworthy AI for sustainable development therefore necessitates a dedicated framework to systematically evaluate and address such biases (Bommasani et al., 2021; Mitchell et al., 2019).

To this end, we propose **SDGBiasBench**, a large-scale benchmark suite for vision–language reasoning over SDG-related indicators, explicitly designed to diagnose sustainability-oriented bias (Yeh et al., 2021; Manvi et al., 2024). The benchmark comprises expert-involved, multi-step, multi-modal questions spanning both qualitative judgments (*e.g.* risk levels, comparative conditions) and quantitative estimation (*e.g.* rates, indices, magnitudes), enabling comprehensive evaluation of model performance on sustainability-oriented prediction tasks (Antol et al., 2015; Hudson and Manning, 2019; Lu et al., 2022; Fu et al., 2025). By enforcing multi-step, multi-modal reasoning and evaluating behavior under controlled evidence conditions (Niu et al., 2021; Zhang et al., 2022, 2025), SDGBiasBench exposes whether predictions track the provided inputs or remain dominated by bias.

We empirically reveal a pronounced **SDG bias** in Vision–Language Models (VLMs) that manifests as pillar-conditioned directional defaults and imbalanced multi-modal grounding (Manvi et al., 2024; Zhou et al., 2022; Ruggeri et al., 2023; Weng et al., 2024a) using SDGBiasBench. Across the three pillars in SDGBiasBench, VLMs exhibit systematic optimistic, conservative, or pessimistic leanings that are triggered differently by health, infrastructure, and human-capital indicators, forming distinctive bias signatures (Zhou et al., 2022; Ruggeri et al., 2023). Meanwhile, even when imagery contains informative spatial cues (*e.g.* settlement patterns, road connectivity, land use), models often under-weight visual evidence relative to textual cues (Zhang et al., 2025; Leng et al., 2024; Li et al., 2023c), leading predictions to follow entrenched geographic correlations rather than grounded inspection (Manvi et al., 2024; Burke et al., 2021).

To mitigate such bias, we introduce a simple yet effective training-free, plug-and-play approach **CADE** (**C**ontrastive **A**daptive **D**ebias **E**nsemble), which detects over-reliance on priors by comparing predictions from different constrained input views, adaptively reweighting outputs to favor visual evidence over context (Niu et al., 2021; Zhang et al., 2025; Leng et al., 2024; Janghorbani and De Melo, 2023; Weng et al., 2024b). Without introducing training overhead during model inference, CADE yields remarkable gains, improving multiple-choice accuracy and reducing regression MAE significantly across multiple VLMs, promoting more fair and evidence-based predictions.

In summary, our contributions are:

1. **Benchmark Suite for Revealing SDG Bias**: We introduce **SDGBiasBench**, the first large-scale benchmark suite for both qualitative judgments and quantitative estimation on diagnosing bias across Sustainable Development Goal (SDG) indicators.

2. **Identification of SDG Bias**: Our analysis reveals pillar-conditioned directional defaults and modality imbalance in SDG reasoning, raising fairness concerns for SDG applications.

3. **Training-free Debiasing Baseline**: We propose **CADE**, a plug-and-play debiasing method that adaptively reweights multiple input views to mitigate SDG bias thus improve performance, serving as a simple-but-effective baseline.

2 SDGBiasBench: An Overview

SDGBiasBench aims to evaluate vision–language reasoning on real-world Sustainable Development Goals (SDGs) related indicators to reveal SDG biases. As illustrated in Fig. 2, we organize the benchmark into three pillars: ***Health & Nutrition***, ***Basic Services & Infrastructure***, and ***Human Capital & Development***, each grouping indicators with shared real-world measurement and evidence patterns. The benchmark comprises two task types: qualitative judgments and quantitative estimation

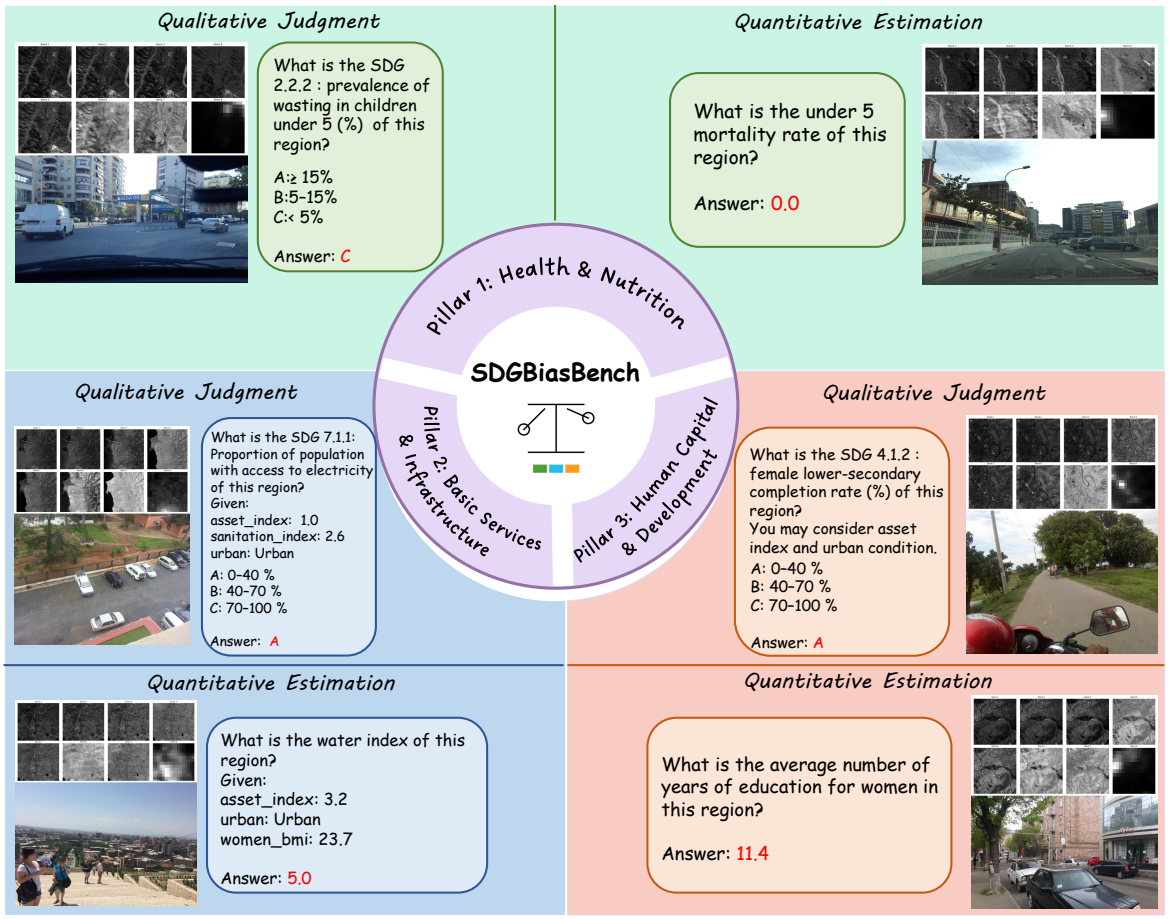

Figure 2: **Overview of *SDGBiasBench***. The three sustainability pillars are each illustrated with one qualitative judgment and one quantitative estimation example, showcasing the multi-modal SDG reasoning tasks used to probe SDG biases in VLMs.

to mirror real-world scenarios. Specifically, we use multiple-choice questions for qualitative judgments and regression tasks for quantitative estimation, respectively. Each task is paired with satellite imagery, structured context variables (*e.g.* sanitation index, women's education level, asset index) and often augmented with street-level photos, requiring multi-step reasoning across modalities.

The three pillars capture distinct but complementary forms of sustainable development reasoning. ***Health & Nutrition*** focuses on inferring latent health conditions, nutritional status, and healthcare access from indirect environmental and socio-economic cues. ***Basic Services & Infrastructure*** centers on access to essential services such as electricity, water, and sanitation, requiring models to relate infrastructure and environmental patterns to underlying service quality and deprivation. ***Human Capital & Development*** targets broader reasoning on education and composite development outcomes, where signals are more diffuse and must be inferred from holistic socio-economic context.

Human experts are involved to provide high-quality answers for diverse reasoning tasks for sustainable development that require grounded multi-modal inference beyond surface-level perception. We provide more details in Sec. A including annotation protocols, dataset statistics, and benchmark composition and quality-control details.

3 Empirical Analysis of SDG Bias

In this section, we use SDGBiasBench as a diagnostic lens to evaluate VLMs' behaviors on SDG-related questions. Although reliance on language priors is a known phenomenon in VLMs, our findings suggest an SDG-specific mechanism, which we define as **SDG Bias**: *when inferring the SDG indicators, models (i) produce* ***pillar-conditioned directional defaults****: systematically being optimistic/conservative/pessimistic in ways that differ across health, infrastructure, and human capital; (ii) exhibit* ***modality-imbalanced evidence use****: treating structured SDG context as largely sufficient and under-utilizing aligned imagery.* These signatures motivate our view-aware debiasing method, which will be further elaborated

in Sec. 4.

3.1 Setup: Views, Models, and Metrics

We evaluate three representative 7B-parameter VLMs: LLaVA-v1.5 (Liu et al., 2024), InstructBLIP (Dai et al., 2023), Qwen2.5-VL (Bai et al., 2025) on SDGBiasBench under four *input views*. For each datapoint i, let I_i denote the image(s) (satellite and/or street-level), C_i the structured context (auxiliary variables/indices), and Q_i the question. We define the four inference prompts:

- **Q-only** (T^{q}): (Q_i)
- **CTX+Q** (T^{ctx}): (C_i, Q_i)
- **IMG+Q** (T^{img}): (I_i, Q_i)
- **Full** (T^{full}): (I_i, C_i, Q_i)

In the main text, we use the multiple-choice questions (MCQs) as our primary diagnostic signal and report **Accuracy** as the metric.[1] A summary of per-view MCQ performance is given in Fig. 3.

3.2 Pillar-conditioned Directional Defaults Form an SDG Bias Signature

SDG bias manifests as **directional defaults** that are **pillar-conditioned**. To isolate **directional SDG bias** from evidence use, we examine the **Q-only** setting, where an evidence-grounded model should have no systematic reason to prefer any particular answer direction (*i.e.* roughly a 33%-each distribution for a three-option question). In our MCQ tasks, the candidate options are *ordinal*, allowing us to map outputs into three directional categories: **Pessimistic** (lower-development choice), **Conservative** (middle choice), and **Optimistic** (higher-development choice). Accordingly, these directional categories do not measure correctness in Q-only, but rather diagnose the direction in which the model tends to guess when no supporting evidence is available.

Figure 4 shows that these priors are not only model-specific but also *pillar-dependent*, forming distinctive bias signatures across Health & Nutrition, Basic Services & Infrastructure, and Human Capital & Development. Concretely, each model exhibits a characteristic triplet of outcome distributions across the three pillars, revealing where it systematically leans pessimistic, anchors to the middle, or defaults optimistic. We discuss each VLM's preference in detail as follows:

[1] More experimental details are provided in Sec. B.1.

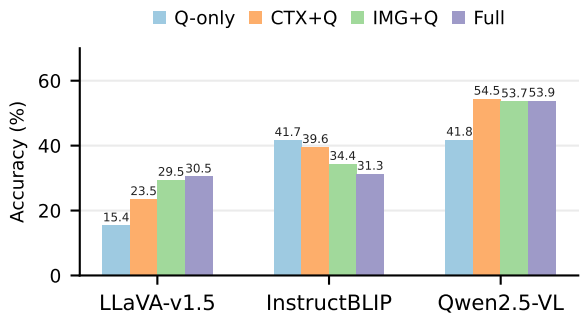

Figure 3: **Per-view MCQ Accuracy.** Accuracy (%) for three VLMs under four evidence views (Q-only, CTX+Q, IMG+Q, Full).

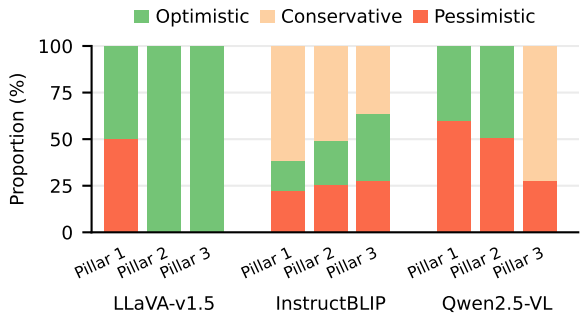

Figure 4: **Distribution of Q-only prediction tendencies across pillars.** Results reveal directional defaults without visual evidence.

LLaVA-v1.5: Polarized Health, Near-Deterministic Optimism Elsewhere. LLaVA-v1.5-7B displays a strongly pillar-conditioned pattern with minimal conservative anchoring. On Pillar 1, it is bimodal: responses split roughly evenly between pessimistic and optimistic options. In contrast, on Pillar 2 and Pillar 3, it becomes overwhelmingly **optimistic**, suggesting that for infrastructure and human-capital questions, the model carries a strong upward default even without supporting evidence.

InstructBLIP: Conservative Anchoring with Moderated Asymmetry. InstructBLIP-7B exhibits a comparatively stable conservative default across pillars, consistently placing the largest mass on the middle option. This conservative anchoring is the strongest on Pillar 1 (majority conservative, with smaller pessimistic and optimistic fractions), remains dominant on Pillar 2, and becomes more mixed on Pillar 3 where the distribution is closer to a three-way split. Overall, InstructBLIP-7B behaves like a "centrist" prior under missing evidence: it tends to hedge toward the midpoint, while still allowing pillar-dependent skews in the tails.

Qwen2.5-VL: Pessimistic on Health and Infrastructure, Conservative on Human Capi-

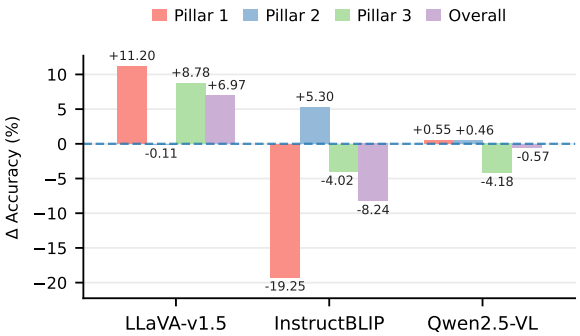

Figure 5: **Performance difference under different input views.** Results of ΔAcc (Full − CTX+Q) show consistent modality imbalance.

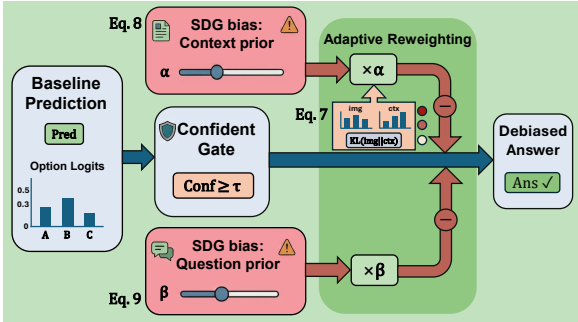

Figure 6: **Overview of *CADE***, which applies a confidence gate and, when triggered, adaptively reweights context and question SDG bias priors, scaling the context prior by the image context divergence, to produce a debiased answer.

tal. QWEN2.5-VL-7B shows a pronounced pessimistic leaning on Pillar 1 and Pillar 2: for both pillars, the pessimistic category constitutes a large portion of outputs (roughly half or more), with the remainder largely optimistic and little reliance on the conservative option. The pattern changes sharply on Pillar 3, where the model shifts to a mostly **conservative** default and optimism nearly disappears. This pillar-sensitive switch indicates that the model activates different directional priors depending on the SDG theme, rather than applying a uniform "cautious" strategy.

3.3 Modality Imbalance in SDG Reasoning

Furthermore, we observe a consistent *modality imbalance* across all three pillars. As shown in Fig. 5, once structured SDG context is provided, models often commit early, and adding aligned imagery yields only marginal gains (sometimes even degrading accuracy). This is notable because SDG-BiasBench depends on *latent* visual proxies: settlement morphology/density, road connectivity, and land-use/environmental cues, yet imagery is frequently under-weighted even when it should refine or correct the inference. We discuss the observed imbalance in each pillar as follows:

Pillar 1: Health & Nutrition. On indicators like stunting prevalence (SDG 2.2.1), predictions largely stabilize under **CTX+Q**; imagery consistent with better living conditions (denser, more connected settlements, stronger access corridors) rarely overturns an activated "high-risk" interpretation.

Pillar 2: Basic Services & Infrastructure. For outcomes like electricity access (SDG 7.1.1), imagery should be highly diagnostic (built-up continuity, road hierarchy), but several VLMs show little change from **CTX+Q** to **Full**, indicating inference driven more by textual priors than spatial inspection.

Pillar 3: Human Capital & Development. For composites like Subnational Human Development Index (SHDI), models often default to the middle bin under context, with weak image-driven shifts despite diffuse proxies (urbanization gradients, corridor structure), suggesting context-conditioned priors dominate multi-modal integration.

4 CADE: Contrastive Adaptive Debiasing Ensemble

To mitigate **SDG Bias**, we introduce **CADE**, a training-free, plug-and-play debiasing method that operates purely on *logits* from a VLM, without modifying the model parameters.[2] We reuse the four views $\{T^{\text{q}}, T^{\text{ctx}}, T^{\text{img}}, T^{\text{full}}\}$ defined in Section 3.1. For each datapoint i, we query the model under these views and obtain per-option logits. We describe CADE for MCQ inference below.

Candidates and logits. Let $\mathcal{O}_i$ denote the finite option set for datapoint i (*e.g.* $\{A, B\}$ or $\{A, B, C\}$). Under each view T, the VLM returns a pre-softmax logit vector over $\mathcal{O}_i$:

$$\mathbf{z}_i^{(T)} = \left(z_{i,a}^{(T)}\right)_{a \in \mathcal{O}_i}.$$

Here we denote the collection of all view-wise logits for datapoint i as $\mathbf{Z}_i = \{\mathbf{z}_i^{(T)}\}_{T \in \{\text{q,ctx,img,full}\}}$.

Step 1: View-wise probabilities computation. We convert logits into probabilities with softmax:

$$p_{i,a}^{(T)} = \frac{\exp\left(z_{i,a}^{(T)}\right)}{\sum_{a' \in \mathcal{O}_i} \exp\left(z_{i,a'}^{(T)}\right)}, \qquad a \in \mathcal{O}_i. \quad (1)$$

[2]We provide further analysis and discussions on the uniqueness, significance, and superiority of CADE in Sec. D.

Step 2: Confidence-gated thresholding. We treat the Full view T^{full} as a conservative baseline:

$$\hat{a}_i^{\text{base}} = \arg\max_{a \in \mathcal{O}_i} p_{i,a}^{(\text{full})}, \quad m_i = \max_{a \in \mathcal{O}_i} p_{i,a}^{(\text{full})}. \quad (2)$$

Given $\tau \in (0, 1)$, we keep the baseline prediction when it is confident (*i.e.* $m_i \geq \tau$). Otherwise (when $m_i < \tau$), we invoke the debiasing ensemble below to produce the final answer.

Step 3: Three streams engagement (image / context / bias). We group the four views into three conceptual streams:

- *Image stream*: combines Full and IMG+Q,
- *Context stream*: CTX+Q,
- *Bias stream*: Q-only (intrinsic prior),

with the following calculations:

$$\tilde{p}_{i,a}^{\text{img}} = p_{i,a}^{(\text{full})} + p_{i,a}^{(\text{img})}, \quad (3)$$

$$\tilde{p}_{i,a}^{\text{ctx}} = p_{i,a}^{(\text{ctx})}, \quad (4)$$

$$\tilde{p}_{i,a}^{\text{q}} = p_{i,a}^{(\text{q})}. \quad (5)$$

Next, we renormalize each stream to obtain the normalized probabilities:

$$p_{i,a}^{s} = \frac{\tilde{p}_{i,a}^{s}}{\sum_{a' \in \mathcal{O}_i} \tilde{p}_{i,a'}^{s}}, \quad s \in \{\text{img}, \text{ctx}, \text{q}\}. \quad (6)$$

Step 4: Contrastive scoring with adaptive disagreement. We quantify image-context disagreement using KL divergence:

$$D_i = \text{KL}\big(p_i^{\text{img}} \,\|\, p_i^{\text{ctx}}\big) = \sum_{a \in \mathcal{O}_i} p_{i,a}^{\text{img}} \log \frac{p_{i,a}^{\text{img}}}{p_{i,a}^{\text{ctx}}}. \quad (7)$$

For each datapoint i, we define an instance-wise coefficient α_i that controls the strength of the context penalty. It is determined by a base value α and a sensitivity factor λ_{KL}, modulating the penalty according to the image-context KL divergence:

$$\alpha_i = \alpha\big(1 + \lambda_{\text{KL}} D_i\big). \quad (8)$$

Finally, we compute a CADE score for each candidate a in $\mathcal{O}_i$ as:

$$score_{i,a} = \log p_{i,a}^{\text{img}} - \alpha_i \log p_{i,a}^{\text{ctx}} - \beta \log p_{i,a}^{\text{q}}, \; a \in \mathcal{O}_i. \quad (9)$$

Here, $\beta \geq 0$ is introduced to control how strongly we penalize question-only priors. Therefore, the debiased prediction $\hat{a}_i^{CADE}$ is given as:

$$\hat{a}_i^{CADE} = \arg\max_{a \in \mathcal{O}_i} score_{i,a}. \quad (10)$$

The algorithm of this plug-and-play **CADE** is outlined in Algorithm 1.

Algorithm 1 CADE

INPUT: Per-view logit vectors over candidates $\mathcal{O}_i$, obtained by querying the model under four input views $\{\mathbf{z}_i^{(\text{full})}, \mathbf{z}_i^{(\text{ctx})}, \mathbf{z}_i^{(\text{img})}, \mathbf{z}_i^{(\text{q})}\}$

1: Convert logits to probabilities $p_{i,\cdot}^{(T)}$ (Eq. 1)
2: Compute baseline confidence m_i (Eq. 2)
3: **if** $m_i \geq \tau$ **then return** $\hat{a}_i^{\text{base}}$
4: Form $p_{i,\cdot}^{\text{img}}, p_{i,\cdot}^{\text{ctx}}, p_{i,\cdot}^{\text{q}}$ (Eqs. 3–6)
5: Compute D_i and α_i (Eqs. 7–8)
6: Compute $score_{i,a}$ and $\hat{a}_i^{\text{CADE}}$ (Eqs. 9–10)

OUTPUT: Final debiased prediction $\hat{a}_i^{\text{CADE}}$

Hyperparameters. We tune τ, λ_{KL}, α and β via random search on a held-out validation set.

Implementation details. More details and the regression variant of CADE are further discussed in Sec. B.

5 Experiments

This section serves two purposes: (i) to benchmark a diverse set of vision–language models (VLMs) on **SDGBiasBench** under controlled information regimes, and (ii) to evaluate **CADE** as a training-free plug-and-play method applied to each VLM.

5.1 Evaluation Protocol

Inputs and views. Each sample contains a satellite image, a street-level image (when available), a structured text context, and a question. We follow the four-view protocol defined in Section 3.1: **Q-only**, **CTX+Q**, **IMG+Q**, and **Full**.[3]

Models. We evaluate representative open-source VLMs across various scales, which include: LLaVA-v1.5 (Liu et al., 2024), InstructBLIP (Dai et al., 2023), LLaVA-OneVision (Li et al., 2024), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025a). Some well-known closed-source VLMs are also included for comprehensive comparison: GPT-4o (OpenAI, 2024), Gemini 2.0 Flash (Google DeepMind, 2024), and Claude 3.5 Sonnet (Anthropic, 2024).

Metrics. For **MCQ**, we report **Accuracy**. For **Regression**, we report **MAE** (Mean Absolute Error) and **Interval Accuracy**. Interval Accuracy counts a correct prediction if it falls within a pre-defined tolerance interval for each question type. We provide detailed interval definitions, per-interval results, and further in-depth analysis in Sec. B.3.

[3]Unless otherwise stated, the "Original" setting corresponds to standard inference with the **Full** input view.

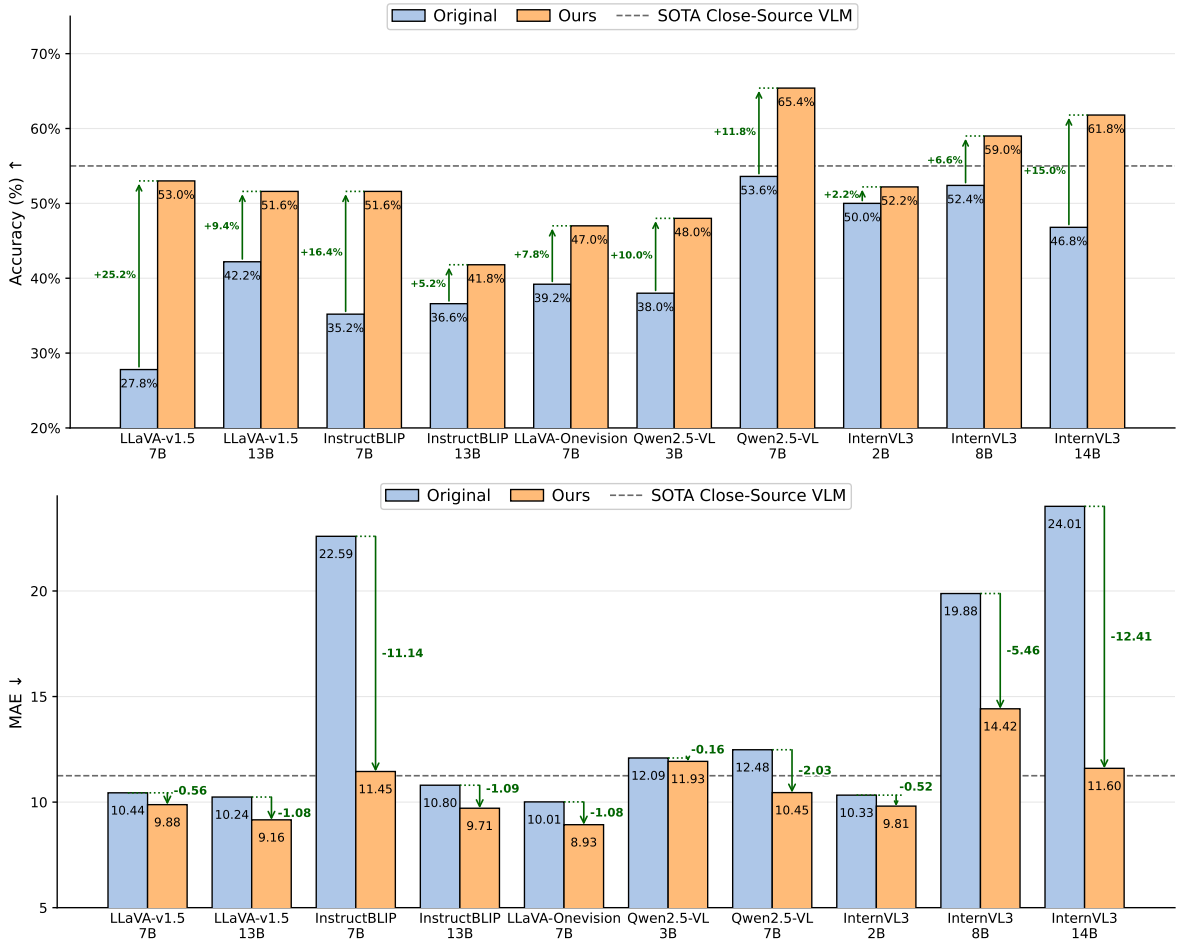

Figure 7: **Performance of Vision–Language Models on: (Upper) multiple-choice questions; (Bottom) regression questions.** "Ours" refers to applying CADE to the same base model, where reductions are highlighted with green.

Table 1: **Comparison with SOTA debiasing methods.**

Model	Original	Post-Hoc	VDD	**Ours**
LLaVA-v1.5-7B	0.28	0.34	0.41	**0.53**
Qwen2.5-VL-7B	0.54	0.57	0.58	**0.65**

5.2 Main Benchmark Results

Fig. 7 shows MCQ accuracy and regression MAE results across VLMs. Across both tasks, the original performance is uniformly weak: MCQ accuracy remains low for most models, with several barely exceeding random-guess levels, while regression errors are consistently large, suggesting that current VLMs still struggle to make precise and dependable predictions on sustainability-oriented tasks.

5.3 CADE Results

We apply CADE to open-source VLMs of different scales and report the same metrics as in the main results in Fig. 7. Results[4] show that CADE consistently benefits all evaluated models, which raises MCQ accuracy across the board, and simultaneously lowers regression MAE for every model. Overall, the gains are substantial and stable across architectures and parameter scales, demonstrating that CADE is an effective and broadly applicable debiasing strategy.

[4]Detailed results and further discussion are in Sec. C.

Table 2: **Progressive ablation of CADE** on LLaVA-v1.5-7B (MCQ, T1). We progressively enable each CADE component and report accuracy.

Variant	**Hyperparameters**				**Acc.**
	α	λ_{KL}	β	τ	
Baseline	–	–	–	–	0.30
+ context penalty	✓	–	–	–	0.45
+ adaptive disagreement	✓	✓	–	–	0.46
+ prior penalty	✓	✓	✓	–	0.52
+ confidence gate (Ours)	✓	✓	✓	✓	**0.53**

5.4 Comparison with SOTA Methods

We additionally provide comparisons against SOTA general debiasing methods, including Post-Hoc and VDD (Zhang et al., 2025) in Table 1. Across models and evaluation settings, CADE consistently delivers the largest and most reliable improvements, showing that it is not only competitive with general-

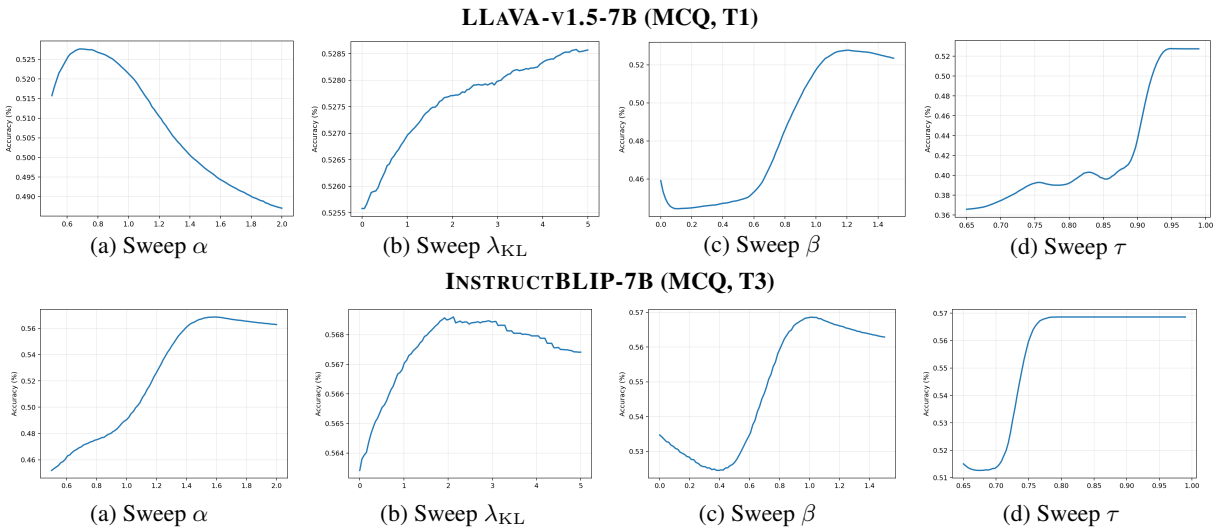

Figure 8: **Hyperparameter sensitivity study on LLaVA-v1.5 and InstructBLIP.** Each subplot varies one specific hyperparameter while fixing the other three. MCQ accuracy is reported as VLM's performance.

Table 3: **Debiasing component selection of CADE.** We compare different image-stream constructions under JS and KL divergence. "Full+Img." denotes the original CADE image stream obtained by normalizing $p^{(\text{full})} + p^{(\text{img})}$.

Model	Original	JS			KL		
		Img.	Full	Full+Img.	Img.	Full	Full+Img.
LLaVA-v1.5-7B	0.28	0.49	0.48	0.50	0.51	0.50	**0.53**
Qwen2.5-VL-7B	0.54	0.59	0.61	0.63	0.58	0.60	**0.65**

purpose debiasing baselines but also particularly effective at mitigating SDG-specific biases.

5.5 Ablation Studies

We conduct the following ablation studies[5] regarding the proposed CADE. Further in-depth analysis and discussions on performance gain and the method's efficiency are provided in Sec. D.

Component analysis. We provide a progressive ablation of CADE on LLaVA-v1.5-7B (MCQ, T1) in Table 2. Results highlight the contribution of each step with consistent gains when components are gradually enabled: adding the context penalty yields the largest performance increase (0.30→0.45), adaptive disagreement provides a small additional boost of +0.01, prior penalty brings a further improvement of +0.06, and the confidence gate gives a final increment of +0.01.

Sensitivity study. We conduct a sensitivity study of CADE as shown in Fig. 8. Results show CADE is not brittle: accuracy peaks at moderate α ($\sim$0.7 for LLaVA, higher for InstructBLIP), rises with λ_{KL} with diminishing returns, and favors larger β (best $\sim$1.0-1.2). τ instead has a sharp threshold ($\sim$0.9 *vs.* $\sim$0.75) then plateaus.

[5]All ablations below are conducted with MCQ on one context test. Further details on these are given in Sec. B.1.

Debiasing component selection. We ablate two CADE design choices: the image stream p_i^{img} and the divergence in Eq. 7. For the image stream, we compare Image, Full, and our default Full+Image reference ($\tilde{p}_{i,a}^{\text{img}} = p_{i,a}^{(\text{full})} + p_{i,a}^{(\text{img})}$), followed by normalization. We also replace KL with Jensen–Shannon (JS) divergence. As shown in Table 3, Full+Image performs best under both JS and KL, suggesting that it gives a more reliable image-present reference than either single-view variant. Since KL is slightly better and more consistent overall, we adopt Full+Image with KL as the default CADE design.

6 Related Work

Geographic benchmarks for SDG monitoring. Earth observation has long supported large-scale estimation of development outcomes from satellite imagery (Henderson et al., 2012; Jean et al., 2016; Head et al., 2017; Yeh et al., 2020; Burke et al., 2021). SustainBench systematizes SDG-relevant tasks and evaluation protocols (Yeh et al., 2021), and subsequent benchmark suites broaden coverage across sensors, tasks, and geographic regions (Marsocci et al., 2024; Bountos et al., 2023). Remote-sensing vision–language benchmarks further study geospatial perception and grounded

dialogue (Kuckreja et al., 2024; Danish et al., 2025; Soni et al., 2025). However, prior geographic benchmarks rarely support both realistic SDG workflows and bias diagnosis on modern VLMs; *SDGBiasBench* addresses these gaps by controlling evidence availability (Q/CTX/IMG/Full) for SDG-indicator reasoning on both qualitative judgments and quantitative estimation.

Social and geographic bias in VLMs. Geographic stereotypes are documented even in text-only LLMs (Manvi et al., 2024), and multi-modal models inherit related skews from data filtering and representation gaps (Pouget et al., 2024). For image geolocation, Huang et al. (2025b) show systematic performance disparities favoring wealthy regions. Complementary analyses measure demographic and social biases in vision–language representations and systems (Ross et al., 2021; Ruggeri et al., 2023; Huang et al., 2025a), building on bias diagnostics developed in NLP (Blodgett et al., 2020; Nangia et al., 2020; Nadeem et al., 2021). Our work operationalizes these concerns in a sustainability setting as SDG-specific, pillar-conditioned *directional defaults*.

Training-free debiasing for VLMs. VLM decisions can be dominated by language priors when visual and textual cues conflict (Lin et al., 2023). A line of training-free methods mitigates such failures by contrasting decoding distributions, including VCD (Leng et al., 2024), instruction-contrastive decoding (Wang et al., 2024; Chuang et al., 2024; Li et al., 2023b; Zhu et al., 2025b), and explicit penalization of language priors (Zhang et al., 2025). CADE follows this training-free framework but targets SDG bias: it detects prior reliance via multi-view disagreement and adaptively reweights logits to reduce systematic directional errors.

7 Conclusion

In this work, we introduce a large-scale **SDGBiasBench** benchmark covering both qualitative MCQ judgments and quantitative regression estimation for SDG-indicator reasoning. Extensive evaluations reveal an intrinsic **SDG bias** across VLMs, characterized by pillar-conditioned directional defaults and persistent prior-driven modality imbalance. To mitigate this, we proposed **CADE**, a plug-and-play debiasing method that reweights outputs from multiple input views. CADE achieves significant gains on the proposed benchmark, supporting more reliable sustainability-oriented predictions.

Limitations

This work focuses on three major pillars and a representative set of indicators, where some lack-of-definition and outdated SDG targets and domains are not covered. Moreover, a small fraction of examples may carry imperfect annotations due to upstream label noise inherited from SUSTAINBENCH; we treat these as part of the realistic evaluation setting rather than artifacts introduced during the benchmark construction. Finally, the proposed CADE is in principle applicable to closed-source VLMs, but it relies on per-candidate logits (or equivalent scores), which are typically not available by current usage of APIs from closed-source VLMs, preventing a direct end-to-end evaluation.

Ethical Considerations

SDG indicators are high-stakes: biased predictions (*e.g.* systematic over/under-estimation across regions) can lead to harmful downstream decisions if used for policy or aid allocation. SDGBiasBench is intended for *evaluation and bias diagnosis* rather than direct deployment, and our design explicitly probes prior reliance and directional error under changing evidence. Following SUSTAINBENCH (Yeh et al., 2021), privacy risks are mitigated by the DHS cluster-level release protocol and randomized geolocation jittering; imagery sources are publicly available and largely low-resolution, while street-level photos (when present) are privacy-protected (*e.g.* automated blurring of faces/plates). We therefore expect limited risk of identifying individuals from the released benchmark, but emphasize that users should follow the original data providers' terms and apply appropriate safeguards when building models.

Acknowledgments

This work was conducted in collaboration with ST Engineering Geo-Insights Pte Ltd. The authors sincerely thank the reviewers, area chairs, and program chairs for their valuable feedback and constructive comments.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2024. Claude 3.5 sonnet model card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Md Atabuzzaman, Ali Asgarov, and Chris Thomas. 2025. Benchmarking and mitigating mcqa selection bias of large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33536–33550.

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. 2023. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. 2021. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Daniel J Corsi, Melissa Neuman, Jocelyn E Finlay, and SV Subramanian. 2012. Demographic and health surveys: a profile. *International journal of epidemiology*, 41(6):1602–1613.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2025. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7132–7142.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Google DeepMind. 2024. Gemini 2.0.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can human development be measured with satellite imagery? In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, pages 1–11.

J Vernon Henderson, Adam Storeygard, and David N Weil. 2012. Measuring economic growth from outer space. *American economic review*, 102(2):994–1028.

Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. 2025a. Visbias: Measuring explicit and implicit social biases in vision language models. *arXiv preprint arXiv:2503.07575*.

Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. 2025b. Ai sees your location—but with a bias toward the wealthy world. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18030–18050.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

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

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*.

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

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, and 1 others. 2024. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-05-13.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496.

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

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008.

Gabriele Ruggeri, Debora Nozza, and 1 others. 2023. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and 1 others. 2025. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14303–14313.

United Nations General Assembly. 2015. Transforming our world: the 2030 agenda for sustainable development. Resolution A/RES/70/1. Adopted on 25 September 2015.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024a. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. *arXiv preprint arXiv:2407.02814*.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024b. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15669–15680.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. 2021. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583.

Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. 2025. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4232–4241.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. Vl-stereoset: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025b. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

SDGBiasBench: Benchmarking and Mitigating Vision–Language Models' Biases in Sustainable Development Goals

Supplementary Material

Contents

1 Introduction **1**

2 SDGBiasBench: An Overview **2**

3 Empirical Analysis of SDG Bias **3**
- 3.1 Setup: Views, Models, and Metrics . . . 4
- 3.2 Pillar-conditioned Directional Defaults Form an SDG Bias Signature . . . 4
- 3.3 Modality Imbalance in SDG Reasoning . . . 5

4 CADE: Contrastive Adaptive Debiasing Ensemble **5**

5 Experiments **6**
- 5.1 Evaluation Protocol . . . 6
- 5.2 Main Benchmark Results . . . 7
- 5.3 CADE Results . . . 7
- 5.4 Comparison with SOTA Methods . . . 7
- 5.5 Ablation Studies . . . 8

6 Related Work **8**

7 Conclusion **9**

A SDGBiasBench: A Comprehensive Description
- A.1 Overview . . .
- A.2 Dataset Statistics and Annotation Protocol . . .
- A.3 Pillar 1: Health & Nutrition . . .
- A.4 Pillar 2: Basic Services & Infrastructure . . .
- A.5 Pillar 3: Human Capital & Development . . .
- A.6 Prompt Templates . . .

B Additional Implementation Details
- B.1 Details of Five Context Tests (T1–T5) for MCQ . . .
- B.2 Details of Regression Variant . . .
- B.3 Details of Interval Accuracy Tolerances . . .
- B.4 Details of Hyperparameter Search . . .

C Additional Experimental Results

D Additional Analysis and Discussions
- D.1 Analysis on SDG Bias and CADE's Relation to Contrastive Debiasing . . .
- D.2 Analysis on Performance Gains . . .
- D.3 Analysis on MCQA Selection Bias . . .
- D.4 Analysis on Efficiency–Effectiveness Tradeoff . . .

E Licenses and Terms of Use

F Intended Use and Access Conditions

G Information on Use of AI Assistants

A SDGBiasBench: A Comprehensive Description

A.1 Overview

SDGBiasBench evaluates vision–language reasoning on real-world SDG-related indicators while revealing SDG biases. It comprises two task types: multiple-choice questions (MCQs) and regression tasks to mirror practical use cases. Each datapoint is paired with satellite imagery and often augmented with street-level photos; each sample also includes structured context variables (*e.g.* sanitation index, women's education level, asset index), encouraging multi-step reasoning across modalities. The benchmark draws from raw data in SUSTAINBENCH (Yeh et al., 2021) and the Demographic and Health Surveys (DHS) (Corsi et al., 2012). We extend them through human-authored high-level reasoning questions and expert-labeled ground truth answers. Below we summarize the three pillars in detail.

Key features.

- **Diverse questions.** SDGBiasBench contains 500,000 multiple-choice questions and 50,000 regression queries targeting SDG-related indicators (*e.g.* poverty, infrastructure access, educational attainment). Together, MCQ and regression tasks reflect real SDG practice, requiring both discrete judgments and accurate prediction of continuous development indicators.
- **Multi-modal context.** Each task is paired with satellite imagery, often complemented by street-level photos, and structured socio-economic context variables (*e.g.* sanitation index, women's education, asset index), requiring joint reasoning over visual and textual modalities.
- **Sustainability pillars.** The benchmark covers three sustainability pillars—Health & Nutrition, Basic Services & Infrastructure, and Human Capital & Development—to support broad assessment across SDG themes.

Design goal. Overall, SDGBiasBench is structured to move beyond surface perception: each task encourages models to synthesize signals across modalities and to revise judgments under changing evidence, making it well-suited for diagnosing SDG-specific prior reliance and directional errors.

Table A: **Number of questions in SDGBiasBench by pillar and task type.**

Dataset Pillar	MCQ	Regression
Health & Nutrition	222,900	19,427
Basic Services & Infrastructure	162,600	16,761
Human Capital & Development	114,500	13,812
Total	**500,000**	**50,000**

Table B: **MCQ label distributions (%) by pillar.**

Dataset Pillar	A	B	C
Health & Nutrition	27.75	42.80	29.45
Basic Services & Infrastructure	40.11	39.11	20.78
Human Capital & Development	32.03	33.87	34.11

A.2 Dataset Statistics and Annotation Protocol

We summarize the benchmark composition and per-pillar MCQ label distributions in Tables A and B, respectively. SDGBiasBench contains both multiple-choice questions (MCQs) and regression tasks across three sustainability pillars.

Construction and annotation pipeline. SDGBiasBench is constructed through a structured expert-involved pipeline to ensure that each example is geographically grounded, semantically meaningful, and quality-controlled. We provide more details on the pipeline below:

- **Source grounding and alignment.** Each example is anchored to a geo-entity. We align satellite imagery, street-level imagery, and structured context variables from SUSTAINBENCH and Demographic and Health Surveys (DHS) using location identifiers, so that all available modalities correspond to the same region.
- **Expert-defined indicator templates.** We invite 9 sustainability-domain experts to define the question templates for each indicator, together with the required evidence sources, option-bin rules, and intended reasoning targets. For each task, the annotation protocol specifies which modalities are needed, what latent construct should be inferred, and how labels or regression targets should be interpreted.
- **Option and bin construction.** For MCQs, we discretize each indicator into ordinal answer bins using either established SDG-related thresholds or indicator-specific distributional cut points. The resulting options are mutually

exclusive and semantically ordered sub-ranges of the underlying indicator, enabling qualitative judgment over value ranges rather than exact numerical prediction. For regression tasks, the targets are continuous ground-truth values obtained directly from SUSTAINBENCH and Demographic and Health Surveys (DHS).

- **Quality control.** We apply both automatic and human quality-control procedures. Automatic checks remove examples with missing modalities, inconsistent metadata, duplicated or near-duplicate items, and invalid option sets. Human quality assurance includes template review and stratified spot-checking across pillars and task types to verify semantic correctness, answer validity, and consistency of evidence alignment. In particular, 3 experts are responsible for this quality-control stage.

- **Human review protocol.** A stratified subset of approximately 5% of the benchmark is human-reviewed across pillars and task types. Each reviewed item is independently checked by two experts, and disagreements are resolved by the third expert. This process validates the expert-defined templates, option construction, answer validity, and modality alignment, while the full benchmark is generated from the validated rules and aligned ground-truth indicators.

A.3 Pillar 1: Health & Nutrition

This pillar evaluates whether VLMs can reason about population health, nutritional status, and healthcare access from indirect, multi-modal evidence. Tasks span child malnutrition, disease burden, physical activity, and health system quality, requiring models to integrate visual cues (*e.g.* settlement density, environmental conditions, infrastructure) with structured socio-economic context. Crucially, both qualitative judgments and quantitative estimation settings emphasize *latent health inference*: target indicators cannot be directly observed from a single modality, but must be inferred through correlated proxies. This pillar therefore probes a model's ability to move beyond surface correlations and performs multi-step reasoning about human well-beings.

MCQ tasks. We define five MCQs:

- Q1: Stunting in children under 5 (SDG 2.2.1)
- Q2: Universal Health Coverage index (SDG 3.8.1)
- Q3: Insufficient physical activity prevalence (WHO)
- Q4: Wasting in children under 5 (SDG 2.2.2)
- Q5: Healthcare Access and Quality (HAQ) Index (IHME)

All MCQs ask models to infer the regional indicator by jointly using imagery and structured context.

Regression tasks. We include four regression questions:

- Q1: Survival Infrastructure (*infer* Under-5 mortality rate)
- Q2: Maternal Capital (*infer* Under-5 mortality rate)
- Q3: Nutritional Wealth (*infer* Women's BMI)
- Q4: Total Development (*infer* Women's BMI)

These regressions require continuous inference from indirect evidence rather than reading the target directly from any single modality.

A.4 Pillar 2: Basic Services & Infrastructure

This pillar focuses on access to essential services such as electricity, water, and sanitation, as well as their downstream health implications. Tasks assess whether models can connect visible infrastructure patterns such as roads, buildings, utilities, and environmental features with structured indices which capture service coverage and deprivation. Beyond direct access estimation, several tasks require *inverse reconstruction*: inferring latent socio-economic status or infrastructure quality from human activity patterns and environmental signals. As such, this pillar stresses spatial reasoning, cross-modal alignment, and the ability to recover unobserved service conditions from partial evidence.

MCQ tasks. Four MCQs cover:

- Q1: Electricity access (SDG 7.1.1)
- Q2: Access to safe drinking water (SDG 6.1.1)
- Q3: Access to sanitation services (SDG 6.2.1)
- Q4: WASH-attributable mortality rate (SDG 3.9.2)

Questions test whether models can connect visible infrastructure cues with structured indices to estimate access or burden.

Table C: **MCQ prompt template.** The **Context** line is included only when structured SDG variables are provided.

Prompt (MCQ)

Question: {QUESTION}
Context: {CONTEXT} *(optional)*
Options: {OPTIONS}
`VALID OPTIONS: {VALID_LETTERS}`
`STRICT OUTPUT: Return EXACTLY ONE uppercase letter from VALID OPTIONS.`
`Output nothing else (no words, numbers, symbols, spaces, or newlines).`

Table D: **Regression prompt template.** The **Context** line is included only when structured SDG variables are provided.

Prompt (Regression)
Question: {QUESTION} **Context:** {CONTEXT} *(optional)* `Answer strictly with a single numerical value. Round to 1 decimal place.`

Regression tasks. Four regression questions target service/wealth reconstruction:

- Q1: Hidden Poverty (*infer* Asset index)
- Q2: Plumbing from People (*infer* Sanitation index)
- Q3: Water Security (*infer* Water index)
- Q4: Sanitation Reconstruction (*infer* Sanitation index)

These tasks emphasize inverse reasoning—recovering infrastructure or SES from human and environmental signals.

A.5 Pillar 3: Human Capital & Development

This pillar examines reasoning about education and broader human development, where signals are diffuse and highly correlated with other aspects of regional context. Models must infer educational attainment and composite development indices by synthesizing multi-modal proxies such as settlement structure, asset distribution, and service availability. The emphasis is on *abstract human-capital inference* such as human development index and education–health alignment. Compared to the other pillars, this setting places greater weight on long-range correlations and holistic interpretation of socio-economic environments.

Table E: **Operational definition of T1–T5 context conditions** used in Table H.

Test	Context form	Example
T1	Value-given (full)	*Given:* asset_index=1.0, water_index=3.4, sanitation_index=2.6, under_5_mortality_rate=0, woman_education_years=9.5, woman_bmi=24.4, urban=Urban
T2	Value-given (reduced)	*Given:* asset_index=1.0, sanitation_index=2.6, urban=Urban
T3	Name-only hint (full)	*You may consider:* asset_index, water_index, sanitation_index, under_5_mortality_rate, woman_education_years, woman_bmi, urban
T4	Name-only hint (reduced)	*You may consider:* asset_index, sanitation_index, urban
T5	No context	Question only (no structured variables provided or hinted).

MCQ tasks. Three MCQs include:

- Q1: Female lower-secondary school completion (SDG 4.1.2)
- Q2: Subnational Human Development Index (SHDI) (UNDP)
- Q3: Education–Health alignment score (custom indicator)

Models must map multi-modal proxies (settlement structure, assets, services) to human-capital indicators.

Regression task. One regression question:

- Q1: Education Access (*infer* Women's education level)

This task reconstructs schooling from correlated wealth/health/service cues, making it a non-trivial, indirect inference.

A.6 Prompt Templates

For all models and evidence views, we construct the textual prompt by concatenating the *question* with (optionally) the structured SDG *context*, followed by a task-specific output constraint. For multiple-choice questions (MCQ), we additionally provide the option set and enforce a *single-letter* response (Table C). For regression queries, we enforce a *single numeric* response rounded to one decimal place

Table F: **Tolerances used to compute regression interval accuracy.**

Indicator	Tolerance δ_k
under 5 mortality rate	$\pm$ 5.0
women bmi	$\pm$ 1.0
asset index	$\pm$ 0.75
sanitation index	$\pm$ 0.75
water index	$\pm$ 0.75
women edu	$\pm$ 1.5

Table G: **Hyperparameter-search ranges for CADE.**

Hyperparameter	MCQ range	Regression range
α	[0.5, 2.0]	[0.0, 7.0]
λ_{KL}	[0.0, 5.0]	[0.0, 7.0]
β	[0.0, 1.5]	[0.0, 7.0]
τ	[0.65, 0.99]	[0.25, 0.99]

(Table D). When structured context is unavailable (*e.g.* question-only or image-only views), the **Context** line is omitted.

B Additional Implementation Details

We provide further implementation details to support the reproducibility of our proposed benchmark with experimental setups below.

B.1 Details of Five Context Tests (T1–T5) for MCQ

We evaluate each multiple-choice question by transforming a single data instance into five context-test variants (T1–T5), which evaluate model's behavior under progressively weaker structured-context conditions. They are designed along two complementary axes: **context explicitness** and **context quantity**. In terms of explicitness, the tests span three levels: **value-given**, **name-only hint**, and **no context**. In terms of quantity, whenever structured context is present, we further distinguish between a **full** set of variables and a **reduced** subset. Accordingly, **T1** and **T2** provide exact structured variable values, with T1 using the full set and T2 using a reduced subset. **T3** and **T4** provide only variable names as soft guidance, again with T3 using the full set and T4 using a reduced subset. Finally, **T5** removes structured context entirely, leaving the model with the question alone. This design enables us to disentangle the effect of **how much** structured context is available (full *vs.* reduced) from **how explicitly** it is presented (value-given *vs.* name-only *vs.* none). Table E summarizes the operational definitions and representative examples of all five tests.

Remark. All reported performance in empirical analysis in Sec. 3 and ablation studies in Sec. 5.5 are obtained only under context test T1 unless otherwise stated. Instead, average accuracy of the five tests for MCQ tasks are reported in Fig. 7.

B.2 Details of Regression Variant

For regression queries, we apply CADE to the *first generated token* only. This design matches the next-token decoding mechanism of current VLMs: once the initial digit is determined, debiasing later tokens is not meaningful because subsequent generation is already conditioned on the first-token choice. Accordingly, we unify the candidate definition across MCQ and regression by specifying $\mathcal{O}_i$ as the option set: for MCQ, $\mathcal{O}_i$ is the set of answer choices (*e.g.* $\{A, B, C\}$); for regression, we discretize the first-token space as $\mathcal{O}_i = \{0, 1, \ldots, 9\}$, where $z_{i,c}^{(T)}$ denotes the logit that the first generated token is digit o under view T. All subsequent steps (view-wise probabilities calculation, confidence-gated threshold, three streams engagement, contrastive scoring with adaptive disagreement) follow the same formulation as in the MCQ case.

B.3 Details of Interval Accuracy Tolerances

We compute interval accuracy using pre-defined indicator-specific tolerances δ_k,and a prediction $\hat{y}$ is counted correct if $|\hat{y} - y| \leq \delta_k$. The used tolerances for each indicator are given in Table F.

B.4 Details of Hyperparameter Search

Hyperparameter ranges. CADE introduces a confidence gate τ, an adaptive exponent base α, a disagreement multiplier λ_{KL}, and a prior-penalty coefficient β (Eq. (8), Eq. (9)). The random-search ranges are listed in Table G.

Implementation details. We tune CADE via random search on a held-out validation split: we sample 10,000 hyperparameter configurations (uniformly within preset ranges) and evaluate each configuration once on 20% of the data. We then re-evaluate the top 100 performing configuration on the full validation set to finalize the hyperparameters used in all reported results. Since all hyperparameters are sampled jointly in each trial, any given pair of hyperparameters is effectively explored through the same 10,000 joint samples (*i.e.* there is no separate pairwise grid sweep).

Table H: **Performance of vision–language models on multiple-choice questions.** "Ours" refers to applying CADE to the same base model, where per-test gains are highlighted with green.

Model	Params	Setting	T1	T2	T3	T4	T5	Avg. Gain	Rel. Gain
Open-source VLMs									
LLaVA-v1.5	7B	Original	0.30	0.32	0.22	0.26	0.29	–	–
		Ours	0.53 $^{+0.23}$	0.54 $^{+0.22}$	0.55 $^{+0.33}$	0.54 $^{+0.28}$	0.49 $^{+0.20}$	**+0.25**	**+90.6%**
	13B	Original	0.42	0.40	0.42	0.44	0.43	–	–
		Ours	0.57 $^{+0.15}$	0.48 $^{+0.08}$	0.52 $^{+0.10}$	0.51 $^{+0.07}$	0.50 $^{+0.07}$	**+0.09**	**+22.3%**
InstructBLIP	7B	Original	0.31	0.42	0.32	0.39	0.32	–	–
		Ours	0.47 $^{+0.16}$	0.63 $^{+0.21}$	0.57 $^{+0.25}$	0.45 $^{+0.06}$	0.46 $^{+0.14}$	**+0.16**	**+46.6%**
	13B	Original	0.34	0.48	0.33	0.34	0.34	–	–
		Ours	0.39 $^{+0.05}$	0.49 $^{+0.01}$	0.38 $^{+0.05}$	0.45 $^{+0.11}$	0.38 $^{+0.04}$	**+0.05**	**+14.2%**
LLaVA-OneVision	7B	Original	0.49	0.49	0.21	0.43	0.34	–	–
		Ours	0.53 $^{+0.04}$	0.52 $^{+0.03}$	0.42 $^{+0.21}$	0.48 $^{+0.05}$	0.40 $^{+0.06}$	**+0.08**	**+19.9%**
Qwen2.5-VL	3B	Original	0.44	0.38	0.35	0.42	0.31	–	–
		Ours	0.51 $^{+0.07}$	0.49 $^{+0.11}$	0.45 $^{+0.10}$	0.49 $^{+0.07}$	0.46 $^{+0.15}$	**+0.10**	**+26.3%**
	7B	Original	0.54	0.54	0.54	0.53	0.53	–	–
		Ours	0.65 $^{+0.11}$	0.67 $^{+0.13}$	0.65 $^{+0.11}$	0.66 $^{+0.13}$	0.64 $^{+0.11}$	**+0.12**	**+22.0%**
InternVL3	2B	Original	0.60	0.59	0.42	0.41	0.48	–	–
		Ours	0.62 $^{+0.02}$	0.60 $^{+0.01}$	0.44 $^{+0.02}$	0.42 $^{+0.01}$	0.53 $^{+0.05}$	**+0.02**	**+4.4%**
	8B	Original	0.59	0.55	0.50	0.51	0.47	–	–
		Ours	0.61 $^{+0.02}$	0.61 $^{+0.06}$	0.58 $^{+0.08}$	0.60 $^{+0.09}$	0.55 $^{+0.08}$	**+0.07**	**+12.6%**
	14B	Original	0.61	0.51	0.42	0.40	0.40	–	–
		Ours	0.69 $^{+0.08}$	0.61 $^{+0.10}$	0.62 $^{+0.20}$	0.59 $^{+0.19}$	0.58 $^{+0.18}$	**+0.15**	**+32.1%**
Close-source VLMs									
GPT-4o	–	API	0.61	0.62	0.56	0.49	0.47	–	–
Gemini 2.0 Flash	–	API	0.62	0.61	0.52	0.45	0.44	–	–
Claude 3.5 Sonnet	–	API	0.56	0.53	0.48	0.43	0.42	–	–

C Additional Experimental Results

We provide comprehensive numerical results as the extension of Fig. 7. Specifically, we report Accuracy across all five context tests (T1–T5) for MCQ in Table H, and report MAE with Interval Accuracy for Regression in Table I.

Main benchmark results: MCQ. The original MCQ results reveal two clear patterns. First, most open-source VLMs show limited robustness when structured context becomes weaker, with performance generally degrading from stronger-context settings to weaker-context ones. Second, there remains a noticeable gap between weaker open-source models and stronger open- or closed-source systems, suggesting that SDG-oriented multiple-choice reasoning is still challenging even for relatively capable VLMs. Overall, the original results indicate substantial sensitivity to context availability and incomplete generalization under reduced guidance.

CADE results: MCQ. CADE consistently improves MCQ performance across architectures and parameter scales, and the improvements are especially evident in the harder weak-context tests. This suggests that CADE does not merely raise performance in already favorable settings, but more importantly improves robustness when explicit structured evidence is partially removed or entirely absent. Another notable trend is that the gains are broader for weaker and mid-tier models, while still remaining stable for stronger ones, showing that CADE is effective both as a recovery mechanism for more biased models and as a refinement mechanism for already stronger baselines.

Main benchmark results: Regression. The original regression results show that continuous-value

Table I: **Performance of vision–language models on regression tasks.** "Ours" refers to applying CADE to the same base model, where per-test gains are highlighted with green.

Model	Params	Setting	MAE ↓	Rel. Reduction ↓	Interval Accuracy ↑	Rel. Gain ↑
Open-source VLMs						
LLaVA-v1.5	7B	Original	10.44	–	0.33	–
		Ours	9.88 $^{-0.56}$	**-5.36%**	0.35 $^{+0.02}$	**+6.06%**
	13B	Original	10.24	–	0.37	–
		Ours	9.16 $^{-1.08}$	**-10.55%**	0.43 $^{+0.06}$	**+16.22%**
InstructBLIP	7B	Original	22.59	–	0.11	–
		Ours	11.45 $^{-11.14}$	**-49.31%**	0.20 $^{+0.09}$	**+81.82%**
	13B	Original	10.80	–	0.33	–
		Ours	9.71 $^{-1.09}$	**10.00%**	0.37 $^{+0.04}$	**+12.12%**
LLaVA-OneVision	7B	Original	10.01	–	0.20	–
		Ours	8.93 $^{-1.08}$	**-10.79%**	0.25 $^{+0.05}$	**+25.00%**
Qwen2.5-VL	3B	Original	12.09	–	0.27	–
		Ours	11.93 $^{-0.16}$	**-1.32%**	0.28 $^{+0.01}$	**+3.70%**
	7B	Original	12.48	–	0.22	–
		Ours	10.45 $^{-2.03}$	**-16.27%**	0.31 $^{+0.09}$	**+40.91%**
InternVL3	2B	Original	10.33	–	0.29	–
		Ours	9.81 $^{-0.52}$	**-5.03%**	0.37 $^{+0.08}$	**+27.59%**
	8B	Original	19.88	–	0.14	–
		Ours	14.42 $^{-5.46}$	**-27.46%**	0.20 $^{+0.06}$	**+42.86%**
	14B	Original	24.01	–	0.19	–
		Ours	11.60 $^{-12.41}$	**-51.69%**	0.25 $^{+0.06}$	**+31.58%**
Close-source VLMs						
GPT-4o	–	API	11.25	–	0.25	–
Gemini 2.0 Flash	–	API	12.35	–	0.12	–
Claude 3.5 Sonnet	–	API	11.46	–	0.15	–

prediction is also difficult for current VLMs. In general, the models exhibit relatively large estimation errors together with low interval-level accuracy, indicating that they struggle not only with precise value prediction but also with coarse range alignment. Similar to MCQ, the original regression setting reflects clear room for improvement rather than near-saturated performance.

CADE results: Regression. CADE brings consistent improvements in the regression setting by reducing prediction error and simultaneously improving interval-level correctness. This joint improvement is important because it indicates that CADE is not simply shifting outputs in a way that benefits one metric while hurting another; instead, it improves overall prediction quality. The gains are particularly pronounced for models with weaker original regression behavior, while stronger models still benefit in a stable manner. This mirrors the MCQ findings and supports the view that CADE improves model calibration and robustness across task formats.

Overall discussion. Taken together, the detailed results show that the original models are consistently vulnerable to SDG-oriented bias and context sensitivity in both discrete and continuous prediction settings. CADE addresses these issues in a broad and stable way: it improves both qualitative and quantitative outcomes without requiring model retraining significantly, and particularly strengthens robustness under weak-context conditions.

D Additional Analysis and Discussions

D.1 Analysis on SDG Bias and CADE's Relation to Contrastive Debiasing

Debiasing SDG-related predictions is a specific instance of debiasing in general, but it differs in both the source and manifestation of bias. In many general debiasing settings, bias is often associated

with socially salient correlations or shortcut features learned from data. In contrast, SDG prediction requires models to infer latent development conditions from noisy, region-dependent multi-modal proxies, such as satellite imagery, street-view imagery, and socio-economic context. When evidence is weak or ambiguous, models therefore tend to rely on generic development priors rather than grounded evidence.

This difference also leads to distinctive bias patterns. As revealed by SDGBiasBench, SDG bias appears as **pillar-conditioned directional defaults** under Q-only inputs and **modality imbalance** when structured context dominates visual evidence. Accordingly, debiasing SDG biases requires not only suppressing prior-driven predictions, but also correcting how textual and visual evidence are balanced. CADE is designed for these task-specific challenges: it penalizes unsupported Q-only defaults and adaptively reweights context-dominated logits when image–context disagreement is large. Thus, while CADE follows the general principle of contrastive debiasing, its formulation is tailored to the distinctive structure of SDG monitoring tasks.

D.2 Analysis on Performance Gains

We observe cases where the improvement brought by CADE is only incremental. This behavior is expected for contrast-style debiasing methods. Specifically, for each instance i with finite candidate set O_i (*e.g.* $\{A, B, C\}$), each view $T \in \{T_q, T_{ctx}, T_{img}, T_{full}\}$ produces per-option logits $z_i^{(T)}$, which are converted by softmax into a candidate distribution $p_{i,\cdot}^{(T)}$ over O_i. Since contrastive methods operate by comparing two such distributions, their corrective effect depends on the divergence between the contrasted views. When the full-view prediction $p_{i,\cdot}^{(T_{full})}$ is already very close to the text-prior distribution $p_{i,\cdot}^{(T_{ctx})}$, for example when the question prior happens to align with the ground truth, the resulting contrast signal is necessarily weak. In that regime, the correction magnitude becomes small and more susceptible to noise, so the final decoding often changes only slightly and may yield only marginal gains. This interpretation is consistent with the analysis in IBD (Zhu et al., 2025b), which similarly notes that small divergence between contrasted predictions leads to limited debiasing effect. More broadly, it also aligns with the same failure mode discussed in earlier contrastive decoding literature, where highly similar contrasted distributions provide only weak or noisy optimization signals (Li et al., 2023b; Chuang et al., 2024).

D.3 Analysis on MCQA Selection Bias

Recent studies show that LVLMs may exhibit MCQA selection bias, where predictions are affected by option positions or option tokens rather than only by question semantics (Janghorbani and De Melo, 2023; Weng et al., 2024b; Atabuzzaman et al., 2025). This issue is particularly relevant to our Q-only analysis, since the model receives no visual or structured evidence. To separate SDG-oriented directional defaults from general MCQA formatting artifacts, we conduct an option-perturbation control experiment on INSTRUCTBLIP-7B.

Specifically, we test two types of perturbations: changing the option order and changing the option-token format. After each perturbation, we map the selected answer back to its semantic direction: optimistic, conservative, or pessimistic. As shown in Table J, MCQA formatting does affect the exact distribution, especially under option-order changes. However, the effect is limited: across both order and token perturbations, the relative tendency remains consistent as optimistic $>$ pessimistic $>$ conservative. This suggests that the observed Q-only directional default is not solely caused by option-position or option-token artifacts.

Importantly, this control does not change the interpretation of CADE. CADE uses Q-only only as one bias probe, and also relies on CTX+Q and image-present views to penalize context dominance and modality imbalance.

D.4 Analysis on Efficiency–Effectiveness Tradeoff

CADE requires four forward passes at inference time, but this cost remains moderate in context. Contrastive debiasing inherently requires multiple predictions, with at least two passes already needed to compare the original and bias-related signals. In CADE, two of the four passes are text-only and are much faster than multi-modal inference. More importantly, this limited overhead yields substantially stronger results than competing methods (as shown in Sec. 5.4) while remaining far cheaper and easier to deploy than fine-tuning-based alternatives, which require additional optimization time, GPU resources, and checkpoint management. Overall, CADE offers a favorable efficiency–effectiveness tradeoff in practice.

Table J: **Option-perturbation control experiment on INSTRUCTBLIP-7B under the Q-only setting.** Predictions are mapped back to semantic directions after perturbation. Values are percentages.

Perturbation type	Perturbed format	Original format	Optimistic	Conservative	Pessimistic
Original	A/B/C	A/B/C	56.2	20.5	23.3
Reverse order	C/B/A	A/B/C	52.8	22.1	25.1
Random order	Random permutation	A/B/C	57.4	19.8	22.8
Lowercase token	a/b/c	A/B/C	53.9	22.1	24.0
Numeric token	1/2/3	A/B/C	58.6	18.9	22.5

Table K: **Inference runtime comparison on A100 GPUs.**

Method	GPU-sec / question	Relative cost
Full inference	0.144	1.00×
CADE	0.324	2.25×

Table L: **Approximate per-view runtime breakdown of CADE.**

View	Runtime proportion
Full	40%
IMG+Q	40%
CTX+Q	10%
Q-only	10%
CADE total	100%

To make this cost explicit, we report the average per-question runtime on A100 GPUs in Table K and per-view runtime breakdown analysis in Table L.

E Licenses and Terms of Use

SDGBiasBench is a derivative benchmark built on the Demographic and Health Surveys (DHS) (Corsi et al., 2012) and *DHS-based* subset of SUSTAINBENCH (Yeh et al., 2021), augmented with our human-authored reasoning questions and expert-labeled answers. We release our added benchmark artifacts (questions, answer keys, splits, and evaluation code) under the CC BY-SA 4.0 license, consistent with SUSTAINBENCH. The underlying modalities follow the original SUSTAINBENCH data licenses/terms: Landsat, DMSP, NAIP, and VIIRS imagery are public domain; Sentinel-2 imagery is under CC BY-SA 3.0 IGO; Sentinel-1 and MODIS imagery are freely reusable under their respective open access terms; and PlanetScope and Mapillary street-level imagery (when used) are provided under CC BY-SA 4.0 as packaged in SUSTAINBENCH. We do not redistribute any DHS micro-level survey records; all DHS-derived labels we use are aggregated at the cluster level and comply with the DHS Program Terms of Use.

F Intended Use and Access Conditions

Our benchmark is intended for non-commercial research on multi-modal reasoning and bias diagnosis in SDG-related settings. We use existing artifacts (*e.g.* pretrained VLMs and publicly available imagery/context sources) in ways consistent with their stated intended use and license/terms, and we do not attempt to recover personal data or identify individuals. Any derivatives we release are provided for research purposes only and inherit upstream restrictions where applicable; users must comply with original licenses/ToS for any underlying data and models. We do not endorse deployment of models trained/evaluated on this benchmark for high-stakes decision-making without additional validation and governance.

G Information on Use of AI Assistants

We used AI assistants in a limited capacity to support manuscript preparation, such as refining wording, improving clarity, and checking grammar and formatting. All technical content (*e.g.* ideas, methods, experiments, results, and conclusions) was produced and verified by the authors, who take full responsibility for the final paper.