ViLBias: Detecting and Reasoning about Bias in Multimodal Content

Shaina Raza[1] et al.,

Vector Institute, Toronto, ON, Canada [1]

Abstract

Detecting bias in multimodal news requires reasoning over text-image pairs rather than simple classification. We present **ViLBias**, a VQA-style benchmark comprising **40,945 text-image pairs**. Results show that incorporating images improves detection accuracy by 3–5% over text-only models.

The Problem: Multimodal Bias

Bias in news extends beyond language to include visual strategies:

- **Textual:** Loaded wording and selective emphasis.
- **Visual:** Image selection, cropping, staging, and emotive imagery.
- **Cross-Modal:** Mismatches between what is said and what is shown.

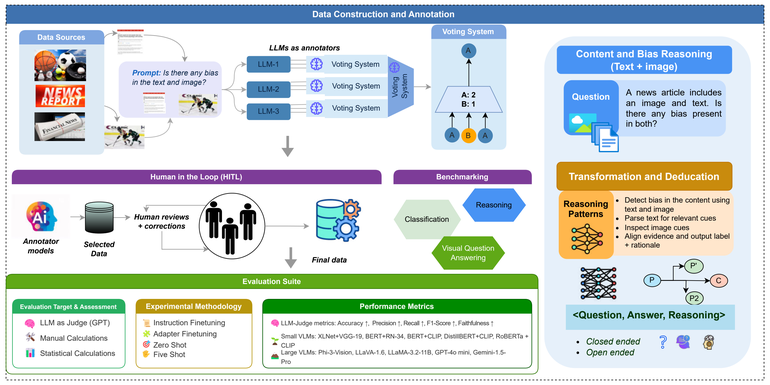
Benchmark	Mod.	Foc.	Ann.	Rat.	MM.	Reason.
BiasLab	T	B	H	✓	–	◐
BIASsist	T	B	LLM	✓	–	✓
MDAM3	T+I	MI	AI+H	✗	◐	✗
BASIL	T	B	H	✓	–	✗
NewsBag	T+I	MI	H	✗	◐	✗
FakeNewsNet	T+I	MI	AI+H	✗	◐	✗
ViLBias (ours)	**T+I**	**B**	**AI+H**	✓	✓	✓

BiasCorpus Curation

- **Sources:** Collected via Google News RSS feeds (May 2023–Sept 2024) to ensure temporal diversity and real-world news coverage.
- **Outlets:** Politically and geographically diverse media sources, including Financial Times, USA TODAY, CNN, BBC, and Al Jazeera, to capture variation in framing and editorial orientation.
- **Scale & Distribution:** 40,945 unique news images paired with articles; 22,974 labeled *Biased* and 17,971 labeled *Not Biased*.
- **Annotation & Verifiability:** All samples reviewed and verified by domain experts using structured labeling guidelines to ensure reliability and consistency.

ViLBias Framework

The pipeline consists of three main stages: (1) Data Collection, (2) Automated/Human Labeling, and (3) Benchmarking.



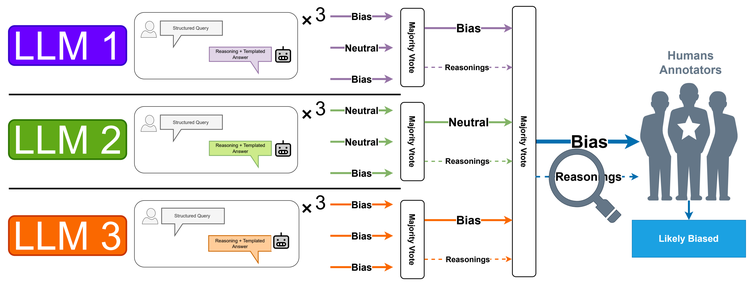
ViLBias Framework. The pipeline comprises three main stages: (1) Data Collection and Preprocessing, (2) Automated and Human-Annotated Labeling, and (3) Final Evaluation.

HITL Hybrid Annotation Pipeline

We use a two-step voting mechanism to maximize reliability:

1. **Intra-Model Voting:** Each LLM is queried 3x; majority vote removes stochasticity.
2. **Inter-Model Voting:** Majority vote across three different LLMs.
3. **Human-In-The-Loop:** 12 expert reviewers (CS, Media Studies, etc.) validate outputs.

Agreement: Cohen's Kappa reached 0.72 (Substantial alignment).



LLM-As-Annotators with Human-In-The-Loop (HITL) Bias Annotation Framework. Each LLM is queried three times, and its responses are majority-voted to remove stochasticity. These majority-voted labels are taken from three LLMs and then majority-voted (across LLMs) again to produce a final label, which is reviewed—along with LLM reasoning—by human annotators.

Main Results

Vision–Language (Text+Image) Performance. Metrics are reported in percentages: Precision (Prec.), Recall, F1, and Accuracy (Acc.). Closed-source models are marked with †. Results are grouped into SLMs and VLMs. For large models (Phi, LLaVA, LLaMA), we report *Instruct* variants. Configurations include fine-tuning (FT), instruction fine-tuning (IFT), and few-shot (0-shot, 5-shot). All models use the same splits and both modalities. Bold means best values in each model family.

Model	Config	Prec. (%)	Recall (%)	F1 Score (%)	Acc. (%)
Small VLMs					
XLNet + VGG-19	FT	72.0	68.2	70.1	77.1
BERT + RN-34	FT	75.8	71.5	73.6	79.4
BERT + CLIP	FT	81.3	73.4	77.2	84.2
DistilBERT + CLIP	FT	68.5	64.0	66.2	74.9
RoBERTa + CLIP	FT	**84.5**	**81.4**	**82.9**	**83.6**
Large VLMs					
Phi-3-Vision	0-shot	70.4	66.0	68.1	69.8
	5-shot	73.2	71.0	72.1	70.5
	IFT	**76.8**	**78.1**	**77.4**	**74.0**
LLaVA-1.6	0-shot	62.5	60.8	61.6	62.7
	5-shot	68.1	67.0	67.5	65.2
	IFT	75.4	74.6	75.0	76.1
LLaMA-3.2-11B	0-shot	65.0	68.9	66.9	68.4
	5-shot	73.4	74.2	73.8	72.1
GPT-4o mini†	0-shot	71.8	74.6	73.2	72.9
	5-shot	**77.9**	**79.8**	**78.8**	**77.2**
Gemini-1.5 Pro†	0-shot	70.9	73.1	72.0	71.5
	5-shot	76.8	78.5	77.6	76.9

- **Multimodal Gain:** Adding images provides an average gain of **+8.22 F1 points** over text-only variants.
- **Reasoning Gap:** Reasoning accuracy (52–79%) lags behind classification accuracy.

Conclusion

This work introduces ViLBias, a multimodal benchmark of over 40,000 text-image pairs designed to detect and reason about news media bias using a hybrid human-AI annotation pipeline. Experiments demonstrate that multimodal grounding enhances detection accuracy and parameter-efficient methods are highly effective, though a gap persists between classification success and reasoning quality.