Addressing Blind Guessing: Calibration of Selection Bias in Multiple-Choice Question Answering by Video Language Models

Anonymous ACL submission

Abstract

Evaluating Video Language Models (VLMs) is a challenging task. Due to its transparency, Multiple-Choice Question Answering (MCQA) is widely used to measure the performance of these models through accuracy. However, existing MCQA benchmarks fail to capture the full reasoning capabilities of VLMs due to *selection bias*, when models disproportionately favor certain answer options based on positional patterns observed during training. In this work, we conduct a comprehensive empirical analysis of several VLM architectures across major datasets designed to assess complex video-focused reasoning. We identify where the bias is most pronounced and demonstrate to what extent model responses reflect genuine understanding of video content and related questions, as opposed to reliance on arbitrary patterns or superficial cues, such as answer position. By decomposing the MCQA task and adapting fairness bias metrics to VLMs, we introduce a post-processing calibration technique **BOLD** to balance this bias. Our results show that reducing selection bias improves not only debiasing metrics but also overall model performance, including Accuracy and F1 Mean score. Our method, by suppressing "blind guessing", offers a more cost- and time-effective approach to mitigating selection bias compared to existing techniques. This study represents the first focused investigation of selection bias in video-to-text LLM-powered models.

1 Introduction

Multiple-choice question answering (MCQA) is a convenient method to assess language models' performance thanks to its simplicity and transparency. Each MCQA instance presents a question with several answer options, and the task is to select the correct one. However, large language models (LLMs) often exhibit *selection bias*, favoring certain answer positions (Zheng et al., 2024a; Wang et al., 2024a). This bias undermines the fairness, reliability, and reasoning capabilities of the models.

While LLMs process text, video-language models (VLMs) handle sequences of frames to capture both spatial and temporal features. Recent advances align pre-training LLMs with video encoders to enhance video understanding (Bordes et al., 2024). Although VLMs inherit much from LLMs, they face unique challenges, including higher computational costs and the need for effective spatial and temporal mapping for dynamic inputs.

Despite their backbone shared with LLMs, VLMs' specific characteristics require separate scrutiny of selection bias. Notably, no research has yet focused on bias in video-language models.

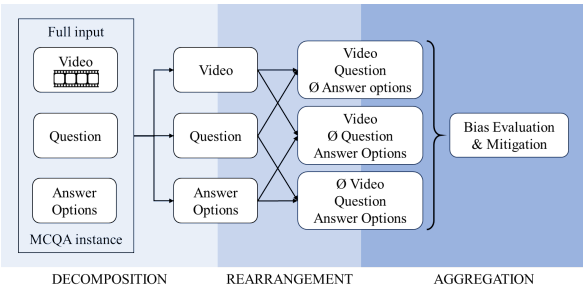


Figure 1: Decomposition approach: we separate the three key components of MCQA, reorganize them into pairs by removing one at a time, and apply probability debiasing to the aggregated data.

We address this gap by proposing the first study of MCQA selection bias in video datasets, focusing on temporal reasoning, spatial object tracking, and camera behavior rather than general plot understanding. This approach highlights the specific challenges of VLMs.

Traditionally, selection bias is mitigated through shuffling (Zheng et al., 2024a; Liu et al., 2023), where answer options are permutated multiple times to yield more impartial results. In contrast, we propose treating it similarly to the fairness bias,

where models exhibit preference toward specific groups, such as gender or race. We adapt fairness bias metrics for VLMs via the decomposition approach showed in Figure 1.

To evaluate our approach, we select three VLMs with different architectures and apply post-processing bias calibration techniques to four datasets. Our method, which adjusts predictions without retraining, improves interpretability and fairness with reduced computational costs. Even though there is no direct correlation between debiasing and performance, our approach generally improves both bias metrics and accuracy, as well as the F1 mean score, signaling improved recall.

In summary, our contributions are:

- A comprehensive study of selection bias in VLMs using 11 dataset modifications focused on video-specific challenges.
- Adaptation of debiasing metrics to assess VLMs' behavior on video MCQA benchmarks.
- An efficient calibration technique that improves the reliability of current MCQA datasets and the VLM performance assessment.

Our work aims to reveal deeper reasoning issues in contemporary VLMs through refined MCQA benchmarks.

2 Related Works

Selection bias in LLMs has been extensively studied (Pezeshkpour and Hruschka, 2023; Zheng et al., 2023; Wang et al., 2024a; Li et al., 2024; Liusie et al., 2024; Wang et al., 2024b). Fewer works are dedicated to image-to-text vision LLMs (Zong et al., 2024; Zhang et al., 2024; Xue et al., 2024).

Zheng et al. (2024a) conducted an empirical analysis of both position and token biases and proposed PriDe, a debiasing method for LLMs that separates prior bias from unbiased results using shuffling.

We also treat the selection bias as subtraction from observed results in order to obtain the debiased outputs, but we move away from shuffling and regard bias as a vector projected onto three planes. Thus, our calibration method aligns more closely with Zhang et al. (2024), who cut off the visual input for the image-to-text MCQA. Differently, we operate on video-to-text, and, going further, decompose the bias vector into several projections rather than one. This way we eliminate the unimodal bias highlighted by Zhang et al. (2023b).

Wang et al. (2024a) and Liu et al. (2023) resolve the selection bias by augmenting datasets with reordered and additional options, increasing number of question-answer pairs in the dataset. We, in contrast, manipulate answer options, as well as videos and questions, not as a solution but as a means of analysis: we identify where the bias is most pronounced and use these modifications as a shortcut for debiasing.

Lastly, Pezeshkpour and Hruschka (2023) propose calibrating LLM predictions by redistributing probabilities to empirically determined options based on the model's uncertainty. However, we find this approach insufficiently robust, as, according to our analysis, the bias depends on both the model and the dataset, and there is no consistent correlation between the model confidence and the presence of bias.

3 Empirical Analysis

3.1 Datasets

We selected four video MCQA datasets that challenge VLMs' temporal and spatial reasoning abilities. These datasets represent everyday situations and span videos from a few seconds to over an hour, most of which last up to one minute.

Table 1: Summary of Question-Answer pairs, Videos, and Number of Options for each dataset.

Dataset	QA pairs	Videos	Options
NExT-QA	8564	1000	5
NExT-GQA	4962	971	5
STAR	7098	914	4
Video-MME	2700	900	4
Perception Test	7656	3926	3

NExT-QA and NExT-GQA NExT-QA (Xiao et al., 2021) is one of the first datasets designed to assess a VLM's ability to capture causal, temporal, and descriptive topics through what/when/why/how and before/after questions. We evaluate the models on the test split.

NExT-GQA is a subset of the NExT-QA test split annotated with timestamp to locate the answer moment in the video.

STAR The question-answer (QA) pairs in STAR (Wu and Yu, 2021) are generated using functional programs based on situation hyper-graphs. They

Table 2: Inference Accuracy (Acc) and F1 Mean Score by Models and Datasets.

Model	NExT-QA		STAR		Perception Test		Video-MME	
	Acc	F1 Mean	Acc	F1 Mean	Acc	F1 Mean	Acc	F1 Mean
Video-LLaMA	40.85	40.85	36.59	31.86	41.59	37.19	32.67	28.15
Video-LLaVA	49.96	49.81	34.71	31.83	40.73	35.69	34.22	30.99
SeViLA	63.78	63.88	46.28	46.14	45.30	45.11	39.85	39.82

cover four categories of situational reasoning: Interaction, Sequence, Prediction, and Feasibility. We use the publicly available validation split.

Perception Test The Perception Test dataset (Patruaucean et al., 2023) tests the memory, abstraction, physics, and semantics reasoning abilities of the models. We use a 40% split of the correct-answer-annotated validation set to save the resources. Sampling is stratified to maintain the original balance of question types and the correct answer distribution.

Video-MME Video-MME (Fu et al., 2024) is a comprehensive MCQA evaluation benchmark with key distinctions: (a) video duration range from 11 seconds to 1 hour, (b) it covers 6 primary visual domains with 30 subfields, (c) includes subtitles, and (d) features meticulously crafted annotations. The released test split has an imbalanced option distribution, so we equalized it to properly assess option bias.

The amount of QA pairs, videos and answer options of each dataset is shown in Table 1. Additional details on the datasets including QA examples are given in Appendix A.

3.2 Models

Three models were selected to represent different approaches to uniting video and language modalities (Bordes et al., 2024). Despite architectural differences, SeViLA, Video-LLaMA, and Video-LLaVA share key characteristics: they sample frames to extract visual features, either by keyframe selection or uniform sampling, and align visual features with textual representations using pre-trained LLMs. Each model uses pre-trained visual encoders and fine-tunes pre-trained components with video-text datasets.

Video-LLaMA Video-LLaMA (Zhang et al., 2023a) aligns a BLIP-2-based visual encoder (Li et al., 2023) with LLaMA (Touvron et al., 2023). It incorporates audio signals using Imagebind (Girdhar et al., 2023). Both the video and audio Q-formers are trained on the Webvid-2M dataset (Bain et al., 2021) to align the modalities. LLaMA is fine-tuned with visual instructional data.

Video-LLaVA Video-LLaVA (Lin et al., 2023a) aligns video and LLMs in a shared representation. It uses the pre-trained CLIP ViT-L/14 vision encoder (Radford et al., 2021) and the Vicuna language model (Chiang et al., 2023), improving multimodal instruction following.

SeViLA SeViLA (Yu et al., 2023) adopts a self-chaining approach with BLIP-2 (Li et al., 2023) for keyframe localization and question answering. It includes a Localizer, which uniformly samples 32 frames and selects the 4 most relevant key ones, and an Answerer that generates responses based on those frames. The Localizer and Answerer both have Flan-T5 (Chung et al., 2022) for LLM.

We provide the Accuracy and F1 Mean Score results of the zero-shot inference on the aforementioned datasets in Table 2. Further details on the model settings and prompts can be found in Appendix B.

3.3 Experimental Settings

Eleven dataset modifications are used to examine the role of each component in MCQA benchmarks: video, question, and answer options.

Video Modifications alter the video input:

- **Correct Frames**: Only frames containing the answer are provided. This tests if reducing extraneous visual information improves performance. The modification applies to NExT-QA (NExT-GQA) and STAR due to the availability of timestamp annotations.

- **Empty Frames**: Black frames are passed instead of video input. This tests the model's response to the absence of visual information within the multimodal setting.

Question Modifications adjust the questions in two ways:

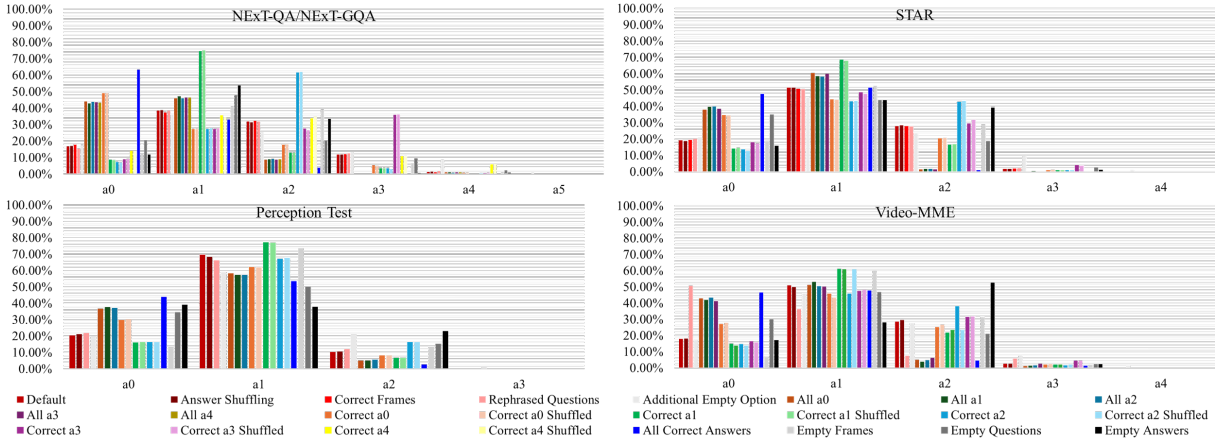


Figure 2: Video-LLaMA option distribution for all settings of NeXT-QA, NExT-GQA, STAR, Perception Test and Video-MME. Correct a_i, Correct a_i Shuffled and All a_i represent Correct Answer, Correct Answer with Shuffling and All Correct Answers in each option, respectively.

- **Rephrased Questions**: Llama3 (Dubey et al., 2024) rephrases the original questions. Five variations are generated, and one is randomly selected.[1]

- **Empty Questions**: Following (Balepur et al., 2024), we replace the question with an empty string to test if models can infer the question from answer choices and video.[2]

Answer Option Modifications examine the role of answer:

- **Answer Shuffling**: Options are randomly shuffled to test robustness against bias through changing the option order.

- **Correct Answer in Each Option**: The correct answer is placed in a fixed position; its content is swapped with another option.

- **Correct Answer with Shuffling**: The correct answer is fixed, and the remaining options are shuffled.

- **Additional Empty Option**: A new empty option is added in the end to check for overfitting to specific options.

- **All Identical Answers**: All options are identical testing for deviation from uniform distribution.

- **All Correct Answers**: Every option is correct testing for deviations in a setting where any answer is valid.

- **Empty Answers**: All options are empty to check the model's ability to choose a non-existent answer.

To reduce token bias, we use less common option names a0, a1, a2, etc. instead of the conventional A, B, C, etc.

Modifications can also be categorized as accuracy-relevant (e.g., correct frames, shuffling, or rephrased questions) and accuracy-irrelevant (e.g., identical answers or empty parts). Further details, including accuracy results and confusion matrices, are provided in the Appendix C.

3.4 Observations

Figures 2– 4 display the answer distributions of each model across options and datasets. High bars concentrated in a specific option indicate bias, while consistently low bars suggest the model disregards that option.

Across all models and datasets, shuffling answer options – either by shifting the correct option or random permutation, results in similar option distributions. This indicates that models rely heavily on answer position rather than content. However, when the correct answer is moved to less favored positions, those positions show increased selection rates, as seen in peaks for *Correct* a_i and *Correct* a_i *Shuffled*. Additionally, rephrased questions yield distributions similar to the default, regardless of correctness.

[1]A manual check of 10% of each rephrased dataset confirmed that meaning was retained, with minimal impact on accuracy due to meaning loss. Notably, very few hallucinations were spotted in the NExT-QA rephrased copy.

[2]The previous research for LLMs has shown surprisingly high accuracy in this scenario. We investigate if the video modality contributes to even more meaningful predictions.

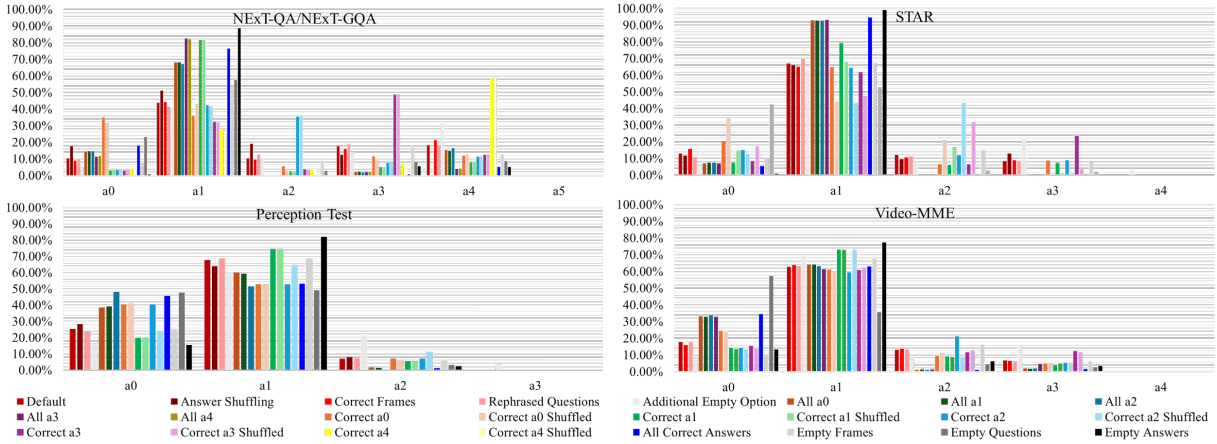


Figure 3: Video-LLaVA option distribution for all settings of NeXT-QA, NExT-GQA, STAR, Perception Test and Video-MME. Correct a_i, Correct a_i Shuffled and All a_i represent Correct Answer, Correct Answer with Shuffling and All Correct Answers in each option, respectively.

Using all correct frames did not significantly alter the distributions, suggesting that the models' localization capabilities are adequate. Remarkably, in scenarios where all answer options are identical (accuracy-irrelevant), the selection distribution is not uniform.

Video-LLaMA Video-LLaMA exposes a significant bias toward option *a1* (Figure 2): when the correct answer is in this position, accuracy exceeds 70% on NExT-QA and Perception Test. This bias is more evident across STAR, Perception Test, and Video-MME, especially with empty frames.

The model tends to ignore options beyond the third one. In Perception Test, which has only three choices, bias against the third option is already noticeable.

Notably, with empty answers, option *a2* receives an unexpectedly high number of selections, even surpassing *a1* in Video-MME.

Video-LLaVA Video-LLaVA's performance (Figure 3) is similar to Video-LLaMA's but shows an even stronger bias toward option *a1*, with higher-numbered options being largely ignored. Unlike Video-LLaMA, Video-LLaVA overwhelmingly favors *a1* when answers are empty. When questions are empty, option *a0* starts competing with *a1*, especially in the Video-MME dataset. In NeXT-QA, moving the correct answer among options leads to a preference for the option containing it.

SeViLA SeViLA is unique in becoming distracted when an empty option is added. This suggests SeViLA may have been trained on answer distributions with a fixed number of options. It shows reduced accuracy for all correct frames in STAR (see Table 13), echoing findings by Goyal et al. (2016) and Lin et al. (2023b), where "blind" language-only models could answer image questions.

Nevertheless, SeViLA demonstrates the greatest robustness on NExT-QA and STAR. Its unstable option distributions occur only when all correct answers are identical (preferring *a0*) and with empty answers (preferring the last option). Figure 4 shows slightly higher saturation for these options across all datasets.

4 Methodology

The bias calibration is based on the following requirements to the MSQA benchmark:

- **Balanced options**: All answer options are presented equally without any inherent bias (e.g., one answer option is not more likely than others just by how it is presented).
- **Independent options**: Each option for a particular question is independent of others (e.g., there are no inherent relationships between them that would influence selection).
- **Equal difficulty**: All questions are equally difficult, and no question skews results toward a particular option group.
- **No logically better or worse options except the correct one**: This ensures that other options are plausible distractors, but only one option is objectively correct.

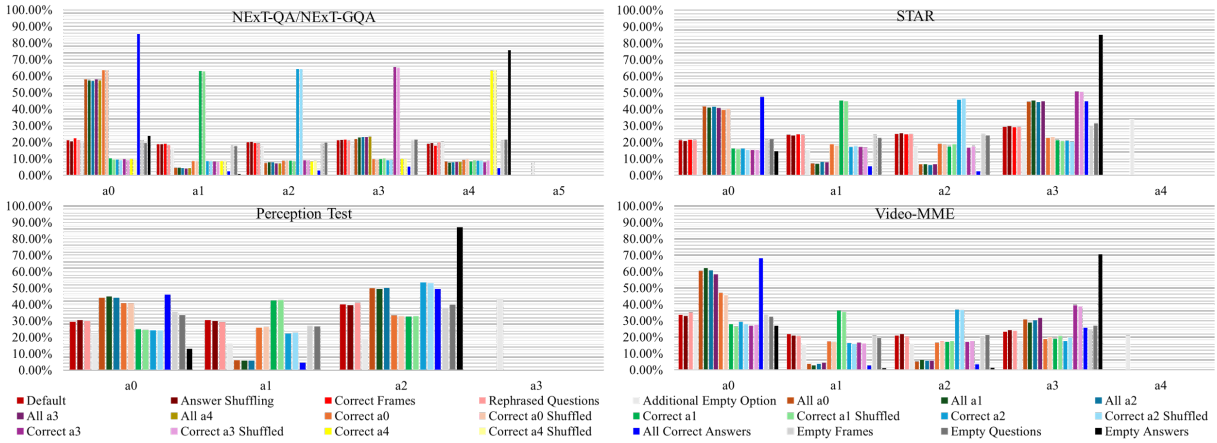


Figure 4: SeViLA option distribution for all settings of NeXT-QA, NExT-GQA, STAR, Perception Test and Video-MME. Correct a_i, Correct a_i Shuffled and All a_i represent Correct Answer, Correct Answer with Shuffling and All Correct Answers in each option, respectively.

4.1 Evaluation Criteria

Our framework both **detects bias** and **evaluates performance**. Bias metrics assess fairness by evaluating the model's decision making on consistency in treating all answer options, while performance metrics evaluate accuracy and reliability.

We monitor the standard deviations of the F1 score (F1_std) and the recall (Recall_std). High F1_std indicates inconsistent performance across options, signaling bias. High Recall_std shows the model is better at correctly identifying answers for certain options over others.

For option probabilities, we use the standard deviation of the symmetric Jensen-Shannon distance (JS_std) to detect inconsistencies in how the predicted probability distributions differ from the true distributions. This metric, bounded between 0 and 1, is more interpretable than Kullback-Leibler (KL) divergence.

By tracking these metrics, we address group fairness, ensuring the model treats all options equitably. Large standard deviations in any metric suggest bias, which we aim to minimize.

4.2 Debiasing Approach

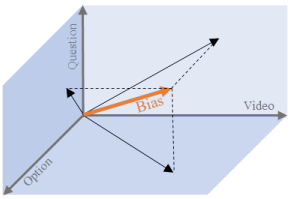


Figure 5: Bias vector and its projections on the decomposition planes.

Every MCQA benchmark comprises essential components that make the answering task feasible. In the case of video-MCQA, these components include the video, the question, and the answer options. We construct decompositions where the task becomes impossible by design, as one of the key components is removed. This allows us to examine the projections of model's behavioral bias on each decomposition plane as shown in Figure 5. In a fair and unbiased model, projecting the dataset by omitting one of the key components would result in a uniform distribution. However, as shown in Section 3, this is far from the case. In the decomposition planes, the bias is more clearly manifested, and since we expect uniformity, the bias becomes easier to detect and mitigate.

4.2.1 Decomposition

Consider a task T that we wish to decompose into n non-overlapping components $C_1, C_2, \ldots, C_n$. This decomposition can be expressed as:

$$T = \bigcup_{i=1}^{n} C_i, \quad \text{where} \quad C_i \cap C_j = \emptyset \quad \text{for all} \quad i \neq j \qquad (1)$$

The task is considered ill-defined if it lacks one of the key components. Additionally, there is a well-defining property, or correctness, $P(T)$ of the task. If any component C_i is removed, the property $P(T)$ becomes ill defined. We assume that all tasks are initially well-defined.

A model M takes a task T as input and outputs an option ID corresponding to the correct answer for a video-based question: $M(T) = ID_{opt}$.

For an ill-defined task, a model that is both reasonable and unbiased should uniformly select

any option ID. Formally, if $P(T)$ is 'ill-defined': $M(T) = \text{RandomChoice}(\{ID_{opt}\})$.

The key insight is that if the model has bias on a given dataset, this bias is likely to be more ubiquitous, hence evident, in the ill-defined versions of the tasks. Thus, the model bias can be decomposed as follows: $Bias_M \supseteq \bigcup Bias_M|_{\hat{D}}$, where $\hat{D}$ represents an ill-defined copy of the dataset D.

Let A be an attack on the well-defined task that makes this task defined insufficiently. Subsequently, there are three types of attacks that make the task ill-defined: $A|_{v=0}$ for all zeroed frames in the video, $A|_{q=0}$ for all questions as an empty string, and $A|_{o=0}$ for all options as IDs only.

We investigate the importance of the following components of the bias: $Bias_M|_{A|_{v=0}(D)}$, $Bias_M|_{A|_{q=0}(D)}$, $Bias_M|_{A|_{o=0}(D)}$, and how efficiently we can handle the general bias by balancing through these components.

4.2.2 Debiasing

The observed prediction distribution P_o over options $\{d_i\}_{i=1}^n$ and conditioned on task T where task T = (video, question, options) can be decomposed as the prior distribution P_p over d_i and the debiased distribution P_d:

$$P_o(d_i|T) = \frac{1}{Z_T} P_p(d_i|T) \times P_d(d_i|T) \quad (2)$$

where Z_T is a normalization factor.

The baseline method **BOLD** (Bias Optimisation Leveraging Decomposition) estimates the prior bias as following:

$$\tilde{P}_p(d_i|T) = softmax\left(\sum_j P_p(d_i|A_j(T))\right) \quad (3)$$

First, we estimate the model bias on the dataset after attack A. We derive the P_p knowing that P_d is uniformly distributed from Equation 2:

$$P_p(d_i \mid A(T)) = P_o(d_i \mid A(T)) \quad (4)$$

To estimate the bias under $A|_{v=0}$, $A|_{q=0}$ and $A|_{o=0}$ attacks, we first compute $P_p(d_i|A|_{v=0}(T))$, $P_p(d_i|A|_{q=0}(T))$ and $P_p(d_i|A|_{o=0})$, and then the prior estimation, respectively: $\tilde{P}_p(d_i|T) = softmax[P_p(d_i|A|_{v=0}(T)) + P_p(d_i|A|_{q=0}(T)) + P_p(d_i|A|_{o=0})]$.

According to Zheng et al. (2024b), biases related to option IDs may transfer effectively across different samples and domains. Therefore, finding the global prior with K samples, we can mitigate the bias for the entire dataset with minimal efforts.

From the dataset D, we take $K = k * ||D||$ samples, denoted as D_k, where k is adjusted coefficient based on estimation budget. Each sample in D_k is subjected to the set of attacks A. For each, we calculate the sample-specific prior $P_p(d_i|A(T))$ by Equation 4 and estimate the adjusted prior $\tilde{P}_p(d_i|T)$ by Equation 3. The global prior, $\tilde{P}_{\text{p}}(d_i)$, is then obtained by averaging the priors for each sample.

Ultimately, the posterior debiased through the global prior is calculated as follows:

$$P_d(d_i|T) = softmax\left(\log P_o(d_i|T) - \log \tilde{P}_p(d_i)\right) \quad (5)$$

In the extended version, **Weighted_BOLD**, we use weights for $P_p(d_i|A(T))$:

$$\tilde{P}_p(d_i|T) = softmax\left(\sum_j w_j P_p(d_i|A_j(T))\right) \quad (6)$$

This allows for a more optimal direction in the latent space of the priors, as we can prioritize certain priors by significance and refine the prior estimation $\tilde{P}_p(d_i|T)$.

To gain the optimal weights, we use a 5-fold cross-validation procedure. In each fold, we optimize weights $\{w_i\}$ using the COBYLA algorithm (Powell, 1994), subject to the constraints $0 \leq w_i \leq 1$ or $|w_i| \leq 1$. While we minimize the standard deviation of Recall_std across the options on the test set, we monitor the same bias metrics on the validation set. After completing the cross-validation we average the estimated global priors across the folds and apply this averaged global prior.

Both algorithms, BOLD and Weighted_BOLD, are outlined in Appendix D.

5 Results

We conducted tests with the k sample parameter set to 0.25, 0.5, 0.75, and 1 (the entire dataset) in combination with positive and negative weights. The configuration with k=0.5 and $w_i \geq 0$ provided the optimal balance between sample size and performance. As shown in Table 3 with results for positive weights and k=0.5,[3] the overall significant improvement in bias metrics, including *Recall_std*, indicates that our method effectively reduces option disparity.

Weighted_BOLD generally outperforms BOLD because the weights capture critical priors in the latent space. Thus, they prove to be effective for

[3] Additional results for k values of 0.25, 0.75, and 1 combined with different weights are provided in Appendix D.

Table 3: Comparison of BOLD and Weighted_BOLD bias mitigation approaches across models and datasets for performance and bias monitoring metrics with $k = 0.5$. Green arrows indicate improvements: upward for Accuracy and F1_mean, and downward for standard deviation metrics. Red arrows represent deterioration, respectively.

Model	Dataset	Configuration	Performance Metrics		Bias Monitoring Metrics		
			Accuracy	F1_mean	Recall_std	F1_std	JS_std
Video-LLaMA	NExT-QA	**BOLD**	45.88 (↑2.43%)	42.15 (↑3.18%)	22.98 (↓3.53%)	15.53 (↓2.11%)	15.55 (↓3.31%)
		Weighted_BOLD	45.91 (↑2.51%)	42.2 (↑3.29%)	22.83 (↓4.19%)	15.51 (↓2.2%)	15.56 (↓3.27%)
	STAR	**BOLD**	37.19 (↑1.82%)	33.02 (↑3.65%)	22.29 (↓7.5%)	14.55 (↓5.75%)	14.58 (↓4.66%)
		Weighted_BOLD	37.34 (↑2.24%)	33.5 (↑5.16%)	21.13 (↓12.3%)	14.05 (↓8.98%)	14.14 (↓7.54%)
	Perception Test	**BOLD**	41.9 (↑1.24%)	38.68 (↑4.01%)	24.05 (↓11.11%)	9.87 (↓13.45%)	14.45 (↓10.01%)
		Weighted_BOLD	42.06 (↑1.62%)	39.36 (↑5.86%)	21.8 (↓19.45%)	9.11 (↓20.06%)	13.02 (↓18.94%)
	Video-MME	**BOLD**	32.73 (↑2.13%)	29.23 (↑3.85%)	19.29 (↓6.45%)	11.96 (↓4.46%)	13.1 (↓4.73%)
		Weighted_BOLD	32.2 (↑0.47%)	28.98 (↑2.94%)	17.65 (↓14.38%)	11.69 (↓6.56%)	12.64 (↓8.04%)
Video-LLaVA	NExT-QA	**BOLD**	51.81 (↑3.72%)	51.71 (↑3.82%)	13.27 (↓18.63%)	2.67 (↓18.42%)	4.58 (↓15.06%)
		Weighted_BOLD	52.15 (↑4.39%)	52.04 (↑4.49%)	12.72 (↓22.02%)	2.62 (↓19.83%)	4.49 (↓16.83%)
	STAR	**BOLD**	37.21 (↑7.01%)	35.76 (↑12.37%)	18.28 (↓26.85%)	4.97 (↓27.15%)	4.54 (↓21.39%)
		Weighted_BOLD	37.53 (↑7.94%)	36.18 (↑13.69%)	17.45 (↓30.15%)	4.91 (↓28.03%)	4.26 (↓26.24%)
	Perception Test	**BOLD**	41.62 (↑2.22%)	38.69 (↑8.4%)	21.75 (↓20.9%)	10.92 (↓25.97%)	3.78 (↓28.4%)
		Weighted_BOLD	41.95 (↑3.02%)	39.46 (↑10.58%)	19.73 (↓28.26%)	10.24 (↓30.53%)	3.4 (↓35.51%)
	Video-MME	**BOLD**	34.7 (↑1.19%)	32.79 (↑5.81%)	18.19 (↓24.51%)	6.16 (↓23.63%)	3.84 (↓19.93%)
		Weighted_BOLD	34.63 (↑0.97%)	32.97 (↑6.38%)	16.36 (↓32.07%)	6.01 (↓25.43%)	3.43 (↓28.5%)
SeViLA	NExT-QA	**BOLD**	63.92 (↑0.02%)	63.89 (↑0.02%)	2.03 (↓5.47%)	1.18 (↓10.4%)	1.99 (↓0.17%)
		Weighted_BOLD	63.93 (↑0.04%)	63.91 (↑0.04%)	1.99 (↓7.64%)	1.19 (↓9.83%)	1.99 (↓0.04%)
	STAR	**BOLD**	46.22 (↓0.12%)	46.1 (↓0.08%)	4.13 (↓7.46%)	2.26 (↓1.78%)	2.1 (↓6.64%)
		Weighted_BOLD	46.2 (↓0.18%)	46.08 (↓0.13%)	4.01 (↓10.17%)	2.2 (↓4.36%)	2.07 (↓8.2%)
	Perception Test	**BOLD**	45.32 (↑0.06%)	45.18 (↑0.16%)	5.18 (↓16.61%)	2.95 (↓2.43%)	1.3 (↓18.42%)
		Weighted_BOLD	45.31 (↑0.03%)	45.2 (↑0.21%)	4.56 (↓26.58%)	2.83 (↓6.43%)	1.08 (↓31.92%)
	Video-MME	**BOLD**	40.19 (↑0.84%)	40.17 (↑0.88%)	4.11 (↓12.03%)	1.41 (↓16.19%)	0.73 (↓12.52%)
		Weighted_BOLD	40.04 (↑0.46%)	40.03 (↑0.54%)	3.78 (↓19.14%)	1.44 (↓14.6%)	0.69 (↓17.22%)

fine-tuning bias correction within the decomposed priors.

Both BOLD and Weighted_BOLD consistently improve performance metrics for Video-LLaMA and Video-LLaVA. SeViLA generally exhibits greater resilience to bias calibration, showing minimal effect on performance on NExT-QA and marginal deterioration on STAR. This may be attributed to the model's training for statistically close to uniform distribution of results over a fixed number of options. Despite decreases in Accuracy and F1 Mean on STAR, the reduction in *Recall_std* across all datasets for this model suggests that its debiased responses are more truthful, reflecting a fairer distribution of correct answers across options.

Video-LLaVA, which exhibits the most pronounced selection bias, benefits the most from bias reduction via prior decomposition across all datasets, achieving a maximum accuracy gain of 7.94% on STAR with the weighted configuration.

6 Conclusion

We performed the first comprehensive analysis on modified MCQA datasets to localize selection bias in specific options across VLMs of various architectures. Our findings reveal that while all models exhibit common behavioral patterns in their answer distributions, each model also has unique characteristics both across and within datasets. Despite these differences, the bias vector consistently manifests more prominently in the projections on the decomposed planes of video-question, video-answer, and question-answer interactions with one MCQA key component removed.

Using bias metrics to detect inconsistencies between the predicted and true probability distributions of the answers and calculating the prior bias, we proposed an effective method to mitigate this bias. Using a unique global prior for each dataset we reinforced the robustness and generalizability. Our BOLD bias removal approach not only enhances fairness, but also proves effective in improving performance metrics. Using weights in Weighted_BOLD further optimizes the performance.

This easily scalable, post-processing, automatic, and unsupervised method allows for a sharper focus on the true reasoning capabilities of the models. Our study contributes to making MCQA benchmarks more transparent and robust against the ordering of options, thereby delivering fairer and more meaningful evaluation results for VLMs.

Furthermore, elevating bias suppression to an abstract level, where tasks are systematically decomposed and key components are removed to identify where bias is most pronounced, can extend our approach to address other bias-related challenges in machine learning models.

7 Limitations

A primary limitation of our approach is the assumption that models rely on indirect cues when lacking sufficient direct information to determine an answer. Our conclusion—that bias can be decomposed across empty videos, questions, and answers—is based on empirical observations where bias is most pronounced in these settings. However, this decomposition may not be exhaustive; there could be more optimal directions in the latent space of decomposed priors. For instance, models often perform markedly different results compared to other settings when all answers are correct. A supervised debiasing method using true labels, such as applying COBYLA optimization directly, might also yield better results. Nonetheless, our goal was to find an easy-to-implement unsupervised solution without using correct answers.

We also assume the validity of the uniform distribution: Jensen-Shannon Divergence operates under the assumption that, in the absence of prior information, the distribution should be uniform. If the test MCQA data or task suggest an uneven distribution of answer options, this may lead to underestimating the significance of certain factors.

8 Ethical Statement

We based our experiments on existing datasets, ensuring compliance with their respective copyrights. Our research uses open-source VLMs, which, while offering accessibility and flexibility, also carry the inherent risks associated with open-ended text generation. Upon acceptance of the paper, we will release our code and dataset to the public for research and reproducibility.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *Preprint*, arXiv:2402.12483.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Huijuan Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An introduction to vision-language modeling. *ArXiv*, abs/2405.17247.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz

Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam,

Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2019. Action genome: Actions as compositions of spatio-temporal scene graphs. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10233–10244.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms? *ArXiv*, abs/2403.17752.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023b. Revisiting the role of language priors in vision-language models. In *International Conference on Machine Learning*.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281.

Adian Liusie, Yassir Fathullah, and Mark J. F. Gales. 2024. Teacher-student training for debiasing: General permutation debiasing for large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *ArXiv*, abs/2308.11483.

M. J. D. Powell. 1994. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*, pages 51–67.

Viorica Puatruaucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan S. Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yezhou Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexander Fréchette, Hanna Klimczak, R. Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *ArXiv*, abs/2305.13786.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. *Proceedings of the 2019 on International Conference on Multimedia Retrieval*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hao Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *ArXiv*, abs/2402.01349.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Annual Meeting of the Association for Computational Linguistics*.

Bo Wu and Shoubin Yu. 2021. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS Datasets and Benchmarks*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4344, Bangkok, Thailand. Association for Computational Linguistics.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. 2017. Temporal dynamic graph lstm for action-driven video object detection. In *ICCV*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. 2023b. Understanding unimodal bias in multimodal deep linear networks. In *International Conference on Machine Learning*.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tien-Ping Tan. 2024. Debiasing multimodal large language models. *ArXiv*, abs/2403.05262.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *ArXiv*, abs/2309.03882.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024b. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Yongshuo Zong, Yist Tingyang YU, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2024. Fool your large (vision and) language models with embarrassingly simple permutations.

A Datasets

A.1 NExT-QA and NExT-GQA

The NExT-QA dataset, released in 2021, comprises videos sourced from the VidOR dataset (Shang et al., 2019), with an average length of 35.73 seconds. The test split primarily contains causal questions (52.57%) addressing *why* and *how*, temporal reasoning questions (31.03%) about preceeding, following, or concurrent actions, and descriptive tasks (16.04%). Videos are manually annotated with questions and correct answers. Distractors are selected from 50 candidates based on a cosine similarity of less than 0.9 to the correct answer, ensuring they are close but not identical. This approach sometimes allows questions to be meaningfully answered without viewing the video (see Figures 6- 8).

Since our models were released after 2021, they might have included NExT-QA in their training or fine-tuning data.

NExT-GQA, released in 2024 as an extension of NExT-QA, focuses on evaluating the localization abilities of VLMs. It includes manually annotated timestamp labels that are key for answering the questions. Descriptive questions are excluded because they pertain to global video content or have answers available throughout the video with no specific temporal grounding. We also exclude QA pairs with multiple annotated time intervals. Consequently, the test set of NExT-GQA comprises 56.69% causal questions and 43.31% temporal questions.

Q: how did the boy in blue made music from the piano?

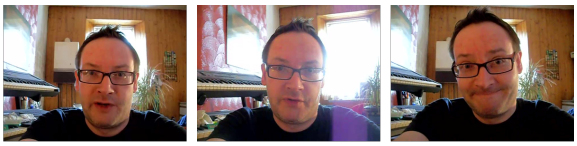
a0: music score sheet. | **a1: press piano keys.** | a2: spread his arms out. | a3: press the handles. | a4: drum on lap using hands.

Figure 6: Example of a QA pair from NExT-QA in the Causal category. The correct answer is in the box.

Q: did the woman from the driving seat get back in the car after dancing?

a0: yes. | a1: help to move the car. | a2: watching the car. | a3: talk to the audience. | a4: touch her feet.

Figure 7: Example of a QA pair from NExT-QA in the Temporal category. The correct answer is in the box.



Q: how many people are in the video?

a0: five. a1: three. **a2: one.** a3: ten. a4: nine.

Figure 8: Example of a QA pair from NExT-QA in the Descriptive category. The correct answer is in the box.

A.2 STAR

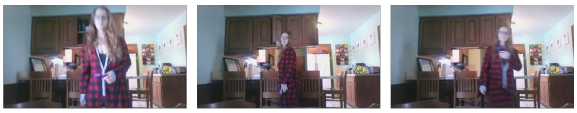
The STAR dataset is derived from the Action Genome dataset (Ji et al., 2019), which itself is based on Charades (Yuan et al., 2017). Charades comprises indoor activity videos with an average length of 30 seconds. Action Genome provides annotations of objects and their relationships within these videos.

Situation hypergraphs are used to extract abstract representations from the videos. Multiple types of questions are then designed using functional programs to cover different levels of difficulty in situated reasoning. There are four types of questions focusing on occurred facts, temporal order, future probability, and feasibility in specific situations:

- **Sequence** (50.53%): What did the person do before/after ...?"
- **Interaction** (33.78%): "What did a person do ...?"
- **Prediction** (8.79%): "What will the person do next with ...?"
- **Feasibility** (6.9%): "What is the person able to do?" or "Which object is possible to be ...?"

All QA pairs are annotated with a timestamp label. To avoid answer distribution bias, each question type is balanced through resampling. To prevent reasoning shortcuts due to frequent action or entity co-occurrences and to ensure models cannot easily answer without genuine understanding, the compositionality of verbs and nouns is controlled.

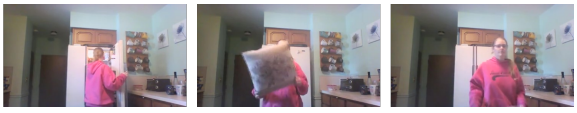
Examples are shown in Figures 9- 12.



Q: What happened after the person held the phone/camera?

a0: Took the picture. **a1: Closed the door.** a2: Put down the dish. a3: Took the cup/glass/bottle.

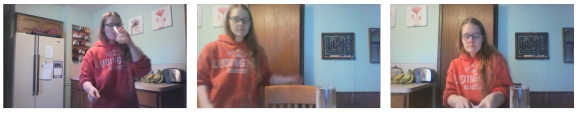
Figure 9: Example of a QA pair from STAR in the Sequence category. The correct answer is in the box.



Q: Which object was taken by the person?

a0: The box. **a1: The pillow.** a2: The blanket. a3: The bag.

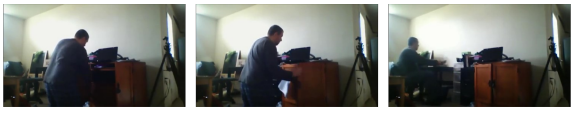
Figure 10: Example of a QA pair from STAR in the Interaction category. The correct answer is in the box.



Q: What will the person do next with the table?

a0: Wash. a1: Tidy up. **a2: Sit at.** a3: Take.

Figure 11: Example of a QA pair from STAR in the Prediction category. The correct answer is in the box.



Q: What is the person able to do after closing the door?

a0: Tidy up the closet/cabinet. a1: Hold the food. a2: Throw the bag. **a3: Put down the paper/notebook.**

Figure 12: Example of a QA pair from STAR in the Feasibility category. The correct answer is in the box.

A.3 Perception Test

The Perception Test dataset consists of four categories designed to evaluate VLMs on different types of reasoning:

- **Physics** (36.26%)
- **Semantics** (30.54%)
- **Abstraction** (29.28%)
- **Memory** (3.92%)

These categories test descriptive, explanatory, predictive, and counterfactual reasoning abilities.

The video scripts are based on human perception screening tests, with an average length of 23 seconds. All answers are crowd-sourced. Distractors are both manually created and automatically generated, particularly for standard closed yes/no questions.

The validation split we use contains approximately 15% closed questions, which poses an additional challenge for models that tend to exhibit affirmation bias (Pezeshkpour and Hruschka, 2023). Notably, for closed questions in this dataset, Video-LLaMA and Video-LLaVA often respond with "yes" or "no" rather than selecting an answer option.

Examples of QA pairs from the Perception Test are shown in Figures 13- 16.

Q: Was the hidden object moved while being occluded?

a0: no a1: I don't know a2: yes

Figure 13: Example of a QA pair from Perception Test in the Descriptive Physics category. The correct answer is in the box.

Q: What is the person preparing?

a0: a sandwich a1: a salad a2: eggs

Figure 14: Example of a QA pair from Perception Test in the Descriptive Semantics category. The correct answer is in the box.

Q: If the person had taken out one more object from the bag, how many objects would the person have taken out in total?

a0: 5 a1: 2 a2: 3

Figure 15: Example of a QA pair from Perception Test in the Counterfactual Abstraction category. The correct answer is in the box.

Q: What changed on the table while the camera was looking away?

a0: The stapler was removed. a1: Nothing changed visibly on the table while the camera was looking away. a2: A paper was added.

Figure 16: Example of a QA pair from Perception Test in the Explanatory Memory category. The correct answer is in the box.

A.4 Video-MME

Video-MME was developed to address the lack of diversity in video MCQA benchmarks. It consists of 900 YouTube videos ranging from 11 seconds to 1 hour in length. The dataset includes 12 task types covering questions on temporal and spatial perception, object and action reasoning, OCR, and counting problems. All QA pairs and distractors are manually created with three questions per video. Additionally, the dataset features videos in languages other than English.

Examples are shown in Figures 17- 20.

Q: What's the correct order of the following events?
1. Creating basic background. 2. Add a few flowers to the painting. 3. Building up the texture of the painting even more. 4. Drawing the pads of the water lilies.

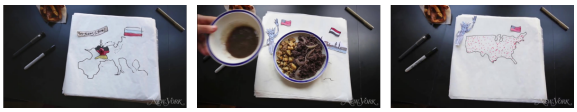
a0: 4, 1, 2, 3. a1: 2, 4, 1, 3. a2: 1, 4, 2, 3. a3: 1, 4, 3, 2.

Figure 17: Example of a QA pair from Video-MME in the Temporal Reasoning category. The correct answer is in the box.

Q: Which of the following is not a sponsor of the race?

a0: Lenovo. a1: BWT. a2: DEKRA. a3: IG.

Figure 18: Example of a QA pair from Video-MME in the OCR Problems category. The correct answer is in the box.



Q: How many national flags appear in the video?

a0: 3. a1: 5. a2: 2. **a3: 4.**

Figure 19: Example of a QA pair from Video-MME in the Counting Problem category. The correct answer is in the box.

Q: Where does the baby bird go at last?

a0: Flying away alone. a1: Staying with the fox. **a2: Flying with its mom to its herd.** a3: Going to find its nest.

Figure 20: Example of a QA pair from Video-MME in the Spatial Perception category. The correct answer is in the box.

B Models' Settings and Prompts

B.1 Video-LLaMA

Video-LLaMA is designed as a conversational model and was likely not specifically trained on MCQA, but rather on open-ended VQA datasets to generate captions, descriptions, and explanatory answers. Consequently, throughout our experiments with different prompt options, it tends to engage in dialogue: it repeats the question, responds with the general phrase "Hello, how can I help you today?", describes the video content, or even asks follow-up questions such as "What do you think about the video?" rather than providing a direct answer from a predefined list of options or option IDs.[4] Making it elicit a clear and concise answer option is quite challenging.

To address this issue, we experimented with various prompts, temperature settings, and token limits, and developed the following strategy:

System prompt:

```
Analyse the video, choose the
correct answer in multiple choice
question answering. Give the ID
of the correct answer option from
the given list of options. Be
concise.
```

User prompt:

```
<question> Options: <options>.
The ID of the correct answer:
```

We set the *number of beams* to 2 and the *temperature* to 0.8. Additionally, we allowed the model up to 30 attempts to produce an answer that includes an answer ID. We used the 7B parameter model due to resource constraints. For answering, the model selects eight uniformly sampled frames.

B.2 Video-LLaVA

Video-LLaVA effectively delivers concise answers as just a specific options based on a short, general prompt. However, in some cases, the model outputs as an answer option only "A" or responds with "yes" or "no" to a closed question.

For this model, we used the following prompt:

```
<video_tag> Question: <question>
Options: <options> The ID of
the correct answer is: Respond
with only the correct answer ID.
ASSISTANT:
```

Notably, on the completely unfamiliar datasets Perception Test and Video-MME adding an empty option increases the number of responses that cannot be parsed as an answer ID (although this increase is statistically insignificant). Combined with the observation that the model tends to disregard options presented later in the prompt, this suggests that the length of the question affects the model's performance, and tokens closer to the end of the prompt have a lower probability.

The model also selects eight frames from the video by uniform sampling. To conserve computational resources, we used Video-LLaVA-7B with half-precision set by Torch's autocast. The temperature is equal to 1.

B.3 SeViLA

SeViLA does not have any unanswered or unparsable responses because it was specifically tailored for the multiple-choice VQA task: internally, it maps any answer option IDs to {A, B, C, D, E}. However, it is limited to observing at most six answer options. In a side experiment, we shifted all answers to options G, H, I, J, K leaving the first six options empty, and obtained identical results to our other experimental setting, *Empty Answers*. This indicates that the model is completely blind to options beyond F and is likely overfitted to the options {A, B, C, D, E}.

[4] It is noteworthy that the model often uses the following template to explain its answer choice: "The ID of the correct answer: a_i. Explanation: . . . "

Furthermore, the shape of the confusion matrices for NExT-QA and STAR in Figures 21 and 22 shows an unusually accurate understanding of these datasets, unlike those for Perception Test and Video-MME. Since NExT-QA and STAR were released before SeViLA, the model might have been fine-tuned on them. On the other hand, this suggests that the positional bias in this model is much less pronounced.

Another vulnerability of SeViLA is that it uses *argmax* calculated on logit probabilities. However, logits without numerical stabilization can produce `NaN` values when activated via softmax. In such cases, according to *argmax*, the model's answer defaults to a_0. Among the settings relevant for debiasing, such behavior was observed in 22 cases on NExT-QA.

For this model, we used the original prompts: **Localizer prompt (when the model selects 32 frames):**

```
Does the information within the frame provide necessary details to accurately answer the given question?
```

Answerer prompt (when the model selects the answer based on 4 key frames):

```
Considering information in frames, select the correct answer from the options.
```

The confusion matrices for all models in the default (unmodified) setting in Figures 21–23 provide a quick insight into positional bias: darker blue cells indicate a greater accumulation of answers.

C Models' Performance on All Settings

Tables 4 through 15 present the distribution of selections across answer options and the accuracy (where applicable) for all settings, along with the target distributions. Notably, as soon as the correct answer aligns with the model's biased option, the performance boosts.

D Debiasing Algorithms and Results

The BOLD algorithm 1 describes estimating and adjusting for prior probabilities derived from decomposed inputs of ill-defined tasks. The Weighted_BOLD algorithm 2 is an extension of Algorithm 1, this method introduces weighting to the prior estimation to enhance debiasing by better capturing critical priors.

We fixed the random seed to 1 and conducted a series of experiments varying the weights and the sampling parameter k. All results including the default values are presented in Tables 16–19.

When considering only positive weights ($w_i \geq 0$), we assume that the debiasing vector is oriented correctly and use the weights to fine-tune the magnitude of its decomposed constituents. Therefore, positive weights generally enhance improvements and aggravate deteriorations. Specifically, in this case, all the primary debiasing metrics improve along with *Recall_std*, albeit at different rates across models and datasets.

When we optimize a single bias metric, i. e. *Recall_std*, by allowing negative weights, we search the latent subspace for the most optimal vector to improve that specific metric. However, this optimization may not correlate with an improvement in the overall bias, as it focuses solely on that metric. Consequently, using negative weights can lead to deterioration in other bias and performance metrics. This observation underscores the importance of evaluating bias using multiple metrics.

We provide the results for both cases, when weights are constrained to $0 \leq w_i \leq 1$ and when $|w_i| \leq 1$, as we believe that the observation of the direction of the bias vectors is crucial for understanding the overall debiasing effect.

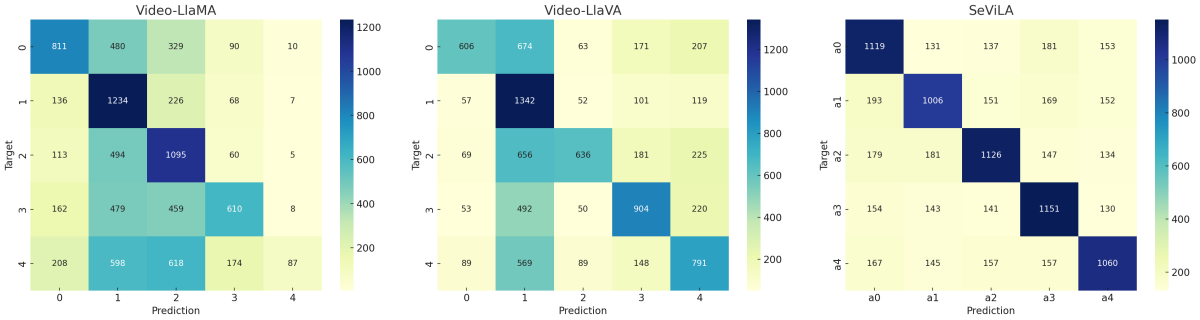


Figure 21: All Models' Confusion Matrices for NeXT-QA.

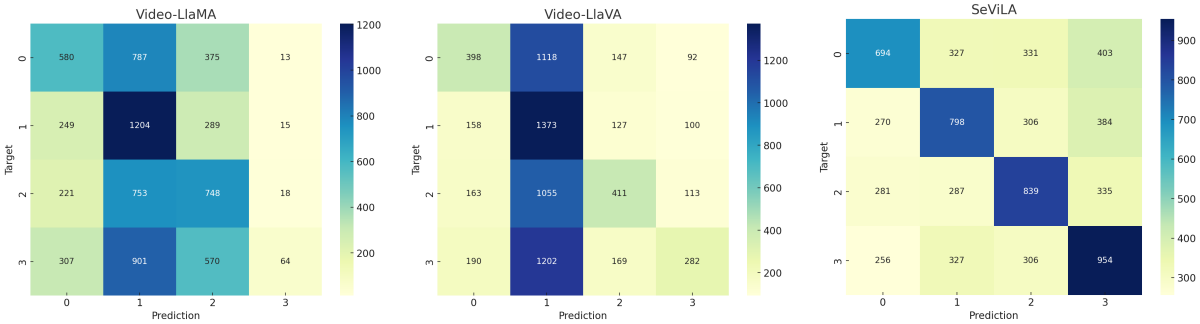


Figure 22: All Models' Confusion Matrices for STAR.

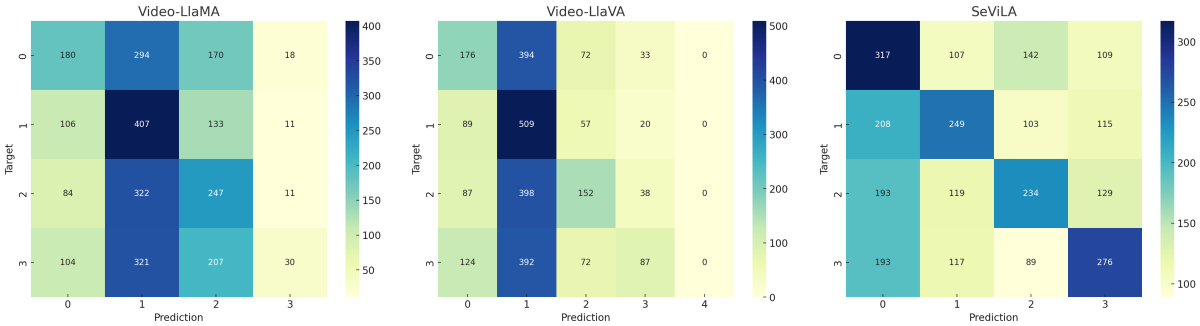


Figure 23: All Models' Confusion Matrices for Video-MME.

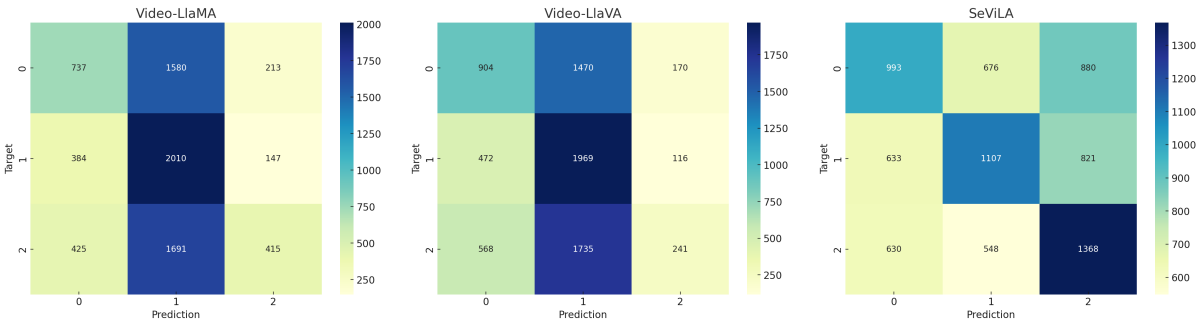


Figure 24: All Models' Confusion Matrices for Perception Test.

Table 4: Video-LLaMA Performance on All Settings of NExT-QA (including NExT-GQA). Each cell shows the count of selections for each option (*a0–a5*), along with the percentage in parentheses. The Accuracy column represents the number of correct answers where applicable. The N/A column indicates the number of cases where the model did not output an answer option after 30 attempts.

Setting	a0	a1	a2	a3	a4	a5	N/A	Accuracy
Target	1,721 (20.10%)	1,671 (19.51%)	1,767 (20.63%)	1,719 (20.07%)	1,686 (19.69%)	–	0 (0.00%)	–
Default	1,430 (16.70%)	3,285 (38.37%)	2,727 (31.85%)	1,002 (11.70%)	117 (1.37%)	–	3 (0.04%)	3,837 (44.82%)
Answer Shuffling	1,443 (16.97%)	3,295 (38.75%)	2,654 (31.21%)	993 (11.68%)	119 (1.40%)	–	60 (0.70%)	3,799 (44.67%)
Correct Frames	869 (17.61%)	1,836 (37.20%)	1,588 (32.18%)	587 (11.89%)	55 (1.11%)	–	27 (0.54%)	2,155 (43.67%)
Rephrased Questions	1,305 (15.36%)	3,270 (38.50%)	2,704 (31.83%)	1,062 (12.50%)	153 (1.80%)	–	70 (0.82%)	3,762 (44.29%)
Additional Empty Option	1,432 (17.17%)	2,987 (35.82%)	2,067 (24.78%)	1,090 (13.07%)	725 (8.69%)	39 (0.47%)	224 (2.62%)	4,086 (48.99%)
All a1	3,537 (42.71%)	3,891 (46.98%)	737 (8.90%)	32 (0.39%)	85 (1.03%)	–	282 (3.29%)	–
All a2	3,614 (43.82%)	3,780 (45.83%)	748 (9.07%)	32 (0.39%)	73 (0.89%)	–	317 (3.70%)	–
All a3	3,585 (43.43%)	3,831 (46.41%)	722 (8.75%)	33 (0.40%)	83 (1.01%)	–	310 (3.62%)	–
All a4	3,590 (43.35%)	3,837 (46.33%)	733 (8.85%)	34 (0.41%)	87 (1.05%)	–	283 (3.30%)	–
Correct a0	4,136 (48.98%)	2,303 (27.27%)	1,489 (17.63%)	453 (5.36%)	64 (0.76%)	–	119 (1.39%)	4,136 (48.98%)
Correct a0 Shuffled	4,136 (49.00%)	2,304 (27.30%)	1,525 (18.07%)	415 (4.92%)	61 (0.72%)	–	123 (1.44%)	4,136 (49.00%)
Correct a1	730 (8.61%)	6,323 (74.61%)	1,102 (13.00%)	292 (3.45%)	28 (0.33%)	–	89 (1.04%)	6,323 (74.61%)
Correct a1 Shuffled	704 (8.30%)	6,348 (74.88%)	1,098 (12.95%)	297 (3.50%)	31 (0.37%)	–	86 (1.00%)	6,348 (74.88%)
Correct a2	622 (7.32%)	2,325 (27.35%)	5,240 (61.64%)	280 (3.29%)	34 (0.40%)	–	63 (0.74%)	5,240 (61.64%)
Correct a2 Shuffled	583 (6.88%)	2,356 (27.81%)	5,240 (61.84%)	254 (3.00%)	40 (0.47%)	–	91 (1.06%)	5,240 (61.84%)
Correct a3	762 (8.99%)	2,315 (27.32%)	2,321 (27.39%)	3,029 (35.74%)	47 (0.55%)	–	90 (1.05%)	3,029 (35.74%)
Correct a3 Shuffled	766 (9.07%)	2,335 (27.65%)	2,235 (26.47%)	3,049 (36.11%)	59 (0.70%)	–	120 (1.40%)	3,049 (36.11%)
Correct a4	1,194 (14.11%)	3,014 (35.63%)	2,857 (33.77%)	909 (10.74%)	486 (5.74%)	–	104 (1.21%)	486 (5.74%)
Correct a4 Shuffled	1,164 (13.77%)	2,912 (34.45%)	2,949 (34.88%)	912 (10.79%)	517 (6.12%)	–	110 (1.28%)	517 (6.12%)
All Correct Answers	5,247 (63.17%)	2,730 (32.87%)	295 (3.55%)	9 (0.11%)	25 (0.30%)	–	258 (3.01%)	–
Empty Frames	1,092 (12.77%)	3,494 (40.87%)	3,344 (39.11%)	544 (6.36%)	76 (0.89%)	–	14 (0.16%)	2,620 (30.64%)
Empty Questions	1,686 (20.31%)	3,975 (47.89%)	1,676 (20.19%)	786 (9.47%)	178 (2.14%)	–	263 (3.07%)	3,082 (37.13%)
Empty Answers	987 (11.66%)	4,543 (53.68%)	2,837 (33.52%)	37 (0.44%)	59 (0.70%)	–	101 (1.18%)	–

Table 5: Video-LLaMA Performance on All Settings of STAR. Each cell shows the count of selections for each option (*a0–a4*), along with the percentage in parentheses. The Accuracy column represents the number of correct answers where applicable. The N/A column indicates the number of cases where the model did not output an answer option after 30 attempts.

Setting	a0	a1	a2	a3	a4	N/A	Accuracy
Target	1,755 (24.73%)	1,758 (24.77%)	1,742 (24.54%)	1,843 (25.97%)	–	0 (0.00%)	–
Default	1,357 (19.13%)	3,645 (51.38%)	1,982 (27.94%)	110 (1.55%)	–	4 (0.06%)	2,596 (36.59%)
Answer Shuffling	1,255 (18.72%)	3,458 (51.57%)	1,893 (28.23%)	99 (1.48%)	–	393 (5.54%)	2,467 (36.79%)
Correct Frames	1,322 (19.43%)	3,455 (50.79%)	1,899 (27.91%)	127 (1.87%)	–	295 (4.16%)	2,524 (37.10%)
Rephrased Questions	1,376 (20.03%)	3,462 (50.39%)	1,891 (27.53%)	141 (2.05%)	–	228 (3.21%)	2,520 (36.68%)
Additional Empty Option	1,314 (20.59%)	2,892 (45.31%)	1,494 (23.41%)	630 (9.87%)	52 (0.81%)	716 (10.09%)	2,528 (39.61%)
All a0	2,484 (38.04%)	3,949 (60.47%)	88 (1.35%)	9 (0.14%)	–	568 (8.00%)	–
All a1	2,695 (39.68%)	3,976 (58.54%)	104 (1.53%)	17 (0.25%)	–	306 (4.31%)	–
All a2	2,711 (39.93%)	3,957 (58.28%)	113 (1.66%)	9 (0.13%)	–	308 (4.34%)	–
All a3	2,512 (38.52%)	3,909 (59.94%)	94 (1.44%)	7 (0.11%)	–	576 (8.11%)	–
Correct a0	2,340 (34.53%)	2,996 (44.21%)	1,378 (20.33%)	63 (0.93%)	–	321 (4.52%)	2,340 (34.53%)
Correct a0 Shuffled	2,313 (34.36%)	2,960 (43.97%)	1,392 (20.68%)	67 (1.00%)	–	366 (5.16%)	2,313 (34.36%)
Correct a1	966 (14.16%)	4,677 (68.54%)	1,124 (16.47%)	57 (0.84%)	–	274 (3.86%)	4,677 (68.54%)
Correct a1 Shuffled	985 (14.48%)	4,622 (67.93%)	1,136 (16.70%)	61 (0.90%)	–	294 (4.14%)	4,622 (67.93%)
Correct a2	918 (13.53%)	2,913 (42.93%)	2,898 (42.71%)	56 (0.83%)	–	313 (4.41%)	2,898 (42.71%)
Correct a2 Shuffled	844 (12.57%)	2,908 (43.32%)	2,901 (43.21%)	60 (0.89%)	–	385 (5.42%)	2,901 (43.21%)
Correct a3	1,226 (18.03%)	3,298 (48.50%)	2,002 (29.44%)	274 (4.03%)	–	298 (4.20%)	274 (4.03%)
Correct a3 Shuffled	1,153 (17.21%)	3,190 (47.60%)	2,127 (31.74%)	231 (3.45%)	–	397 (5.59%)	231 (3.45%)
All Correct Answers	3,129 (47.60%)	3,376 (51.35%)	63 (0.96%)	6 (0.09%)	–	524 (7.38%)	–
Empty Frames	1,323 (18.64%)	3,721 (52.42%)	2,047 (28.84%)	7 (0.10%)	–	0 (0.00%)	1,854 (26.12%)
Empty Questions	2,365 (34.84%)	2,973 (43.80%)	1,275 (18.78%)	175 (2.58%)	–	310 (4.37%)	2,427 (35.75%)
Empty Answers	1,121 (15.84%)	3,093 (43.72%)	2,778 (39.27%)	83 (1.17%)	–	23 (0.32%)	–

Table 6: Video-LLaMA Performance on All Settings of Perception Test. Each cell shows the count of selections for each option (*a0–a3*), along with the percentage in parentheses. The Accuracy column represents the number of correct answers where applicable. The N/A column indicates the number of cases where the model did not output an answer option after 30 attempts.

Setting	a0	a1	a2	a3	N/A	Accuracy
Target	2,549 (33.29%)	2,561 (33.45%)	2,546 (33.25%)	–	0 (0.00%)	–
Default	1,546 (20.34%)	5,281 (69.47%)	775 (10.19%)	–	54 (0.71%)	3,162 (41.59%)
Answer Shuffling	1,593 (21.10%)	5,165 (68.42%)	791 (10.48%)	–	107 (1.40%)	3,146 (41.67%)
Rephrased Questions	1,652 (21.83%)	5,004 (66.13%)	911 (12.04%)	–	89 (1.16%)	3,161 (41.77%)
Additional Empty Option	1,448 (19.56%)	4,356 (58.86%)	1,549 (20.93%)	48 (0.65%)	255 (3.33%)	3,004 (40.59%)
All a0	2,768 (36.60%)	4,414 (58.36%)	381 (5.04%)	–	93 (1.21%)	–
All a1	2,842 (37.56%)	4,334 (57.28%)	390 (5.15%)	–	90 (1.18%)	–
All a2	2,806 (37.15%)	4,326 (57.28%)	421 (5.57%)	–	103 (1.35%)	–
Correct a0	2,245 (29.76%)	4,675 (61.98%)	623 (8.26%)	–	113 (1.48%)	2,245 (29.76%)
Correct a0 Shuffled	2,271 (30.12%)	4,654 (61.73%)	614 (8.14%)	–	117 (1.53%)	2,271 (30.12%)
Correct a1	1,204 (15.95%)	5,833 (77.29%)	510 (6.76%)	–	109 (1.42%)	5,833 (77.29%)
Correct a1 Shuffled	1,233 (16.32%)	5,820 (77.06%)	500 (6.62%)	–	103 (1.35%)	5,820 (77.06%)
Correct a2	1,239 (16.40%)	5,077 (67.18%)	1,241 (16.42%)	–	99 (1.29%)	1,241 (16.42%)
Correct a2 Shuffled	1,235 (16.36%)	5,082 (67.30%)	1,234 (16.34%)	–	105 (1.37%)	1,234 (16.34%)
All Correct Answers	3,330 (43.87%)	4,053 (53.40%)	207 (2.73%)	–	66 (0.86%)	–
Empty Frames	990 (12.93%)	5,641 (73.68%)	1,025 (13.39%)	–	0 (0.00%)	3,005 (39.25%)
Empty Questions	2,564 (34.61%)	3,707 (50.03%)	1,138 (15.36%)	–	247 (3.23%)	3,021 (40.77%)
Empty Answers	2,969 (39.11%)	2,876 (37.88%)	1,747 (23.01%)	–	64 (0.84%)	–

Table 7: Video-LLaMA Performance on All Settings of Video-MME. Each cell shows the count of selections for each option (*a0–a4*), along with the percentage in parentheses. The Accuracy column represents the number of correct answers where applicable. The N/A column indicates the number of cases where the model did not output an answer option after 30 attempts.

Setting	a0	a1	a2	a3	a4	N/A	Accuracy
Target	675 (25.00%)	675 (25.00%)	675 (25.00%)	675 (25.00%)	–	0 (0.00%)	–
Default	474 (17.92%)	1,344 (50.81%)	757 (28.62%)	70 (2.65%)	–	55 (2.04%)	864 (32.67%)
Answer Shuffling	451 (18.11%)	1,242 (49.86%)	733 (29.43%)	65 (2.61%)	–	209 (7.74%)	785 (31.51%)
Rephrased Questions	1,367 (50.65%)	979 (36.27%)	198 (7.34%)	155 (5.74%)	–	1 (0.04%)	873 (32.35%)
Additional Empty Option	391 (17.29%)	1,040 (45.98%)	624 (27.59%)	178 (7.87%)	29 (1.28%)	438 (16.22%)	746 (32.98%)
All a0	1,042 (42.81%)	1,244 (51.11%)	120 (4.93%)	28 (1.15%)	–	266 (9.85%)	–
All a1	1,027 (42.00%)	1,295 (52.97%)	91 (3.72%)	32 (1.31%)	–	255 (9.44%)	–
All a2	1,061 (43.36%)	1,232 (50.35%)	116 (4.74%)	38 (1.55%)	–	253 (9.37%)	–
All a3	1,001 (41.06%)	1,223 (50.16%)	149 (6.11%)	65 (2.67%)	–	262 (9.70%)	–
Correct a0	678 (27.19%)	1,141 (45.75%)	624 (25.02%)	51 (2.04%)	–	206 (7.63%)	678 (27.19%)
Correct a0 Shuffled	699 (27.90%)	1,085 (43.31%)	664 (26.51%)	57 (2.28%)	–	195 (7.22%)	699 (27.90%)
Correct a1	376 (14.98%)	1,532 (61.04%)	547 (21.79%)	55 (2.19%)	–	190 (7.04%)	1,532 (61.04%)
Correct a1 Shuffled	344 (13.78%)	1,519 (60.83%)	584 (23.39%)	50 (2.00%)	–	203 (7.52%)	1,519 (60.83%)
Correct a2	366 (14.63%)	1,147 (45.86%)	948 (37.90%)	40 (1.60%)	–	199 (7.37%)	948 (37.90%)
Correct a2 Shuffled	355 (14.27%)	1,144 (45.98%)	942 (37.86%)	47 (1.89%)	–	212 (7.85%)	942 (37.86%)
Correct a3	408 (16.47%)	1,178 (47.56%)	777 (31.37%)	114 (4.60%)	–	223 (8.26%)	114 (4.60%)
Correct a3 Shuffled	386 (15.55%)	1,199 (48.31%)	786 (31.67%)	111 (4.47%)	–	218 (8.07%)	111 (4.47%)
All Correct Answers	1,195 (46.46%)	1,226 (47.67%)	114 (4.43%)	37 (1.44%)	–	128 (4.74%)	–
Empty Frames	188 (6.98%)	1,612 (59.81%)	843 (31.28%)	52 (1.93%)	–	5 (0.19%)	776 (28.79%)
Empty Questions	753 (29.89%)	1,180 (46.84%)	527 (20.92%)	59 (2.34%)	–	181 (6.70%)	798 (31.68%)
Empty Answers	440 (17.07%)	725 (28.13%)	1,352 (52.46%)	60 (2.33%)	–	123 (4.56%)	–

Table 8: Video-LLaVA Performance on All the Settings of NExT-QA (including NExT-GQA). Each cell shows the count of selections for each option (*a0–a5*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	a5	N/A	Accuracy
Target	1,721 (20.10%)	1,671 (19.51%)	1,767 (20.63%)	1,719 (20.07%)	1,686 (19.69%)	–	0 (0.00%)	–
Default	874 (10.21%)	3,733 (43.59%)	890 (10.39%)	1,505 (17.57%)	1,562 (18.24%)	–	0 (0.00%)	4,279 (49.96%)
Answer Shuffling	1,507 (17.60%)	4,366 (50.98%)	1,630 (19.03%)	1,061 (12.39%)	0 (0.00%)	–	0 (0.00%)	3,653 (42.66%)
Correct Frames	440 (8.87%)	2,194 (44.22%)	482 (9.71%)	791 (15.94%)	1,055 (21.26%)	–	0 (0.00%)	2,407 (48.51%)
Rephrased Questions	836 (9.79%)	3,531 (41.36%)	1,054 (12.34%)	1,613 (18.89%)	1,504 (17.62%)	–	26 (0.30%)	4,287 (50.21%)
Additional Empty Option	418 (4.88%)	3,449 (40.27%)	847 (9.89%)	1,208 (14.11%)	2,641 (30.84%)	1 (0.01%)	0 (0.00%)	4,264 (49.79%)
All a0	1,232 (14.39%)	5,839 (68.18%)	9 (0.11%)	175 (2.04%)	1,309 (15.28%)	–	0 (0.00%)	–
All a1	1,254 (14.64%)	5,839 (68.18%)	9 (0.11%)	187 (2.18%)	1,275 (14.89%)	–	0 (0.00%)	–
All a2	1,238 (14.46%)	5,755 (67.21%)	12 (0.14%)	158 (1.85%)	1,400 (16.35%)	–	1 (0.01%)	–
All a3	966 (11.28%)	7,063 (82.47%)	14 (0.16%)	178 (2.08%)	343 (4.01%)	–	0 (0.00%)	–
All a4	996 (11.63%)	7,021 (81.98%)	20 (0.23%)	176 (2.06%)	351 (4.10%)	–	0 (0.00%)	–
Correct a0	2,966 (34.87%)	3,054 (35.91%)	484 (5.69%)	975 (11.46%)	1,026 (12.06%)	–	59 (0.69%)	2,966 (34.87%)
Correct a0 Shuffled	2,702 (31.55%)	3,680 (42.97%)	341 (3.98%)	759 (8.86%)	1,082 (12.63%)	–	0 (0.00%)	2,702 (31.55%)
Correct a1	251 (2.93%)	6,993 (81.66%)	209 (2.44%)	431 (5.03%)	680 (7.94%)	–	0 (0.00%)	6,993 (81.66%)
Correct a1 Shuffled	270 (3.15%)	6,981 (81.52%)	203 (2.37%)	424 (4.95%)	686 (8.01%)	–	0 (0.00%)	6,981 (81.52%)
Correct a2	274 (3.20%)	3,619 (42.26%)	3,032 (35.41%)	669 (7.81%)	969 (11.32%)	–	1 (0.01%)	3,032 (35.41%)
Correct a2 Shuffled	282 (3.29%)	3,561 (41.58%)	3,056 (35.68%)	707 (8.26%)	958 (11.19%)	–	0 (0.00%)	3,056 (35.68%)
Correct a3	258 (3.01%)	2,765 (32.29%)	319 (3.73%)	4,161 (48.59%)	1,060 (12.38%)	–	1 (0.01%)	4,161 (48.59%)
Correct a3 Shuffled	278 (3.25%)	2,746 (32.06%)	271 (3.16%)	4,182 (48.83%)	1,087 (12.69%)	–	0 (0.00%)	4,182 (48.83%)
Correct a4	324 (3.78%)	2,374 (27.72%)	300 (3.50%)	557 (6.50%)	5,009 (58.49%)	–	0 (0.00%)	5,009 (58.49%)
Correct a4 Shuffled	336 (3.92%)	2,305 (26.91%)	313 (3.65%)	579 (6.76%)	5,031 (58.75%)	–	0 (0.00%)	5,031 (58.75%)
All Correct Answers	1,536 (17.94%)	6,531 (76.27%)	4 (0.05%)	59 (0.69%)	433 (5.06%)	–	1 (0.01%)	–
Empty Frames	634 (7.40%)	4,625 (54.01%)	740 (8.64%)	1,499 (17.50%)	1,066 (12.45%)	–	0 (0.00%)	3,025 (35.32%)
Empty Questions	1,976 (23.07%)	4,925 (57.51%)	228 (2.66%)	696 (8.13%)	739 (8.63%)	–	0 (0.00%)	2,979 (34.79%)
Empty Answers	45 (0.53%)	7,590 (88.63%)	12 (0.14%)	484 (5.65%)	433 (5.06%)	–	0 (0.00%)	–

Table 9: Video-LLaVA Performance on All the Settings of STAR. Each cell shows the count of selections for each option (*a0–a4*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	N/A	Accuracy
Target	1,755 (24.73%)	1,758 (24.77%)	1,742 (24.54%)	1,843 (25.97%)	–	0 (0.00%)	–
Default	909 (12.81%)	4,748 (66.89%)	854 (12.03%)	587 (8.27%)	–	0 (0.00%)	2,464 (34.71%)
Answer Shuffling	825 (11.62%)	4,688 (66.05%)	675 (9.51%)	910 (12.82%)	–	0 (0.00%)	2,521 (35.52%)
Correct Frames	1,115 (15.71%)	4,611 (64.96%)	743 (10.47%)	629 (8.86%)	–	0 (0.00%)	1,881 (26.50%)
Rephrased Questions	759 (10.70%)	4,945 (69.70%)	799 (11.26%)	592 (8.34%)	–	3 (0.04%)	2,455 (34.60%)
Additional Empty Option	346 (4.87%)	4,472 (63.00%)	545 (7.68%)	1,541 (21.71%)	194 (2.73%)	0 (0.00%)	2,490 (35.08%)
All a0	499 (7.03%)	6,596 (92.93%)	1 (0.01%)	2 (0.03%)	–	0 (0.00%)	–
All a1	525 (7.40%)	6,573 (92.60%)	0 (0.00%)	0 (0.00%)	–	0 (0.00%)	–
All a2	523 (7.37%)	6,571 (92.58%)	1 (0.01%)	3 (0.04%)	–	0 (0.00%)	–
All a3	485 (6.83%)	6,610 (93.12%)	2 (0.03%)	1 (0.01%)	–	0 (0.00%)	–
Correct a0	1,438 (20.26%)	4,598 (64.78%)	459 (6.47%)	603 (8.50%)	–	0 (0.00%)	1,438 (20.26%)
Correct a0 Shuffled	1,424 (20.06%)	4,578 (64.51%)	453 (6.38%)	642 (9.05%)	–	1 (0.01%)	1,424 (20.06%)
Correct a1	533 (7.51%)	5,621 (79.19%)	426 (6.00%)	518 (7.30%)	–	0 (0.00%)	5,621 (79.19%)
Correct a1 Shuffled	566 (7.98%)	5,560 (78.34%)	449 (6.33%)	522 (7.36%)	–	1 (0.01%)	5,560 (78.34%)
Correct a2	1,055 (14.86%)	4,561 (64.26%)	849 (11.96%)	633 (8.92%)	–	0 (0.00%)	849 (11.96%)
Correct a2 Shuffled	586 (8.26%)	4,486 (63.20%)	1,381 (19.46%)	645 (9.09%)	–	0 (0.00%)	1,381 (19.46%)
Correct a3	588 (8.28%)	4,387 (61.81%)	458 (6.45%)	1,665 (23.46%)	–	0 (0.00%)	1,665 (23.46%)
Correct a3 Shuffled	1,061 (14.95%)	4,503 (63.44%)	741 (10.44%)	793 (11.17%)	–	0 (0.00%)	793 (11.17%)
All Correct Answers	374 (5.27%)	6,710 (94.53%)	1 (0.01%)	13 (0.18%)	–	0 (0.00%)	–
Empty Frames	709 (9.99%)	4,772 (67.23%)	1,048 (14.76%)	569 (8.02%)	–	0 (0.00%)	1,891 (26.64%)
Empty Questions	3,013 (42.45%)	3,748 (52.80%)	194 (2.73%)	143 (2.01%)	–	0 (0.00%)	2,119 (29.85%)
Empty Answers	63 (0.89%)	7,033 (99.08%)	0 (0.00%)	2 (0.03%)	–	0 (0.00%)	–

Table 10: Video-LLaVA Performance on All the Settings of Perception Test. Each cell shows the count of selections for each option (*a0–a3*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	N/A	Accuracy
Target	2,549 (33.29%)	2,561 (33.45%)	2,546 (33.25%)	–	0 (0.00%)	–
Default	1,944 (25.43%)	5,174 (67.68%)	527 (6.89%)	–	11 (0.14%)	3,114 (40.73%)
Answer Shuffling	2,164 (28.28%)	4,887 (63.86%)	602 (7.87%)	–	3 (0.04%)	3,259 (42.58%)
Rephrased Questions	1,836 (24.04%)	5,253 (68.79%)	547 (7.16%)	–	20 (0.26%)	3,154 (41.30%)
Additional Empty Option	1,133 (15.01%)	4,548 (60.24%)	1,634 (21.64%)	235 (3.11%)	106 (1.38%)	3,233 (42.82%)
All a0	2,942 (38.43%)	4,586 (59.90%)	128 (1.67%)	–	0 (0.00%)	–
All a1	3,001 (39.20%)	4,538 (59.27%)	117 (1.53%)	–	0 (0.00%)	–
All a2	3,684 (48.13%)	3,951 (51.61%)	20 (0.26%)	–	1 (0.01%)	–
Correct a0	3,086 (40.32%)	4,040 (52.79%)	527 (6.89%)	–	3 (0.04%)	3,086 (40.32%)
Correct a0 Shuffled	3,121 (40.78%)	4,038 (52.76%)	494 (6.45%)	–	3 (0.04%)	3,121 (40.78%)
Correct a1	1,524 (19.91%)	5,712 (74.62%)	419 (5.47%)	–	1 (0.01%)	5,712 (74.62%)
Correct a1 Shuffled	1,542 (20.15%)	5,712 (74.63%)	400 (5.23%)	–	2 (0.03%)	5,712 (74.63%)
Correct a2	3,085 (40.32%)	4,040 (52.80%)	527 (6.89%)	–	4 (0.05%)	527 (6.89%)
Correct a2 Shuffled	1,839 (24.03%)	4,944 (64.60%)	870 (11.37%)	–	3 (0.04%)	870 (11.37%)
All Correct Answers	3,493 (45.64%)	4,078 (53.28%)	83 (1.08%)	–	2 (0.03%)	–
Empty Frames	1,920 (25.08%)	5,265 (68.77%)	471 (6.15%)	–	0 (0.00%)	2,957 (38.62%)
Empty Questions	3,649 (47.66%)	3,767 (49.20%)	240 (3.13%)	–	0 (0.00%)	3,007 (39.28%)
Empty Answers	1,183 (15.45%)	6,284 (82.09%)	188 (2.46%)	–	1 (0.01%)	–

Table 11: Video-LLaVA Performance on All the Settings of Video-MME. Each cell shows the count of selections for each option (*a0–a4*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	N/A	Accuracy
Target	675 (25.00%)	675 (25.00%)	675 (25.00%)	675 (25.00%)	–	0 (0.00%)	–
Default	476 (17.63%)	1,693 (62.70%)	353 (13.07%)	178 (6.59%)	–	0 (0.00%)	924 (34.22%)
Answer Shuffling	427 (15.81%)	1,728 (64.00%)	373 (13.81%)	172 (6.37%)	–	0 (0.00%)	841 (31.15%)
Rephrased Questions	478 (17.70%)	1,705 (63.15%)	352 (13.04%)	165 (6.11%)	–	0 (0.00%)	911 (33.74%)
Additional Empty Option	208 (7.84%)	1,833 (69.12%)	191 (7.20%)	420 (15.84%)	0 (0.00%)	48 (1.78%)	909 (34.28%)
All a0	891 (33.00%)	1,730 (64.07%)	27 (1.00%)	52 (1.93%)	–	0 (0.00%)	–
All a1	888 (32.90%)	1,733 (64.21%)	33 (1.22%)	45 (1.67%)	–	1 (0.04%)	–
All a2	914 (33.86%)	1,703 (63.10%)	27 (1.00%)	55 (2.04%)	–	1 (0.04%)	–
All a3	888 (32.90%)	1,657 (61.39%)	32 (1.19%)	122 (4.52%)	–	1 (0.04%)	–
Correct a0	660 (24.44%)	1,651 (61.15%)	258 (9.56%)	131 (4.85%)	–	0 (0.00%)	660 (24.44%)
Correct a0 Shuffled	642 (23.78%)	1,620 (60.00%)	293 (10.85%)	145 (5.37%)	–	0 (0.00%)	642 (23.78%)
Correct a1	384 (14.22%)	1,974 (73.11%)	241 (8.93%)	101 (3.74%)	–	0 (0.00%)	1,974 (73.11%)
Correct a1 Shuffled	364 (13.48%)	1,964 (72.74%)	240 (8.89%)	132 (4.89%)	–	0 (0.00%)	1,964 (72.74%)
Correct a2	386 (14.30%)	1,606 (59.48%)	566 (20.96%)	142 (5.26%)	–	0 (0.00%)	566 (20.96%)
Correct a2 Shuffled	352 (13.04%)	1,623 (60.11%)	589 (21.81%)	136 (5.04%)	–	0 (0.00%)	589 (21.81%)
Correct a3	416 (15.41%)	1,642 (60.81%)	311 (11.52%)	331 (12.26%)	–	0 (0.00%)	331 (12.26%)
Correct a3 Shuffled	357 (13.23%)	1,682 (62.32%)	345 (12.78%)	315 (11.67%)	–	1 (0.04%)	315 (11.67%)
All Correct Answers	928 (34.38%)	1,702 (63.06%)	29 (1.07%)	40 (1.48%)	–	1 (0.04%)	–
Empty Frames	267 (9.89%)	1,828 (67.70%)	437 (16.19%)	168 (6.22%)	–	0 (0.00%)	764 (28.30%)
Empty Questions	1,545 (57.24%)	960 (35.57%)	119 (4.41%)	75 (2.78%)	–	1 (0.04%)	790 (29.27%)
Empty Answers	359 (13.32%)	2,082 (77.23%)	165 (6.12%)	90 (3.34%)	–	4 (0.15%)	–

Table 12: SeViLA Performance on All the Settings of NExT-QA (including NExT-GQA). Each cell shows the count of selections for each option (*a0–a5*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	a5	Accuracy
Target	1,721 (20.10%)	1,671 (19.51%)	1,767 (20.63%)	1,719 (20.07%)	1,686 (19.69%)	–	–
Default	1,812 (21.16%)	1,606 (18.75%)	1,712 (19.99%)	1,805 (21.08%)	1,629 (19.02%)	–	5,462 (63.78%)
Answer Shuffling	1,740 (20.32%)	1,598 (18.66%)	1,737 (20.28%)	1,824 (21.30%)	1,665 (19.44%)	–	5,481 (64.00%)
Correct Frames	1,108 (22.33%)	939 (18.92%)	961 (19.37%)	1,069 (21.54%)	885 (17.84%)	–	2,932 (59.09%)
Rephrased Questions	1,825 (21.31%)	1,571 (18.34%)	1,657 (19.35%)	1,793 (20.94%)	1,718 (20.06%)	–	5,160 (60.25%)
Additional Empty Option	1,669 (19.49%)	1,401 (16.36%)	1,457 (17.01%)	1,563 (18.25%)	1,758 (20.53%)	716 (8.36%)	5,054 (59.01%)
All a0	4,972 (58.06%)	381 (4.45%)	631 (7.37%)	1,865 (21.78%)	715 (8.35%)	–	–
All a1	4,906 (57.29%)	381 (4.45%)	675 (7.88%)	1,954 (22.82%)	648 (7.57%)	–	–
All a2	4,887 (57.06%)	359 (4.19%)	681 (7.95%)	1,968 (22.98%)	669 (7.81%)	–	–
All a3	4,961 (57.93%)	337 (3.94%)	601 (7.02%)	1,981 (23.13%)	684 (7.99%)	–	–
All a4	4,904 (57.26%)	370 (4.32%)	615 (7.18%)	1,987 (23.20%)	688 (8.03%)	–	–
Correct a0	5,434 (63.45%)	732 (8.55%)	761 (8.89%)	844 (9.86%)	793 (9.26%)	–	5,434 (63.45%)
Correct a0 Shuffled	5,443 (63.56%)	736 (8.59%)	745 (8.70%)	814 (9.50%)	826 (9.65%)	–	5,443 (63.56%)
Correct a1	870 (10.16%)	5,399 (63.04%)	752 (8.78%)	825 (9.63%)	718 (8.38%)	–	5,399 (63.04%)
Correct a1 Shuffled	809 (9.45%)	5,379 (62.81%)	740 (8.64%)	866 (10.11%)	770 (8.99%)	–	5,379 (62.81%)
Correct a2	813 (9.49%)	741 (8.65%)	5,495 (64.16%)	772 (9.01%)	743 (8.68%)	–	5,495 (64.16%)
Correct a2 Shuffled	777 (9.07%)	734 (8.57%)	5,504 (64.27%)	796 (9.29%)	753 (8.79%)	–	5,504 (64.27%)
Correct a3	827 (9.66%)	704 (8.22%)	778 (9.08%)	5,591 (65.28%)	664 (7.75%)	–	5,591 (65.28%)
Correct a3 Shuffled	782 (9.13%)	684 (7.99%)	783 (9.14%)	5,566 (64.99%)	749 (8.75%)	–	5,566 (64.99%)
Correct a4	860 (10.04%)	730 (8.52%)	708 (8.27%)	853 (9.96%)	5,413 (63.21%)	–	5,413 (63.21%)
Correct a4 Shuffled	810 (9.46%)	727 (8.49%)	760 (8.87%)	831 (9.70%)	5,436 (63.48%)	–	5,436 (63.48%)
All Correct Answers	7,300 (85.24%)	209 (2.44%)	237 (2.77%)	448 (5.23%)	370 (4.32%)	–	–
Empty Frames	1,833 (21.40%)	1,495 (17.46%)	1,626 (18.99%)	1,820 (21.25%)	1,790 (20.90%)	–	3,921 (45.78%)
Empty Questions	1,671 (19.51%)	1,513 (17.67%)	1,685 (19.68%)	1,851 (21.61%)	1,844 (21.53%)	–	4,485 (52.37%)
Empty Answers	2,026 (23.66%)	52 (0.61%)	5 (0.06%)	0 (0.00%)	6,481 (75.68%)	–	–

Table 13: SeViLA Performance on All the Settings of STAR. Each cell shows the count of selections for each option (*a0–a4*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	Accuracy
Target	1,755 (24.73%)	1,758 (24.77%)	1,742 (24.54%)	1,843 (25.97%)	–	–
Default	1,501 (21.15%)	1,739 (24.50%)	1,782 (25.11%)	2,076 (29.25%)	–	3,285 (46.28%)
Answer Shuffling	1,473 (20.75%)	1,713 (24.13%)	1,807 (25.46%)	2,105 (29.66%)	–	3,260 (45.93%)
Correct Frames	1,526 (21.50%)	1,757 (24.75%)	1,759 (24.78%)	2,056 (28.97%)	–	3,248 (45.76%)
Rephrased Questions	1,507 (21.23%)	1,736 (24.46%)	1,771 (24.95%)	2,084 (29.36%)	–	3,153 (44.42%)
Additional Empty Option	1,003 (14.13%)	1,077 (15.17%)	1,214 (17.10%)	1,395 (19.65%)	2,409 (33.94%)	2,351 (33.12%)
All a0	2,951 (41.58%)	510 (7.19%)	480 (6.76%)	3,157 (44.48%)	–	–
All a1	2,909 (40.98%)	492 (6.93%)	480 (6.76%)	3,217 (45.32%)	–	–
All a2	2,935 (41.35%)	572 (8.06%)	442 (6.23%)	3,149 (44.36%)	–	–
All a3	2,892 (40.74%)	550 (7.75%)	472 (6.65%)	3,184 (44.86%)	–	–
Correct a0	2,806 (39.53%)	1,342 (18.91%)	1,351 (19.03%)	1,599 (22.53%)	–	2,806 (39.53%)
Correct a0 Shuffled	2,864 (40.35%)	1,274 (17.95%)	1,312 (18.48%)	1,648 (23.22%)	–	2,864 (40.35%)
Correct a1	1,149 (16.19%)	3,216 (45.31%)	1,231 (17.34%)	1,502 (21.16%)	–	3,216 (45.31%)
Correct a1 Shuffled	1,113 (15.68%)	3,181 (44.82%)	1,335 (18.81%)	1,469 (20.70%)	–	3,181 (44.82%)
Correct a2	1,141 (16.07%)	1,223 (17.23%)	3,253 (45.83%)	1,481 (20.87%)	–	3,253 (45.83%)
Correct a2 Shuffled	1,090 (15.36%)	1,252 (17.64%)	3,301 (46.51%)	1,455 (20.50%)	–	3,301 (46.51%)
Correct a3	1,088 (15.33%)	1,215 (17.12%)	1,189 (16.75%)	3,606 (50.80%)	–	3,606 (50.80%)
Correct a3 Shuffled	1,074 (15.13%)	1,194 (16.82%)	1,242 (17.50%)	3,588 (50.55%)	–	3,588 (50.55%)
All Correct Answers	3,363 (47.38%)	391 (5.51%)	171 (2.41%)	3,173 (44.70%)	–	–
Empty Frames	1,501 (21.15%)	1,739 (24.50%)	1,782 (25.11%)	2,076 (29.25%)	–	3,285 (46.28%)
Empty Questions	1,563 (22.02%)	1,608 (22.65%)	1,700 (23.95%)	2,227 (31.38%)	–	3,496 (49.25%)
Empty Answers	1,031 (14.53%)	25 (0.35%)	0 (0.00%)	6,042 (85.12%)	–	–

Table 14: SeViLA Performance on All the Settings of Perception Test. Each cell shows the count of selections for each option (*a0–a3*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	Accuracy
Target	2,549 (33.29%)	2,561 (33.45%)	2,546 (33.25%)	–	–
Default	2,256 (29.47%)	2,331 (30.45%)	3,069 (40.09%)	–	3,468 (45.30%)
Answer Shuffling	2,323 (30.34%)	2,300 (30.04%)	3,033 (39.62%)	–	3,519 (45.96%)
Rephrased Questions	2,286 (29.86%)	2,231 (29.14%)	3,139 (41.00%)	–	3,369 (44.00%)
Additional Empty Option	1,674 (21.87%)	1,246 (16.27%)	1,436 (18.76%)	3,300 (43.10%)	2,223 (29.04%)
All a0	3,380 (44.15%)	464 (6.06%)	3,812 (49.79%)	–	–
All a1	3,427 (44.76%)	449 (5.86%)	3,780 (49.37%)	–	–
All a2	3,377 (44.11%)	449 (5.86%)	3,830 (50.03%)	–	–
Correct a0	3,128 (40.86%)	1,978 (25.84%)	2,550 (33.31%)	–	3,128 (40.86%)
Correct a0 Shuffled	3,112 (40.65%)	2,022 (26.41%)	2,522 (32.94%)	–	3,112 (40.65%)
Correct a1	1,909 (24.93%)	3,249 (42.44%)	2,498 (32.63%)	–	3,249 (42.44%)
Correct a1 Shuffled	1,869 (24.41%)	3,268 (42.69%)	2,519 (32.90%)	–	3,268 (42.69%)
Correct a2	1,858 (24.27%)	1,714 (22.39%)	4,084 (53.34%)	–	4,084 (53.34%)
Correct a2 Shuffled	1,846 (24.11%)	1,750 (22.86%)	4,060 (53.03%)	–	4,060 (53.03%)
All Correct Answers	3,531 (46.12%)	353 (4.61%)	3,772 (49.27%)	–	–
Empty Frames	2,684 (35.06%)	2,058 (26.88%)	2,914 (38.06%)	–	2,766 (36.13%)
Empty Questions	2,574 (33.62%)	2,040 (26.65%)	3,042 (39.73%)	–	3,139 (41.00%)
Empty Answers	985 (12.87%)	0 (0.00%)	6,671 (87.13%)	–	–

Table 15: SeViLA Performance on All the Settings of Video-MME. Each cell shows the count of selections for each option (*a0–a4*) along with the percentage in parentheses. Accuracy represents the number of correct answers where applicable.

Setting	a0	a1	a2	a3	a4	Accuracy
Target	675 (25.00%)	675 (25.00%)	675 (25.00%)	675 (25.00%)	–	–
Default	911 (33.74%)	592 (21.93%)	568 (21.04%)	629 (23.30%)	–	1,076 (39.85%)
Answer Shuffling	887 (32.85%)	566 (20.96%)	591 (21.89%)	656 (24.30%)	–	1,041 (38.56%)
Rephrased Questions	947 (35.07%)	556 (20.59%)	552 (20.44%)	645 (23.89%)	–	1,041 (38.56%)
Additional Empty Option	817 (30.26%)	413 (15.30%)	428 (15.85%)	472 (17.48%)	570 (21.11%)	875 (32.41%)
All a0	1,631 (60.41%)	97 (3.59%)	140 (5.19%)	832 (30.81%)	–	–
All a1	1,681 (62.26%)	76 (2.81%)	164 (6.07%)	779 (28.85%)	–	–
All a2	1,637 (60.63%)	100 (3.70%)	148 (5.48%)	815 (30.19%)	–	–
All a3	1,573 (58.26%)	117 (4.33%)	153 (5.67%)	857 (31.74%)	–	–
Correct a0	1,271 (47.07%)	469 (17.37%)	451 (16.70%)	509 (18.85%)	–	1,271 (47.07%)
Correct a0 Shuffled	1,229 (45.52%)	464 (17.19%)	469 (17.37%)	538 (19.93%)	–	1,229 (45.52%)
Correct a1	752 (27.85%)	977 (36.19%)	460 (17.04%)	511 (18.93%)	–	977 (36.19%)
Correct a1 Shuffled	714 (26.44%)	951 (35.22%)	474 (17.56%)	561 (20.78%)	–	951 (35.22%)
Correct a2	793 (29.37%)	443 (16.41%)	990 (36.67%)	474 (17.56%)	–	990 (36.67%)
Correct a2 Shuffled	755 (27.96%)	430 (15.93%)	984 (36.44%)	531 (19.67%)	–	984 (36.44%)
Correct a3	725 (26.85%)	451 (16.70%)	463 (17.15%)	1,061 (39.30%)	–	1,061 (39.30%)
Correct a3 Shuffled	746 (27.63%)	433 (16.04%)	475 (17.59%)	1,046 (38.74%)	–	1,046 (38.74%)
All Correct Answers	1,840 (68.15%)	75 (2.78%)	93 (3.44%)	692 (25.63%)	–	–
Empty Frames	912 (33.78%)	569 (21.07%)	555 (20.56%)	664 (24.59%)	–	773 (28.63%)
Empty Questions	876 (32.44%)	523 (19.37%)	571 (21.15%)	730 (27.04%)	–	1,020 (37.78%)
Empty Answers	727 (26.93%)	29 (1.07%)	37 (1.37%)	1,907 (70.63%)	–	–

Algorithm 1 BOLD: Debiasing with Prior Estimation by Decomposition

Require: Video-Language model M, dataset $\mathcal{D} = \{T_i\}$, number of options n, estimation budget K
Ensure: Model predictions $\mathcal{Y}$

Initialize prediction set $\mathcal{Y} \leftarrow \emptyset$ and prior set $\mathcal{P} \leftarrow \emptyset$ ▷ Initialization
Sample estimation set $\mathcal{D}_k$ of size K from $\mathcal{D}$, set $\mathcal{D}_r \leftarrow \mathcal{D} \setminus \mathcal{D}_k$
for $T \in \mathcal{D}_k$ **do**
  Apply attacks to get observed predictions:
  $P_o(d_i \mid A|_{v=0}(T))$, $P_o(d_i \mid A|_{q=0}(T))$, $P_o(d_i \mid A|_{\text{opts}=0}(T))$
  Compute decomposed priors using Equation 4:
  $P_{\text{pri}}(d_i \mid A|_{v=0}(T))$, $P_{\text{pri}}(d_i \mid A|_{q=0}(T))$, $P_{\text{pri}}(d_i \mid A|_{\text{opts}=0}(T))$
  Estimate sample-specific prior $\tilde{P}_{\text{prior}}(d_i \mid T)$ using Equation 3; add to $\mathcal{P}$
end for
Estimate global prior $\tilde{P}_{\text{prior}}(d_i)$ by averaging $\mathcal{P}$ ▷ Prior Estimation
for $T \in \mathcal{D}$ **do**
  Debias model prediction using the global prior:
  $P_{\text{deb}}(d_i \mid T) = \text{softmax}\left(\log P_o(d_i \mid T) - \log \tilde{P}_{\text{prior}}(d_i)\right)$
  Add the predicted answer to $\mathcal{Y}$
end for
return $\mathcal{Y}$

Algorithm 2 Weighted_BOLD: Debiasing with Prior Estimation by Decomposition with Weights

Require: Model M, dataset $\mathcal{D} = \{T_i\}$, estimation budget K, number of folds $F = 5$
Ensure: Debiased model predictions $\mathcal{Y}$

Initialize prediction set $\mathcal{Y} \leftarrow \emptyset$ and prior set $\mathcal{P} \leftarrow \emptyset$ ▷ Initialization
Sample estimation set $\mathcal{D}_k$ of size K from $\mathcal{D}$; set remaining samples $\mathcal{D}_r \leftarrow \mathcal{D} \setminus \mathcal{D}_k$
for each fold $f = 1$ to F **do** ▷ Cross-validation
  Split $\mathcal{D}_k$ into test set $\mathcal{D}^f_{\text{test}}$ and validation set $\mathcal{D}^f_{\text{val}}$
  for $T \in \mathcal{D}^f_{\text{test}}$ **do**
    Apply attacks to obtain observed predictions: $P_o(d_i \mid A(T))$
    Compute decomposed priors for each attack using Equation 4: $P_{\text{pri}}(d_i \mid A(T))$
    Estimate sample-specific prior $\tilde{P}_{\text{pri}}(d_i \mid T)$ using Equation 6; add to $\mathcal{P}^f$
  end for
  Optimize weights $\{w_i\}$ using COBYLA, subject to constraints $0 \leq w_i \leq 1$ or $|w_i| \leq 1$, to minimize *recall_std* bias on $\mathcal{D}^f_{\text{test}}$
end for
Compute global prior by averaging over folds:

$$\tilde{P}_{\text{prior}}(d_i) = \frac{1}{F} \sum_{f=1}^{F} \tilde{P}^f_{\text{pri}}(d_i)$$

▷ Global Prior Estimation
for $T \in \mathcal{D}$ **do** ▷ Debiasing
  Debias the model prediction using the global prior:

$$P_{\text{deb}}(d_i \mid T) = \text{softmax}\left(\log P_o(d_i \mid T) - \log \tilde{P}_{\text{prior}}(d_i)\right)$$

  Add the predicted answer to $\mathcal{Y}$
end for
return $\mathcal{Y}$

Table 16: Comparison of BOLD and Weighted_BOLD bias mitigation approaches across models and datasets for performance and bias monitoring metrics with $k = 0.25$. Green arrows indicate improvements: upward for Accuracy and F1_mean, and downward for standard deviation metrics. Red arrows represent deterioration, respectively.

Model	Dataset	Configuration	Performance Metrics		Bias Monitoring Metrics		
			Accuracy	F1_mean	Recall_std	F1_std	JS_std
Video-LLaVA	NExT-QA	**Default**	49.95	49.81	16.31	3.27	5.4
		BOLD	51.81 (↑3.72%)	51.71 (↑3.82%)	13.29 (↓18.54%)	2.66 (↓18.59%)	4.58 (↓15.12%)
		+Weighted_BOLD	51.89 (↑3.88%)	51.77 (↑3.95%)	13.28 (↓18.57%)	2.73 (↓16.57%)	4.56 (↓15.39%)
		-/+Weighted_BOLD	52.28 (↑4.65%)	52.25 (↑4.9%)	10.9 (↓33.18%)	2.23 (↓31.89%)	3.9 (↓27.63%)
	STAR	**Default**	34.77	31.83	24.99	6.82	5.78
		BOLD	37.45 (↑7.7%)	36.14 (↑13.54%)	17.43 (↓30.23%)	4.84 (↓29.02%)	4.52 (↓21.74%)
		+Weighted_BOLD	37.81 (↑8.75%)	36.67 (↑15.22%)	16.15 (↓35.38%)	4.71 (↓30.94%)	4.15 (↓28.15%)
		-/+Weighted_BOLD	39.87 (↑14.67%)	39.65 (↑24.57%)	6.85 (↓72.56%)	2.34 (↓65.67%)	2.08 (↓63.98%)
	Perception Test	**Default**	40.72	35.69	27.5	14.75	5.28
		BOLD	41.62 (↑2.22%)	38.68 (↑8.38%)	21.77 (↓20.85%)	10.93 (↓25.88%)	3.77 (↓28.54%)
		+Weighted_BOLD	41.95 (↑3.02%)	39.46 (↑10.58%)	19.73 (↓28.26%)	10.24 (↓30.53%)	3.39 (↓35.79%)
		-/+Weighted_BOLD	42.01 (↑3.18%)	40.48 (↑13.43%)	14.69 (↓46.59%)	8.48 (↓42.52%)	2.91 (↓44.9%)
	Video-MME	**Default**	34.3	30.99	24.09	8.06	4.79
		BOLD	34.63 (↑0.97%)	32.71 (↑5.54%)	18.21 (↓24.4%)	6.18 (↓23.36%)	3.86 (↓19.47%)
		+Weighted_BOLD	34.59 (↑0.86%)	33.01 (↑6.51%)	15.92 (↓33.93%)	5.91 (↓26.68%)	3.38 (↓29.56%)
		-/+Weighted_BOLD	33.93 (↓1.08%)	33.21 (↑7.16%)	10.1 (↓58.07%)	4.09 (↓49.26%)	2.33 (↓51.39%)
Video-LLaMA	NExT-QA	**Default**	44.79	40.85	23.83	15.86	16.09
		BOLD	45.88 (↑2.43%)	42.15 (↑3.18%)	22.98 (↓3.53%)	15.53 (↓2.11%)	15.56 (↓3.3%)
		+Weighted_BOLD	45.96 (↑2.61%)	42.27 (↑3.46%)	22.78 (↓4.4%)	15.46 (↓2.55%)	15.52 (↓3.55%)
		-/+Weighted_BOLD	47.56 (↑6.19%)	46.63 (↑14.14%)	13.79 (↓42.11%)	8.47 (↓46.62%)	12.14 (↓24.53%)
	STAR	**Default**	36.52	31.86	24.1	15.44	15.29
		BOLD	37.19 (↑1.82%)	33.02 (↑3.65%)	22.29 (↓7.5%)	14.55 (↓5.75%)	14.58 (↓4.6%)
		+Weighted_BOLD	37.21 (↑1.89%)	33.32 (↑4.59%)	21.45 (↓10.97%)	14.03 (↓9.1%)	14.22 (↓7.0%)
		-/+Weighted_BOLD	38.63 (↑5.76%)	38.07 (↑19.52%)	10.16 (↓57.84%)	4.78 (↓69.05%)	8.93 (↓41.61%)
	Perception Test	**Default**	41.39	37.19	27.06	11.4	16.06
		BOLD	41.92 (↑1.27%)	38.7 (↑4.06%)	24.04 (↓11.16%)	9.85 (↓13.61%)	14.44 (↓10.12%)
		+Weighted_BOLD	42.06 (↑1.62%)	39.36 (↑5.86%)	21.8 (↓19.45%)	9.11 (↓20.06%)	13.05 (↓18.75%)
		-/+Weighted_BOLD	42.05 (↑1.59%)	41.35 (↑11.2%)	11.13 (↓58.86%)	4.71 (↓58.65%)	8.43 (↓47.51%)
	Video-MME	**Default**	32.05	28.15	20.62	12.52	13.75
		BOLD	32.69 (↑2.01%)	29.2 (↑3.73%)	19.28 (↓6.5%)	11.95 (↓4.54%)	13.13 (↓4.53%)
		+Weighted_BOLD	32.2 (↑0.47%)	28.98 (↑2.94%)	17.65 (↓14.38%)	11.69 (↓6.56%)	12.6 (↓8.39%)
		-/+Weighted_BOLD	32.92 (↑2.72%)	30.39 (↑7.97%)	15.65 (↓24.08%)	10.21 (↓18.39%)	11.68 (↓15.04%)
SeViLA	NExT-QA	**Default**	63.91	63.88	2.15	1.32	1.99
		BOLD	63.92 (↑0.02%)	63.89 (↑0.03%)	2.0 (↓6.81%)	1.18 (↓10.43%)	1.97 (↓0.91%)
		+Weighted_BOLD	63.88 (↓0.04%)	63.86 (↓0.03%)	2.08 (↓3.47%)	1.28 (↓2.87%)	1.9 (↓4.64%)
		-/+Weighted_BOLD	63.95 (↑0.07%)	63.95 (↑0.11%)	1.44 (↓32.8%)	1.16 (↓12.17%)	1.63 (↓18.3%)
	STAR	**Default**	46.28	46.14	4.47	2.3	2.25
		BOLD	46.17 (↓0.24%)	46.05 (↓0.2%)	4.09 (↓8.42%)	2.24 (↓2.93%)	2.11 (↓6.25%)
		+Weighted_BOLD	46.07 (↓0.46%)	45.95 (↓0.42%)	4.12 (↓7.63%)	2.28 (↓1.24%)	2.13 (↓5.17%)
		-/+Weighted_BOLD	46.2 (↓0.18%)	46.08 (↓0.13%)	4.04 (↓9.49%)	2.23 (↓3.15%)	2.16 (↓3.9%)
	Perception Test	**Default**	45.3	45.11	6.21	3.02	1.59
		BOLD	45.28 (↓0.03%)	45.15 (↑0.09%)	5.1 (↓17.85%)	2.93 (↓3.09%)	1.26 (↓21.04%)
		+Weighted_BOLD	45.34 (↑0.09%)	45.24 (↑0.29%)	4.26 (↓31.39%)	2.65 (↓12.43%)	1.03 (↓35.41%)
		-/+Weighted_BOLD	45.22 (↓0.17%)	45.15 (↑0.09%)	3.69 (↓40.51%)	2.38 (↓21.06%)	0.85 (↓46.67%)
	Video-MME	**Default**	39.85	39.82	4.67	1.68	0.84
		BOLD	40.04 (↑0.46%)	40.04 (↑0.56%)	3.79 (↓18.89%)	1.48 (↓12.03%)	0.74 (↓11.46%)
		+Weighted_BOLD	40.07 (↑0.56%)	40.08 (↑0.66%)	3.64 (↓22.05%)	1.48 (↓11.82%)	0.67 (↓20.11%)
		-/+Weighted_BOLD	40.04 (↑0.46%)	40.06 (↑0.63%)	2.31 (↓50.5%)	1.36 (↓18.83%)	0.3 (↓63.83%)

Table 17: Comparison of BOLD and Weighted_BOLD bias mitigation approaches across models and datasets for performance and bias monitoring metrics with k = 0.5. Green arrows indicate improvements: upward for Accuracy and F1_mean, and downward for standard deviation metrics. Red arrows represent deterioration, respectively.

Model	Dataset	Configuration	Performance Metrics		Bias Monitoring Metrics		
			Accuracy	F1_mean	Recall_std	F1_std	JS_std
Video-LLaMA	NExT-QA	**Default**	44.79	40.85	23.83	15.86	16.09
		BOLD	45.88 (↑2.43%)	42.15 (↑3.18%)	22.98 (↓3.53%)	15.53 (↓2.11%)	15.55 (↓3.31%)
		+Weighted_BOLD	45.91 (↑2.51%)	42.2 (↑3.29%)	22.83 (↓4.19%)	15.51 (↓2.2%)	15.56 (↓3.27%)
		-/+Weighted_BOLD	48.25 (↑7.73%)	46.52 (↑13.86%)	16.6 (↓30.32%)	10.6 (↓33.17%)	13.27 (↓17.48%)
	STAR	**Default**	36.52	31.86	24.1	15.44	15.29
		BOLD	37.19 (↑1.82%)	33.02 (↑3.65%)	22.29 (↓7.5%)	14.55 (↓5.75%)	14.58 (↓4.66%)
		+Weighted_BOLD	37.34 (↑2.24%)	33.5 (↑5.16%)	21.13 (↓12.3%)	14.05 (↓8.98%)	14.14 (↓7.54%)
		-/+Weighted_BOLD	38.61 (↑5.72%)	38.06 (↑19.47%)	10.18 (↓57.75%)	4.79 (↓68.99%)	8.58 (↓43.89%)
	Perception Test	**Default**	41.39	37.19	27.06	11.4	16.06
		BOLD	41.9 (↑1.24%)	38.68 (↑4.01%)	24.05 (↓11.11%)	9.87 (↓13.45%)	14.45 (↓10.01%)
		+Weighted_BOLD	42.06 (↑1.62%)	39.36 (↑5.86%)	21.8 (↓19.45%)	9.11 (↓20.06%)	13.02 (↓18.94%)
		-/+Weighted_BOLD	42.27 (↑2.13%)	41.24 (↑10.9%)	13.55 (↓49.93%)	5.64 (↓50.54%)	9.73 (↓39.41%)
	Video-MME	**Default**	32.05	28.15	20.62	12.52	13.75
		BOLD	32.73 (↑2.13%)	29.23 (↑3.85%)	19.29 (↓6.45%)	11.96 (↓4.46%)	13.1 (↓4.73%)
		+Weighted_BOLD	32.2 (↑0.47%)	28.98 (↑2.94%)	17.65 (↓14.38%)	11.69 (↓6.56%)	12.64 (↓8.04%)
		-/+Weighted_BOLD	32.88 (↑2.6%)	31.21 (↑10.87%)	12.77 (↓38.06%)	8.02 (↓35.91%)	10.42 (↓24.25%)
Video-LLaVA	NExT-QA	**Default**	49.95	49.81	16.31	3.27	5.4
		BOLD	51.81 (↑3.72%)	51.71 (↑3.82%)	13.27 (↓18.63%)	2.67 (↓18.42%)	4.58 (↓15.06%)
		+Weighted_BOLD	52.15 (↑4.39%)	52.04 (↑4.49%)	12.72 (↓22.02%)	2.62 (↓19.83%)	4.49 (↓16.83%)
		-/+Weighted_BOLD	53.65 (↑7.41%)	53.53 (↑7.48%)	8.49 (↓47.97%)	3.15 (↓3.61%)	3.78 (↓29.87%)
	STAR	**Default**	34.77	31.83	24.99	6.82	5.78
		BOLD	37.21 (↑7.01%)	35.76 (↑12.37%)	18.28 (↓26.85%)	4.97 (↓27.15%)	4.54 (↓21.39%)
		+Weighted_BOLD	37.53 (↑7.94%)	36.18 (↑13.69%)	17.45 (↓30.15%)	4.91 (↓28.03%)	4.26 (↓26.24%)
		-/+Weighted_BOLD	38.31 (↑10.17%)	37.39 (↑17.47%)	13.26 (↓46.92%)	4.59 (↓32.59%)	3.25 (↓43.67%)
	Perception Test	**Default**	40.72	35.69	27.5	14.75	5.28
		BOLD	41.62 (↑2.22%)	38.69 (↑8.4%)	21.75 (↓20.9%)	10.92 (↓25.97%)	3.78 (↓28.4%)
		+Weighted_BOLD	41.95 (↑3.02%)	39.46 (↑10.58%)	19.73 (↓28.26%)	10.24 (↓30.53%)	3.4 (↓35.51%)
		-/+Weighted_BOLD	41.62 (↑2.22%)	40.88 (↑14.56%)	11.01 (↓59.96%)	5.44 (↓63.11%)	2.59 (↓50.97%)
	Video-MME	**Default**	34.3	30.99	24.09	8.06	4.79
		BOLD	34.7 (↑1.19%)	32.79 (↑5.81%)	18.19 (↓24.51%)	6.16 (↓23.63%)	3.84 (↓19.93%)
		+Weighted_BOLD	34.63 (↑0.97%)	32.97 (↑6.38%)	16.36 (↓32.07%)	6.01 (↓25.43%)	3.43 (↓28.5%)
		-/+Weighted_BOLD	34.0 (↓0.86%)	32.6 (↑5.21%)	14.11 (↓41.43%)	5.79 (↓28.18%)	2.92 (↓39.01%)
SeViLA	NExT-QA	**Default**	63.91	63.88	2.15	1.32	1.99
		BOLD	63.92 (↑0.02%)	63.89 (↑0.02%)	2.03 (↓5.47%)	1.18 (↓10.4%)	1.99 (↓0.17%)
		+Weighted_BOLD	63.93 (↑0.04%)	63.91 (↑0.04%)	1.99 (↓7.64%)	1.19 (↓9.83%)	1.99 (↓0.04%)
		-/+Weighted_BOLD	64.01 (↑0.16%)	64.0 (↑0.19%)	1.14 (↓47.18%)	1.17 (↓11.43%)	2.01 (↑1.1%)
	STAR	**Default**	46.28	46.14	4.47	2.3	2.25
		BOLD	46.22 (↓0.12%)	46.1 (↓0.08%)	4.13 (↓7.46%)	2.26 (↓1.78%)	2.1 (↓6.64%)
		+Weighted_BOLD	46.2 (↓0.18%)	46.08 (↓0.13%)	4.01 (↓10.17%)	2.2 (↓4.36%)	2.07 (↓8.2%)
		-/+Weighted_BOLD	46.38 (↑0.21%)	46.3 (↑0.34%)	3.41 (↓23.66%)	1.94 (↓15.89%)	1.93 (↓14.33%)
	Perception Test	**Default**	45.3	45.11	6.21	3.02	1.59
		BOLD	45.32 (↑0.06%)	45.18 (↑0.16%)	5.18 (↓16.61%)	2.95 (↓2.43%)	1.3 (↓18.42%)
		+Weighted_BOLD	45.31 (↑0.03%)	45.2 (↑0.21%)	4.56 (↓26.58%)	2.83 (↓6.43%)	1.08 (↓31.92%)
		-/+Weighted_BOLD	45.25 (↓0.12%)	45.2 (↑0.2%)	3.25 (↓47.69%)	2.28 (↓24.52%)	0.71 (↓55.13%)
	Video-MME	**Default**	39.85	39.82	4.67	1.68	0.84
		BOLD	40.19 (↑0.84%)	40.17 (↑0.88%)	4.11 (↓12.03%)	1.41 (↓16.19%)	0.73 (↓12.52%)
		+Weighted_BOLD	40.04 (↑0.46%)	40.03 (↑0.54%)	3.78 (↓19.14%)	1.44 (↓14.6%)	0.69 (↓17.22%)
		-/+Weighted_BOLD	39.81 (↓0.09%)	39.83 (↑0.04%)	2.99 (↓35.93%)	1.43 (↓14.97%)	0.45 (↓45.74%)

Table 18: Comparison of BOLD and Weighted_BOLD bias mitigation approaches across models and datasets for performance and bias monitoring metrics with $k = 0.75$. Green arrows indicate improvements: upward for Accuracy and F1_mean, and downward for standard deviation metrics. Red arrows represent deterioration for Accuracy and F1_mean.

Model	Dataset	Configuration	Performance Metrics		Bias Monitoring Metrics		
			Accuracy	F1_mean	Recall_std	F1_std	JS_std
Video-LLaMA	NExT-QA	**Default**	44.79	40.85	23.83	15.86	16.09
		BOLD	45.87 (↑2.4%)	42.16 (↑3.19%)	22.96 (↓3.63%)	15.49 (↓2.37%)	15.56 (↓3.28%)
		+Weighted_BOLD	45.91 (↑2.51%)	42.2 (↑3.29%)	22.83 (↓4.19%)	15.51 (↓2.2%)	15.57 (↓3.23%)
		-/+Weighted_BOLD	47.56 (↑6.19%)	46.63 (↑14.14%)	13.36 (↓43.91%)	8.42 (↓46.91%)	11.79 (↓26.73%)
	STAR	**Default**	36.52	31.86	24.1	15.44	15.29
		BOLD	37.19 (↑1.82%)	33.02 (↑3.65%)	22.29 (↓7.5%)	14.55 (↓5.75%)	14.59 (↓4.59%)
		+Weighted_BOLD	37.34 (↑2.24%)	33.5 (↑5.16%)	21.13 (↓12.3%)	14.05 (↓8.98%)	14.13 (↓7.56%)
		-/+Weighted_BOLD	38.74 (↑6.07%)	38.14 (↑19.74%)	10.58 (↓56.11%)	4.94 (↓68.03%)	9.64 (↓36.94%)
	Perception Test	**Default**	41.39	37.19	27.06	11.4	16.06
		BOLD	41.9 (↑1.24%)	38.67 (↑3.98%)	24.09 (↓11.0%)	9.88 (↓13.29%)	14.46 (↓9.97%)
		+Weighted_BOLD	42.06 (↑1.62%)	39.36 (↑5.86%)	21.8 (↓19.45%)	9.11 (↓20.06%)	13.12 (↓18.33%)
		-/+Weighted_BOLD	42.29 (↑2.16%)	41.32 (↑11.13%)	12.96 (↓52.13%)	5.53 (↓51.52%)	9.56 (↓40.45%)
	Video-MME	**Default**	32.05	28.15	20.62	12.52	13.75
		BOLD	32.69 (↑2.01%)	29.2 (↑3.75%)	19.24 (↓6.7%)	11.94 (↓4.63%)	13.09 (↓4.8%)
		+Weighted_BOLD	32.2 (↑0.47%)	28.97 (↑2.94%)	17.65 (↓14.38%)	11.7 (↓6.54%)	12.56 (↓8.62%)
		-/+Weighted_BOLD	32.65 (↑1.89%)	30.55 (↑8.55%)	13.94 (↓32.38%)	9.24 (↓26.17%)	11.04 (↓19.73%)
Video-LLaVA	NExT-QA	**Default**	49.95	49.81	16.31	3.27	5.4
		BOLD	51.81 (↑3.72%)	51.71 (↑3.82%)	13.27 (↓18.63%)	2.67 (↓18.42%)	4.58 (↓15.14%)
		+Weighted_BOLD	51.96 (↑4.02%)	51.86 (↑4.12%)	12.92 (↓20.76%)	2.64 (↓19.44%)	4.48 (↓17.02%)
		-/+Weighted_BOLD	53.75 (↑7.6%)	53.66 (↑7.73%)	8.51 (↓47.82%)	2.58 (↓21.07%)	3.6 (↓33.28%)
	STAR	**Default**	34.77	31.83	24.99	6.82	5.78
		BOLD	37.21 (↑7.01%)	35.76 (↑12.37%)	18.28 (↓26.85%)	4.97 (↓27.15%)	4.62 (↓20.03%)
		+Weighted_BOLD	37.52 (↑7.9%)	36.17 (↑13.64%)	17.45 (↓30.14%)	4.91 (↓28.02%)	4.27 (↓26.1%)
		-/+Weighted_BOLD	37.86 (↑8.87%)	36.87 (↑15.85%)	13.12 (↓47.51%)	4.68 (↓31.37%)	3.09 (↓46.45%)
	Perception Test	**Default**	40.72	35.69	27.5	14.75	5.28
		BOLD	41.62 (↑2.22%)	38.68 (↑8.38%)	21.77 (↓20.85%)	10.93 (↓25.88%)	3.79 (↓28.21%)
		+Weighted_BOLD	41.95 (↑3.02%)	39.46 (↑10.58%)	19.73 (↓28.26%)	10.24 (↓30.53%)	3.41 (↓35.4%)
		-/+Weighted_BOLD	42.21 (↑3.66%)	41.64 (↑16.67%)	9.47 (↓65.56%)	5.14 (↓65.14%)	1.8 (↓65.84%)
	Video-MME	**Default**	34.3	30.99	24.09	8.06	4.79
		BOLD	34.63 (↑0.97%)	32.64 (↑5.33%)	18.59 (↓22.84%)	6.19 (↓23.2%)	3.85 (↓19.65%)
		+Weighted_BOLD	34.48 (↑0.54%)	32.89 (↑6.13%)	15.94 (↓33.83%)	5.86 (↓27.32%)	3.37 (↓29.66%)
		-/+Weighted_BOLD	33.78 (↓1.51%)	33.2 (↑7.12%)	9.23 (↓61.7%)	3.35 (↓58.46%)	1.42 (↓70.35%)
SeViLA	NExT-QA	**Default**	63.91	63.88	2.15	1.32	1.99
		BOLD	63.91 (↑0.0%)	63.88 (↑0.01%)	2.03 (↓5.45%)	1.18 (↓10.77%)	1.98 (↓0.34%)
		+Weighted_BOLD	63.93 (↑0.04%)	63.91 (↑0.04%)	2.0 (↓6.83%)	1.2 (↓8.97%)	1.98 (↓0.23%)
		-/+Weighted_BOLD	64.11 (↑0.31%)	64.09 (↑0.34%)	1.2 (↓44.36%)	1.09 (↓17.41%)	2.15 (↑8.2%)
	STAR	**Default**	46.28	46.14	4.47	2.3	2.25
		BOLD	46.27 (↓0.03%)	46.15 (↑0.02%)	4.07 (↓8.91%)	2.22 (↓3.53%)	2.08 (↓7.72%)
		+Weighted_BOLD	46.17 (↓0.24%)	46.05 (↓0.19%)	4.01 (↓10.17%)	2.2 (↓4.43%)	2.04 (↓9.14%)
		-/+Weighted_BOLD	46.17 (↓0.24%)	46.1 (↓0.09%)	2.96 (↓33.63%)	1.78 (↓22.67%)	1.64 (↓27.25%)
	Perception Test	**Default**	45.3	45.11	6.21	3.02	1.59
		BOLD	45.31 (↑0.03%)	45.16 (↑0.12%)	5.25 (↓15.42%)	2.98 (↓1.26%)	1.3 (↓18.44%)
		+Weighted_BOLD	45.35 (↑0.12%)	45.25 (↑0.32%)	4.27 (↓31.19%)	2.65 (↓12.18%)	1.03 (↓34.92%)
		-/+Weighted_BOLD	45.15 (↓0.32%)	45.09 (↓0.03%)	3.48 (↓43.88%)	2.3 (↓24.02%)	0.78 (↓51.02%)
	Video-MME	**Default**	39.85	39.82	4.67	1.68	0.84
		BOLD	40.04 (↑0.46%)	40.03 (↑0.53%)	3.84 (↓17.85%)	1.46 (↓13.42%)	0.69 (↓18.0%)
		+Weighted_BOLD	39.89 (↑0.09%)	39.87 (↑0.15%)	3.7 (↓20.9%)	1.43 (↓14.96%)	0.58 (↓31.13%)
		-/+Weighted_BOLD	39.7 (↓0.37%)	39.71 (↓0.26%)	2.69 (↓42.53%)	1.42 (↓15.75%)	0.15 (↓81.72%)

Table 19: Comparison of BOLD and Weighted_BOLD bias mitigation approaches across models and datasets for performance and bias monitoring metrics $k = 1$ (equal to the entire dataset). Green arrows indicate improvements: upward for Accuracy and F1_mean, and downward for standard deviation metrics. Red arrows represent deterioration for Accuracy and F1_mean.

Model	Dataset	Configuration	Performance Metrics		Bias Monitoring Metrics		
			Accuracy	F1_mean	Recall_std	F1_std	JS_std
Video-LLaMA	NExT-QA	**Default**	44.79	40.85	23.83	15.86	16.09
		BOLD	45.87 (↑2.4%)	42.16 (↑3.19%)	22.96 (↓3.63%)	15.49 (↓2.37%)	15.56 (↓3.29%)
		+Weighted_BOLD	45.91 (↑2.51%)	42.2 (↑3.29%)	22.83 (↓4.19%)	15.51 (↓2.2%)	15.57 (↓3.23%)
		-/+Weighted_BOLD	47.53 (↑6.11%)	46.62 (↑14.1%)	13.25 (↓44.39%)	8.46 (↓46.68%)	11.78 (↓26.76%)
	STAR	**Default**	36.52	31.86	24.1	15.44	15.29
		BOLD	37.19 (↑1.82%)	33.02 (↑3.65%)	22.29 (↓7.5%)	14.55 (↓5.75%)	14.59 (↓4.58%)
		+Weighted_BOLD	37.34 (↑2.24%)	33.5 (↑5.16%)	21.13 (↓12.3%)	14.05 (↓8.98%)	14.13 (↓7.55%)
		-/+Weighted_BOLD	38.72 (↑6.03%)	38.18 (↑19.85%)	10.0 (↓58.49%)	4.85 (↓68.6%)	8.53 (↓44.23%)
	Perception Test	**Default**	41.39	37.19	27.06	11.4	16.06
		BOLD	41.89 (↑1.21%)	38.67 (↑3.98%)	24.04 (↓11.18%)	9.86 (↓13.52%)	14.45 (↓10.04%)
		+Weighted_BOLD	42.08 (↑1.65%)	39.38 (↑5.89%)	21.8 (↓19.46%)	9.11 (↓20.03%)	12.93 (↓19.5%)
		-/+Weighted_BOLD	42.25 (↑2.07%)	41.53 (↑11.69%)	11.21 (↓58.59%)	4.76 (↓58.2%)	8.83 (↓45.01%)
	Video-MME	**Default**	32.05	28.15	20.62	12.52	13.75
		BOLD	32.65 (↑1.89%)	29.13 (↑3.49%)	19.29 (↓6.44%)	12.03 (↓3.86%)	13.09 (↓4.82%)
		+Weighted_BOLD	32.2 (↑0.47%)	28.97 (↑2.94%)	17.65 (↓14.38%)	11.7 (↓6.54%)	12.57 (↓8.56%)
		-/+Weighted_BOLD	32.92 (↑2.72%)	31.32 (↑11.27%)	12.6 (↓38.89%)	7.85 (↓37.25%)	10.36 (↓24.67%)
Video-LLaVA	NExT-QA	**Default**	49.95	49.81	16.31	3.27	5.4
		BOLD	51.81 (↑3.72%)	51.71 (↑3.82%)	13.27 (↓18.63%)	2.67 (↓18.42%)	4.58 (↓15.18%)
		+Weighted_BOLD	51.83 (↑3.76%)	51.77 (↑3.93%)	12.79 (↓21.58%)	2.5 (↓23.66%)	4.43 (↓17.81%)
		-/+Weighted_BOLD	52.94 (↑5.98%)	53.0 (↑6.41%)	8.49 (↓47.93%)	1.98 (↓39.48%)	3.02 (↓43.96%)
	STAR	**Default**	34.77	31.83	24.99	6.82	5.78
		BOLD	37.14 (↑6.81%)	35.65 (↑12.02%)	18.53 (↓25.84%)	5.01 (↓26.55%)	4.6 (↓20.38%)
		+Weighted_BOLD	37.33 (↑7.37%)	35.98 (↑13.04%)	17.62 (↓29.49%)	4.83 (↓29.09%)	4.35 (↓24.73%)
		-/+Weighted_BOLD	38.56 (↑10.9%)	37.86 (↑18.96%)	11.35 (↓54.56%)	3.69 (↓45.92%)	2.91 (↓49.67%)
	Perception Test	**Default**	40.72	35.69	27.5	14.75	5.28
		BOLD	41.62 (↑2.22%)	38.69 (↑8.4%)	21.75 (↓20.9%)	10.92 (↓25.97%)	3.79 (↓28.14%)
		+Weighted_BOLD	41.95 (↑3.02%)	39.46 (↑10.58%)	19.73 (↓28.26%)	10.24 (↓30.53%)	3.41 (↓35.37%)
		-/+Weighted_BOLD	41.57 (↑2.09%)	41.11 (↑15.2%)	8.95 (↓67.47%)	4.3 (↓70.81%)	2.11 (↓59.94%)
	Video-MME	**Default**	34.3	30.99	24.09	8.06	4.79
		BOLD	34.56 (↑0.76%)	32.63 (↑5.3%)	18.24 (↓24.28%)	6.08 (↓24.59%)	3.86 (↓19.41%)
		+Weighted_BOLD	34.41 (↑0.32%)	32.78 (↑5.77%)	16.1 (↓33.17%)	5.89 (↓26.93%)	3.37 (↓29.69%)
		-/+Weighted_BOLD	34.22 (↓0.22%)	33.56 (↑8.31%)	9.8 (↓59.31%)	3.71 (↓54.0%)	1.84 (↓61.61%)
SeViLA	NExT-QA	**Default**	63.91	63.88	2.15	1.32	1.99
		BOLD	63.91 (↑0.0%)	63.88 (↑0.01%)	2.03 (↓5.45%)	1.18 (↓10.77%)	1.98 (↓0.66%)
		+Weighted_BOLD	64.0 (↑0.15%)	63.98 (↑0.15%)	1.95 (↓9.2%)	1.24 (↓6.37%)	1.97 (↓0.96%)
		-/+Weighted_BOLD	64.07 (↑0.26%)	64.06 (↑0.29%)	0.84 (↓60.74%)	1.07 (↓18.78%)	2.08 (↑4.64%)
	STAR	**Default**	46.28	46.14	4.47	2.3	2.25
		BOLD	46.25 (↓0.06%)	46.14 (↓0.01%)	4.05 (↓9.25%)	2.22 (↓3.75%)	2.06 (↓8.29%)
		+Weighted_BOLD	46.15 (↓0.27%)	46.04 (↓0.22%)	4.01 (↓10.23%)	2.2 (↓4.52%)	2.05 (↓9.07%)
		-/+Weighted_BOLD	46.18 (↓0.21%)	46.1 (↓0.1%)	3.48 (↓22.16%)	2.04 (↓11.64%)	1.85 (↓17.64%)
	Perception Test	**Default**	45.3	45.11	6.21	3.02	1.59
		BOLD	45.4 (↑0.23%)	45.25 (↑0.33%)	5.24 (↓15.53%)	3.04 (↑0.72%)	1.3 (↓18.45%)
		+Weighted_BOLD	45.34 (↑0.09%)	45.24 (↑0.3%)	4.16 (↓33.06%)	2.62 (↓13.42%)	1.03 (↓34.99%)
		-/+Weighted_BOLD	45.22 (↓0.17%)	45.14 (↑0.08%)	3.85 (↓38.0%)	2.43 (↓19.68%)	0.9 (↓43.13%)
	Video-MME	**Default**	39.85	39.82	4.67	1.68	0.84
		BOLD	40.0 (↑0.37%)	39.99 (↑0.44%)	3.81 (↓18.5%)	1.48 (↓11.8%)	0.66 (↓21.49%)
		+Weighted_BOLD	39.81 (↓0.09%)	39.81 (↓0.02%)	3.57 (↓23.49%)	1.45 (↓13.53%)	0.53 (↓36.44%)
		-/+Weighted_BOLD	39.85 (↑0.0%)	39.85 (↑0.08%)	2.66 (↓43.14%)	1.47 (↓12.42%)	0.16 (↓80.65%)