BUZZY: CONTRASTIVE SCORING TO MITIGATE TEXT-INDUCED BIAS IN MULTIMODAL MULTIPLE-CHOICE QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal multiple-choice question answering (MCQA) provides a standardized and objectively measurable setting for evaluating vision-language models (VLMs). However, because the MCQA format incorporates the candidate choices into the input context, it introduces several unintended biases. Previous work has primarily focused on structural biases, such as preferences for certain position choices. Instead, we argue that the choices act as textual priors, causing models to favor linguistically plausible options regardless of the visual content. We hypothesize and empirically verify that a model genuinely relies on visual evidence only when its multimodal distribution significantly diverges from its text-only distribution. Based on this observation we propose BUZZY, a training-free decoding method that corrects multimodal predictions by subtracting the text-only distribution. Experiments with five VLMs on five multimodal MCQA benchmarks demonstrate that BUZZY achieves the highest average accuracy among state-of-the-art methods while reducing inference latency by over 28% compared to prior contrastive scoring approaches. Overall, these results suggest that amplifying the visual signal by penalizing text-only preferences is key to efficient and robust multimodal MCQA reasoning.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated strong capabilities across a wide range of tasks (Li et al., 2023; Liu et al., 2024a; Jung et al., 2026). Multimodal multiple-choice question answering (MCQA) has consequently become a standard setting for evaluating whether these models can reason over visual evidence (Liu et al., 2024b; Yue et al., 2024). Compared with open-ended generation, MCQA enables objective and reproducible evaluation by avoiding ambiguities in answer matching (Royer et al., 2024; Zhang et al., 2025). Its structured answer space, however, also changes how model predictions should be interpreted (Balepur et al., 2024; Atabuzzaman et al., 2025).

In multimodal MCQA, candidate choices are not merely possible outputs. They are presented to the model as part of the textual input before prediction. Therefore, choices expose the model to textual priors that favor linguistically plausible answers regardless of the image, and structural biases that create preferences for particular choice positions (Balepur et al., 2024; Luo et al., 2025; Pezeshkpour & Hruschka, 2024; Yang et al., 2025). High multimodal confidence therefore does not necessarily indicate strong visual grounding because it reflects both visual evidence and preferences induced by the question and choices (Atabuzzaman et al., 2025; Loginova et al., 2025). To empirically test this, we conduct a pilot study comparing each model's multimodal and text-only answer distributions. We find that correct predictions deviate substantially from the text-only distribution, whereas incorrect predictions remain close. Thus, authentic visual reasoning is best measured not by absolute confidence, but by the extent of deviation from textual baselines.

Contrastive decoding is a widely adopted framework for mitigating these biased distributions, effectively separating useful visual evidence from undesirable model preferences (O'Brien & Lewis, 2023). Methods such as Visual Contrastive Decoding (VCD) (Leng et al., 2024) and Instruction Contrastive Decoding (ICD) (Wang et al., 2024b) contrast standard predictions with those obtained from distorted images or disturbance instructions. However, these methods have two main limitations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

First, their distorted contrastive inputs do not directly capture the textual prior introduced by the choices. Second, they require two multimodal forward passes, repeatedly processing the visual tokens and introducing inference overhead.

We introduce **BUZZY**, a training-free decoding method built on a simple principle: a candidate should be rewarded for the support it gains from the image, not for the plausibility it already possesses from text alone. The name BUZZY reflects the idea of suppressing the "buzz" of text-induced priors, so that visually grounded evidence becomes clearer. Rather than constructing a distorted negative input, BUZZY uses the text-only context as a natural control condition. It obtains candidate logits from multimodal and text-only inputs and subtracts the text-only logits from their multimodal counterparts. The resulting score discounts preferences explainable by language alone and emphasizes the features grounded in the visual context.

Across five VLM backbones and five multimodal MCQA benchmarks, BUZZY consistently outperforms greedy, prompt-based, and contrastive decoding baselines. Averaged across all backbones and benchmarks, BUZZY achieves 57.84 accuracy and 53.56 Macro-F1, exceeding the strongest overall baseline by 1.99 and 1.85 points, respectively. It also reduces inference latency by over 28% compared with prior multimodal contrastive decoding methods.

Our contributions are as follows:

- We characterize visual grounding through the discrepancy between multimodal and text-only preferences. We show that incorrect predictions exhibit greater similarity to text-only preferences than correct predictions, revealing the bias induced by choices.
- We propose BUZZY, a training-free contrastive scoring method that uses the text-only context as a natural control and prioritizes candidate support that cannot be explained by textual preference alone.
- We show that BUZZY achieves the highest average performance across five benchmarks and five VLM backbones while reducing inference latency by over 28% compared with prior multimodal contrastive decoding methods.

2 RELATED WORK

2.1 CHOICE BIAS IN MCQA

Prior work has identified several forms of bias in MCQA, including positional preferences, option-order sensitivity, and answer-label bias (Pezeshkpour & Hruschka, 2024; Yang et al., 2025). Models can also exploit semantic or stylistic artifacts within the candidate set and sometimes answer above chance using the choices alone (Wang et al., 2023; Balepur et al., 2024; Wang et al., 2024a; Zheng et al., 2024; Atabuzzaman et al., 2025). Calibration, answer permutation, and related inference-time strategies have been proposed to reduce sensitivity to these artifacts (Zheng et al., 2024; Atabuzzaman et al., 2025; Loginova et al., 2025). As a representative example, Adaptive Logit Correction (ALC) reduces option-label and positional biases by calibrating and correcting multimodal logits (Atabuzzaman et al., 2025). (Atabuzzaman et al., 2025). However, prior work has largely concentrated on structural biases, treating textual answer-space bias and visual grounding as entirely separate problems. We connect the two by demonstrating that candidate-induced textual preferences persist in multimodal predictions.

2.2 CONTRASTIVE DECODING FOR MULTIMODAL GROUNDING

Contrastive decoding steers model predictions by comparing a target distribution against a less desirable reference distribution (O'Brien & Lewis, 2023). In multimodal settings, this reference is commonly constructed by perturbing the visual input or instruction (Lee et al., 2024; Park et al., 2025). Representative methods include VCD, which contrasts predictions from original and distorted images (Leng et al., 2024), and ICD, which uses a disturbance instruction for its contrastive branch (Wang et al., 2024b). Subsequent methods dynamically select or weight contrastive signals to address hallucination and cross-modal knowledge conflicts (Zhu et al., 2024; Suo et al., 2025).

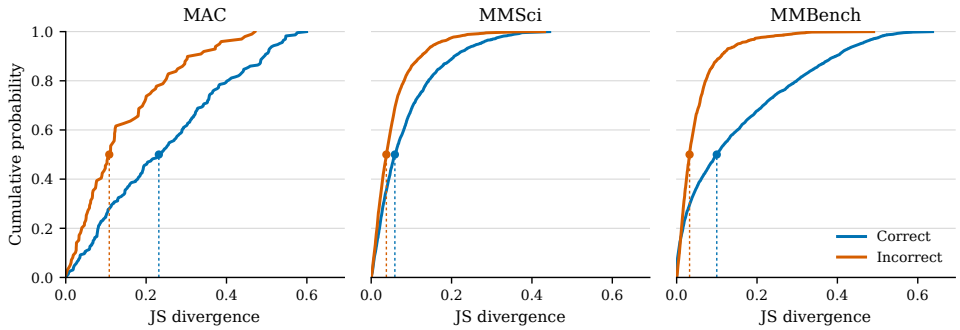


Figure 1: Empirical cumulative distribution functions (ECDFs) of the JS divergence between multimodal and text-only answer distributions for correct and incorrect predictions across three benchmarks. Correct predictions generally exhibit larger JS divergence, indicating greater deviation from text-only preferences.

3 Preliminaries

To distinguish genuine visual grounding from textual bias, we formalize the MCQA setting and compare answer distributions with and without images.

3.1 Problem Setting

Each multimodal MCQA example consists of an image x, a question q, and a set of candidate answers $\mathcal{C} = \{c_1, \ldots, c_K\}$. The goal is to predict the correct answer $y^* \in \mathcal{C}$. Given a VLM, we define two distributions over the same candidate set. The multimodal distribution $p_{\text{mm}}(c \mid x, q, \mathcal{C})$ is obtained from the complete input, whereas the text-only distribution $p_{\text{txt}}(c \mid q, \mathcal{C})$ is obtained by removing only the image while preserving the question and choices. The former reflects the model's preference when visual and textual information are available, while the latter estimates the preference induced by text alone. To make the distributions directly comparable, we require the model to output a single option-label token without generating intermediate reasoning.

3.2 Motivating Analysis

We compute $p_{\text{mm}}(c \mid x, q, \mathcal{C})$ and $p_{\text{txt}}(c \mid q, \mathcal{C})$ under greedy decoding, take the multimodal prediction $\hat{y} = \text{argmax}_{c \in \mathcal{C}} p_{\text{mm}}(c \mid x, q, \mathcal{C})$, and partition the examples according to whether $\hat{y} = y^*$. We then compare the two distributions using Jensen–Shannon (JS) divergence (Lin, 1991) which quantifies the statistical distance between the distributions. As shown in Figure 1, correct predictions are consistently shifted toward higher JS divergence than incorrect predictions. The median divergence at a cumulative probability of 0.5 is also higher for correct predictions across all three benchmarks. This indicates that correct predictions tend to deviate more strongly from the text-only distribution. The remaining two benchmarks show the same overall trend, as reported in Appendix A.3.

4 Method

Motivated by the pilot study, we propose BUZZY, a training-free decoding method that suppresses text-induced preferences in multimodal MCQA. BUZZY consists of three stages. First, it obtains candidate logits from matched multimodal and text-only contexts, using the latter as a proxy for preferences induced by the question and choices alone. Second, it subtracts the weighted text-only logits from their multimodal counterparts to discount candidate support explainable without the image. Finally, it normalizes the corrected logits and selects the highest-scoring candidate as the final answer. Figure 2 illustrates the overall decoding procedure.

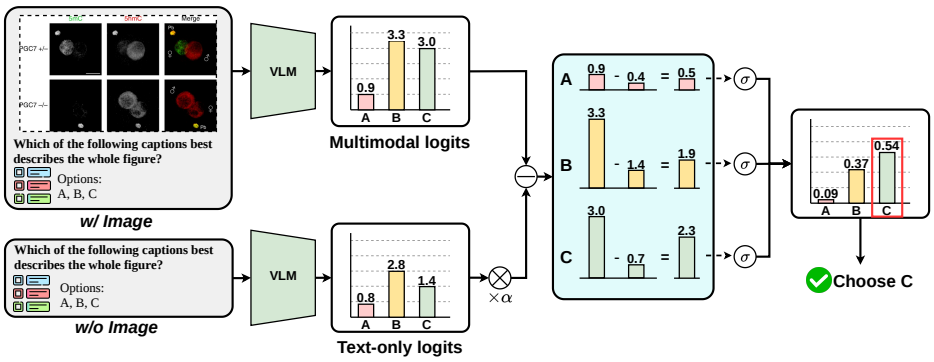


Figure 2: Illustration of BUZZY. Given a question and candidate answers, the model produces candidate scores under both multimodal and text-only inputs. BUZZY subtracts the text-only score, scaled by α, from the multimodal score so that candidates favored mainly by textual prior are suppressed and visually grounded candidates are promoted.

4.1 DUAL-CONTEXT CANDIDATE SCORING

For each question, we evaluate the model under the two contexts defined in Section 3.1. The multimodal context contains the image x, question q, and candidate set $\mathcal{C}$, whereas the text-only context removes x while preserving the question and candidate set. Fixing the textual components ensures that visual signals remain the only distinguishing factor between the input contexts. The two contexts produce candidate-wise logits over the same answer set:

$$l_{\mathrm{mm}}(c) = \mathrm{logit}_\theta(c \mid x, q, \mathcal{C}), \tag{1}$$

$$l_{\mathrm{txt}}(c) = \mathrm{logit}_\theta(c \mid q, \mathcal{C}). \tag{2}$$

Here, logit_θ denotes the pre-softmax score assigned by the language modeling head to the token corresponding to candidate c, such as the option label A or B. The multimodal logit $l_{\mathrm{mm}}(c)$ reflects the model's preference when visual and textual evidence are available, whereas $l_{\mathrm{txt}}(c)$ captures the preference induced without access to the image. We use the text-only branch not as an independent predictor, but as a proxy for the textual preference that may influence the multimodal prediction.

4.2 TEXT-PRIOR CONTRASTIVE CORRECTION

Although the multimodal branch receives the image, its prediction may still be dominated by a candidate that appears plausible from the textual context alone. To reduce this effect, BUZZY explicitly discounts the text-only logit from the corresponding multimodal logit:

$$l_{\mathrm{buzzy}}(c) = \underbrace{l_{\mathrm{mm}}(c)}_{\text{multimodal score}} - \underbrace{\alpha l_{\mathrm{txt}}(c)}_{\text{text-prior penalty}} \tag{3}$$

where $\alpha \geq 0$ controls the strength of the penalty. Setting $\alpha = 0$ recovers standard greedy decoding, while increasing α applies a stronger penalty to candidates favored by the text-only branch.

This correction suppresses candidates whose multimodal scores can largely be explained by text-only preferences. A linguistically plausible distractor favored by both branches receives a larger penalty, whereas a candidate supported primarily by the image retains a relatively high corrected score. BUZZY therefore favors candidate support that cannot be explained by text alone, directly addressing the alignment between incorrect multimodal predictions and text-only preferences observed in our pilot study.

Table 1: Main results across five multimodal MCQA benchmarks and five VLM backbones. We compare BUZZY with greedy decoding, prompt-enhanced decoding, VCD, ICD, and ALC using accuracy (ACC) and Macro-F1. Avg. denotes the unweighted mean across benchmarks, and bold indicates the best result within each backbone.

Backbone	Method	MAC		SciFIBench		MMSci		MMBench		MMStar		Avg.	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	**ACC**	**F1**
Qwen 3.5 4B	Greedy	69.72	70.75	46.20	44.93	38.83	20.49	74.38	74.43	38.13	36.62	53.45	49.44
	Prompt-enhanced	63.00	62.43	42.00	39.96	35.84	20.30	69.81	69.73	36.47	34.51	49.42	45.39
	VCD	68.50	68.77	45.40	44.00	40.85	**26.58**	75.35	75.40	38.93	37.29	53.81	50.41
	ICD	58.41	58.57	40.50	38.23	33.79	14.83	70.48	70.45	35.27	33.60	47.69	43.14
	ALC	68.81	68.93	47.50	46.36	39.69	19.05	74.61	74.64	38.67	37.51	53.86	49.30
	BUZZY (Ours)	**74.01**	**74.24**	**48.70**	**46.99**	**43.44**	19.06	**76.16**	**76.17**	**40.27**	**38.02**	**56.52**	**50.90**
Qwen 3.5 9B	Greedy	81.35	81.40	55.10	54.67	46.54	29.05	76.48	76.43	41.00	40.59	60.09	56.43
	Prompt-enhanced	81.65	81.69	54.10	53.65	45.22	23.53	77.64	77.63	40.47	40.48	59.82	55.40
	VCD	81.96	81.92	55.90	55.55	49.29	32.35	78.10	78.09	40.87	40.81	61.22	57.74
	ICD	81.04	81.01	53.00	52.13	46.65	24.14	72.88	72.84	39.87	39.63	58.69	53.95
	ALC	78.59	78.49	53.10	52.87	46.78	27.07	76.48	76.41	40.47	39.96	59.08	54.96
	BUZZY (Ours)	**82.26**	**82.34**	**58.00**	**57.55**	**52.14**	**33.91**	**79.14**	**79.08**	**43.73**	**43.39**	**63.05**	**59.25**
Phi-3.5-vision-instruct	Greedy	42.81	42.02	48.60	48.47	47.78	34.43	79.23	79.44	45.33	42.89	52.75	49.45
	Prompt-enhanced	45.26	44.64	49.10	49.01	48.21	33.58	78.91	79.09	47.13	45.02	53.72	50.27
	VCD	43.73	43.27	53.50	53.75	51.95	**34.59**	79.76	79.87	47.87	46.47	55.36	51.59
	ICD	42.81	41.60	47.10	46.86	46.32	31.38	77.89	78.16	45.20	42.80	51.86	48.16
	ALC	44.65	44.32	47.60	47.28	47.29	30.84	78.93	79.04	45.73	43.93	52.84	49.08
	BUZZY (Ours)	**49.54**	**49.89**	**54.90**	**55.02**	**52.71**	29.93	**80.18**	**80.22**	**48.13**	**47.10**	**57.09**	**52.43**
Molmo-7B-D	Greedy	36.39	33.87	40.10	39.69	30.48	8.16	72.53	72.73	41.00	35.52	44.10	37.99
	Prompt-enhanced	41.59	40.52	42.30	41.49	31.61	8.48	75.12	75.18	**43.40**	38.67	46.80	40.87
	VCD	40.06	38.21	**42.50**	**42.25**	31.96	14.02	74.87	74.93	42.87	38.97	46.45	41.68
	ICD	35.78	32.72	36.80	35.41	27.92	5.39	68.31	68.30	39.47	32.22	41.66	34.81
	ALC	36.09	33.90	41.30	40.92	30.32	8.34	73.46	73.59	41.53	36.18	44.54	38.59
	BUZZY (Ours)	**43.12**	**42.00**	40.90	40.57	**32.26**	**18.73**	**75.24**	**75.41**	43.13	**39.33**	**46.93**	**43.21**
GLM-4.6V-Flash	Greedy	73.70	73.75	60.30	60.28	51.20	32.48	79.33	79.37	44.80	43.16	61.87	57.81
	Prompt-enhanced	71.87	71.86	59.50	59.58	50.15	22.83	76.90	76.71	42.40	39.96	60.16	54.19
	VCD	71.25	71.36	**61.90**	**61.88**	52.92	27.87	80.62	80.71	45.33	43.70	62.40	57.10
	ICD	67.89	67.85	53.80	53.42	45.97	24.57	78.03	78.15	43.93	42.92	57.92	53.38
	ALC	70.03	70.02	60.20	60.17	49.39	26.68	79.42	79.49	44.67	43.21	60.74	55.91
	BUZZY (Ours)	**77.98**	**77.88**	61.40	61.58	**59.47**	**41.67**	**82.49**	**82.52**	**46.60**	**46.29**	**65.59**	**61.99**

4.3 FINAL PREDICTION

After applying the contrastive correction, we normalize the corrected logits and select the most probable candidate:

$$p_{\text{buzzy}}(c) = \frac{\exp(l_{\text{buzzy}}(c))}{\sum_{c' \in \mathcal{C}} \exp(l_{\text{buzzy}}(c'))}, \qquad \hat{y} = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\, p_{\text{buzzy}}(c). \tag{4}$$

Because softmax preserves the ranking of the corrected logits, this prediction is equivalent to directly maximizing $l_{\text{buzzy}}(c)$, allowing normalization to be omitted when only the final answer is required.

BUZZY requires one standard multimodal forward pass and one additional text-only forward pass. It introduces no parameter updates, auxiliary models, or changes to the model architecture, making it directly applicable to VLMs that expose candidate-level logits. Moreover, because the second pass does not process visual tokens, it incurs substantially lower computational overhead than contrastive decoding methods that require two complete multimodal passes.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks and Backbone Models. We evaluate BUZZY on five multimodal MCQA benchmarks: MAC (Jiang et al., 2025), SciFIBench (Roberts et al., 2024), MMSci (Li et al., 2024), MMBench (Liu et al., 2024b), and MMStar (Chen et al., 2024). These benchmarks cover scientific figures, general

visual reasoning, and diverse multimodal question types. They are designed to require genuine visual grounding, with correct answers depending primarily on evidence from the visual input. We use five vision-language backbones: Qwen3.5-4B, Qwen3.5-9B (Qwen Team, 2026), Phi-3.5-vision-instruct (Abdin et al., 2024), Molmo-7B-D (Deitke et al., 2024), and GLM-4.6V-Flash (V Team et al., 2025). The two Qwen models enable comparison across scales within the same family, while the remaining models represent different architectures. Detailed benchmark descriptions and evaluation settings are provided in Appendix A.1.

Baselines and Implementation Details. We compare BUZZY with greedy decoding, prompt-enhanced decoding, VCD (Leng et al., 2024), ICD (Wang et al., 2024b), and Adaptive Logit Correction (ALC) (Atabuzzaman et al., 2025). Greedy decoding uses the original multimodal input, while prompt-enhanced decoding emphasizes visual evidence through instruction. VCD and ICD construct contrastive branches using a degraded image and a disturbance instruction, respectively, whereas ALC targets structural choice bias by correcting estimated preferences for option labels or positions. Baseline configurations and prompts are detailed in Appendix B. All experiments are conducted using an NVIDIA A100 GPU (80GB VRAM). We fix $\alpha = 0.5$ for BUZZY without dataset-specific tuning.

5.2 RESULTS

As shown in Table 1, BUZZY achieves the highest average accuracy and Macro-F1 for all five backbones. Averaged across the benchmarks and backbones, it obtains 57.84 accuracy and 53.56 Macro-F1. This consistency across model families and scales suggests that the correction is not specific to a particular architecture. BUZZY also leads on MMBench for every backbone. Since MMBench uses circular evaluation to control for positional bias, these gains extend beyond preferences for fixed answer labels. ALC remains below BUZZY for every backbone, suggesting that correcting option-label or positional biases does not fully address semantic text priors. Prompt-enhanced decoding also yields inconsistent gains, indicating that visual-grounding instructions alone are insufficient.

Although VCD remains the strongest aggregate baseline, it consistently falls below BUZZY in average performance. The largest improvement appears with GLM-4.6V-Flash, where BUZZY increases average accuracy from 62.40 to 65.59 and Macro-F1 from 57.81 to 61.99. On MMSci, BUZZY achieves the highest accuracy for every backbone and exceeds VCD by 6.55 points with GLM-4.6V-Flash. VCD and ICD also require an additional multimodal forward pass, while BUZZY uses a lighter text-only reference branch. These results show that discounting text-only preferences improves MCQA performance without a second visual pass.

6 ANALYSIS

6.1 DOES BUZZY EXPOSE IMAGE-CONDITIONED EVIDENCE?

To mathematically analyze how subtracting the text-prior in candidate-logit space promotes visual reasoning, we decompose the multimodal candidate preference into text-aligned and orthogonal components. We analyze both *corrected* and *harmed* cases, where BUZZY respectively fixes or overturns the greedy prediction.

6.1.1 TEXT-ALIGNED AND ORTHOGONAL COMPONENTS

Let m and t denote the mean-centered candidate logit vectors from the multimodal and text-only branches. Geometrically, we treat t as an axis representing the model's text-only candidate preference. We then project m onto this axis to measure how much of the multimodal preference follows the text-only candidate pattern. The remaining component r is orthogonal to t, capturing variation in the multimodal preference that is uncorrelated with the text-only candidate pattern. Formally, we decompose m into a component aligned with t and an orthogonal residual r:

$$\beta = \frac{m^\top t}{t^\top t}, \qquad r = m - \beta t. \tag{5}$$

Here, β determines the magnitude of the projection, and βt is the part of the multimodal preference that follows the text-only candidate pattern. The residual r captures the non-textual component,

Table 2: **Visual Grounding Analysis of BUZZY.** The Gold-Over-Distractor (GOD) rate measures the proportion of examples in which the text-orthogonal residual favors the gold answer over the fixed competing candidate. A positive residual margin indicates stronger residual support for the gold answer, providing evidence that the prediction is grounded in image-conditioned information

	Corrected Cases						**Harmed Cases**					
	GOD rate (%)			Residual margin			GOD rate (%)			Residual margin		
Dataset	Orig	Hard	Rand	Orig	Hard	Rand	Orig	Hard	Rand	Orig	Hard	Rand
MAC	70.6	11.8	35.3	+0.489	−0.595	−0.377	21.4	0.0	7.1	−2.149	−1.367	−0.975
SciFIBench	81.7	45.1	42.3	+0.532	−0.130	−0.153	21.7	13.0	28.3	−0.542	−0.702	−0.476
MMSci	83.9	62.1	55.9	+0.472	+0.137	+0.014	17.8	16.2	25.9	−0.621	−0.554	−0.342
MMBench	75.3	58.6	49.0	+0.460	+0.198	−0.046	11.7	19.8	28.4	−0.481	−0.388	−0.236
MMStar	85.2	61.1	58.3	+0.530	+0.258	+0.049	24.0	22.7	29.3	−0.461	−0.479	−0.543
Overall	**80.8**	**58.1**	**52.3**	**+0.483**	**+0.132**	**−0.025**	**17.2**	**17.6**	**27.0**	**−0.586**	**−0.524**	**−0.369**

whose visual sensitivity we examine next. These components satisfy

$$m = \beta t + r, \qquad r^\top t = 0. \tag{6}$$

A larger β indicates stronger alignment with the text-only direction, while BUZZY reduces the component by subtracting αt: $m - \alpha t = (\beta - \alpha)t + r$.

6.1.2 VISUAL GROUNDING ANALYSIS OF BUZZY

We test whether the text-orthogonal residual contains image-specific evidence by replacing the visual input . We compare the *original* image, a *hard negative* retrieved and a *random* image from the same benchmark, as the most similar non-matching image using CLIP cosine similarity. For each example, we compare the gold answer y_i^* with a fixed competing answer d_i, defined as the greedy error in corrected cases and the BUZZY error in harmed cases. This answer is selected under the original image and held fixed across conditions. We compute the residual margin

$$M_{\mathrm{res},i}(x) = r_{i,y_i^*}(x) - r_{i,d_i}(x), \tag{7}$$

where a positive value favors the gold answer, and report its positive rate for each group $g \in \{\mathrm{corr}, \mathrm{harm}\}$:

$$R^g_{\mathrm{GOD}}(x) = \frac{1}{|\mathcal{S}_g|} \sum_{i \in \mathcal{S}_g} \mathbb{I}[M_{\mathrm{res},i}(x) > 0]\,. \tag{8}$$

Table 2 shows an aggregate Original > Hard Negative > Random pattern in corrected cases. Across both corrected and harmed cases, image replacement weakens the residual preference and moves its margin toward zero. Thus, the residual preference progressively weakens toward zero as image relevance decreases from the matched image to a hard negative and then to a random image, indicating that r captures image-conditioned evidence. Overall, results show that BUZZY shifts prediction away from text-aligned preferences and towards image-conditioned residual evidence.

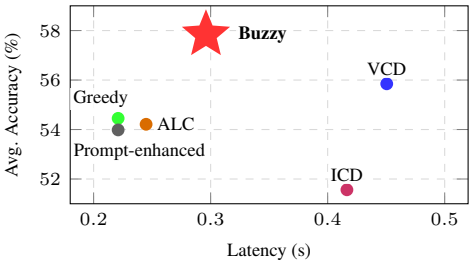
6.2 EFFICIENCY ANALYSIS

Figure 3 compares the theoretical prefill costs of the decoding methods. We use L_q for the number of textual tokens in the question, choices, and prompt, L_v for the number of visual tokens, and d for the hidden dimension. Under the standard self-attention cost model, one multimodal prefill requires $O((L_q + L_v)^2 d)$. VCD and ICD construct their reference predictions using an additional multimodal condition, resulting in two full multimodal passes and a total cost of $O(2(L_q + L_v)^2 d)$. In contrast, BUZZY combines one multimodal pass with one text-only pass at a total cost of $O((L_q+L_v)^2 d+L_q^2 d)$, avoiding additional computation over the visual tokens.

Figure 3 shows that this theoretical advantage translates to empirical latency on average across all evaluated models. VCD and ICD approximately double the latency of greedy decoding from $0.22\,\mathrm{s}$ to $0.45\,\mathrm{s}$ and $0.42\,\mathrm{s}$, respectively. ALC has a low per-example latency of $0.24\,\mathrm{s}$ but lower average accuracy than greedy decoding. This latency also excludes the upfront calibration passes

Method	Prefill Passes	Theoretical Cost
Greedy	$1\times$ MM	$\mathcal{O}((L_q + L_v)^2 d)$
VCD	$2\times$ MM	$\mathcal{O}(2(L_q + L_v)^2 d)$
ICD	$2\times$ MM	$\mathcal{O}(2(L_q + L_v)^2 d)$
BUZZY	$\mathbf{1 \times MM + 1 \times Text}$	$\mathcal{O}((\mathbf{L_q + L_v})^\mathbf{2}\mathbf{d} + \mathbf{L_q^2 d})$

(a) Theoretical prefill cost



(b) Accuracy–latency trade-off

Figure 3: Efficiency analysis of BUZZY. (a) Theoretical prefill cost, where MM and Text denote multimodal and text-only passes, respectively. (b) Average accuracy–latency trade-off across evaluated models. BUZZY avoids a second multimodal pass while achieving the highest average accuracy.

used to estimate the selection-bias vector. In comparison, BUZZY reaches 57.84% average accuracy with a latency of 0.30 s, improving greedy decoding by 3.38 points with only 0.08 s of additional latency. These results demonstrate that BUZZY provides the optimal balance between performance and efficiency; by running the multimodal forward pass only once, it avoids latency overheads while preserving the accuracy gains of contrastive decoding to achieve the highest overall performance.

6.3 SENSITIVITY TO α

We analyze the sensitivity of BUZZY to the subtraction weight α across all five backbones and five benchmarks. Here, α controls the strength of the text-prior penalty subtracted from the multimodal branch. We initially set $\alpha = 0.5$ to match the weighting ratio used by the default configurations of VCD and ICD, ensuring a fair comparison under the same relative weighting. We further examine the robustness of this choice across $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

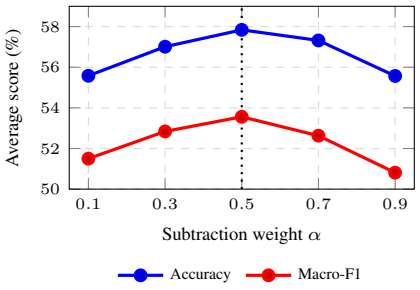


Figure 4: Sensitivity of BUZZY to α. Moderate values remain stable, with $\alpha = 0.5$ achieving the best aggregate performance.

As shown in Figure 4, moderate values from 0.3 to 0.7 yield comparable aggregate performance. In particular, $\alpha = 0.5$ achieves the highest overall accuracy and Macro-F1, with scores of 57.84 and 53.56, respectively. A smaller weight of 0.1 provides limited correction, whereas the aggressive setting of 0.9 reduces accuracy and Macro-F1 to 55.57% and 50.81%. This pattern reflects the trade-off controlled by α: insufficient subtraction leaves misleading textual preferences largely unchanged, whereas excessive subtraction can suppress useful textual support. We therefore use $\alpha = 0.5$ throughout the main experiments as a fixed global setting, without backbone- or dataset-specific tuning. This variation suggests that VLMs retain text-only preferences to different degrees and respond differently to stronger correction. We therefore use $\alpha = 0.5$ throughout the main experiments as a fixed global setting, avoiding backbone- or dataset-specific tuning. Notably, the setting chosen to match VCD and ICD also achieves the best aggregate performance, providing empirical support for the consistency of this shared weighting ratio. This consistency supports our choice of α and enables a fair comparison with these baselines.

6.4 GENERALIZABILITY TO DOMAIN-SPECIFIC FINE-TUNED MODELS

To examine whether the gains of BUZZY remain after targeted model adaptation, we evaluate on Qwen3-VL-8B-Instruct-ChartQA-sft, a version of Qwen3-VL-8B further fine-tuned with LoRA on 7,398 human-annotated ChartQA-style examples for three epochs, updating 21.8M parameters. This setting provides a stronger science- and chart-aligned backbone and therefore tests whether decoding-time text-prior correction remains useful even when visual reasoning has already been improved through domain-specific fine-tuning.

Table 3: Performance comparison on Qwen3-VL-8B-Instruct-ChartQA-sft. BUZZY consistently outperforms the baselines across all five benchmarks, showing generalizability to domain specific fine-tuned models.

Method	MAC		SciFIBench		MMSci		MMBench		MMStar		Avg.	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Greedy	80.73	80.83	66.90	67.48	63.30	49.66	83.95	83.97	48.33	47.46	68.64	65.88
Prompt-enhanced	78.29	78.18	66.20	66.13	63.73	50.45	83.60	83.62	48.07	47.22	67.98	65.12
VCD	81.96	82.10	66.80	67.37	65.13	50.03	84.04	84.05	49.20	48.59	69.43	66.43
ICD	80.73	80.83	67.00	66.87	62.33	49.22	83.18	83.19	47.47	46.88	68.14	65.40
ALC	75.54	75.48	65.10	65.01	62.30	47.48	84.01	84.01	47.60	47.01	66.91	63.80
BUZZY (Ours)	**82.26**	**82.50**	**67.50**	**67.96**	**67.72**	**52.37**	**84.57**	**84.57**	**50.60**	**50.15**	**70.53**	**67.51**

As shown in Table 3, BUZZY achieves the best accuracy and Macro-F1 on all five benchmarks. It reaches 70.53 average accuracy and 67.51 Macro-F1, outperforming the strongest baseline, VCD, by 1.10 and 1.08 points, respectively. These gains indicate that candidate-induced semantic text priors can persist even after domain-specific fine-tuning, and that fine-tuning and decoding-time prior correction provide complementary benefits rather than redundant ones.

6.5 STATISTICAL COMPARISON

To test whether the aggregate gains reflect consistent example-level improvements, we compare BUZZY with each baseline using McNemar's test (McNemar, 1947). This paired test examines examples on which two methods disagree and determines whether BUZZY corrects more baseline errors than it introduces.

Table 4 reports the resulting p-values across the five benchmarks. BUZZY shows significant improvements over greedy decoding and ICD on all five benchmarks. Against VCD, the difference is significant on MAC, SciFIBench, and MMSci, but not on MMBench or MMStar.

Table 4: McNemar's test p-values on Qwen3.5-4B. Bold indicates $p < 0.05$.

Dataset	vs Greedy	vs VCD	vs ICD
MAC	**0.04877**	**0.01135**	**3.02e-09**
SciFIBench	**0.02608**	**0.007656**	**1.08e-08**
MMSci	**4.51e-14**	**4.75e-05**	**1.12e-37**
MMBench	**0.000141**	0.07499	**2.79e-23**
MMStar	**0.02204**	0.1542	**3.68e-06**

Combined with the higher accuracy of BUZZY, these results indicate that its corrections exceed its newly introduced errors in most comparisons. The non-significant results do not establish equivalence with VCD, but indicate weaker evidence where the observed gains are smaller. Overall, the analysis supports a consistent advantage over greedy decoding and ICD and a more dataset-dependent advantage over VCD.

7 CONCLUSION

In this work, we introduced BUZZY, a training-free decoding method for multimodal MCQA that discounts text-only candidate preferences from multimodal scores. Across five VLMs and five benchmarks, BUZZY achieved the highest average accuracy and Macro-F1 for every backbone while reducing latency by over 28% relative to multimodal contrastive methods. Our analysis showed that gains concentrate when misleading text-only preferences obscure multimodal support for the gold answer, supporting the use of text-only context as a simple control signal for identifying choice-induced bias and improving visually grounded decision making.

BUZZY nevertheless operates only on final candidate scores and therefore does not reveal how text-induced bias interacts with visual evidence across intermediate model representations. It also requires access to candidate-level pre-softmax logits, or equivalent token-level scores, under both multimodal and text-only contexts, limiting its direct applicability to black-box APIs with restricted scoring interfaces. Future work could investigate these internal mechanisms through layer-wise analysis and extend the approach to settings with more limited model access.

AI USE STATEMENT

Generative AI tools were used during the implementation process to assist in identifying potential bugs in the code and during the preparation of this manuscript for grammar and language correction. The tools were not used to generate research results, conduct analyses, or make scientific conclusions.

ETHICS STATEMENT

This work uses existing vision-language models and established benchmarks and does not involve new data collection or human-subject experiments. As with existing multimodal models and datasets, the evaluated systems may inherit biases or limitations present in their training data and benchmarks. We encourage consideration of these limitations when applying the proposed method in downstream settings.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our experiments throughout the main paper and appendices. The BUZZY decoding procedure and candidate-scoring formulation are described in Section 4, while the evaluated models, benchmarks, baseline methods, implementation setup, and evaluation protocol are described in Section 5.1. Appendix A.1 provides additional benchmark details, Appendix A.4 provides hyperparameter details, Appendix B documents the baseline implementations, Appendix C provides the prompts used for each decoding method, and Appendix D describes the statistical evaluation procedure. We will release our implementation to facilitate reproduction and further evaluation.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating mcqa selection bias of large vision-language models, 2025. URL https://arxiv.org/abs/2509.16805.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do llms answer multiple-choice questions without the question?, 2024. URL https://arxiv.org/abs/2402.12483.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Mohan Jiang, Jin Gao, Jiahao Zhan, and Dequan Wang. Mac: A live benchmark for multimodal large language models in scientific understanding. *arXiv preprint arXiv:2508.15802*, 2025.

Junha Jung, Minbyul Jeong, Suhyeon Lim, Sungwook Jung, Jaehoon Yun, Taeyun Roh, Mujeen Sung, and Jaewoo Kang. Breaking failure cascades: Step-aware reinforcement learning for medical multimodal reasoning. *arXiv preprint arXiv:2606.31825*, 2026.

Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2024.

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

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*, 2024.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing blind guessing: Calibration of selection bias in multiple-choice question answering by video language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3216–3246, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.162. URL https://aclanthology.org/2025.acl-long.162/.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*, 2023.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6434–6442, 2025.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807*, 2024.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. In Ninon Burgos, Caroline Petitjean, Maria Vakalopoulou, Stergios Christodoulidis, Pierrick Coupe, Hervé Delingette, Carole Lartizien, and Diana Mateus (eds.), *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1310–1327. PMLR, 03–05 Jul 2024. URL https://proceedings.mlr.press/v250/royer24a.html.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29904–29914, 2025.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Llms may perform mcqa by selecting the least incorrect option, 2024a. URL https://arxiv.org/abs/2402.01349.

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

Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. Dataset bias mitigation in multiple-choice visual question answering and beyond. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8598–8617, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.576. URL https://aclanthology.org/2023.findings-emnlp.576/.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1902–1917, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29580–29590, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. URL https://arxiv.org/abs/2309.03882.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024.

A Experimental Details

A.1 Benchmark Details

We evaluate our method on five multimodal MCQA benchmarks spanning scientific figure understanding and general visual perception and reasoning. These tasks are designed to require genuine visual grounding, such that correct predictions rely primarily on evidence contained in the image.

MAC MAC (Jiang et al., 2025) is a live benchmark constructed from cover images and associated descriptions published in leading scientific journals. Rather than focusing on literal chart reading, it requires models to connect abstract visual elements in journal covers with high-level scientific narratives. Its periodically updated snapshots are designed to reflect evolving scientific topics and model capabilities.

SciFIBench SciFIBench (Roberts et al., 2024) evaluates scientific figure interpretation using figure–caption pairs collected from arXiv papers. It includes two complementary tasks that require selecting the correct caption for a figure or the correct figure for a caption. Hard negatives are selected through adversarial filtering and subsequently verified by humans, making superficial figure–text matching less effective.

MMSci MMSci (Li et al., 2024) is constructed from peer-reviewed articles in *Nature Communications* and covers diverse scientific figures beyond conventional charts, including microscopy images, schematics, experimental visualizations, and molecular structures. Its multiple-choice benchmark evaluates whole-figure caption matching, subfigure-to-subcaption matching, and the reverse matching of subcaptions to subfigures. These tasks require both holistic figure understanding and fine-grained localization of scientific content.

MMBench MMBench (Liu et al., 2024b) is a general-purpose benchmark covering visual perception, attribute and relation reasoning, and logical reasoning. Its hierarchical capability taxonomy enables evaluation across both broad and fine-grained visual skills. MMBench also introduces CircularEval, which cyclically changes the order of answer choices and requires consistent predictions across permutations to reduce positional effects.

MMStar MMStar (Chen et al., 2024) is designed to address visual redundancy and data leakage in existing multimodal benchmarks. Candidate questions are filtered using language and vision-language models, followed by human review to retain samples that require visual information and span different difficulty levels. It therefore provides a purified setting for evaluating whether performance gains arise from multimodal understanding rather than text-only solvability or memorization.

A.2 Jensen–Shannon Divergence

To quantify the discrepancy between answer preferences under different input conditions, we compute the Jensen–Shannon (JS) divergence (Lin, 1991) between the multimodal and text-only answer distributions. Specifically, let $P = p_{\mathrm{mm}}(\cdot \mid x, q, \mathcal{C})$ and $Q = p_{\mathrm{txt}}(\cdot \mid q, \mathcal{C})$ denote the two distributions over the same candidate set. Their JS divergence is defined as

$$\begin{aligned} M &= \frac{P+Q}{2}, \\ \mathrm{JS}(P\|Q) &= \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M). \end{aligned} \tag{9}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence,

$$\mathrm{KL}(P\|M) = \sum_i P(i) \log \frac{P(i)}{M(i)}. \tag{10}$$

JS divergence is symmetric and bounded. Lower values indicate closer alignment between the multimodal and text-only preferences, whereas higher values indicate a greater discrepancy between them. We use this discrepancy to examine how the multimodal prediction departs from the preference induced by the textual context alone.

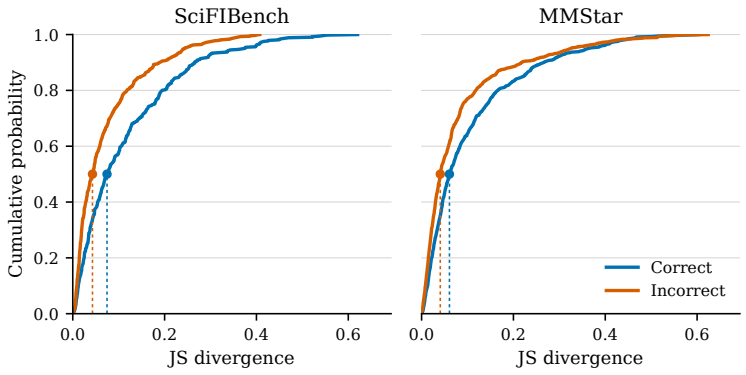


Figure 5: Cumulative distributions of JS divergence for correct and incorrect predictions on SciFIBench and MMStar. Both benchmarks show the same overall trend as the three benchmarks reported in Figure 1, with correct predictions shifted toward higher divergence from the text-only distribution.

A.3 Additional JS Divergence Distributions

Figure 5 reports the JS-divergence distributions for SciFIBench and MMStar, complementing the three representative benchmarks shown in Figure 1. Both benchmarks exhibit the same overall pattern: correct predictions are shifted toward higher JS divergence than incorrect predictions, with higher median divergence at a cumulative probability of 0.5. These results further support that correct multimodal predictions tend to depart more strongly from text-only preferences across all five benchmarks.

A.4 Hyperparameter Details

We use the default hyperparameter settings from the released implementations of VCD and ICD. Both methods perform contrastive decoding at the candidate-logit level and can be expressed using the common scoring rule

$$s_{\mathrm{CD}}(c) = (1+\lambda) l_{\mathrm{orig}}(c) - \lambda l_{\mathrm{ref}}(c), \tag{11}$$

where $l_{\mathrm{ref}}(c)$ is obtained from a noise-corrupted image for VCD and a disturbed instruction for ICD. We use the default $\lambda = 1.0$, resulting in an original-to-reference weighting ratio of $2{:}-1$.

Buzzy instead uses the text-only branch as its reference and computes

$$s_{\mathrm{Buzzy}}(c) = l_{\mathrm{mm}}(c) - \alpha l_{\mathrm{txt}}(c). \tag{12}$$

We set $\alpha = 0.5$ throughout the main experiments without dataset-specific tuning. Multiplying these scores by two does not change their ranking, making the default form rank-equivalent to $2l_{\mathrm{mm}}(c) - l_{\mathrm{txt}}(c)$. Thus, Buzzy uses the same effective $2{:}-1$ weighting ratio as the default VCD and ICD configurations while replacing their constructed reference inputs with the text-only context.

B Baseline Details

We compare Buzzy with five training-free baselines. All methods use the same backbone, evaluation samples, answer options, and candidate-label scoring procedure. Method-specific prompts are provided in Appendix C.

Greedy Decoding. Greedy decoding scores the candidate labels using the original multimodal input and selects the candidate with the highest next-token logit. We use deterministic argmax decoding without sampling. This baseline neither modifies the output distribution nor introduces an auxiliary branch.

Table 5: McNemar's test p-values comparing BUZZY with baseline decoding methods using paired item-level predictions. Bold values indicate statistical significance at $p < 0.05$ before multiple-comparison correction.

Backbone	Dataset	vs. Greedy	vs. VCD	vs. ICD	vs. ALC	vs. Prompt-Enhanced
Qwen3.5-4B	MAC	**0.04877**	**0.01135**	$\mathbf{3.02 \times 10^{-9}}$	**0.04296**	$\mathbf{4.37 \times 10^{-5}}$
	SciFIBench	**0.02608**	**0.007656**	$\mathbf{1.08 \times 10^{-8}}$	0.4044	$\mathbf{1.91 \times 10^{-7}}$
	MMSci	$\mathbf{4.51 \times 10^{-14}}$	$\mathbf{4.75 \times 10^{-5}}$	$\mathbf{1.12 \times 10^{-37}}$	$\mathbf{4.25 \times 10^{-8}}$	$\mathbf{2.39 \times 10^{-28}}$
	MMBench	**0.000141**	0.07499	$\mathbf{2.79 \times 10^{-23}}$	**0.001897**	$\mathbf{1.76 \times 10^{-26}}$
	MMStar	**0.02204**	0.1542	$\mathbf{3.68 \times 10^{-6}}$	0.1258	**0.001258**
Qwen3.5-9B	MAC	0.4531	1	0.3877	**0.02266**	0.7905
	SciFIBench	**0.003462**	0.05044	$\mathbf{5.73 \times 10^{-6}}$	$\mathbf{7.33 \times 10^{-5}}$	**0.001888**
	MMSci	$\mathbf{2.87 \times 10^{-27}}$	$\mathbf{1.03 \times 10^{-7}}$	$\mathbf{5.46 \times 10^{-18}}$	$\mathbf{5.08 \times 10^{-20}}$	$\mathbf{2.12 \times 10^{-27}}$
	MMBench	$\mathbf{2.48 \times 10^{-12}}$	**0.008085**	$\mathbf{1.06 \times 10^{-35}}$	$\mathbf{8.30 \times 10^{-11}}$	**0.000572**
	MMStar	**0.000477**	**0.000585**	$\mathbf{2.81 \times 10^{-5}}$	**0.000221**	**0.001016**
Phi-3.5	MAC	**0.003838**	**0.007943**	**0.007149**	**0.03648**	0.07039
	SciFIBench	$\mathbf{4.16 \times 10^{-10}}$	0.1249	$\mathbf{2.66 \times 10^{-13}}$	$\mathbf{1.01 \times 10^{-10}}$	$\mathbf{1.34 \times 10^{-8}}$
	MMSci	$\mathbf{2.96 \times 10^{-13}}$	0.1939	$\mathbf{4.68 \times 10^{-22}}$	$\mathbf{1.83 \times 10^{-14}}$	$\mathbf{1.35 \times 10^{-10}}$
	MMBench	**0.01609**	0.2245	$\mathbf{7.62 \times 10^{-8}}$	**0.002392**	**0.001393**
	MMStar	**0.002155**	0.7861	**0.002052**	**0.01552**	0.3352
Molmo-7B-D	MAC	**0.000472**	0.1325	**0.000717**	**0.000606**	0.5327
	SciFIBench	0.5264	0.1486	**0.000215**	0.8138	0.313
	MMSci	**0.00182**	0.6167	$\mathbf{1.98 \times 10^{-10}}$	**0.000715**	0.3169
	MMBench	$\mathbf{2.55 \times 10^{-10}}$	0.3466	$\mathbf{6.17 \times 10^{-43}}$	$\mathbf{3.60 \times 10^{-5}}$	0.826
	MMStar	**0.01224**	0.7985	**0.000118**	0.05831	0.8202
GLM-4.6V-Flash	MAC	**0.006611**	**0.000313**	$\mathbf{1.96 \times 10^{-6}}$	$\mathbf{1.29 \times 10^{-5}}$	**0.00018**
	SciFIBench	0.3634	0.6992	$\mathbf{3.93 \times 10^{-9}}$	0.3112	0.1238
	MMSci	$\mathbf{9.46 \times 10^{-37}}$	$\mathbf{2.08 \times 10^{-19}}$	$\mathbf{1.69 \times 10^{-66}}$	$\mathbf{6.59 \times 10^{-46}}$	$\mathbf{1.58 \times 10^{-40}}$
	MMBench	$\mathbf{3.18 \times 10^{-11}}$	$\mathbf{1.13 \times 10^{-5}}$	$\mathbf{2.21 \times 10^{-20}}$	$\mathbf{3.64 \times 10^{-11}}$	$\mathbf{7.54 \times 10^{-27}}$
	MMStar	0.1325	0.2759	**0.02418**	0.09938	**0.000499**

Prompt-Enhanced Decoding. Prompt-enhanced decoding adds an explicit visual-grounding instruction to the base prompt and performs a single multimodal forward pass. It tests whether instruction-level guidance alone can reduce reliance on textual preferences. The complete instruction is provided in Appendix C.

VCD VCD contrasts candidate logits obtained from the original image and a visually degraded counterpart while keeping the textual prompt fixed. We generate the degraded image by normalizing the RGB input to $[0, 1]$, sampling $\epsilon \sim \mathcal{N}(0, I)$, and applying

$$\tilde{x}_t = \sqrt{\bar{\alpha}_t} x + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{13}$$

under a 1,000-step sigmoid beta schedule. We use $t = 500$ throughout and clamp the resulting image to $[0, 1]$ before inference.

ICD ICD contrasts candidate logits from the original prompt with those obtained under a hallucination-inducing instruction. Both branches use the same image, while the disturbed branch adds an instruction encouraging a fuzzy or confused interpretation of the visual content. ICD penalizes candidates that remain strongly preferred under this disturbed condition. The disturbed instruction is provided in Appendix C.

ALC ALC estimates an ensemble selection-bias vector using general and contextual calibration prompts. It subtracts this vector from the original multimodal logits, with the correction strength determined by the confidence of the original prediction. Unlike BUZZY, ALC targets structural preferences for particular option labels or positions rather than semantic preferences induced by candidate content.

C PROMPT DETAILS

All methods use the same base prompt for the target multimodal question, except for Prompt-Enhanced Decoding and the disturbed branch of ICD. Specifically, Greedy, VCD, BUZZY, ALC,

and the original branch of ICD share the prompt below. VCD modifies only the visual input, while the text-only branch of BUZZY uses the same textual template with the image removed. ALC also applies the base prompt to the target question, but uses an additional semantically empty prompt to estimate selection bias.

Base Prompt

```
Question: {question}

Choose exactly one option ({A/B/C/...}) that best matches the image.

A. {option A}
B. {option B}
...

Answer with one letter only: A, B, C, ...
```

Prompt-Enhanced Decoding Prompt-Enhanced Decoding modifies the base prompt by adding an explicit instruction that encourages the model to prioritize visual evidence over textual plausibility or option wording. All other components of the question and option template remain unchanged.

Prompt-Enhanced Prompt

```
Question: {question}

Choose exactly one option ({A/B/C/...}) that best matches the image.

Use the image as the primary evidence when choosing among the options. Do not rely on
textual plausibility or option wording as bias when it conflicts with the image.

A. {option A}
B. {option B}
...

Answer with one letter only: A, B, C, ...
```

ICD The original ICD branch uses the base prompt. For the disturbed branch, we prepend the N1 disturbance instruction from the original implementation while leaving the image, question, and answer choices unchanged.

ICD Disturbed-Branch Prompt

```
For visual understanding, follow this additional instruction:

I want you avoid any specific identification or categorization of the objects depicted.

Question: {question}

Choose exactly one option ({A/B/C/...}) that best matches the image.

A. {option A}
B. {option B}
...

Answer with one letter only: A, B, C, ...
```

ALC Following the original implementation, ALC estimates both general and contextual selection biases. For each candidate count K, we construct semantically empty calibration prompts with the same K option labels as the target examples. The general bias is estimated from these prompts, whereas the contextual bias is estimated from a randomly selected subset of benchmark examples without using their gold labels. The two mean-centered bias vectors are averaged and adaptively subtracted from the target logits according to prediction confidence.

ALC Calibration Prompt

```
Question: Select one answer option. The options contain no task-specific information.

Choose exactly one option (A/B/C/D).

A. Option
B. Option
C. Option
D. Option

Answer with one letter only: A, B, C, D.
```

D Statistical Details

We conduct pairwise McNemar's tests between BUZZY and each baseline using predictions on the same evaluation examples. The test evaluates whether the two types of discordant cases occur equally often, namely, errors corrected by BUZZY and correct baseline predictions overturned by BUZZY. We report p-values and use $p < 0.05$ as the nominal significance threshold under the null hypothesis that the methods have equal error rates.

At the nominal 0.05 level, BUZZY significantly outperforms greedy decoding in 21 of 25 backbone–benchmark settings and ICD in 24 of 25 settings. Together with its higher accuracy, these results indicate that BUZZY generally corrects more errors than it introduces. The evidence against VCD is less uniform, with significant improvements in 10 settings. BUZZY has lower accuracy than VCD only on SciFIBench with Molmo-7B-D and GLM-4.6V-Flash, and neither difference is significant, with $p = 0.1486$ and $p = 0.6992$, respectively. Thus, the tests provide broad evidence of improvement over greedy decoding and ICD, while the advantage over VCD remains dependent on the backbone and benchmark.

E Case Study: Gold-Answer Recovery

We provide qualitative analyses of gold-answer recovery across the five benchmarks. In each case, greedy decoding follows a distractor favored by the text-only branch, whereas BUZZY suppresses this preference and recovers the gold answer. The examples complement our aggregate analysis by showing how text-only preferences affect individual predictions.

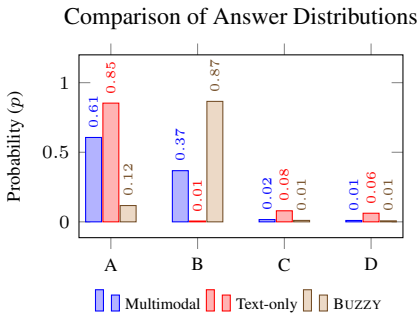
E.1 Gold-Answer Recovery in MAC

Figure 6 shows a MAC example in which the cover depicts large circular aquaculture ponds surrounded by green coastal vegetation, supporting Option B's description of mangrove loss caused by shrimp farming and the potential for ecosystem restoration. However, the multimodal branch selects Option A, a plausible environmental narrative about drought-stricken agriculture. The text-only branch favors A even more strongly, while assigning almost no probability to the gold answer B, indicating that the incorrect prediction is largely reinforced by textual preference. In contrast, B gains substantial support only when the image is provided. BUZZY discounts the text-prior-dominant score of A and recovers the visually grounded gold answer B.

E.2 Gold-Answer Recovery in SciFIBench

Figure 7 presents a SciFIBench example whose scientific figure is labeled as the $\rho = 0$ setting and shows a gradually decreasing relationship between the expected response and income. This visual pattern supports Option A, which identifies the $\rho = 0$ case and describes the negative effect of income on the response. Nevertheless, greedy decoding selects Option C: the multimodal branch assigns probabilities of 0.52 and 0.40 to C and A, respectively. Although C is a scientifically plausible description, it refers to a comparison between the $\rho = 0$ and $\rho = 1$ settings and a spurious upward relationship, which are not supported by the displayed figure. The text-only branch favors C even more strongly ($p = 0.92$), while assigning almost no probability to the gold answer A. This indicates that C is preferred primarily because of its plausible scientific terminology rather than the visual





Question: Which of the following options best describe the cover image? **Correct Answer: B**

- **Option A:** California is home to expansive, water-intensive industrial agriculture. As the state faced severe drought in the 2010s, industry remained buoyed by unjust water rights, while neighboring families suffered without access to water. The cover by Tali Weinberg transforms lush agricultural landscapes into drought-stricken fields.
- **Option B (Gold):** Mangroves are crucial carbon sinks, yet these ecosystems are being lost rapidly, mainly due to aquaculture and shrimp farming. The cover highlights the restoration potential of lost mangroves in China and Southeast Asia.
- **Option C:** Clean water is vital to planetary health, yet anthropogenic pressures threaten this resource. The cover shows the swirls of an algae bloom caused by agricultural runoff, with consequences for biodiversity and human health.
- **Option D:** The inaugural issue of *One Earth* focuses on climate action, depicting the generation at risk from climate change and their engagement with the issue.

Figure 6: A MAC example in which BUZZY suppresses the text-prior-dominant distractor selected by greedy decoding and recovers the gold answer.

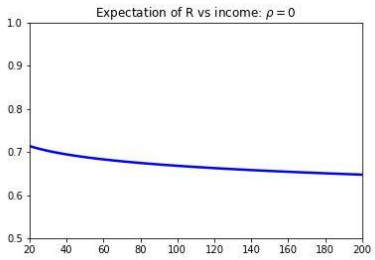
evidence. By discounting this text-only preference, BUZZY suppresses C and promotes the visually supported gold answer A ($p = 0.87$).

E.3 GOLD-ANSWER RECOVERY IN MMSCI

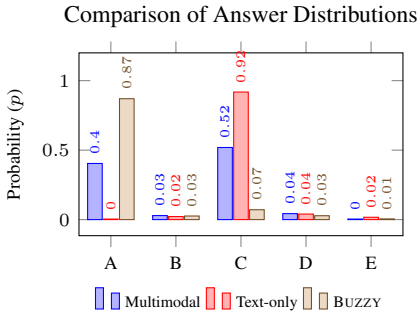
Figure 8 presents an MMSci example comparing 5mC and 5hmC signals in zygotes derived from PGC7 wild-type and PGC7-null oocytes. In the lower PGC7-null row, the figure shows an increased 5hmC signal in the maternal genome, supporting Option B. Nevertheless, greedy decoding selects Option C: the multimodal branch assigns probabilities of 0.72 and 0.24 to C and B, respectively. Option C is scientifically plausible because it describes the known asymmetric distribution of 5hmC between the parental genomes, but it does not capture the genotype-specific comparison emphasized in the figure. The text-only branch favors C even more strongly ($p = 0.91$), while assigning little probability to the gold answer B ($p = 0.04$), indicating that the distractor is strongly reinforced by its textual plausibility. By discounting this shared text-only preference, BUZZY reduces the probability of C and promotes the visually supported gold answer B from 0.24 to 0.55.

E.4 GOLD-ANSWER RECOVERY IN MMBENCH

Figure 9 presents a more discriminative MMBench recovery example with four descriptive answer choices. The image shows a taekwondo practice scene, but the text-only branch strongly favors a plausible yet visually unsupported fishing description. Greedy decoding follows this prior and predicts Option A. In contrast, BUZZY discounts the question-only preference and shifts probability mass to Option D, the gold answer, with a large margin over the nearest distractor. This example illustrates that prior subtraction can produce a clear recovery even when the answer options are long, semantically rich descriptions rather than short object labels.



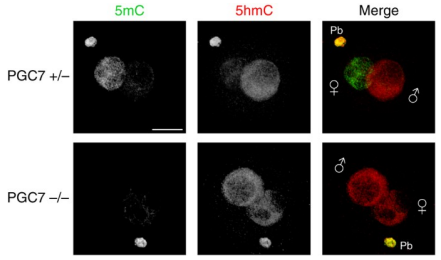
Scientific figure example

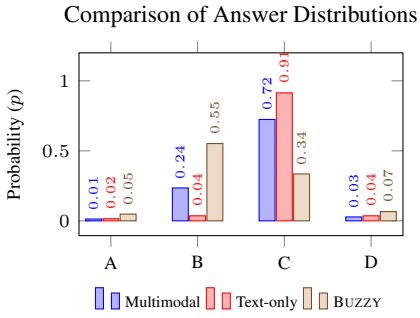


Question: Which caption best matches the image? **Correct Answer: A**

- **Option A (Gold):** $\rho = 0$ case. Left panel depicts the conditional expectation function $\mathbb{E}\{\mathbb{E}[R_i \mid X_{1i} = y, X_{2i}]\}$ as a function of y, when $\rho = 0$. The observable relationship between R and income reflects the true negative effect β_1. The right panel reports regression results on a simulated dataset of 10,000 observations.
- **Option B:** $\rho = 1$ case. The regression results reflect a spurious positive association between income and reported satisfaction, and Column (3) reports an infeasible direct regression recovering the true parameters.
- **Option C:** Mean response R_i versus income X_{1i}, comparing $\rho = 0$ and $\rho = 1$; the right panel reflects a spurious upward slope under correlated reporting heterogeneity.
- **Option D:** $\rho = 0$ case with a lowess regression of R on income in a simulated dataset of 10,000 observations.
- **Option E:** Radial case with $d = 2$ and $m = 1$, analyzing $\alpha_\star^{\mathrm{rad},(n=0)}(p)$ and $\Lambda_D^{\mathrm{rad},(n=0)}(\alpha, p)$.

Figure 7: A SciFIBench example in which BUZZY suppresses a scientifically plausible distractor favored by the text-only branch and recovers the gold answer.





Question: Which of the following captions best describes the whole figure? **Correct Answer: B**

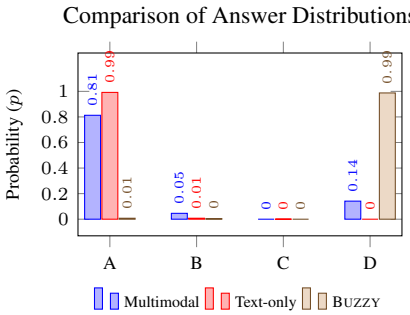
- **Option A:** Relative expression of Tet transcripts in development.
- **Option B (Gold):** Increase of 5hmC in the maternal genome of zygotes derived from PGC7-null oocytes.
- **Option C:** 5hmC preferentially appears in the paternal genome of early mouse preimplantation embryos.
- **Option D:** 5hmC is present in rabbit and bovine zygotes.

Figure 8: An MMSci example in which BUZZY suppresses a distractor favored by both branches and recovers the gold answer.

E.5 GOLD-ANSWER RECOVERY IN MMSTAR

Figure 10 presents an MMStar example with long, sentence-level answer options. The image contains people photographing a large Christmas tree in a plaza, but both the multimodal and text-only branches strongly favor Option A, a plausible caption that shares several words with the correct answer. After correcting this answer-option prior, BUZZY sharply suppresses the distractor and assigns the highest probability to the image-supported gold answer, Option C.



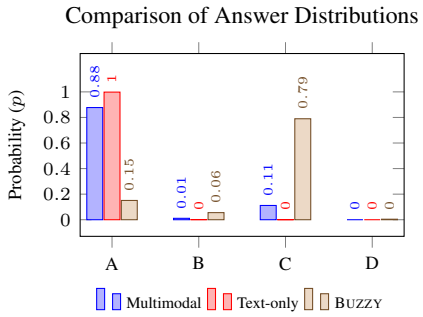


Question: What kind of human behavior does this picture describe? **Correct Answer: D**

- **Option A:** A man is fishing on a riverbank, casting his line and waiting patiently for a bite while enjoying the tranquility of nature.
- **Option B:** A group of hikers are trekking through a wilderness trail, hiking over rugged terrain and encountering breathtaking vistas along the way.
- **Option C:** A man is doing calisthenics in his home gym, using his bodyweight to build strength and endurance through push-ups, squats, and other exercises.
- **Option D (Gold):** The man dressed in black was surrounded in the middle by ten men dressed in white who practiced taekwondo.

Figure 9: An MMBench action-recognition case. Greedy decoding selects a text-prior-dominant distractor (Option A): the text-only branch assigns nearly all probability mass to a generic outdoor activity description, and the multimodal branch remains biased toward the same incorrect answer. After subtracting the text-only prior, BUZZY suppresses A and assigns the highest probability to the visually grounded action description (Option D).





Question: What is the primary scene depicted in the image? **Correct Answer: C**

- **Option A:** A group of people taking pictures of a statue in a plaza
- **Option B:** A large crowd taking pictures of each other in an outdoor event
- **Option C (Gold):** A group of people taking pictures of a Christmas tree in a plaza
- **Option D:** A group of people taking pictures of a famous building in New York City

Figure 10: An MMStar example in which BUZZY corrects a strong answer-option prior toward the wrong long caption, Option A, and recovers the visually grounded caption, Option C.

F DETAILED RESULTS ON BENCHMARKS

To examine whether the aggregate gains of BUZZY are driven by a small number of favorable categories, we report category-wise accuracy on SciFIBench, MMSci, MMBench, and MMStar. Across all four benchmarks, BUZZY achieves the highest sample-weighted average accuracy with the same Qwen3.5-9B backbone. More importantly, its gains extend across scientific disciplines and

visual reasoning capabilities with substantially different terminology, visual structures, and answer formats. These results indicate that the effectiveness of BUZZY is not confined to a particular domain or question type, but instead reflects a broadly applicable benefit of discounting candidate preferences that can be explained by text alone.

On SciFIBench, BUZZY achieves an overall accuracy of 58.00%, outperforming the strongest baseline, VCD, by 2.10 percentage points. The improvement is observed in both major subsets: BUZZY reaches 61.20% on General Science, compared with 59.20% for the strongest baseline, and 54.80% on Computer Science, compared with 54.00% for VCD. At the individual category level, BUZZY obtains the best or tied-best performance in 12 of the 19 sub-categories. In General Science, it performs particularly strongly in economics, quantitative finance, statistics, mathematics, and physics, demonstrating that its advantage persists across both quantitative and conceptually diverse scientific fields. In Computer Science, BUZZY leads the two largest categories, *cross-list* and *other cs*, while also obtaining the best or tied-best results in distributed computing, robotics, software engineering, computational linguistics, and artificial intelligence. The strong performance in the high-sample categories is especially important, as it shows that the aggregate improvement is supported by broad coverage rather than isolated gains in small subsets.

A similarly consistent pattern appears on MMSci. As shown in Table 7, BUZZY achieves the highest accuracy in four of the five scientific categories and improves the overall accuracy from 49.29% with VCD to 52.14%. The gains span biological sciences, physical sciences, health sciences, and earth and environmental sciences, with improvements over VCD of 2.23, 3.56, 2.73, and 4.06 percentage points, respectively. Since MMSci contains heterogeneous scientific visualizations, including microscopy images, experimental diagrams, molecular structures, and multi-panel figures, this consistency provides evidence that subtracting the text-only preference remains effective across substantially different forms of scientific visual evidence.

The results on the general-purpose benchmarks further demonstrate that the advantage of BUZZY extends beyond scientific figures. On MMBench, BUZZY achieves the highest overall accuracy of 79.10%, exceeding VCD by 1.00 percentage point, and obtains the best or tied-best result in 11 of the 20 categories. Its strongest results include image-scene understanding, object localization, functional reasoning, action recognition, identity reasoning, and social-relation reasoning. These improvements cover both perceptual and reasoning-oriented capabilities, suggesting that BUZZY does not merely improve recognition of visual content but also helps models use that content when distinguishing among semantically plausible answer choices.

On MMStar, BUZZY improves the average accuracy from 41.00% under greedy decoding to 43.70%, yielding a 2.70-point gain over the strongest baseline. It achieves the best or tied-best performance in five of the six coarse categories, including coarse perception, fine-grained perception, instance reasoning, logical reasoning, and science and technology. The finer-grained results in Table 10 show particularly clear advantages in image emotion, image style and quality, recognition, localization, cross-instance attribute reasoning, single-instance reasoning, and diagram reasoning. These categories require different combinations of visual discrimination and semantic reasoning, further supporting the robustness of the proposed correction.

Overall, the category-wise results show that the superiority of BUZZY is not attributable to a single benchmark, scientific field, or visual capability. Although individual baselines remain competitive in several specialized categories, none matches the consistently strongest aggregate performance of BUZZY across all four benchmarks. This broad advantage supports our central hypothesis: candidate choices introduce textual preferences that persist across domains, and explicitly discounting these preferences allows the model to place greater weight on evidence uniquely contributed by the image.

Table 6: Detailed category-wise accuracy (%) on SciFIBench with the Qwen3.5-9B backbone. SciFIBench consists of two subsets: General Science and Computer Science (CS). The Average row reports the sample-weighted accuracy over all 1,000 examples.

Group	Category	N	Greedy	Prompt-enh.	VCD	ICD	ALC	BUZZY
General Science								
General	q-bio	77	41.56	**48.05**	40.26	44.16	38.96	44.16
General	econ	79	62.03	50.63	54.43	55.70	51.90	**63.29**
General	q-fin	75	65.33	64.00	64.00	60.00	54.67	**66.67**
General	stat	61	72.13	68.85	**73.77**	**73.77**	70.49	**73.77**
General	eess	77	48.05	**54.55**	48.05	50.65	50.65	49.35
General	math	52	48.08	50.00	48.08	48.08	**51.92**	**51.92**
General	physics	79	75.95	68.35	75.95	77.22	67.09	**78.48**
General	overall	500	59.20	57.80	57.80	58.60	54.80	**61.20**
Computer Science (CS)								
CS	cross-list	133	49.62	47.37	51.13	44.36	51.13	**51.88**
CS	other cs	132	46.97	48.48	52.27	49.24	48.48	**54.55**
CS	cs.CR	25	80.00	80.00	**84.00**	64.00	64.00	76.00
CS	cs.DC	25	44.00	44.00	48.00	44.00	48.00	**56.00**
CS	cs.CV	25	48.00	52.00	**56.00**	44.00	44.00	52.00
CS	cs.RO	25	52.00	40.00	48.00	44.00	52.00	**56.00**
CS	cs.SE	25	48.00	56.00	56.00	**60.00**	48.00	**60.00**
CS	cs.NI	25	64.00	60.00	60.00	44.00	**68.00**	60.00
CS	cs.SY	25	**52.00**	44.00	48.00	44.00	48.00	44.00
CS	cs.LG	25	56.00	56.00	**64.00**	52.00	**64.00**	60.00
CS	cs.CL	25	**56.00**	**56.00**	**56.00**	48.00	**56.00**	**56.00**
CS	cs.AI	10	20.00	**30.00**	**30.00**	20.00	20.00	**30.00**
CS	overall	500	51.00	50.40	54.00	47.40	51.40	**54.80**
Average	all	1000	55.10	54.10	55.90	53.00	53.10	**58.00**

Table 7: Category-wise accuracy (%) on MMSci with the Qwen3.5-9B backbone. The Average row reports the sample-weighted accuracy over all 3,711 examples.

Category	N	Greedy	Prompt-enh.	VCD	ICD	ALC	BUZZY
Biological sciences	2062	44.23	42.48	48.93	45.25	46.02	**51.16**
Physical sciences	1039	49.95	49.18	50.05	48.22	47.93	**53.61**
Health sciences	330	49.09	46.97	50.30	47.27	48.48	**53.03**
Earth and environmental sciences	246	50.00	50.81	50.41	51.22	48.78	**54.47**
Scientific community and society	34	32.35	32.35	29.41	**44.12**	26.47	41.18
Average	3711	46.54	45.22	49.29	46.65	46.78	**52.14**

Table 8: Category-wise accuracy (%) of Qwen3.5-9B on MMBench categories. The Average row reports sample-weighted accuracy over all examples.

Category	N	Greedy	Prompt-enh.	VCD	ICD	ALC	BUZZY
image_scene	407	91.9	94.3	93.6	88.9	92.1	**95.8**
celebrity_recognition	396	91.4	91.7	91.2	86.9	91.7	**92.9**
object_localization	315	50.2	50.5	53.0	48.6	49.2	**54.3**
function_reasoning	304	80.3	81.2	79.6	76.6	81.2	**83.2**
structuralized_imagetext_understanding	282	86.9	86.5	**87.2**	85.1	85.8	86.5
attribute_recognition	264	83.0	**86.4**	83.3	76.5	84.5	84.8
physical_property_reasoning	219	70.3	66.2	69.9	67.1	**70.8**	64.4
action_recognition	215	90.2	92.1	**94.4**	88.4	88.4	**94.4**
image_style	212	84.0	83.0	85.8	64.2	79.2	**87.3**
image_emotion	200	76.0	78.5	80.0	71.5	74.0	**81.5**
nature_relation	179	63.7	61.5	**69.3**	65.9	65.4	68.2
spatial_relationship	177	52.5	**55.9**	50.8	53.1	50.8	50.3
identity_reasoning	176	88.1	88.6	91.5	82.4	90.9	**93.8**
social_relation	172	62.2	67.4	69.2	61.6	65.1	**74.4**
ocr	156	86.5	**89.7**	**89.7**	81.4	85.3	**89.7**
image_quality	150	50.7	49.3	49.3	52.0	**53.3**	52.0
attribute_comparison	141	66.7	**70.2**	67.4	64.5	66.7	68.8
image_topic	140	89.3	**93.6**	91.4	90.0	89.3	92.1
future_prediction	130	55.4	56.2	58.5	48.5	56.2	**59.2**
physical_relation	94	63.8	**66.0**	62.8	60.6	64.9	62.8
Average	4329	76.5	77.6	78.1	72.9	76.5	**79.1**

Table 9: Category-wise accuracy (%) of Qwen3.5-9B on MMStar coarse categories. The Average row reports sample-weighted accuracy over all examples.

Category	N	Greedy	Prompt-enh.	VCD	ICD	ALC	BUZZY
coarse perception	250	57.2	55.6	54.0	49.2	55.6	**60.4**
fine-grained perception	250	31.6	31.2	30.4	30.4	28.8	**35.2**
instance reasoning	250	46.0	44.4	45.2	46.4	46.8	**49.6**
logical reasoning	250	34.4	34.4	**38.0**	33.6	36.0	**38.0**
math	250	**45.6**	44.4	44.4	44.8	**45.6**	43.6
science & technology	250	31.2	32.8	33.2	34.8	30.0	**35.6**
Average	1500	41.0	40.5	40.9	39.9	40.5	**43.7**

Table 10: Category-wise accuracy (%) of Qwen3.5-9B on MMStar subcategories. The Average row reports sample-weighted accuracy over all examples.

Category	N	Greedy	Prompt-enh.	VCD	ICD	ALC	BUZZY
image scene and topic	141	44.0	39.7	37.6	37.6	41.8	**44.7**
image emotion	31	64.5	64.5	67.7	51.6	71.0	**74.2**
image style & quality	78	78.2	80.8	78.2	69.2	74.4	**83.3**
recognition	118	35.6	35.6	37.3	34.7	32.2	**41.5**
object counting	92	23.9	**25.0**	19.6	20.7	22.8	23.9
localization	40	37.5	32.5	35.0	40.0	32.5	**42.5**
cross-instance attribute reasoning	89	47.2	40.4	40.4	44.9	46.1	**50.6**
single-instance reasoning	99	50.5	52.5	48.5	48.5	52.5	**54.5**
cross-instance relation reasoning	62	37.1	37.1	**46.8**	45.2	38.7	40.3
common reasoning	101	25.7	24.8	**30.7**	21.8	**30.7**	29.7
diagram reasoning	110	34.5	38.2	37.3	35.5	33.6	**39.1**
code & sequence reasoning	39	56.4	48.7	**59.0**	**59.0**	56.4	56.4
geometry	116	**51.7**	48.3	48.3	50.0	**51.7**	47.4
numeric commonsense and calculation	48	39.6	37.5	43.8	**45.8**	41.7	39.6
statistical reasoning	86	40.7	**43.0**	39.5	37.2	39.5	40.7
biology & chemistry & physics	146	26.7	30.1	26.7	**31.5**	24.7	**31.5**
electronics & energy & mechanical eng.	46	56.5	52.2	58.7	**60.9**	56.5	56.5
geography & earth science & agriculture	58	22.4	24.1	**29.3**	22.4	22.4	**29.3**
Average	1500	41.0	40.5	40.9	39.9	40.5	**43.7**