This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FRAMES-VQA: Benchmarking Fine-Tuning Robustness across Multi-Modal Shifts in Visual Question Answering

Chengyue Huang* Brisa Maneechotesuwan* Shivang Chopra Zsolt Kira
Georgia Institute of Technology
{chuang475,bmaneech3,shivangchopra11,zkira}@gatech.edu

Abstract

*Visual question answering (VQA) systems face significant challenges when adapting to real-world data shifts, especially in multi-modal contexts. While robust fine-tuning strategies are essential for maintaining performance across in-distribution (ID) and out-of-distribution (OOD) scenarios, current evaluation settings are primarily unimodal or particular to some types of OOD, offering limited insight into the complexities of multi-modal contexts. In this work, we propose a new benchmark FRAMES-VQA (**F**ine-Tuning **R**obustness **A**cross **M**ulti-Modal **S**hifts in VQA) for evaluating robust fine-tuning for VQA tasks. We utilize ten existing VQA benchmarks, including VQAv2, IV-VQA, VQA-CP, OK-VQA and others, and categorize them into ID, near and far OOD datasets covering uni-modal, multi-modal and adversarial distribution shifts. We first conduct a comprehensive comparison of existing robust fine-tuning methods. We then quantify the distribution shifts by calculating the Mahalanobis distance using uni-modal and multi-modal embeddings extracted from various models. Further, we perform an extensive analysis to explore the interactions between uni- and multi-modal shifts as well as modality importance for ID and OOD samples. These analyses offer valuable guidance on developing more robust fine-tuning methods to handle multi-modal distribution shifts. The code is available at https://github.com/chengyuehuang511/FRAMES-VQA.*

1. Introduction

Robust fine-tuning methods aim to adapt pre-trained models to downstream tasks while retaining resilience to distribution shifts [38, 52]. Distribution shifts in the image modality are widely explored, with various datasets designed to evaluate a model's generalization across diverse visual conditions. For example, DomainNet [37] spans multiple visual domains including real images, sketches, paintings, and clipart, challenging models to generalize across different styles and representations. Similarly, various ImageNet variants [10, 22, 39, 50] introduce shifts through image variations, adversarial examples, rendering transformations, and changes in texture or background. Collectively, these datasets provide a comprehensive framework for assessing how well models withstand visual distribution changes.

While robust fine-tuning algorithms are widely examined under distribution shifts in a single modality (images), few studies have explored robust fine-tuning for VQA tasks, where distribution shifts are multi-modal and models must adapt to variations across both visual and textual inputs. Apart from visual shift [1], there are question shifts [15, 41] involving variations in phrasing, structure, or vocabulary, as well as answer shifts [2] with changes in answer distributions such as frequency and formatting. Beyond uni-modal shift, these variations may occur simultaneously across visual, question, and answer inputs [9, 31, 42, 44, 49], posing an even greater challenge as models must generalize across complex, combined shifts.

Therefore, we build upon our preliminary exploration [25] and propose a benchmark FRAMES-VQA (**F**ine-Tuning **R**obustness **A**cross **M**ulti-Modal **S**hifts in VQA) to systematically evaluate the robustness of fine-tuning in VQA task. We leverage ten existing VQA datasets and categorize distribution shifts into uni-modal and multi-modal types, quantified by Mahalanobis distance across various backbones to capture both near and far OOD scenarios. We conduct a comprehensive comparison of the existing robust fine-tuning baselines on ID and OOD performance using the benchmark. Furthermore, we analyze shift scores and modality importance across fine-tuning methods. To summarize, our contributions are:

- We propose FRAMES-VQA for evaluating robust fine-tuning in VQA, including ten VQA datasets categorized by uni-modal (e.g., image, question) and multi-modal shifts. We quantify dataset shifts under different modalities using Mahalanobis distance and embeddings from different backbones.
- We perform an in-depth comparison of robust fine-tuning

*Equal contribution.

methods using the benchmark. We find that FTP [47] has the best far OOD performance while SPD [48] outperforms others on ID, near and average OOD.

- We further provide insightful analyses on the shift scores and modality importance of each baseline across ID and OOD samples. Key observations include: (i) Fine-tuning amplifies question-joint shift correlation, indicating strong influence of question shifts towards multi-modal representations; (ii) more robust fine-tuning methods exhibit low correlation between uni-modal and multimodal shifts; (iii) question-to-image attention rises for OOD samples, implying potential shortcuts; and (iv) robust methods emphasize intra- over inter-modality attention, underscoring intra-modality's role in robustness. These findings offer ways to improve fine-tuning robustness under multi-modal shifts.

2. Related Work

Distribution Shift and OOD Robustness in VQA. DomainNet [37] and ImageNet [10] along with its four variants [21, 22, 39, 50] are commonly used to assess model robustness under distribution shifts. Prior work [38, 52] show that while vanilla fine-tuning enhances ID performance, it can degrade results on OOD datasets compared to the pre-trained model. Beyond traditional image classification with distribution shifts in the image modality, VQA datasets such as VQAv2 [19] and its variants [1, 2, 9, 15, 31, 41, 42, 44, 49] introduce both uni-modal (image, question, answer) and multi-modal shifts, posing a greater challenge for models compared to shifts in image-only classification tasks. Various frameworks have been proposed to assess robustness in VQA. [3] explores cross-dataset evaluations across four VQA datasets, while [30, 34] expand this scope by incorporating VQAv2 variants and distinguishing different distribution shift types. [53] quantifies uni-modal shifts for image and question modalities. Building on these studies, we introduce finer categorization and distinctions between near and far OOD distributions, along with quantifying both uni- and multi-modal distribution distances. Crucially, while prior work has focused on testing different discriminative models [34] or adaptation methods [8], our study focuses on comparing robust fine-tuning algorithms within the same backbone and extends to generative models.

Robust Fine-Tuning of Foundation Models. Robust fine-tuning methods aim to adapt foundation models to new tasks while retaining their pre-trained robustness. LP-FT [28] introduces a two-step approach of linear probing followed by full fine-tuning to mitigate feature distortion. WiSE-FT [52] blends pre-trained and fine-tuned weights through interpolation, balancing the strengths of both embedding spaces. L2-SP [32] and MARS-SP [17] add penalties on the deviation between fine-tuned and pre-trained weights, exploring different norm types. More recent methods, such as TPGM [46], frame regularization as a constraint through bi-level optimization, learning tailored constraints for each layer. FTP [47] enhances TPGM's efficiency by leveraging prior training steps, while SPD [48] selectively regularizes layers with consistent loss reduction, projecting corresponding layers within the constraint.

3. FRAMES-VQA: Fine-Tuning Robustness across Multi-Modal Shifts in VQA

We propose a new setting FRAMES-VQA for benchmarking fine-tuning robustness across multi-modal shifts in VQA. We conduct a comprehensive categorization and experimentation on various VQA datasets, chosen to span different types of distribution shifts. We then evaluate the model on nine OOD datasets which are not included in the training with image, question, answer, multi-modal and adversarial distribution shift. We further distinguish between near and far OOD, concepts initially defined in OOD detection [14, 51], where near OOD represents data that is perceptually similar but semantically dissimilar to the training distribution, while far OOD refers to data that is both perceptually and semantically dissimilar. In FRAMES-VQA, we categorize six near OOD datasets that exhibit various types of distribution shifts relative to VQAv2, and three far OOD datasets where both image and text sources differ from those in VQAv2. Additional results using GQA [26] and GQA-OOD [27] can be found in Suppl. 14.

3.1. Datasets

Fig. 1 and Tab. 1 provide an overview and statistics of all ID and OOD VQA datasets with vision, question, answer, multi-modal, adversarial and far distribution shift.

ID Dataset. VQAv2 [19] contains open-ended questions about images with an emphasis on reducing answer biases through balanced pairs of question-image examples. We choose this as our ID datasets as it is widely used as a benchmark for popular vision-language models (VLMs), and most other OOD datasets are derived from it.

OOD Datasets. 1) *Distribution Shifts to Images.* IV-VQA [1] and CV-VQA [1] remove objects irrelevant and relevant to answering the question, resulting in unchanged and changed answers respectively. 2) *Distribution Shifts to Questions.* VQA-Rephrasings [41] provides three alternative phrasings for each question. 3) *Distribution Shifts to Answers.* VQA-CP [2] disrupts the usual correlation between question types and answers, creating a shift in answer patterns. 4) *Distribution Shifts to Multi-modalities.* VQA-CE [9] selects counterexample instances from VQAv2 validation set that highlight potential multi-modal shortcuts. 5) *Adversarial Distribution Shifts.* AdVQA [42] includes human-adversarial examples where the model's initial an-

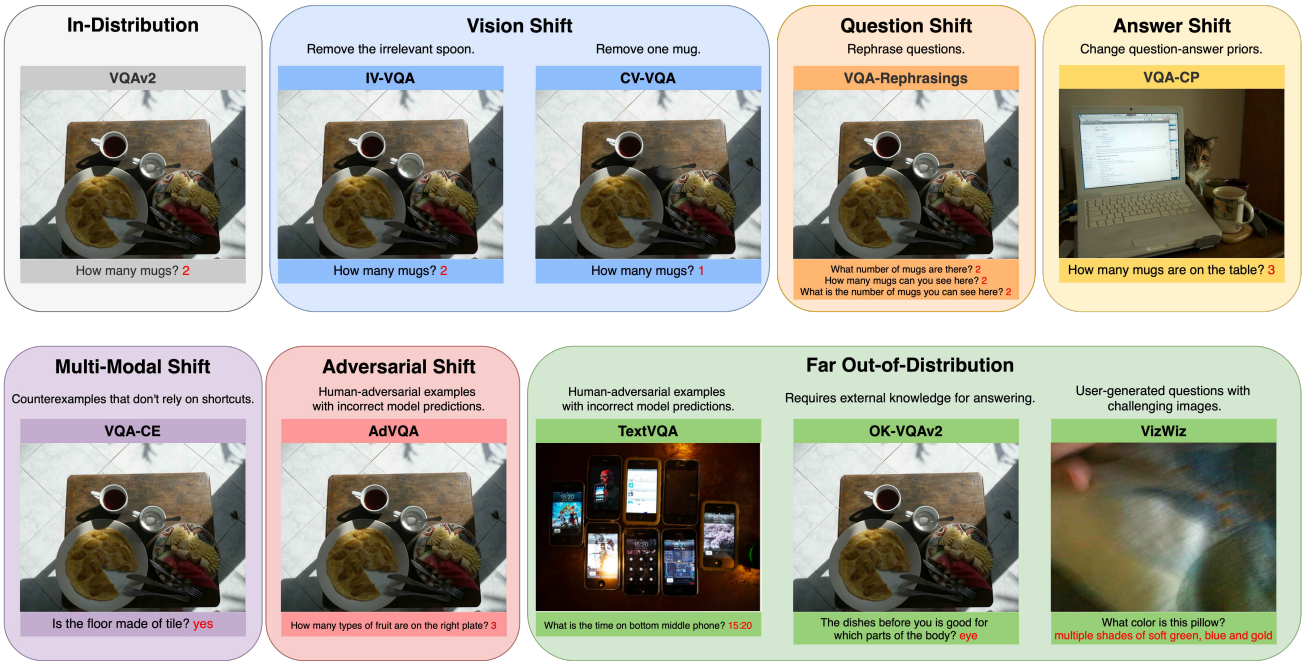


Figure 1. ID and OOD VQA datasets with Uni-modal (Vision, Question, Answer), Multi-modal, Adversarial and Far Distribution Shifts.

swer is incorrect. 6) *Far OODs.* TextVQA [45] requires textual comprehension within images to answer questions. VizWiz [6] features user-generated images with quality and relevance challenges. OK-VQAv2 [40] involves questions that require external knowledge beyond the image content. Tab. 1 shows the statistics of each dataset.

Shift Type	Dataset	# Samples
Source	VQAv2 [19]	Train: 443752 Val: 214354
Image	IV-VQA [1]	119907
	CV-VQA [1]	4141
Question	VQA-Rep. [41]	162020
Answer	VQA-CP [2]	219928
Multi-modal	VQA-CE [9]	63298
Adversarial	AdVQA [42]	10000
Far OOD	TextVQA [45]	5000
	VizWiz [6]	3173
	OK-VQAv2 [40]	5046

Table 1. ID and OOD VQA datasets. We use VQAv2 train for fine-tuning and VQAv2 val for model selection. For other OOD datasets, we only use the test splits for evaluation.

Evaluation and Metrics. We adopt the evaluation metric from VQAv2 [19], which measures accuracy by comparing predicted answers to ground truth human-annotated answers. Each question is paired with 10 human-provided answers, and the accuracy is computed as: Accuracy $= \min\left(\frac{\text{number of humans who gave the answer}}{3}, 1\right)$.

3.2. Measuring Distribution Shifts across Datasets

We aim to quantify distribution shift to more precisely measure model robustness (e.g., relationship between performance and the degree of shift). In VQA, the presence of both uni-modal and multi-modal shifts complicates qualitative analysis of the types of shifts affecting performance. By quantifying these shifts individually, we can better understand how each modality—both independently and in combination—impacts model robustness.

Mahalanobis Distance. We use the Mahalanobis distance to measure the distribution shift following procedures similar to typical feature-based OOD detection methods [43]. Further analysis using Maximum Mean Discrepancy [20] is shown in Suppl. 15. Specifically, given our input training split $X_{\text{in}}^{\text{train}}$, we compute feature representations z of the training samples to estimate the empirical mean μ_{train} and sample covariance matrix Σ_{train}. For each test split, we compute the test set shift relative to the training domain using the Mahalanobis distance defined in Eq. 1. The overall shift score for each test dataset, denoted as S_{maha}, is calculated as the average S_{Maha} across all samples.

$$S_{\text{Maha}}(z_{\text{test}}) = \sqrt{(z_{\text{test}} - \mu_{\text{train}})^\top \Sigma_{\text{train}}^{-1} (z_{\text{test}} - \mu_{\text{train}})} \quad (1)$$

We use the VQAv2 training set as our ID dataset. The average Mahalanobis score provides an overall measure of

how distant the test set is from the ID set, with higher values indicating more shift.

Embedding Extraction. Let q denote the question and v the image. The input features used in measuring shifts include uni-modal embeddings $f(v)$ and $f(q)$ and joint embeddings $f(v,q)$. We first leverage ViT [13] to get $f(v)$ and BERT [11] to get $f(q)$ for pure uni-modal embeddings. Both representations are derived by mean-pooling the last layer of their respective pre-trained models. We use the pre-trained and fine-tuned PaliGemma from various methods to extract both uni-modal and joint embeddings, i.e., $f(v)$, $f(q)$, and $f(v,q)$. For the image embedding $f(v)$, we obtain it via masking out the question input tokens and mean-pooling the image portion from the final layer of the model before the language model head. Similarly, to get $f(q)$, we mask out the image tokens and extract the question portion from the final layer. To obtain $f(v,q)$, we pass in both image and text tokens as input, compute the average embedding for both modalities and then taking the overall mean. All embeddings are summarized in Tab. 2. We display the histograms for the Mahalanobis score distribution between test datasets with the ID set in Suppl. 9. Results of the distances and further analysis will be presented in the following sections.

Embeddings	Model	Modality	PT/FT
$f(v)$	ViT	V	PT
$f(q)$	BERT	Q	PT
$f_{\text{pt}}(v)$	PaliGemma	V	PT
$f_{\text{pt}}(q)$	PaliGemma	Q	PT
$f_{\text{pt}}(v,q)$	PaliGemma	V,Q	PT
$f_{\text{ft_method}}(v)$	PaliGemma	V	FT
$f_{\text{ft_method}}(q)$	PaliGemma	Q	FT
$f_{\text{ft_method}}(v,q)$	PaliGemma	V,Q	FT

Table 2. Embedding extractions from different layers and backbone models.

4. Robust Fine-Tuning

In this section, we summarize several existing robust fine-tuning methods and evaluate their performance on our proposed benchmark FRAMES-VQA.

4.1. Robust Fine-Tuning Baselines

We include pre-training [16], Vanilla Fine-Tuning, Linear Probing, LP-FT [28], WiSE-FT [52], FTP [47] and SPD [48] as baselines. Fig. 2 provides an overview of all methods.

Vanilla fine-tuning updates model parameters to adapt to a new task without constraint, minimizing task-specific loss $\mathcal{L}$ with regularization on the L2 norm of the parameters with a strength of λ:

$$\min_{\lambda|(x,y)\in\mathcal{D}_{val}} \mathcal{L}(x,y;\ \arg\min_{\theta_t|(x,y)\in\mathcal{D}_{tr}} \mathcal{L}(x,y;\theta_t)+\lambda\|\theta_t\|_2^2,\lambda) \tag{2}$$

However, such free fine-tuning can cause feature distortion by excessively altering pre-trained representations. Linear Probing mitigates this by training only the final linear head while keeping other layers frozen, preserving the learned features. LP-FT [28] combines these two approaches and further proposes a two-step strategy of linear probing then full fine-tuning to achieve better adaptation while maintaining pre-trained features.

WiSE-FT [52] introduces an interpolation technique that combines the strengths of pre-trained and fine-tuned embeddings by linearly blending their weights, with $\alpha \in [0,1]$ to control the balance:

$$\tilde{\theta} = \alpha\theta_t + (1-\alpha)\theta_0, \alpha \in [0,1] \tag{3}$$

L2-SP [32] applies an L2 penalty on the deviation between fine-tuned and pre-trained weights, rather than regularizing the L2 norm of the parameters themselves. This explicit penalty encourages the fine-tuned model to stay closer to the pre-trained weights:

$$\min_{\lambda|(x,y)\in\mathcal{D}_{val}} \mathcal{L}(x,y;\ \arg\min_{\theta_t|(x,y)\in\mathcal{D}_{tr}} \mathcal{L}(x,y;\theta_t)+\frac{\lambda}{2}\|\theta_t-\theta_0\|_2^2,\lambda) \tag{4}$$

TPGM [46] approaches the regularization term in L2-SP as a constraint, reformulating the problem as a *bi-level optimization* and enforcing the model to stay within a distance of γ from the pre-trained model, as shown in Eq. 5:

$$\min_{\lambda,\gamma|(x,y)\in\mathcal{D}_{val}} \mathcal{L}(x,y;\ \arg\min_{\theta_t|(x,y)\in\mathcal{D}_{tr}} \mathcal{L}(x,y;\theta_t,\lambda),\lambda), \quad \text{s.t.} \quad \|\theta_t-\theta_0\|_2 \le \gamma, \tag{5}$$

To solve the constrained optimization problem, TPGM utilizes the *Projected Gradient Method* (PGM) to projects the updated model weights to be within the constraint, illustrated in Eq. 6:

$$\Pi_{l2}(\theta_0,\theta_t,\gamma): \tilde{\theta} = \theta_0 + \frac{1}{\max\left(1, \frac{\|\theta_t-\theta_0\|_2}{\gamma}\right)}(\theta_t-\theta_0) \tag{6}$$

FTP [47] further improves the efficiency of TPGM [46] by learning the constraint from training set of previous step instead of the current validation set, which is highlighted in blue in Eq. 7:

$$\min_{\lambda,\gamma|(x,y)\in\mathcal{D}_{tr}} \mathcal{L}(x,y;\ \arg\min_{\theta_t|(x,y)\in\mathcal{D}_{tr}} \mathcal{L}(x,y;\theta_t,\lambda),\lambda), \quad \text{s.t.} \quad \|\theta_t-\theta_0\|_2 \le \gamma, \tag{7}$$

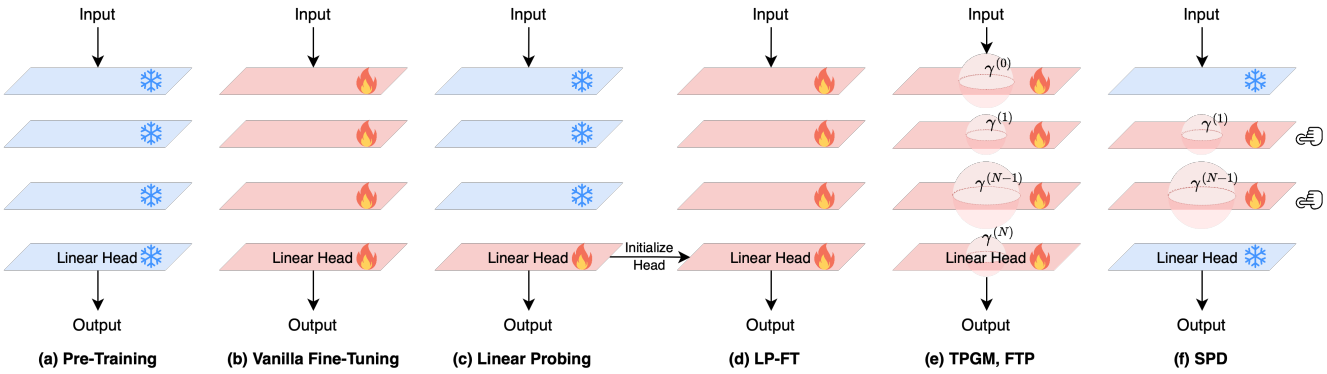


Figure 2. Overview of Robust Fine-Tuning Methods. The flame icon represents tunable layers, while the snowflake icon represents frozen layers. γ in TPGM, FTP and SPD is the constraint for each layer.

SPD [48] selectively imposes a strong penalty on certain layers while allowing others to change freely, with a selecting condition expressed in Eq. 8, where $g_{t+1} = \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}$ represents the gradient in step t:

$$c_t := -g_{t+1}^{\intercal}(\theta_t - \theta_0) \tag{8}$$

Intuitively, SPD expands and contracts the parameter search space for layers with consistent and inconsistent loss reduction between the descent direction $(-g_{t+1})$ and the current progress direction $(\theta_{t-1} - \theta_0)$, respectively. For the layers that meet the condition Eq. 8, SPD projects the corresponding layers using PGM in Eq. 6.

4.2. ID, Near & Far OOD Performance

We fine-tune Google's recently released PaliGemma-3B [4] model on the VQAv2 dataset and evaluate on the other OOD datasets. PaliGemma-3B is lightweight and one of the state-of-the-art models on VQAv2, making it a practical option for benchmarking. We apply LoRA [23], a parameter-efficient fine-tuning method, to reduce the computational and memory overhead associated with fine-tuning large models, as demonstrated in prior work [24, 54]. While LoRA limits excessive parameter updates to maintain robustness [5], as shown in Tab. 3 there is still a loss of robustness compared to vanilla fine-tuning and other robust fine-tuning methods. The results of the robust fine-tuning methods are shown in Tab. 3. Training details including configurations for different methods and additional results using full fine-tuning and LLaVA [33] can be found in Suppl. 8 and 13. Below we discuss our observations of the results.

Vanilla fine-tuning improves zero-shot performance across ID, near OOD, and far OOD datasets. As shown in Tab. 3, we observe no degradation in OOD performance compared to zero-shot performance following vanilla fine-tuning. This differs from the observation in [52] that vanilla fine-tuning for image classification degrades OOD performance compared to zero-shot. In the image classification task, fine-tuning typically involves using only the image encoder from the backbone model with a linear classification head, removing the text encoder and employing a cross-entropy loss that differs from the pre-training objective. However, [18] notes that fine-tuning is more robust when retaining the same objective as pre-training, which may explain why vanilla fine-tuning performs robustly in VQA tasks, as the model structure and loss function remain consistent. Model architecture and task characteristics could also contribute to the observed robustness.

WiSE-FT [52] decreases both ID and OOD performance in VQA. [52] demonstrate that ensembling the weights of zero-shot and fine-tuned models can benefit from the robustness of zero-shot and the adaptability of fine-tuning in image classification tasks. However, WiSE-FT is highly dependent on a reduction in robustness after vanilla fine-tuning and the linear connectivity of the model, which does not occur in our VQA setting. In Tab. 3, we see a significant gap between zero-shot and fine-tuned VQA performance, which limits the effectiveness of WiSE-FT. As a result, combining pre-trained and fine-tuned weights reduces the model's robustness across all shifts, making WiSE less effective than vanilla fine-tuning for VQA.

SPD [48] achieves the highest performance on ID, near and overall OOD. As shown in Tab. 3, SPD achieves the highest scores on ID, near OOD, and overall OOD, demonstrating robustness across various types of distribution shifts, including vision, question, answer, and multimodal combinations, as well as adversarial shifts.

FTP [47] underfits ID data but excels on far OOD datasets. FTP [46] underfits the ID dataset, showing significantly lower ID performance than vanilla fine-tuning, even

	ID	Near OOD							Far OOD				
		Vis.		Ques.	Ans.	M.M.	Adv.						
	VQAv2 (val)	IV-VQA	CV-VQA	VQA-Rep.	VQA-CP	VQA-CE	AdVQA	Near OOD Avg.	TextVQA	VizWiz	OK-VQA	Far OOD Avg.	OOD Avg.
Zero-Shot [16]	54.42	63.95	44.72	50.10	54.29	30.68	30.46	45.70	14.86	16.84	28.60	20.10	37.17
Vanilla FT_{LoRA} [23]	86.29	94.43	**69.36**	78.90	86.21	71.73	49.82	75.08	42.08	22.92	48.30	37.77	62.64
Linear $Prob_{LoRA}$	78.24	87.83	63.87	69.61	78.48	61.66	42.90	67.39	29.61	18.80	42.27	30.23	55.00
LP-FT_{LoRA} [28]	85.97	93.30	65.93	76.49	86.16	72.73	45.68	73.38	31.41	19.01	43.27	31.23	59.33
WiSE-FT_{LoRA} [52]	71.36	85.06	64.55	66.42	70.89	48.74	43.95	63.27	36.98	22.41	42.35	33.91	53.48
FTP_{LoRA} [47]	81.77	92.61	67.93	76.66	81.41	64.14	**50.99**	72.29	**49.12**	**25.67**	**51.07**	**41.95**	62.18
SPD_{LoRA} [48]	**87.39**	**95.25**	68.85	**79.48**	**87.27**	**73.52**	50.90	**75.88**	43.56	23.05	50.11	38.91	**63.55**

Table 3. Visual Question Answering Fine-Tuning Results for ID, Near OOD and Far OOD datasets. **Bold**: best. Underline: second best.

with a minimal positive gradient annealing factor ($\kappa = 0$), indicating the weakest regularization. However, FTP performs exceptionally well on far OOD tasks, achieving the highest scores across the three far OOD datasets. This outcome may be due to FTP's stronger regularization, as the projection constraints are non-decreasing throughout training, providing consistent regularization. The FTP authors also suggest that $\kappa = 0$ is necessary for optimal performance when underfitting is observed. Compared to SPD, which applies a weaker regularization and performs better on ID and near OOD, FTP's strict regularization might uniquely favor far OOD performance. The substantial improvement in far OOD, despite a weaker zero-shot baseline, presents an interesting area for future investigation.

5. Analysis on the Shift Scores

In this section, we aim to deepen our understanding of how multi-modal distribution shifts impact model robustness by analyzing 1) shift distances across different embeddings, datasets and fine-tuning methods, 2) interactions between uni- and multi-modal shifts.

We display the Mahalanobis distance of different datasets, embeddings and fine-tuning methods and its heatmap in Fig. 3. The darker blue represents higher distance score which indicates larger distribution shifts.

5.1. Shift Distance across Embeddings, Datasets & Fine-Tuning Techniques

As shown in Fig. 3, shift scores increase from left to right, with a corresponding darkening in color, which reflects our intuitive understanding of relative dataset shifts. An outlier is CV-VQA, which has the lowest shift scores, likely due to its small sample size or inherent dataset issues. Additionally, question, image, and joint shifts exhibit high negative correlations with performance, with values of -0.63, -0.74, -0.78 under Vanilla-FT, respectively. Full details on shift-performance correlations for various fine-tuning methods are provided in Suppl. 10, supporting the intuition that larger shifts in all modalities degrade VQA performance.

As we move from near to far OOD datasets, question shifts increase significantly, while image shifts increase at a lower rate, suggesting strong adaptivity to visual changes. This indicates that far OOD datasets introduce stronger task-guided and contextual shifts, and VLM tends to capture shifts more strongly in question representations.

Additionally, PaliGemma's visual, question and joint embeddings show similar variability on near OODs, whilst far OODs reveal greater sensitivity in question and joint embeddings, evidenced by higher Mahalanobis distances. In comparison, pure uni-modal embeddings from ViT and BERT remain more stable. This suggests that question and joint embeddings from multi-modal models are more sensitive to significant distribution shifts, likely due to their dependence on contextual and multi-modal interactions.

5.2. Correlation between Uni- & Multi-Modal Shifts

The degree of joint shift can be influenced by both visual and question shifts. We aim to reveal how much direct influence each modality has towards the overall joint shift by computing the Pearson correlation coefficient between $\{f(v), f(v,q)\}$ and $\{f(q), f(v,q)\}$ for each test set. Results are shown in Tab. 4. Full correlation breakdown for each dataset can be viewed in Suppl. 11.

All fine-tuning methods has a significantly higher question-joint correlation than image-joint correlation, making the joint modality more sensitive to question shifts. However, pre-trained method maintains similar levels of correlation. Further, more robust fine-tuning methods show smaller image-joint and question-joint correlations, indicating that robust methods learn representations less sensitive to specific image-question pairings, making them less impacted by uni-modal shifts.

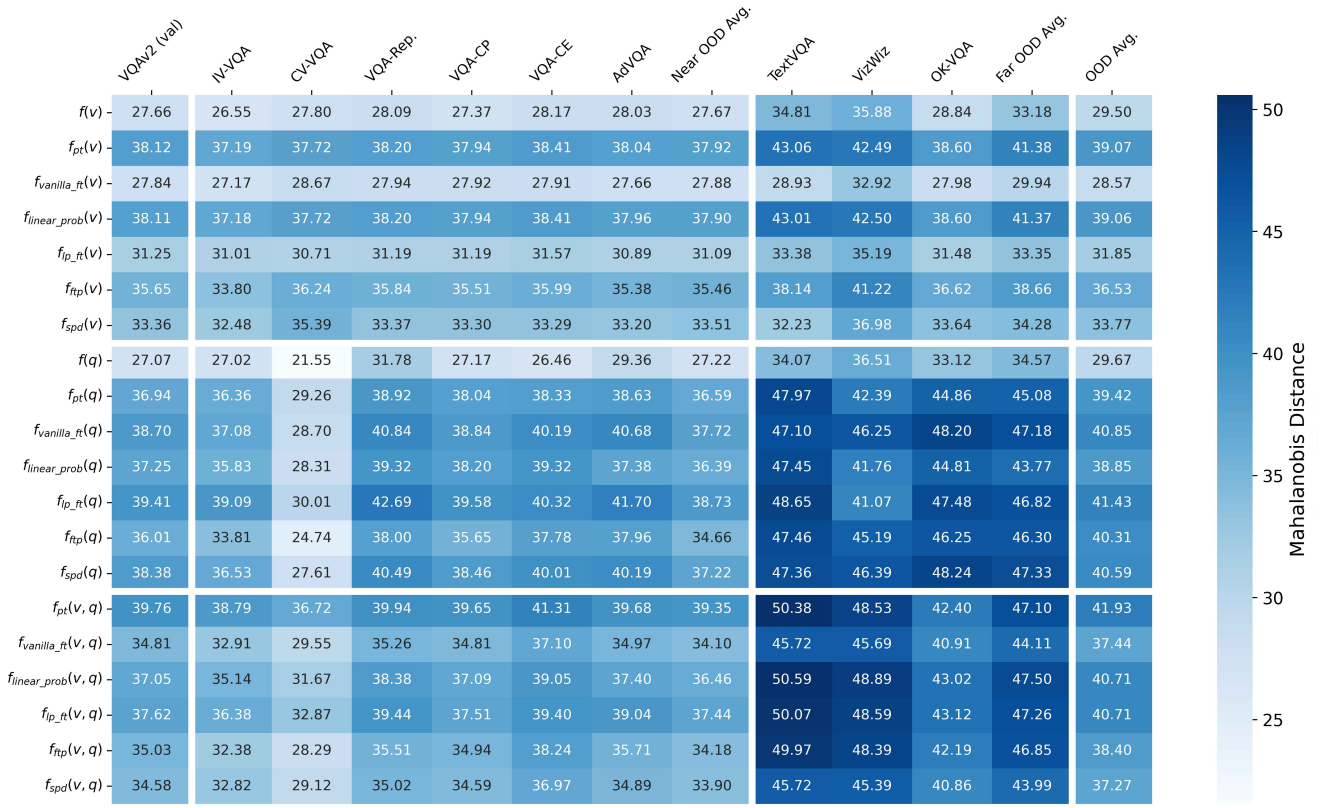


Figure 3. Mahalanobis Distance Heatmap of Different Datasets, Embeddings and Fine-Tuning Methods. Darker blue represents higher distance score. From top to bottom the three groups are image, question and joint shift scores respectively.

	$\{f(v), f(v,q)\}$	$\{f(q), f(v,q)\}$
Pre-Train [16]	0.32	0.34
Vanilla FT_{LoRA} [23]	0.29	0.48
Linear $\text{Prob}_{\text{LoRA}}$	0.26	0.61
LP-FT_{LoRA} [28]	0.34	0.63
FTP_{LoRA} [47]	0.34	0.58
SPD_{LoRA} [48]	0.17	0.51

Table 4. Uni- and multi-modal shifts correlation averaged across datasets for different fine-tuning methods.

6. Analysis on the Modality Importance

We further explore how modality importance impacts model robustness by analyzing intra- versus inter-modality attention, shifts in modality focus between ID and OOD settings, and differences across fine-tuning methods. This analysis reveals how attention to specific modalities influences the model's ability to generalize under distribution shifts.

6.1. Intra- & Inter-Modality Attention

We introduce the following metrics to quantitatively analyze the modality importance, inspired by [7]. Denote $v_1, v_2, \ldots, v_N$ and $q_1, q_2, \ldots, q_M$ as the image and question tokens respectively. The *Modality Importance* (MI) of a token tkn is defined as the ratio of its total attention to all question tokens to its total attention to all image tokens,

$$\text{MI}(tkn) = \frac{\sum_{j=1}^{M} \text{Attn}(tkn, q_j)}{\sum_{i=1}^{N} \text{Attn}(tkn, v_i)} \tag{9}$$

where $\text{Attn}(tkn_1, tkn_2)$ represents the average attention weights of all heads from tkn_1 to tkn_2. Note that the sum of the attention weights that a token attends to all the tokens in both modalities equals one.

Further, we want to explore the modality importance of the tokens in different modalities. We take the average MI of the tokens in each modality, which is expressed in Eq. 10,

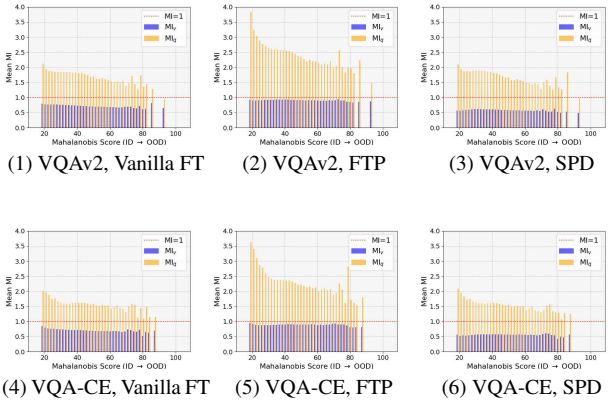
$$\text{MI}_v = \frac{1}{N}\sum_{i=1}^{N}\text{MI}(v_i), \text{MI}_q = \frac{1}{M}\sum_{j=1}^{M}\text{MI}(q_j) \tag{10}$$

.

$\text{MI}_m > 1$ indicates that the text modality is more influential than the image modality for tokens in modality m.

In Fig. 4, we present the variation of MI_v and MI_q w.r.t. shift score under vanilla FT, FTP and SPD across VQAv2

and VQA-CE. Other datasets and baselines can be found in Suppl. 12. In Tab. 5, we further separate ID and OOD samples from all datasets with a Mahalanobis Distance of 60, chosen as the relative median of shift scores to illustrate the distinction between closer and more distant samples, and show MI_v, MI_q for different fine-tune methods.



(1) VQAv2, Vanilla FT (2) VQAv2, FTP (3) VQAv2, SPD

(4) VQA-CE, Vanilla FT (5) VQA-CE, FTP (6) VQA-CE, SPD

Figure 4. Variation of MI_v and MI_q w.r.t. shift score under vanilla FT, FTP and SPD across VQAv2 and VQA-CE. The blue and orange bars represent MI_v and MI_q respectively. The red dotted line marks a reference MI of 1.

	ID MI		OOD MI		Overall MI	
	MI_v	MI_q	MI_v	MI_q	MI_v	MI_q
Pre-Train [16]	0.90	2.37	0.85	1.93	0.89	2.37
Vanilla FT_{LoRA} [23]	0.75	1.90	0.66	1.42	0.75	1.89
Linear $\text{Prob}_{\text{LoRA}}$	0.71	2.10	0.66	1.60	0.71	2.10
$\text{LP-FT}_{\text{LoRA}}$ [28]	0.94	2.40	0.78	1.51	0.94	2.39
FTP_{LoRA} [47]	0.93	2.72	0.89	1.97	0.92	2.72
SPD_{LoRA} [48]	0.61	1.93	0.57	1.42	0.61	1.93

Table 5. MI_v and MI_q of ID and OOD samples for different fine-tuning methods. We use the Mahalanobis Distance of 60 to distinguish between ID and OOD samples.

6.2. Comparison between Intra- & Inter-Modality Attention

According to Tab. 5 and Fig. 4, $\text{MI}_v < 1$ while $\text{MI}_q > 1$, implying dominant image and question attention for image and tokens respectively. The dominance of intra-modality attention is expected, as attention mechanisms tend to focus within the same modality to reinforce contextual coherence. However, Intra-modality attention of text tokens is greater than that of image tokens, since MI_v is close to 1 while MI_q is significantly greater than 1. This suggests that the ratio of question-to-question attention relative to question-to-image attention is higher than that of image-to-image attention relative to image-to-question, indicating stronger intra-modality attention for text tokens.

6.3. Comparison of MIs of ID & OOD

As illustrated in Fig. 4 and Tab. 5, across each dataset and fine-tuning method, MI_v remains relatively stable under distribution shift, whereas MI_q shows a marked decrease as the distribution shift intensifies. Such pattern suggest that, in OOD samples, text tokens increasingly attend to image tokens. Such a shift could indicate a model bias, where ID samples rely heavily on intra-modality shortcuts for text, potentially at the expense of robust cross-modal integration. This reliance on shortcuts and lack of image grounding may reduce the model's ability to generalize effectively.

6.4. Comparison of Fine-Tuning Methods

According to Tab. 5, SPD, which performs best on ID and near-OOD datasets, has the lowest MI_v, while FTP, which excels on far-OOD datasets, shows the highest MI_q. One possible hypothesis is that for optimal performance across ID, near-OOD, and far-OOD settings, a model might benefit from a high MI_q and low MI_v, suggesting that each modality should prioritize intra-modality attention over cross-modality attention. Stronger intra-modality focus—where each modality focuses on internal coherence before cross-modal integration—could potentially yield improved robustness across distribution shifts.

Analyses in Sec. 5 and 6 suggest that future work should aim to enhance robustness under language shifts and implement ways to dynamically handle modality importance. Our findings show that more robust methods exhibit higher intra-modality attention, highlighting the potential of adaptive attention mechanisms and modality-specific regularization to better balance modality contributions.

7. Conclusion

In conclusion, our paper introduces FRAMES-VQA, a new VQA benchmark to evaluate fine-tuning robustness under various distribution shifts. We categorize the shift types and quantify the degree, and conduct comparisons of robust fine-tuning baselines on the proposed benchmark. Our comprehensive analysis reveals the interactions between uni- and multi-modal shifts and how VLMs perform cross-modal learning in ID/OOD scenarios using the modality importance metric. We observe that fine-tuning increases the influence of question shifts on multi-modal representations, with more robust methods showing a lower correlation between uni- and multi-modal shifts. Question-to-image attention rises for OOD samples, while robust methods prioritize intra-modality attention. Our findings lay the foundation and highlight promising directions for future work: (1) Developing training methods to mitigate uni- and multi-modal correlations, and (2) Enhancing techniques for detecting and adapting to different types of distribution shifts, both of which proved crucial in our analysis.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9695, Seattle, WA, USA, 2020. IEEE. 1, 2, 3

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering, 2018. arXiv:1712.00377 [cs]. 1, 2, 3

[3] Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization, 2023. arXiv:2205.12191 [cs]. 2

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer, 2024. arXiv:2407.07726 [cs] version: 1. 5

[5] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024. 5

[6] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. VizWiz: nearly real-time answers to visual questions. 3

[7] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models, 2020. 7

[8] Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models, 2023. 2

[9] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering, 2021. arXiv:2104.03149 [cs]. 1, 2, 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 4

[12] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and H. T. Kung. Neural mean discrepancy for efficient out-of-distribution detection, 2022. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4

[14] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection, 2021. 2

[15] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. VQA-LOL: Visual Question Answering under the Lens of Logic, 2020. arXiv:2002.08325 [cs]. 1, 2

[16] Google. paligemma-3b-pt-224. `https://huggingface.co/google/paligemma-3b-pt-224`, 2024. 4, 6, 7, 8, 1

[17] Henry Gouk, Timothy M. Hospedales, and Massimiliano Pontil. Distance-Based Regularisation of Deep Networks for Fine-Tuning, 2021. arXiv:2002.08253 [cs, stat]. 2

[18] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023. 5

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, 2017. arXiv:1612.00837 [cs]. 2, 3

[20] Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A kernel method for the two-sample problem, 2008. 3, 2

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1, 2

[23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. arXiv:2106.09685 [cs]. 5, 6, 7, 8, 1

[24] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models, 2023. 5

[25] Chengyue Huang, Junjiao Tian, Brisa Maneechotesuwan, Shivang Chopra, and Zsolt Kira. Directional gradient projection for robust fine-tuning of foundation models, 2025. 1

[26] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 2

[27] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to?, 2021. 2

[28] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution, 2022. arXiv:2202.10054 [cs]. 2, 4, 6, 7, 8, 1

[29] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 1

[30] Linjie Li, Zhe Gan, and Jingjing Liu. A Closer Look at the Robustness of Vision-and-Language Pre-trained Models, 2021. arXiv:2012.08673 [cs]. 2

[31] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models, 2021. arXiv:2106.00245 [cs]. 1, 2

[32] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks, 2018. arXiv:1802.01483 [cs]. 2, 4

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5, 1

[34] Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. Robust Visual Question Answering: Datasets, Methods, and Future Challenges, 2024. arXiv:2307.11471 [cs]. 2

[35] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, Ziwei Liu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Generalized out-of-distribution detection and beyond in vision language model era: A survey, 2024. 2

[36] Liwen Ouyang and Aaron Key. Maximum mean discrepancy for generalization in the presence of distribution and missingness shift, 2022. 2

[37] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 1, 2

[38] 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. Learning Transferable Visual Models From Natural Language Supervision, 2021. arXiv:2103.00020 [cs]. 1, 2

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1, 2

[40] Benjamin Z. Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. Outside Knowledge Visual Question Answering Version 2.0. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[41] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-Consistency for Robust Visual Question Answering, 2019. arXiv:1902.05660 [cs]. 1, 2, 3

[42] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-Adversarial Visual Question Answering, 2021. arXiv:2106.02280 [cs]. 1, 2, 3

[43] Xiangxi Shi and Stefan Lee. Benchmarking out-of-distribution detection in visual question answering, 2024. 3

[44] Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA, 2022. arXiv:2210.04692 [cs]. 1, 2

[45] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read, 2019. arXiv:1904.08920 [cs]. 3

[46] Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, and Zsolt Kira. Trainable Projected Gradient Method for Robust Fine-tuning, 2023. arXiv:2303.10720 [cs]. 2, 4, 5

[47] Junjiao Tian, Yen-Cheng Liu, James Seale Smith, and Zsolt Kira. Fast Trainable Projection for Robust Fine-Tuning, 2023. arXiv:2310.19182 [cs]. 2, 4, 5, 6, 7, 8, 1

[48] Junjiao Tian, Chengyue Huang, and Zsolt Kira. Rethinking weight decay for robust fine-tuning of foundation models, 2024. 2, 4, 5, 6, 7, 8, 1

[49] Suraj Jyothi Unni, Raha Moraffah, and Huan Liu. VQA-GEN: A Visual Question Answering Benchmark for Domain Generalization, 2023. arXiv:2311.00807 [cs]. 1, 2

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[51] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection, 2020. 2

[52] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022. arXiv:2109.01903 [cs]. 1, 2, 4, 5, 6

[53] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels, 2021. 2

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 5