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

Supplementary Material

8. Training Details

We use the model pretrained with $224 * 224$ input images and 128 token input/output text sequences and fine-tune with the precision of bfloat16. We use the LAVIS [29] public repository to fine-tune all methods. Standard hyper-parameters are used for all: learning rate ($1e-3$), weight-decay ($1e-4$), optimizer (AdamW), scheduler (Linear Warmup With Cosine Annealing), warm-up learning rate ($1e-4$), minimum learning rate ($1e-4$), accumulation steps (2), beam size (5). The model is trained for 10 epochs with a batch size of 128 for Tab. 3. For LoRA [23], we limit our study to only adapting the attention weights and freeze the MLP modules for parameter-efficiency, specifically apply LoRA to W_q, W_k, W_v, W_o with $r = 8$ in Tab. 3. The regularization hyper-parameter is found through cross-validation, and the model with the best ID validation accuracy is taken. We use 8 A40 GPU for each experiment. The best training configurations for different methods are listed in Tab. 6.

	lr	wd	others
Vanilla FT	$1e-3$	$1e-4$	-
Linear Prob	$1e-3$	$1e-4$	-
LP-FT	$1e-3$	$1e-4$	-
WiSE-FT	-	-	$\alpha = 0.5$
FTP	$1e-3$	$1e-4$	$\kappa = 0$
SPD	$1e-3$	0.5	-

Table 6. **Best Training Configurations for Robust Fine-Tuning Methods.** lr and wd stands for learning rate and weight decay.

9. Histograms of Shift Scores

We display the histograms for the Mahalanobis score distribution between test datasets with the ID set. Fig. 10, 11 and 12 are visual, question and joint shifts from vanilla FT repectively.

The histograms show that under Vanilla FT, visual shifts are minimal across most VQA datasets except for VizWiz, while question shifts are greater for further OOD datasets. Combined visual and question shifts exhibit the largest deviations across all test splits.

10. Correlation between Shift & Performance

Tab. 7 shows the correlation between shift and performance for different embeddings under different fine-tuning methods. Overall, visual and joint shifts exhibit the strongest correlation with performance across all types of methods. The negative correlation supports the intuition that larger shifts in all modalities degrade VQA performance.

Method	V	Q	Joint
Pre-Train [16]	-0.80	-0.66	-0.80
Vanilla FT$_{\text{LoRA}}$ [23]	-0.74	-0.63	-0.78
Linear Prob$_{\text{LoRA}}$	-0.82	-0.55	-0.81
LP-FT$_{\text{LoRA}}$ [28]	-0.75	-0.58	-0.80
FTP$_{\text{LoRA}}$ [47]	-0.86	-0.63	-0.75
SPD$_{\text{LoRA}}$ [48]	-0.52	-0.61	-0.79

Table 7. Correlation between Shift Score vs. Performance for different embeddings under various fine-tuning methods.

11. Correlation between Uni- & Multi-Modal Shifts per Dataset

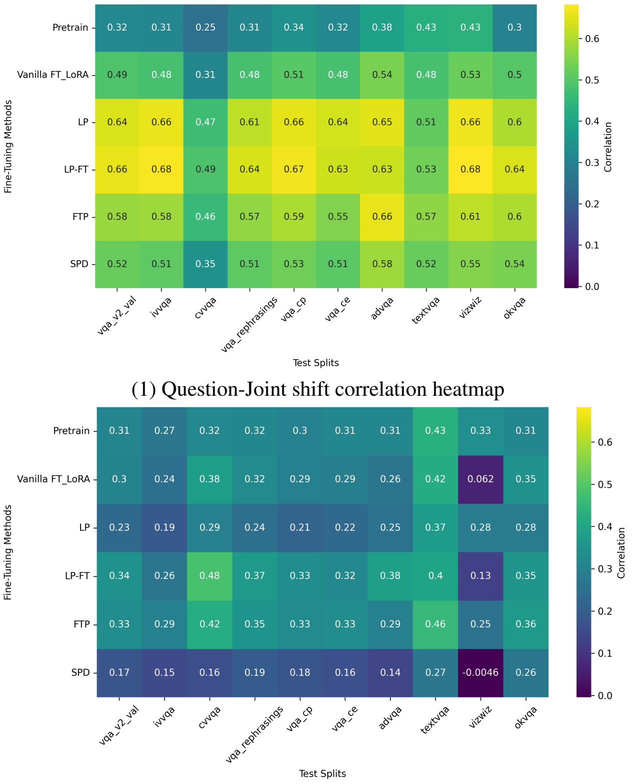
Fig. 5 shows the heatmap of the correlation between uni-modal and multi-modal shifts per dataset. Question-joint shift correlations are higher than image-joint shift correlations across all VQA datasets and fine-tuning methods. However, pre-train model maintains similar correlation between both modalities. Vanilla FT and SPD exhibits the lowest question-joint shift correlation shown by the darkest row color across all fine-tuning methods in Fig. 51. Whilst, SPD shows the lowest image-joint shift correlation across the datasets in Fig. 52.

12. Modality Importance of different Datasets and Fine-Tuning Methods

Fig. 13 and 14 show the variation of MI$_v$ and MI$_q$ w.r.t. shift score under all datasets and fine-tuning methods. Overall, intra-modality attention is more dominant than inter-modality attention. There is a stronger intra-modality attention for text tokens than image tokens. In OOD samples, text tokens increasingly attend to image tokens. A more robust model tends to have higher MI$_q$ and lower MI$_v$.

13. Additional Results using Full Fine-Tuning and LLaVA

We present additional results in Tab. 8, including LLaVA-7B [33] with LoRA and PaliGemma-3B with full fine-tuning. These results are consistent with PaliGemma with



(1) Question-Joint shift correlation heatmap

(2) Image-Joint shift correlation heatmap

Figure 5. Heatmap of correlation between uni-modal and multi-modal shifts per dataset.

LoRA: FTP and SPD remain relatively robust models, which validates the credibility of our analysis.

	VQAv2 val	Near OOD Avg.	Far OOD Avg.	OOD Avg.
(a) LLaVA-7B with LoRA under 10% of VQAv2 (train & val)				
Zero-Shot	3.27	3.44	0.68	2.52
Vanilla FT	72.49	60.07	28.67	49.60
LP-FT	53.01	28.63	7.64	21.63
WiSE-FT	60.47	43.33	9.07	31.98
FTP	67.95	58.49	26.21	47.73
SPD	**73.59**	**61.98**	**29.98**	**51.31**
(b) PaliGemma-3B with Full Fine-Tuning under 10% of VQAv2 (train & val)				
Zero-Shot	54.42	45.70	20.10	37.17
Vanilla FT	95.80	60.73	26.56	49.34
Linear Prob	86.80	59.61	24.17	47.80
LP-FT	94.44	57.13	21.03	45.10
FTP	95.40	**64.33**	**32.55**	**53.74**
SPD	**95.84**	63.92	32.46	53.43

Table 8. **LLaVA-7B (LoRA) and PaliGemma-3B (Full Fine-Tuning) Fine-Tuned on VQAv2.** We sample 10% of the VQAv2 training and validation set. **Bold**: best. Underline: second best.

14. Fine-Tuning Results on GQA

We use VQAv2 as the ID dataset since most OOD VQA datasets, covering various shifts, are built on it. The only exception, GQA-OOD [27] (based on GQA [26]), has only answer shifts. To further validate our findings, we fine-tune PaliGemma-3B on GQA as ID and evaluate it on GQA-OOD and VQAv2 variants (Tab. 9). The results follow the same trend: FTP and SPD remain relatively robust, with SPD excelling on Near OOD (GQA-OOD) and FTP on Far OOD (VQAv2 and its variants). This consistency reinforces the generalizability of our analysis.

	ID	Near OOD	Far OOD		
	GQA	GQA-OOD	VQAv2	VQAv2 Near OOD Avg.	VQAv2 Far OOD Avg.
Zero-Shot	41.44	29.33	54.42	45.70	20.10
Vanilla FT	**67.00**	53.97	64.97	57.08	23.42
Linear Prob	61.70	50.32	54.27	39.64	14.43
LP-FT	61.51	50.72	55.72	43.89	14.95
FTP	64.97	53.15	**66.40**	**58.38**	**25.26**
SPD	66.80	**54.04**	65.27	57.53	24.55

Table 9. **PaliGemma-3B Fine-tuned on GQA with LoRA and Evaluated on GQA-OOD, VQAv2 and its variants.** We sample 10% of the GQA training set. **Bold**: best. Underline: second best.

15. Quantifying Shifts using Maximum Mean Discrepancy

Mahalanobis distance is a dominant metric for measuring distribution shifts [35]. We further compare shifts using Maximum Mean Discrepancy (MMD) [12, 20, 36] with RBF kernel in Tab. 10. We observe similar trends as with Mahalanobis distance (i.e., higher scores indicate greater shifts), reinforcing the reliability of our shift scores.

	ID	IVVQA	VQA-REP	VQA-CE	TextVQA	VizWiz
$f_{\text{vanilla\_ft}}(q)$	20.02	20.12	20.10	20.18	21.92	23.18
$f_{\text{vanilla\_ft}}(v)$	20.17	20.20	20.13	20.28	21.98	23.07
$f_{\text{vanilla\_ft}}(v, q)$	20.10	20.12	20.12	20.20	22.44	23.02
$f_{\text{pt}}(q)$	20.16	20.18	20.22	20.28	21.98	23.32
$f_{\text{pt}}(v)$	20.15	20.18	20.20	20.25	22.47	23.75
$f_{\text{pt}}(v, q)$	20.08	20.06	20.16	20.10	22.28	23.26

Table 10. **Maximum Mean Discrepancy metric with LoRA and Pretrained embeddings** on VQAv2 and its variants. We sample 1000 instances per dataset. Gamma=1.0, scale-up factor=10^4

16. Qualitative Analysis: Inspect via Sampling

In order to investigate the types of ID and OOD samples under different modalities, we perform sampling on the various regions of the histogram to inspect how the model represents ID/OOD embeddings. This also serves as a verification of the reliability in quantifying shifts via feature-based representations. We select the vanilla fine-tuned (with LoRA) model as our model of choice and consider various regions in the distribution between both train and test splits.

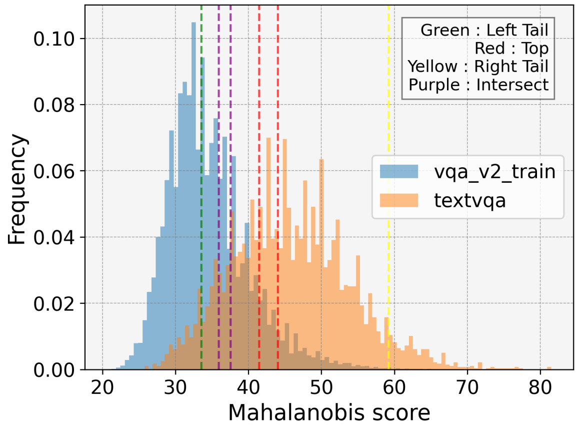


Figure 6. Sampling region in histogram

Under V, Q and V+Q, we sample 50 instances from 4 regions as shown in Fig. 6 including:

- **Left tail (top %5):** ID samples.
- **Top region:** top occurring samples in the test set.
- **Intersect region:** similar samples between train & test split.
- **Right tail (bottom %5):** clear outlier concepts and exhibit uncommon & hard instances (e.g. an image object that barely appears in left tail/peak region).

16.1. Image Shift $f_{\text{ft}}(v)$

We observe that the ID images involve commonly occuring objects shown in Tab. 11. Under image shifts, we have the following observations and potential hypothesis.

Category	Examples
Animals	cats, dogs, giraffe
Sports	baseball, skateboard, frisbee
Objects	kite, fire hydrant, pizza
Vehicles	bus, car, airplane

Table 11. Commonly Occurring Objects

We observe that 1) There are distinct differences between left tail/peak and right tail (OOD) samples in terms of object categories and compositions (e.g., # objects). Some examples are shown in Fig. 7. 2) Intersecting regions have similar type concepts. 3) Tail samples seem to have less number of objects and contain more close up images. 4) There are still some instances where similar objects appear in significantly different regions, i.e., a commonly occuring object appears in the right region (OOD tail). These instances can be shown in Fig. 8.

We hypothesize that 1) There are significant visible shifts under the image domain with barely overlapping image objects between ID and OOD regions. However, some odd samples depicted in Fig. 8 suggests some weakness of the fine-tuned model in robustly representing image embeddings. It fails to represent closer embeddings between images with similar objects. 2) Weight updates under joint inputs causes the two image embeddings of similar objects to steer in different directions. The image embeddings are indirectly conditioned on different questions and answers.

16.2. Question Shift $f_{\text{ft}}(q)$

Tab. 13 details the ID and OOD question examples. The ID questions are much more straightforward where it involves simple identification of colours, activity, yes/no questions etc. OOD questions tend to incorporate more outside knowledge, which explains the drastic question shift in OKVQA, and more complex visual grounding such as OCR (Optical Character Recognition) tasks.

16.3. Joint Shift $f_{\text{ft}}(v, q)$

Under the joint shift, there is an added complexity to this since we must consider the possible combinations of shifts under mixed modalities.

Region	Examples
Left Tail Peak Intersect	ID Object + ID Question
Right Tail	ID Object + OOD Question OOD Object + ID Question OOD Object + OOD Question

Table 12. Region Samples

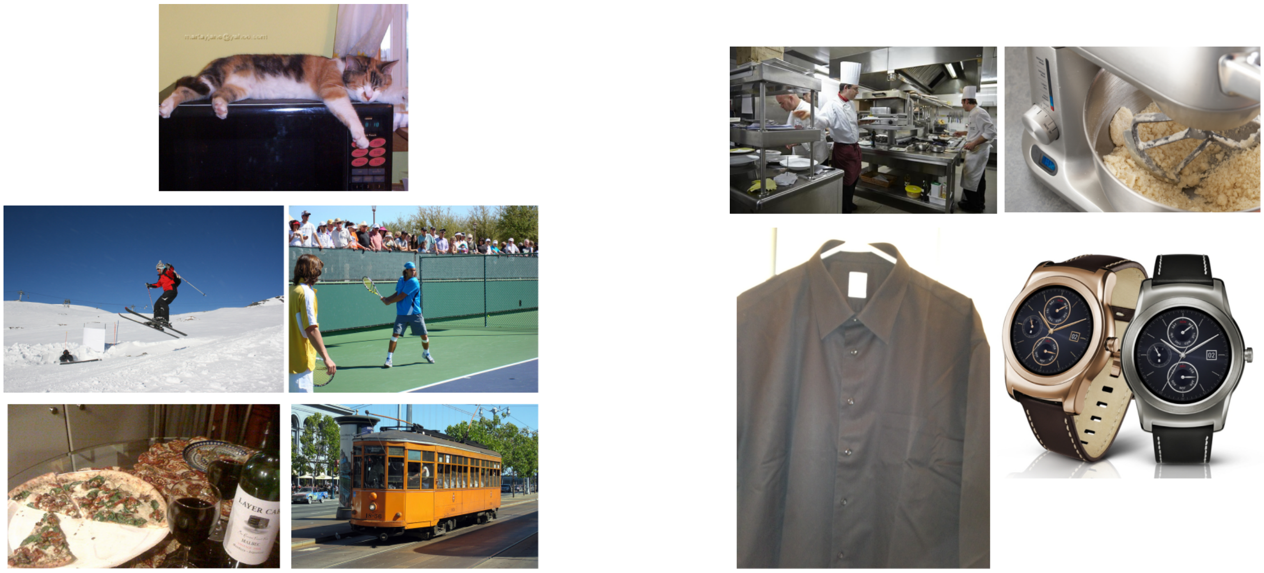
Tab. 12 outlines the types of VQA samples we expect to see under different regions. Intuitively, we would expect samples with OOD Question + OOD Object to be found in the right tail region. Whilst finding OOD Question + ID Object may indicate that the specific dataset has more samples with prominent OOD questions with ID images or that the joint shifts are more heavily influenced by text modality, causing samples with OOD questions to have a significant push to the joint embedding towards the right tail region.

Similarly, for ID Question + OOD Object, if those samples are found in the right tail, then it may indicate dataset having more OOD samples with ID Question + OOD Object or the visual modality has a greater influenced in steering the joint embedding.

ID and OOD joint samples can be viewed in Fig. 9. Most of the ID samples have objective questions and images with easier to view objects. On the other hand, the right tail samples seem to involve harder questions that are more sub-

Category	**ID Examples**	**OOD Examples**
Color	What color is the cat?	What color is on the inside of the speaker?
Activity	What is this man doing?	What is the person washing?
Counting	How many {ID objects}?	How many are chocolate-covered donuts?
Outside knowledge	-	What type of feed does this breed of horse need? What is the name of the knot used on this tie? If you add the two visible numbers, on the jerseys, what is the total sum?
Text-in-image	What is the number in lights on the bus?	What is the name the package delivery company?
Brand/Species	What brand of bike is this? What species of animal are we looking at?	What brand of watch is shown? What species of fish is being served?

Table 13. Common questions with examples for ID and OOD cases.



(1) ID images

(2) OOD images

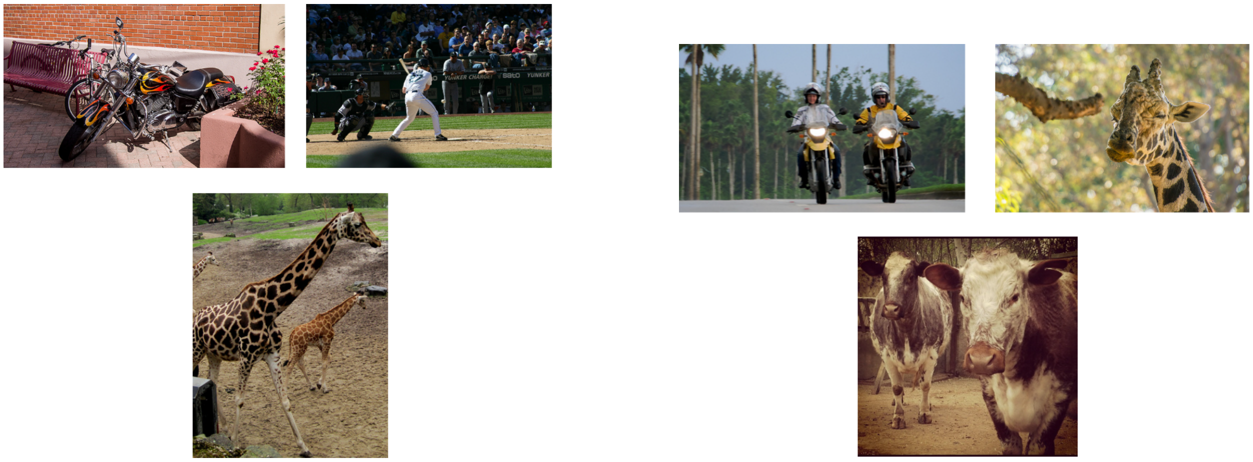
Figure 7. Comparison set of ID and OOD images in terms of object categories and compositions (e.g., # objects)

jective, require outside knowledge, and more reasoning including VQA that involves reading text from an image. Interestingly, most of the right tail samples attribute to OOD questions regardless of whether its image pair is ID/OOD, i.e., the majority of inspected OOD samples are either OOD Question + OOD Object or OOD Question + ID Object. This suggests that amongst the 10 VQA datasets there is a much higher question shift than there is to images as well as a higher sensitivity to question shifts, since farther OODs are OOD Question + ID Object > ID Question + OOD Object.

We also measure the composition of right tail samples under the joint modality by filtering samples that fall under each respective category and using an OOD threshold cutoff for each modality. The percentage of samples that make up the right tail region are shown in Tab. 14. This aligns with our qualitative analysis that OOD questions make up a significant portion of Joint OOD regardless of whether they are ID/OOD images.

	OOD V + ID Q	ID V + OOD Q	OOD V + OOD Q
% composition in Joint OOD	9.68	45.83	14.76

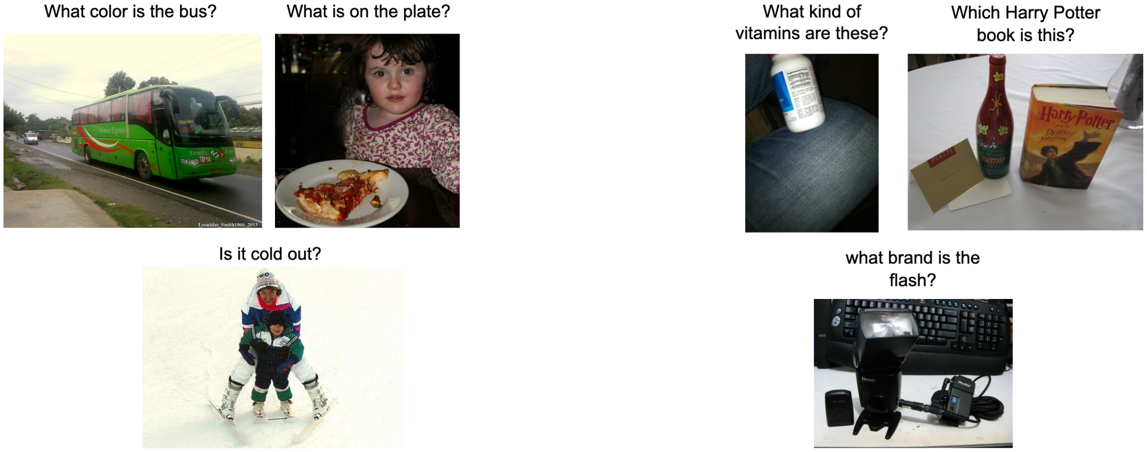
Table 14. Percentage of each OOD type in Joint OOD. We use a threshold cutoff of 45 for visual OOD, 50 for question OOD, and 60 for joint OOD.



(1) ID images

(2) OOD images

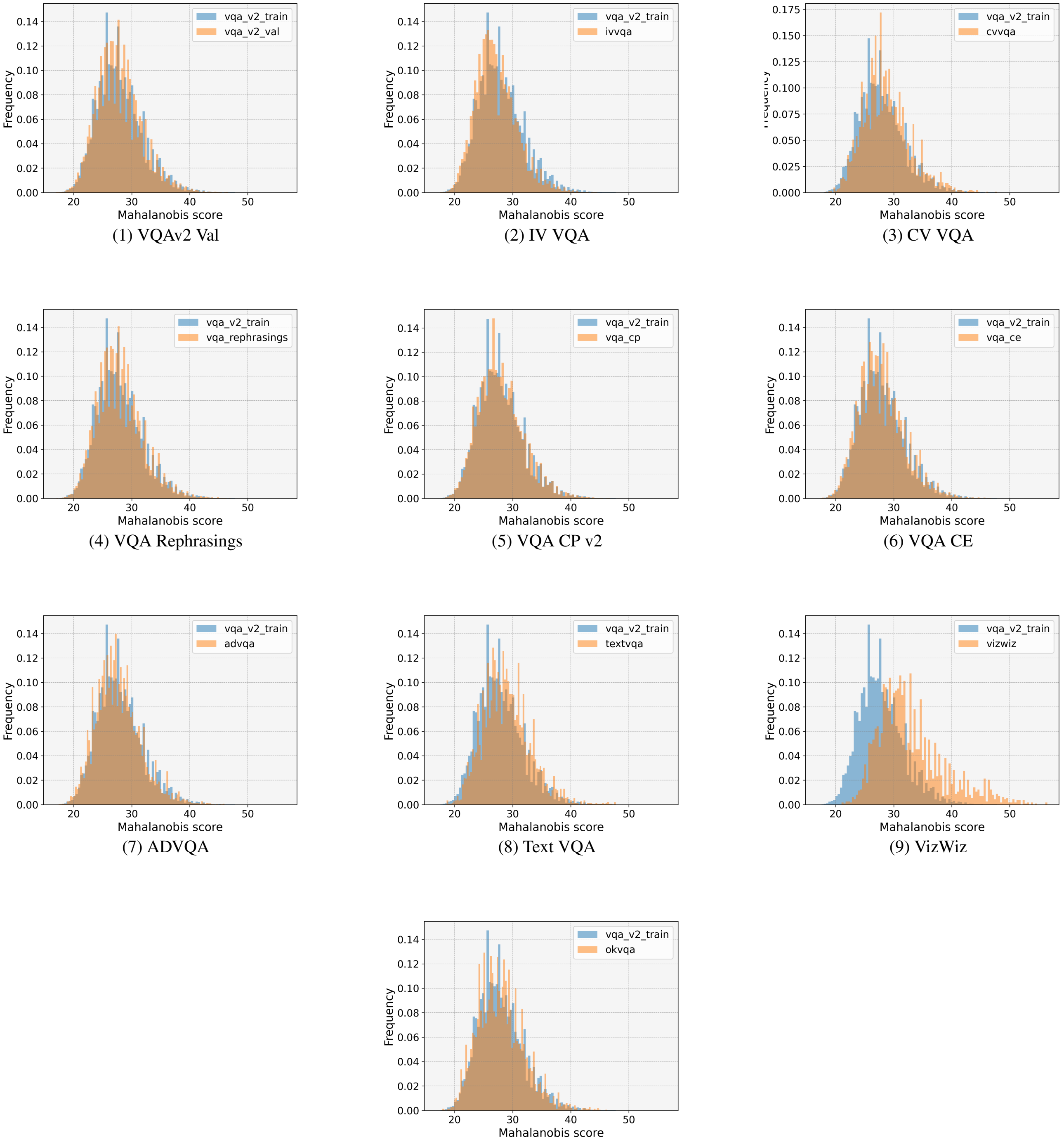
Figure 8. Comparison of ID and OOD images but with similar objects.



(1) ID samples

(2) OOD samples

Figure 9. Comparison set of ID and OOD joint samples.



(1) VQAv2 Val

(2) IV VQA

(3) CV VQA

(4) VQA Rephrasings

(5) VQA CP v2

(6) VQA CE

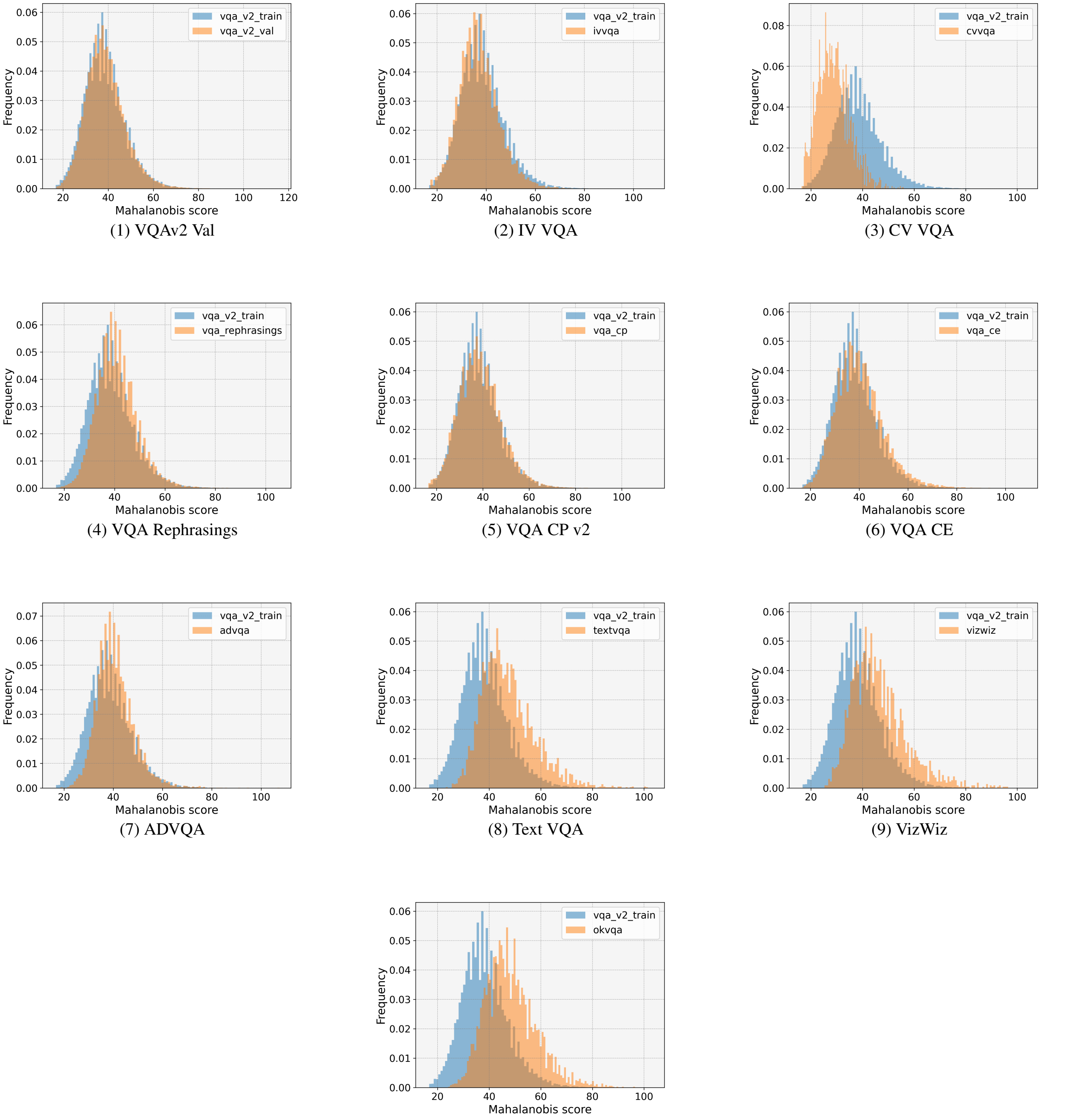
(7) ADVQA

(8) Text VQA

(9) VizWiz

(10) OK VQA

Figure 10. Histogram for Vanilla FT Visual Shifts: We depict the S_{Maha} score on the visual modality for each sample in the VQAv2 train split in blue and the corresponding test samples in orange. There's minimal visual shifts for all VQA datasets from the VQAv2 train, except for Figure (9) which shows evidence of greater shifts between the orange distribution and the blue distribution.



(1) VQAv2 Val

(2) IV VQA

(3) CV VQA

(4) VQA Rephrasings

(5) VQA CP v2

(6) VQA CE

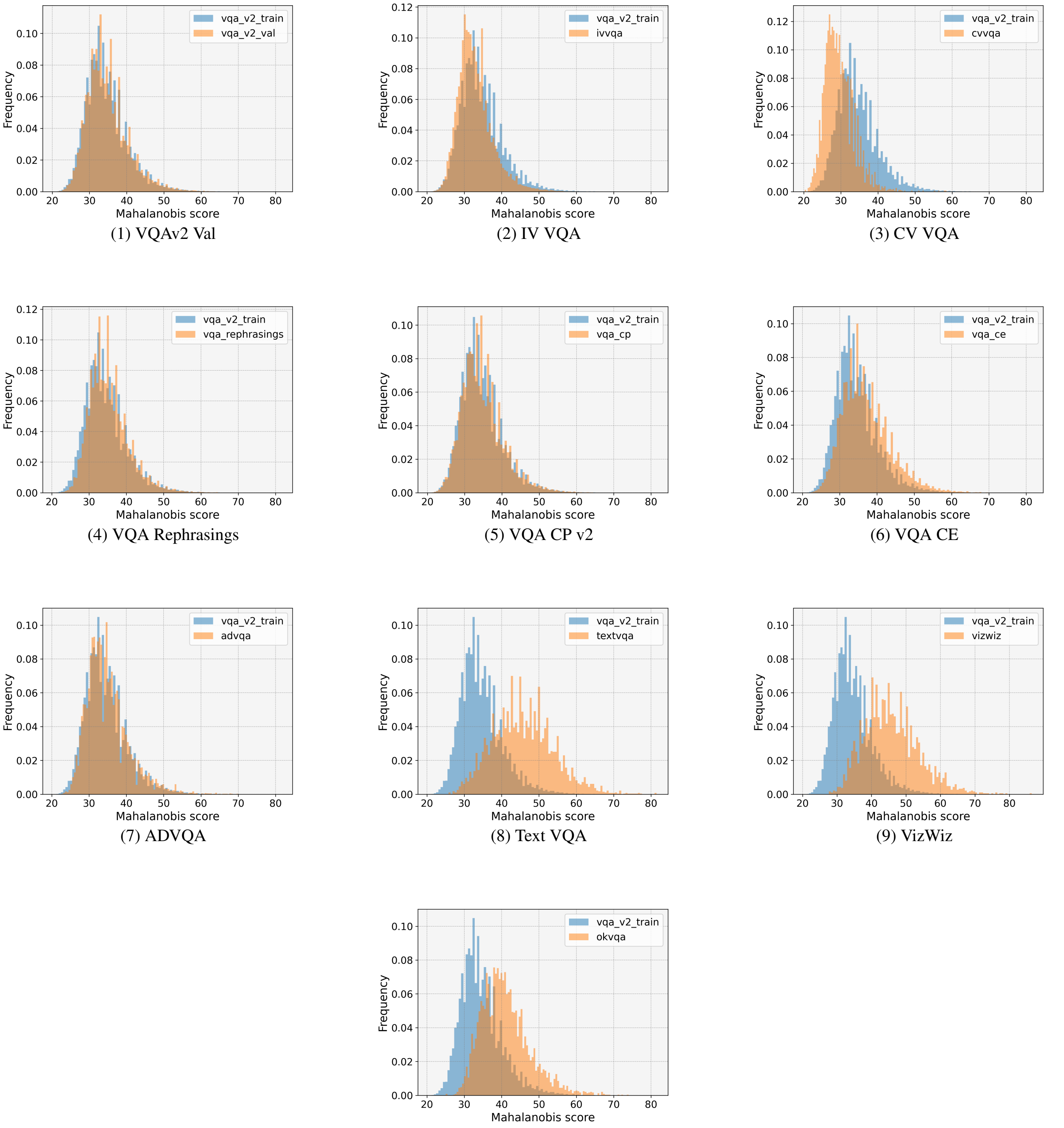
(7) ADVQA

(8) Text VQA

(9) VizWiz

(10) OK VQA

Figure 11. Histogram for Vanilla FT Question Shifts: We depict the S_{Maha} score on the question modality for each sample in the VQAv2 train split in blue and the corresponding test samples in orange. Similar to the visual shift histograms, far OODs (Figures (8), (9), (10)) also show evidence of greater shifts between the orange distribution and the blue distribution than near OODs.



(1) VQAv2 Val

(2) IV VQA

(3) CV VQA

(4) VQA Rephrasings

(5) VQA CP v2

(6) VQA CE

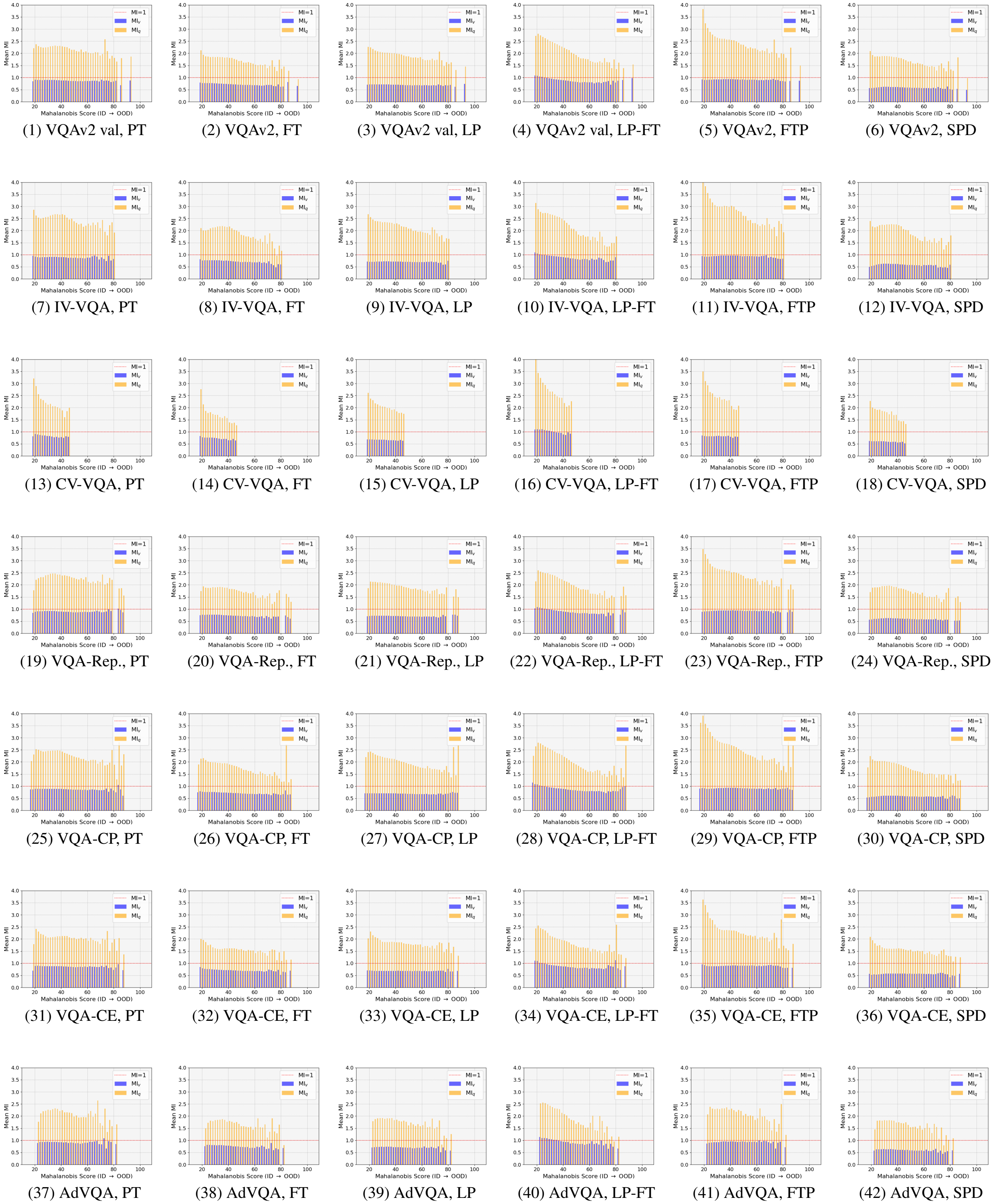
(7) ADVQA

(8) Text VQA

(9) VizWiz

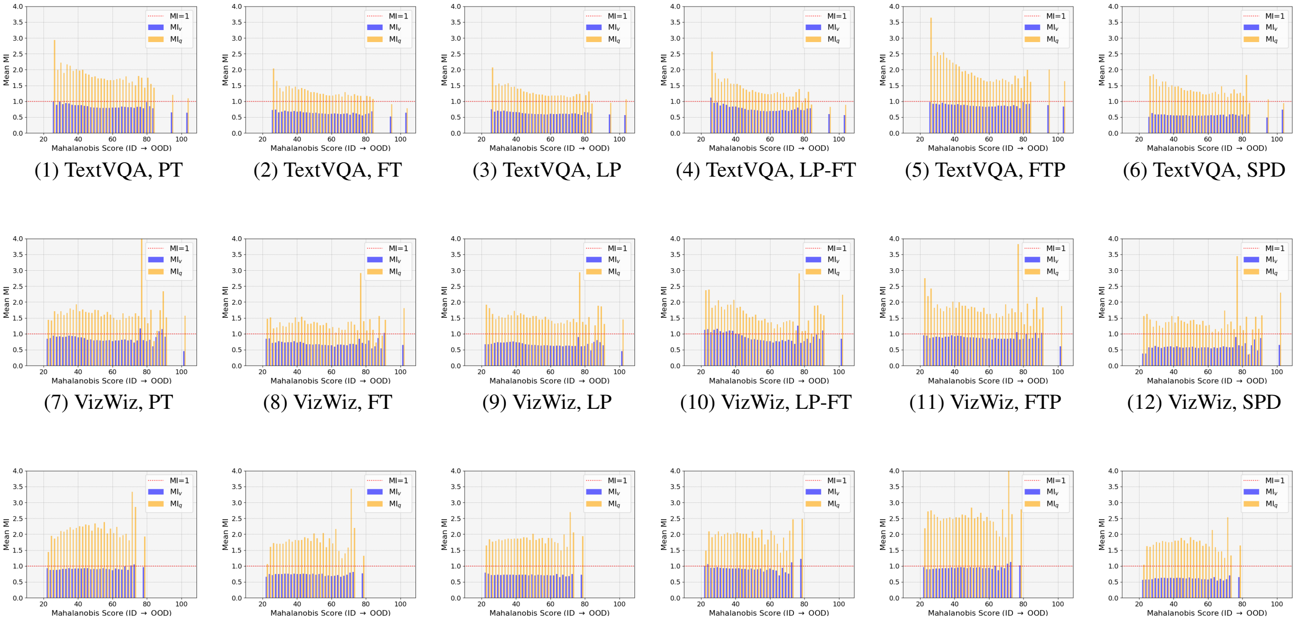
(10) OKVQA

Figure 12. Histogram for Vanilla FT V+Q Shifts : We depict the S_{Maha} score on the V+Q shift for each sample in the VQAv2 train split in blue and the corresponding test samples in orange. For all test splits, V+Q shifts show a greater degree of shift compared to the corresponding visual and question shift.



(1) VQAv2 val, PT
(2) VQAv2, FT
(3) VQAv2 val, LP
(4) VQAv2 val, LP-FT
(5) VQAv2, FTP
(6) VQAv2, SPD
(7) IV-VQA, PT
(8) IV-VQA, FT
(9) IV-VQA, LP
(10) IV-VQA, LP-FT
(11) IV-VQA, FTP
(12) IV-VQA, SPD
(13) CV-VQA, PT
(14) CV-VQA, FT
(15) CV-VQA, LP
(16) CV-VQA, LP-FT
(17) CV-VQA, FTP
(18) CV-VQA, SPD
(19) VQA-Rep., PT
(20) VQA-Rep., FT
(21) VQA-Rep., LP
(22) VQA-Rep., LP-FT
(23) VQA-Rep., FTP
(24) VQA-Rep., SPD
(25) VQA-CP, PT
(26) VQA-CP, FT
(27) VQA-CP, LP
(28) VQA-CP, LP-FT
(29) VQA-CP, FTP
(30) VQA-CP, SPD
(31) VQA-CE, PT
(32) VQA-CE, FT
(33) VQA-CE, LP
(34) VQA-CE, LP-FT
(35) VQA-CE, FTP
(36) VQA-CE, SPD
(37) AdVQA, PT
(38) AdVQA, FT
(39) AdVQA, LP
(40) AdVQA, LP-FT
(41) AdVQA, FTP
(42) AdVQA, SPD

Figure 13. Variation of MI_v and MI_q w.r.t. shift score under PT, FT, LP, LP-FT, FTP and SPD across all ID and near OOD datasets. The blue and orange bars represent MI_v and MI_q respectively. The red dotted line marks a reference MI of 1.



(1) TextVQA, PT (2) TextVQA, FT (3) TextVQA, LP (4) TextVQA, LP-FT (5) TextVQA, FTP (6) TextVQA, SPD

(7) VizWiz, PT (8) VizWiz, FT (9) VizWiz, LP (10) VizWiz, LP-FT (11) VizWiz, FTP (12) VizWiz, SPD

(13) OK-VQA, PT (14) OK-VQA, FT (15) OK-VQA, LP (16) OK-VQA, LP-FT (17) OK-VQA, FTP (18) OK-VQA, SPD

Figure 14. Variation of MI$_v$ and MI$_q$ w.r.t. shift score under PT, FT, LP, LP-FT, FTP and SPD across all far OOD datasets. The blue and orange bars represent MI$_v$ and MI$_q$ respectively. The red dotted line marks a reference MI of 1.