

## A APPENDIX

### A.1 TRAINING DETAILS

**Image Classification.** For *DiGraP*, we fine-tune the model using SGD with a learning rate of  $1e-2$  and  $\mu = 0.1$  with a batchsize of 256. The regularization hyper-parameter is found through cross-validation, and the model with the best ID validation accuracy is taken. We use 4 RTX 2080 GPUs for each experiment.

**Fine-Tuning DomainNet-oVQA.** 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 (Li et al., 2022) 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. 2. For LoRA (Hu et al., 2021), 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. 2. We use  $\lambda = 0.5$  for all *DiGraP* results in Tab. 2. 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.

**Fine-tuning VQA.** 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 (Li et al., 2022) 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 (Hu et al., 2021), 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. We use  $\lambda = 0.5$  for all *DiGraP* results 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.

### A.2 MEASURING OOD DISTANCE

We follow procedures similar to typical feature-based OOD detection methods (Shi & Lee, 2024). Specifically, given our input training split  $X_{in}^{train}$ , we compute feature representations  $z$  of the training samples to estimate the empirical mean  $\mu$  and covariance matrix  $\Sigma$ . For each test split, we compute the test set shift relative to the training domain using the negative Mahalanobis distance metric defined in Eq. 9. The overall shift score for each test dataset, denoted as  $S_{maha}$ , is calculated as the average  $S_{Maha}$  across all samples. Let  $q$  denote the question,  $v$  the image (vision input), and  $a$  the answer. The input features used in measuring shifts include uni-modal embeddings  $f(v)$  and joint embeddings  $f(q, v), f(q, v, a)$ .

$$S_{Maha}(z_{test}) = -\sqrt{(z_{test} - \mu)^T \Sigma^{-1} (z_{test} - \mu)} \quad (9)$$

We utilize the vanilla fine-tuned Paligemma model on the VQAv2 training dataset as our feature encoder. The embedding  $f(v)$  is obtained from the image encoder, taking the mean output across 256 hidden states. For the joint embeddings, we extract the last layer output of the Paligemma model, with  $f(q, v)$  obtained via mean pooling. The embedding  $f(q, v, a)$  follows the same procedure, with  $\{\text{Question} : q \text{ Answer} : a\}$  as the input for text encoder.



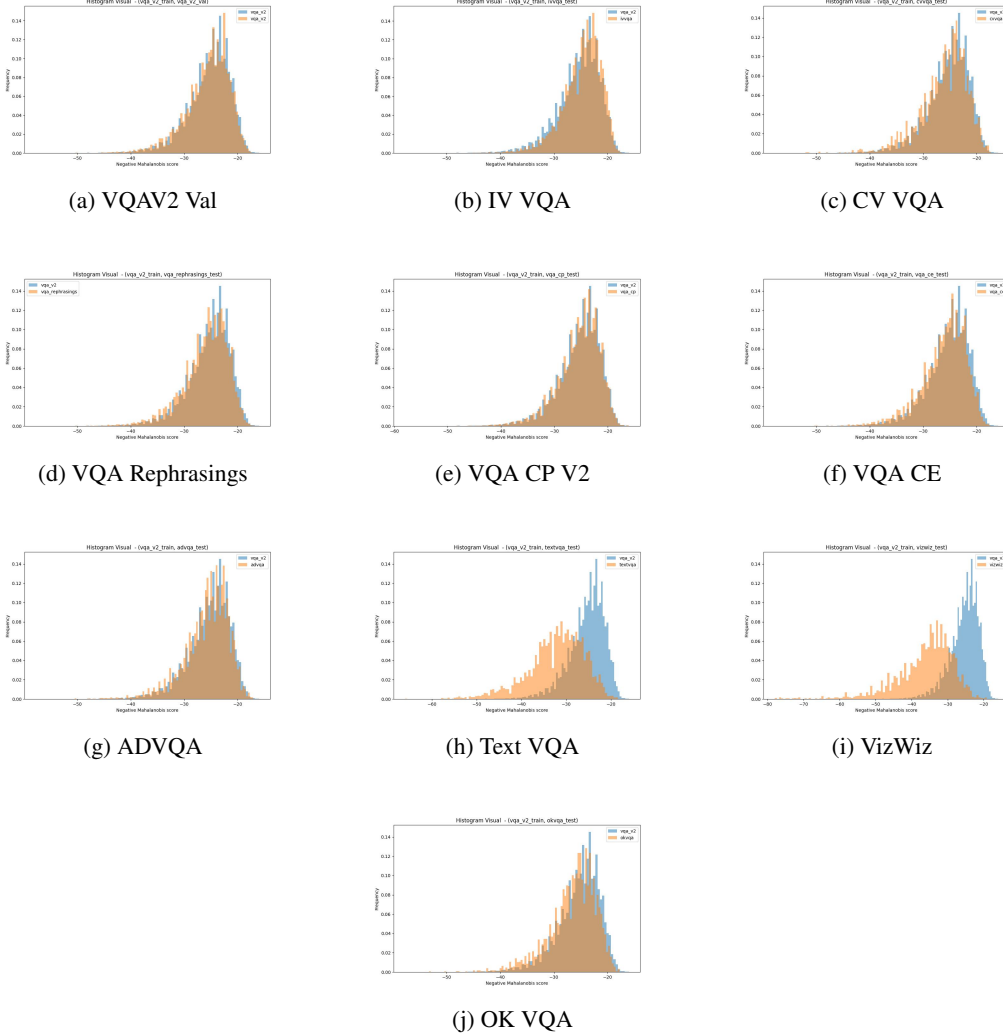


Figure 6: Histogram for Visual Shifts: We depict the  $S_{\text{Maha}}$  score on the visual modality for each sample in the VQAV2 train split in blue and the corresponding test samples in orange. Far OODs (Figures h, i, j) show evidence of greater shifts between the orange distribution and the blue distribution than Near OODs.

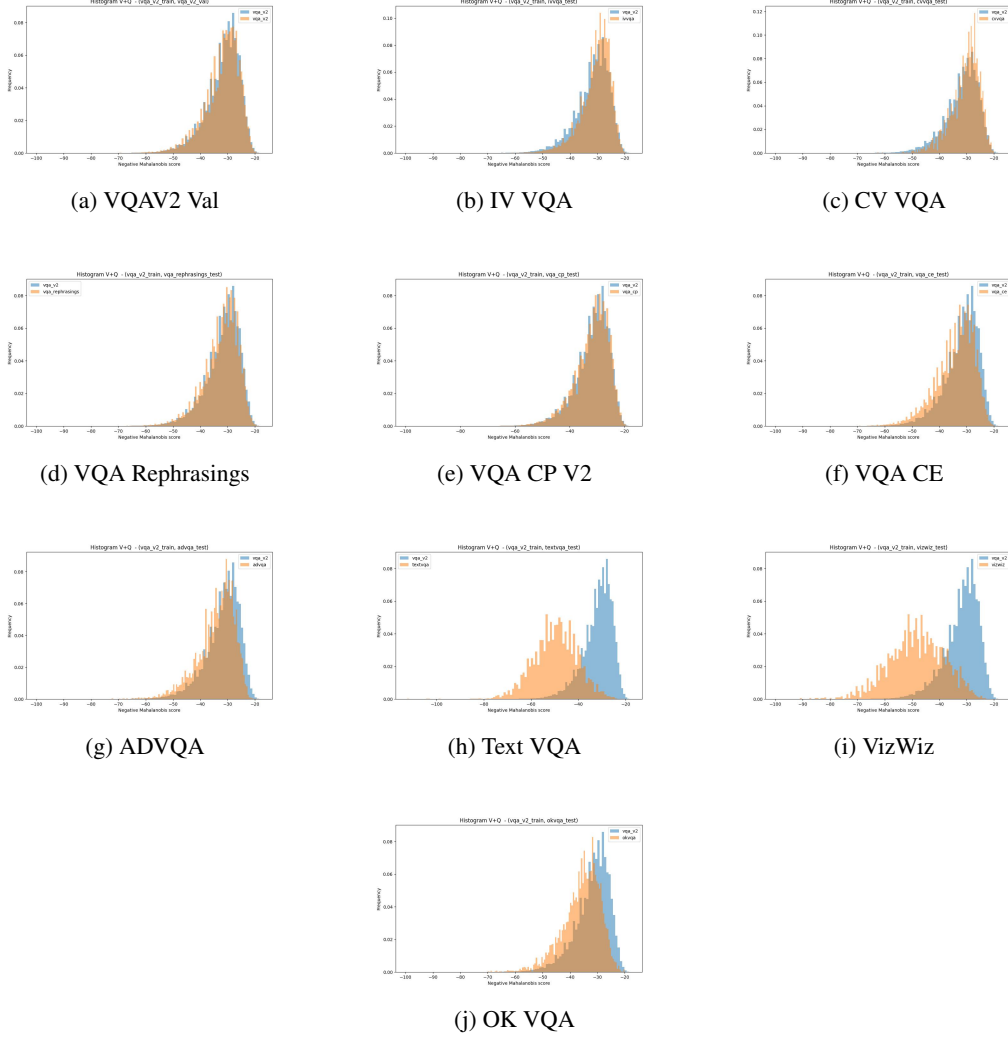


Figure 7: Histogram for V+Q Shifts: We depict the  $S_{\text{Maha}}$  score on the V+Q modality for each sample in the VQAV2 train split in blue and the corresponding test samples in orange. Similar to the Visual shift histograms, far OODs (Figures h, i, j) also show evidence of greater shifts between the orange distribution and the blue distribution than Near OODs. Further, for all test splits, V+Q show a greater degree of shift compared to the corresponding Visual shift

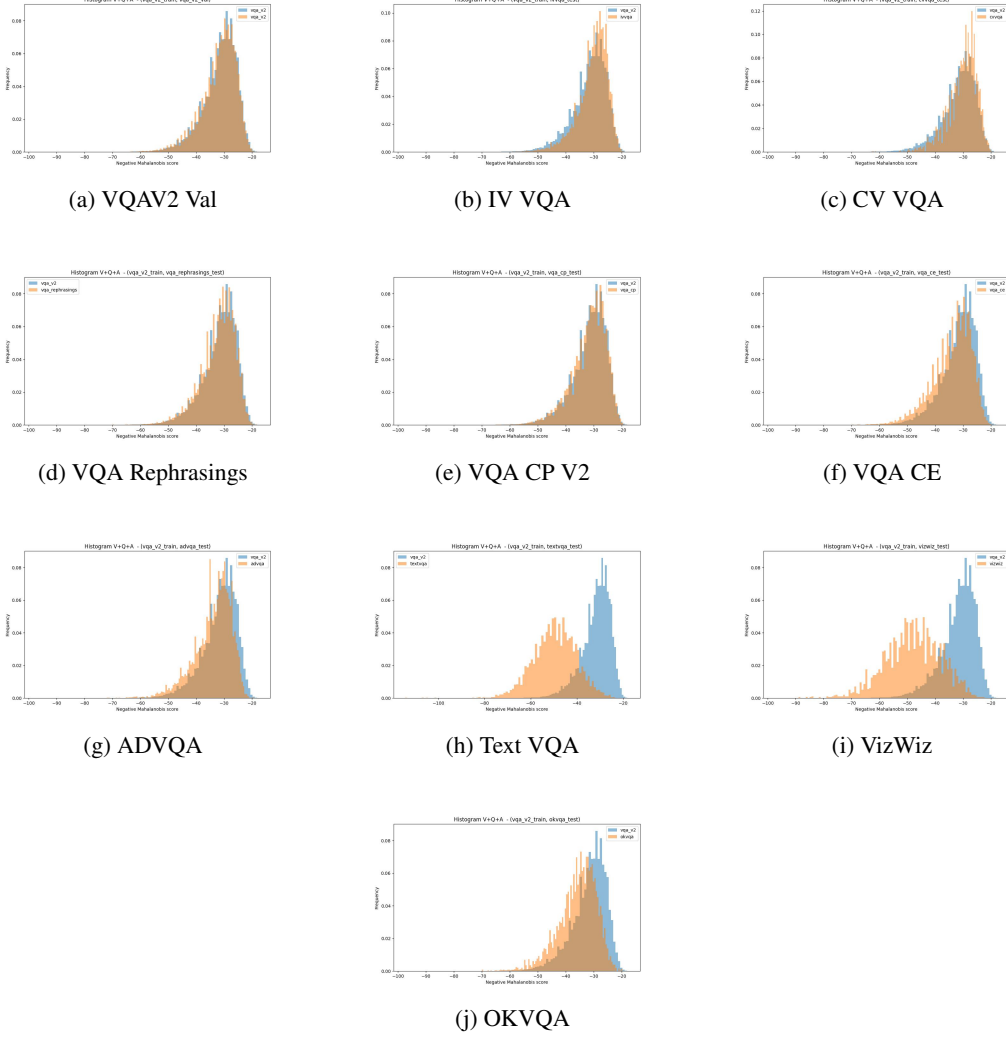


Figure 8: Histogram for V+Q+A Shifts : We depict the  $S_{Maha}$  score on the V+Q+A shift for each sample in the VQAV2 train split in blue and the corresponding test samples in orange.