Appendix

A Updated Results for VTAB

Our BayesTune training for the VTAB benchmark has been in progress, and we report the latest results here in Table 3, which can replace our older version Table 2 in the main paper. Now, we can see that the number of datasets where BayesTune achieves Rank 1 is increased from 6 to 7, so becoming the best method among the competing approaches.

B Chosen Hyperparameters

We grid-search hyperparameters on validation, where the two key hyperparameters are: the effective training data size $\hat{N}$ and the noise discount factor γ (re: Sec. 4.1). The candidate sets are formed as: $\hat{N} \in \{10^8, 10^9, \cdots, 10^{12}\}$, $\gamma \in \{10^{-4}, 10^{-2}, 10^0\}$ for NLP, and $\hat{N} \in \{10^6, 10^7, \cdots, 10^{12}\}$, $\gamma \in \{10^{-4}, 10^{-3}, \cdots, 10^0\}$ for VTAB. The chosen hyperparameters are as follows $(\hat{N}, \gamma)$: (NLP) cola $= (11, 10^{-4})$, stsb $= (12, 10^{-4})$, mrpc $= (12, 10^0)$, rte $= (8, 10^{-4})$, cb $= (10, 10^{-4})$, copa $= (8, 10^{-2})$, wsc $= (10, 10^{-4})$; (VTAB) cifar100 $= (7, 10^{-1})$, caltech101 $= (9, 10^{-2})$, dtd $= (12, 10^0)$, flower102 $= (12, 10^{-2})$, pets $= (12, 10^0)$, svhn $= (10, 10^0)$, sun397 $= (7, 10^{-1})$, camelyon $= (6, 10^0)$, eurosat $= (7, 10^{-1})$, resisc45 $= (12, 10^{-2})$, retinopathy $= (7, 10^{-2})$, clevr-count $= (7, 10^{-3})$, clevr-dist $= (7, 10^{-3})$, dmlab $= (8, 10^0)$, kitti $= (7, 10^0)$, dsprite-loc $= (12, 10^{-4})$, dsprite-ori $= (12, 10^{-3})$, snorb-azim $= (7, 10^{-2})$, snorb-ele $= (6, 10^{-1})$.

C More Analysis

(NLP) Test accuracies at other sparsity levels. Although $p = 0.005$ is recognized as the optimal sparsity level overall for the GLUE and SuperGLUE tasks, we evaluate the test performance of our BayesTune for different sparsity levels: $p \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$. The average test accuracies are shown in Fig. 4. We see that overall there is less significant change in test performance so long as the sparsity level p is small enough, and the resulting sparse updates selected by our BayesTune lead to equally good performance as those with the default value. However, increasing p further (e.g., $p = 0.5$) considerably degrades the performance, which signifies the importance of sparse fine-tuning to avoid potential overfitting.

(VTAB) Scale posterior mean $\hat{\lambda}$ vs. sparsity level p. We visualize the plots that relate the sorted scale posterior means $\hat{\lambda}$ to the sparsity levels p in Fig. 5. The plots are aligned with the the test accuracy plots analyzed in the main paper. As the plots are grouped along the optimal sparsity values, we see certain trends: for the *sparse group* (`sun397` and `cifar100`), the scale $\hat{\lambda}$ values are overall small scaled (in the range of $[0, 0.2]$) with sharp drops at small $\hat{\lambda}$; for the *dense group*: (`camelyon` and `dmlab`), $\hat{\lambda}$ scale is even larger (in the range of $[0, 0.5]$) with relatively smooth decaying at small values; lastly for the *in-between group* (`clever-dist`, `dspr-ori`, `kitti`, and `snorb-ele`), we have much narrower $\hat{\lambda}$ ranges in between 0 and 0.1 except for `kitti`.

Method	#param (M)	Cifar100	Caltech101	DTD	Flower102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Avg Rank (↓)	# Rank 1 (↑)
Full update	85.8	68.9	87.7	64.3	87.2	86.9	87.4	38.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1		
Linear	0.04	64.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.5	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2		
VPT [17]	0.64	**78.8**	90.8	65.8	98.0	88.3	78.1	49.6	81.8	**96.1**	83.4	68.4	68.5	60.0	46.5	72.8	73.6	47.9	**32.9**	37.8	4.16	3
Adapter [15]	0.16	69.2	90.1	68.0	98.8	89.9	82.8	**54.3**	84.0	94.9	81.9	75.5	80.9	65.3	48.6	78.3	74.8	48.5	29.9	41.6	3.68	1
LoRA [16]	0.29	67.1	91.4	69.4	98.8	90.4	85.3	54.0	**84.9**	95.3	**84.4**	73.6	**82.9**	**69.2**	49.8	78.5	75.7	47.1	31.0	44.0	2.68	4
NOAH [36]	0.43	69.6	**92.7**	**70.2**	**99.1**	90.4	86.1	53.7	84.4	95.4	83.9	75.8	82.8	68.9	**49.9**	81.7	81.8	48.3	32.8	**44.2**	**1.95**	5
BayesTune	Avg 0.37	68.9 (.07)	92.6 (.37)	69.5 (.04)	**99.1** (.37)	**90.8** (.15)	**88.1** (.67)	50.0 (.04)	84.6 (.60)	95.8 (.60)	82.8 (.37)	**76.0** (.07)	82.6 (.30)	67.4 (.22)	49.6 (.52)	**82.3** (.60)	**81.9** (.60)	**49.9** (.30)	22.6 (.52)	39.3 (.67)	2.37	**7**

Table 3: (**Latest**) VTAB-1K results. The accuracies at the optimal sparsity levels are reported for our BayesTune. For BayesTune, the optimal number of the updated parameters is dataset-dependent, and these optimal numbers are depicted in the parentheses. The figures of the competing methods are exerpted from [17, 15, 16, 36].

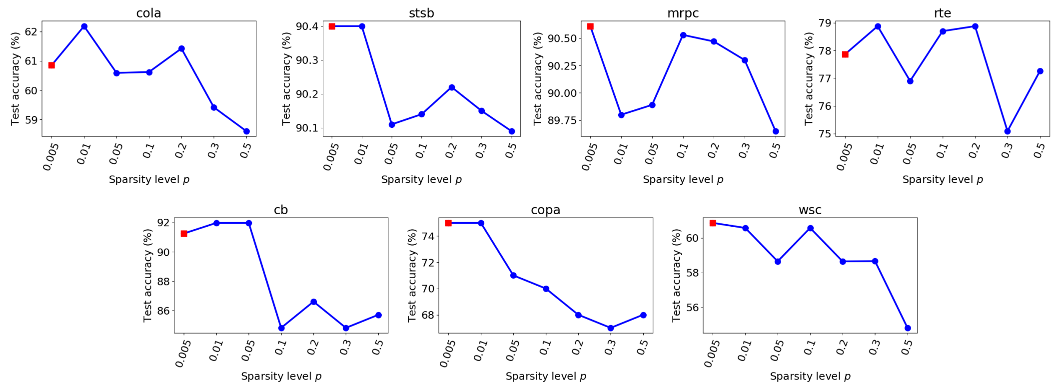
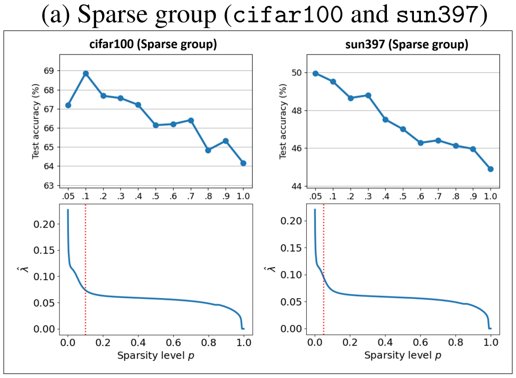
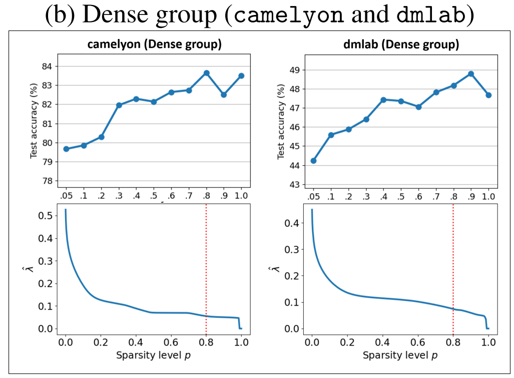


Figure 4: (NLP benchmarks) Test accuracies at sparsity levels other than the default $p = 0.005$. We evaluate the BayesTune sparse update models with $p \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$, where the default ones $p = 0.005$ are shown as red square markers.

(a) Sparse group (`cifar100` and `sun397`)



(b) Dense group (`camelyon` and `dmlab`)



(c) In-between group (`clever-dist`, `dspr-ori`, `kitti`, and `snorb-ele`)

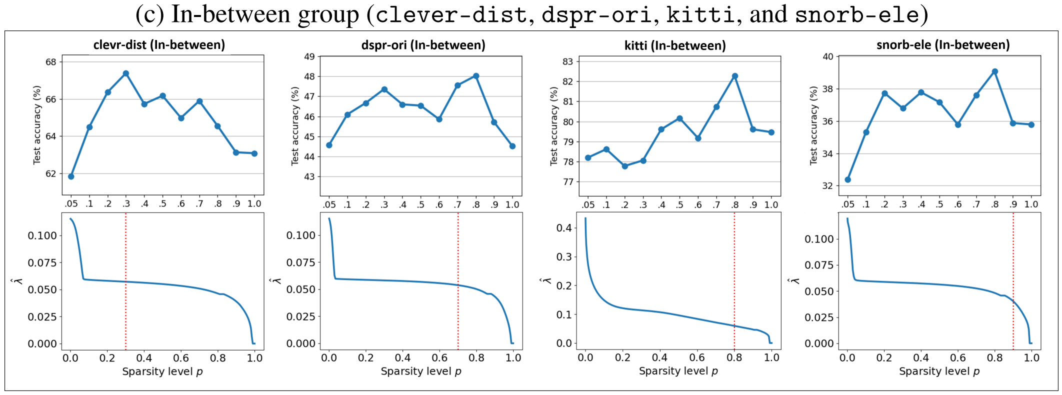


Figure 5: (VTAB benchmarks) The plots of the sorted scale posterior means $\hat{\lambda}$ vs. sparsity levels p, each of which is aligned with the the corresponding test accuracy plot. The plots are grouped along the optimal sparsity values where each group exhibits similar trends.

D Layer-wise and Module-wise Sparsity Patterns of BayesTuned Networks

Sparsity patterns of RoBERTa-base on NLP tasks. We visualize the module-wise and layer-wise sparsity patterns of the BayesTuned RoBERTa-base networks on 7 NLP tasks in Fig. 6–19. First, for the layer-wise sparsity pattern: (Except for `mrpc`) The proportions of the selected updatable parameters are more or less uniformly distributed across the 12 layers of the Transformer, while the first word embedding layer and the last classification layer are significantly less and more selected, respectively. This is intuitively appealing as the task-specific features may tend to be determined at the higher, more global levels in texts/sentences, to account for longer-range dependency. Next, looking at the module-wise sparsity patterns, the proportions are highly non-uniform, layer-specific, and also task/dataset-dependent. For instance, the bias modules in some layers are very densely selected, while they are very sparsely selected in other layers. This shows clear discrepancy to the heuristic strategies like BitFit [34] in which the bias modules are selected 100% for all layers.

Sparsity patterns of ViT-B/16 on VTAB vision tasks. The module-wise and layer-wise sparsity patterns of the BayesTuned ViT-B/16 networks on VTAB benchmark datasets are shown in Fig. 20–38. We also superimpose the optimal p values (dataset dependent). The resulting patterns are quite similar to the NLP case: Except for a few cases, the lowest level visual prompt layers are selected far less, sometimes ignored, compared to the later layers. The last linear classification head, although not shown here in the sparsity diagrams, is selected 100%. Overall the layer-wise selection patterns are nearly uniform while the module-wise selection patterns are highly non-uniform and dataset dependent.

`cola` (**Module-wise** sparsity pattern)

Figure 6: Sparsity pattern of the modules in RoBERTa-base on `cola`. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

`cola` (**Layer-wise** sparsity pattern)

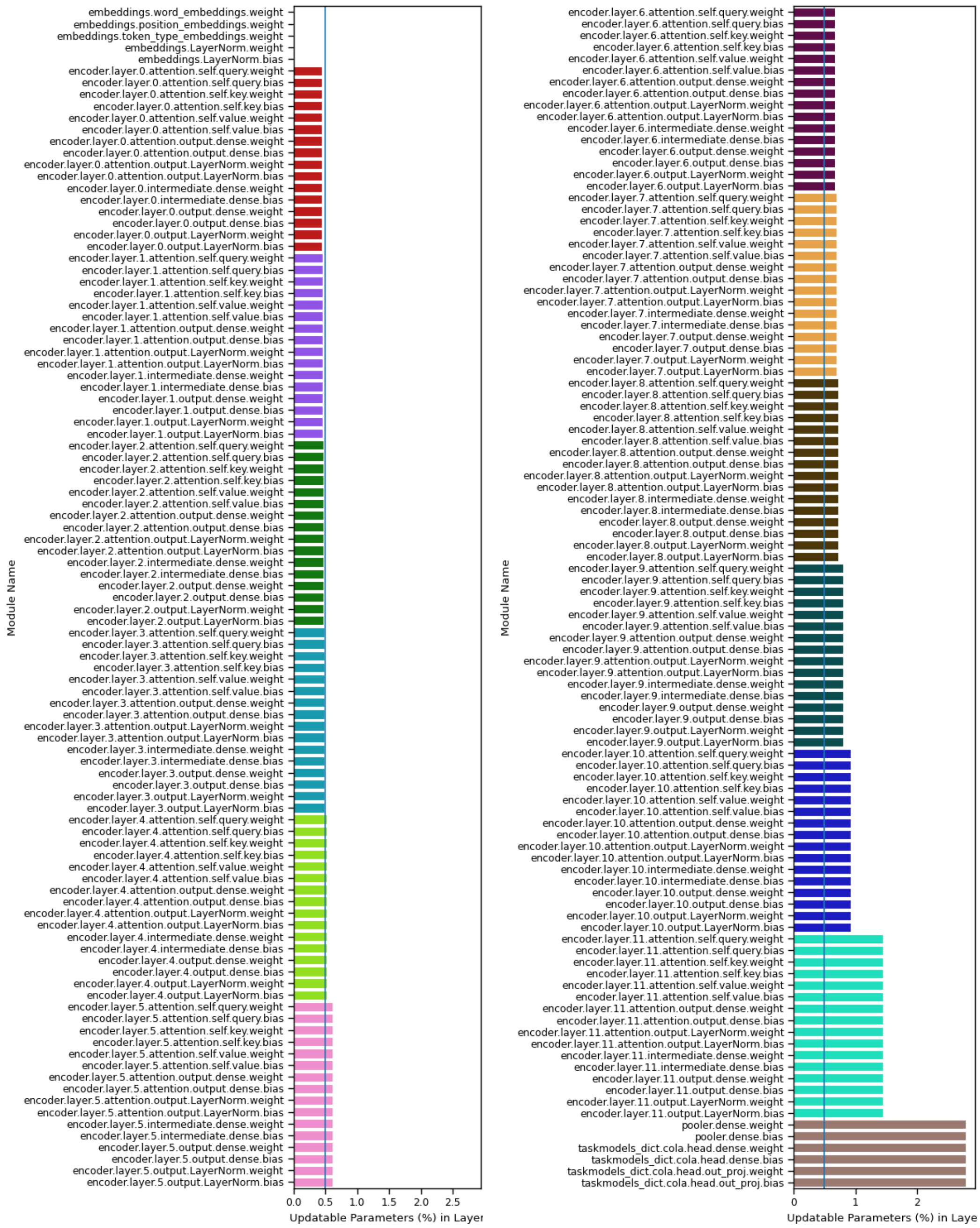


Figure 7: Sparsity pattern of the layers in RoBERTa-base on `cola`. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

stsb (**Module-wise** sparsity pattern)

Figure 8: Sparsity pattern of the modules in RoBERTa-base on stsb. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

`stsb` (**Layer-wise** sparsity pattern)

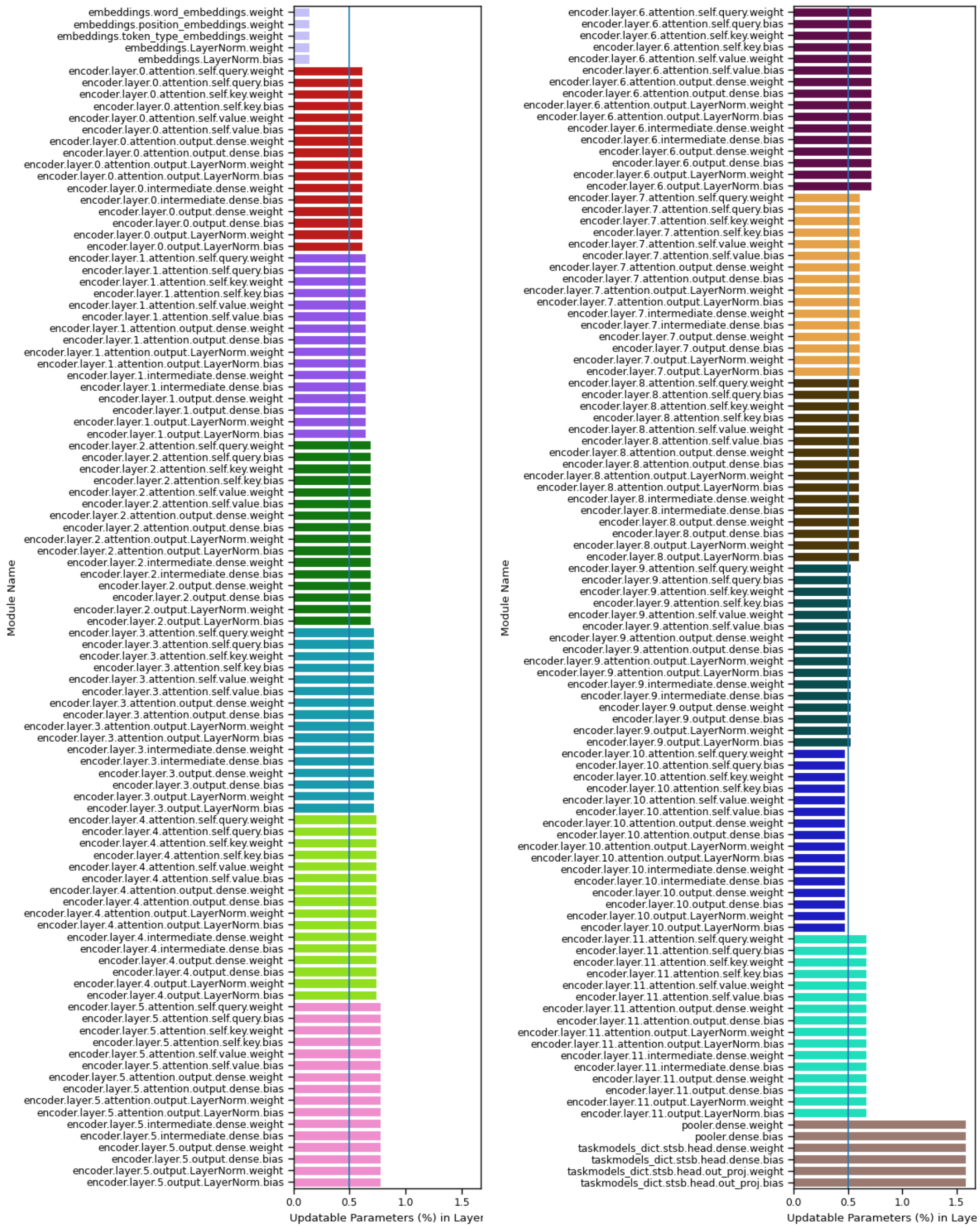


Figure 9: Sparsity pattern of the layers in RoBERTa-base on `stsb`. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

mrpc (**Module-wise** sparsity pattern)

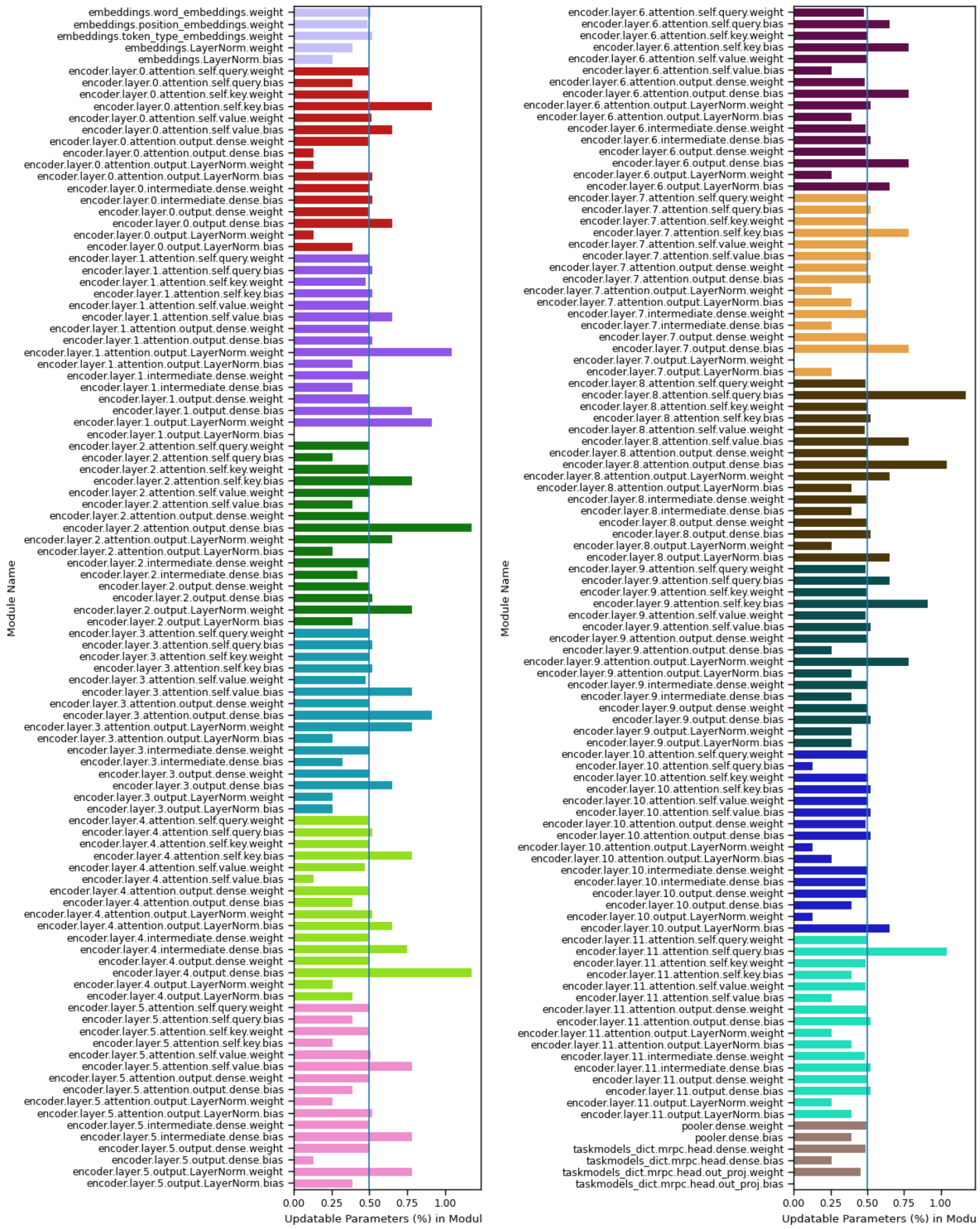


Figure 10: Sparsity pattern of the modules in RoBERTa-base on mrpc. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

`mrpc` (**Layer-wise** sparsity pattern)

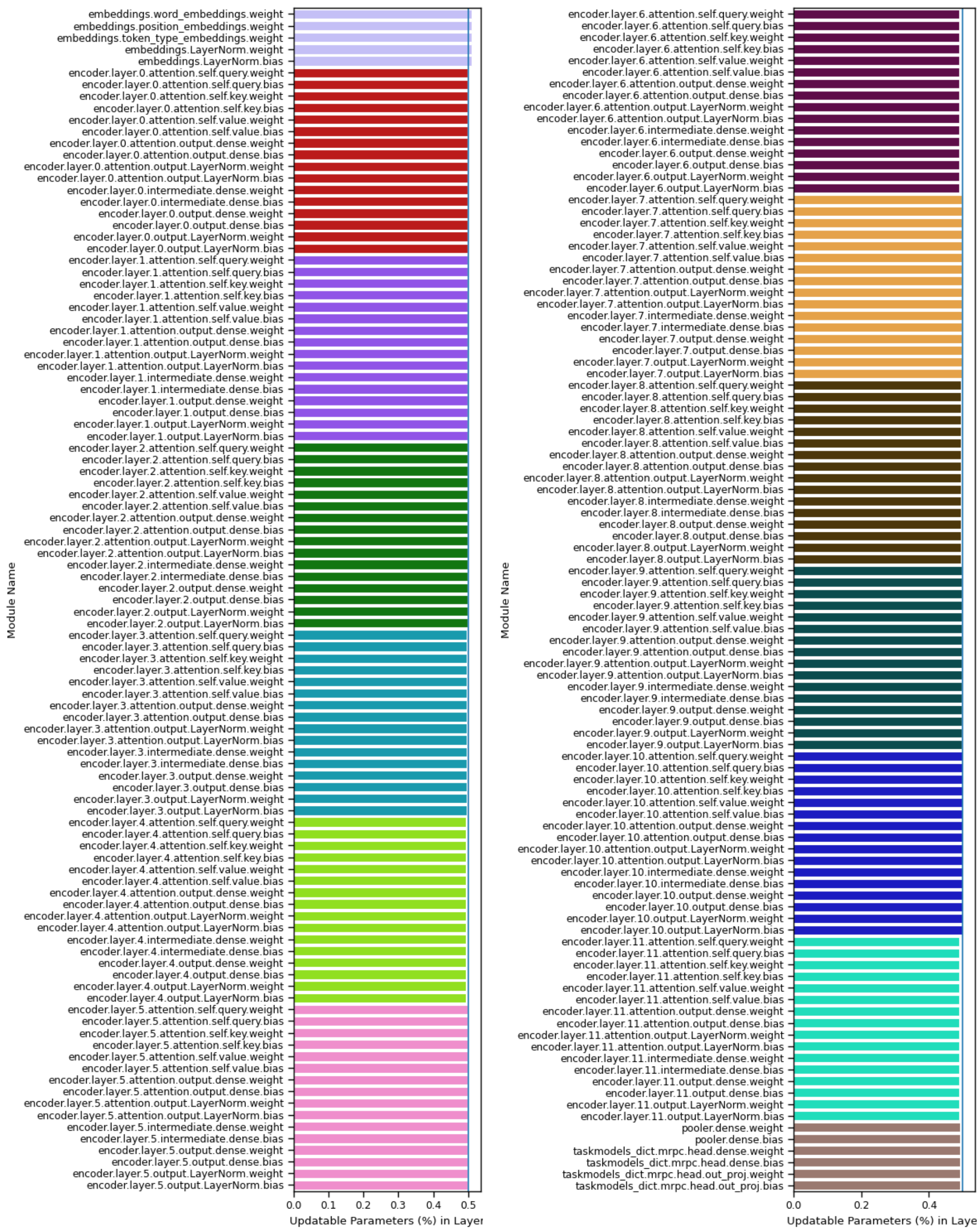


Figure 11: Sparsity pattern of the layers in RoBERTa-base on `mrpc`. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

Figure 12: Sparsity pattern of the modules in RoBERTa-base on `rte`. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

`rte` (**Layer-wise** sparsity pattern)

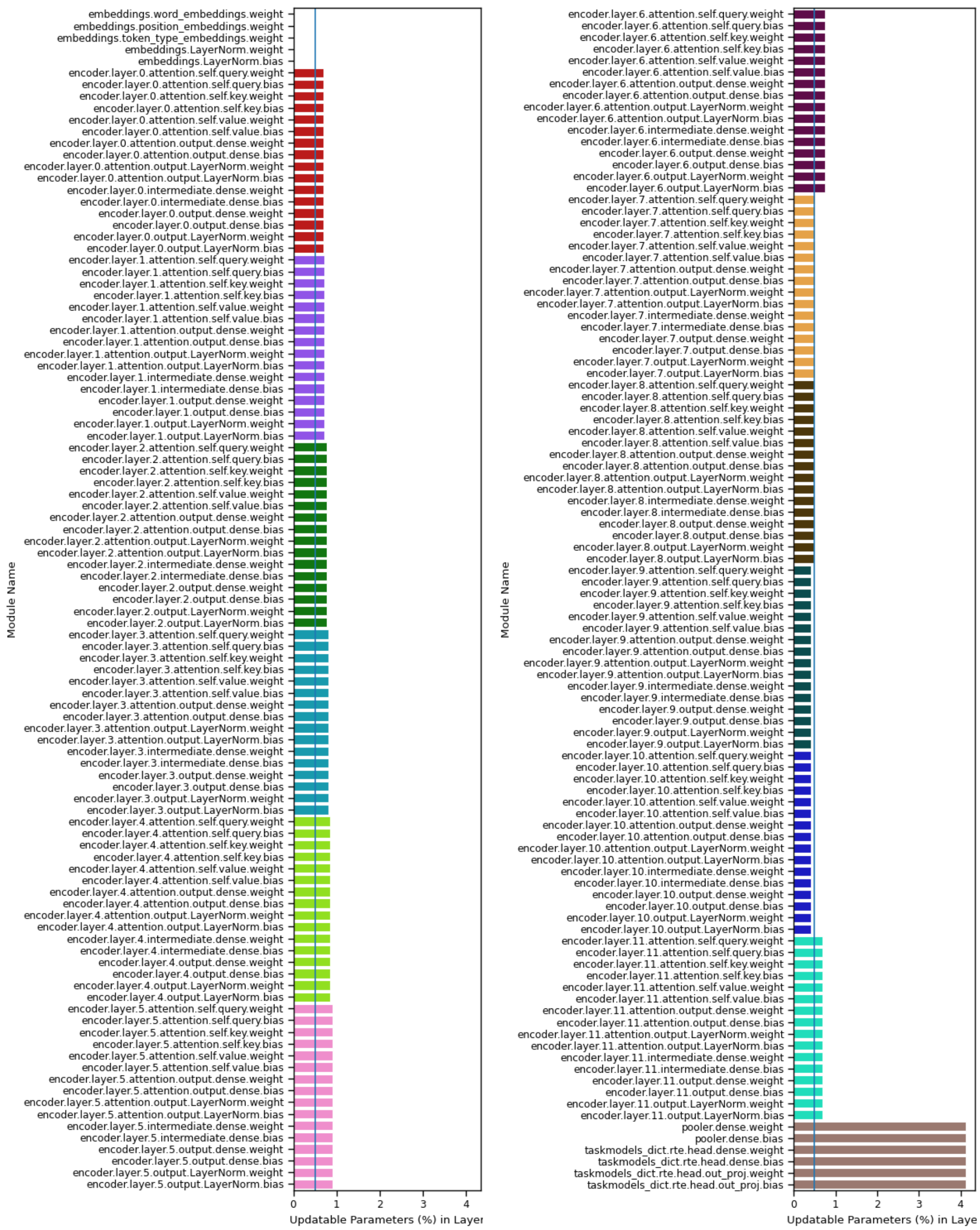


Figure 13: Sparsity pattern of the layers in RoBERTa-base on `rte`. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

`cb` (**Module-wise** sparsity pattern)

Figure 14: Sparsity pattern of the modules in RoBERTa-base on `cb`. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

cb (**Layer-wise** sparsity pattern)

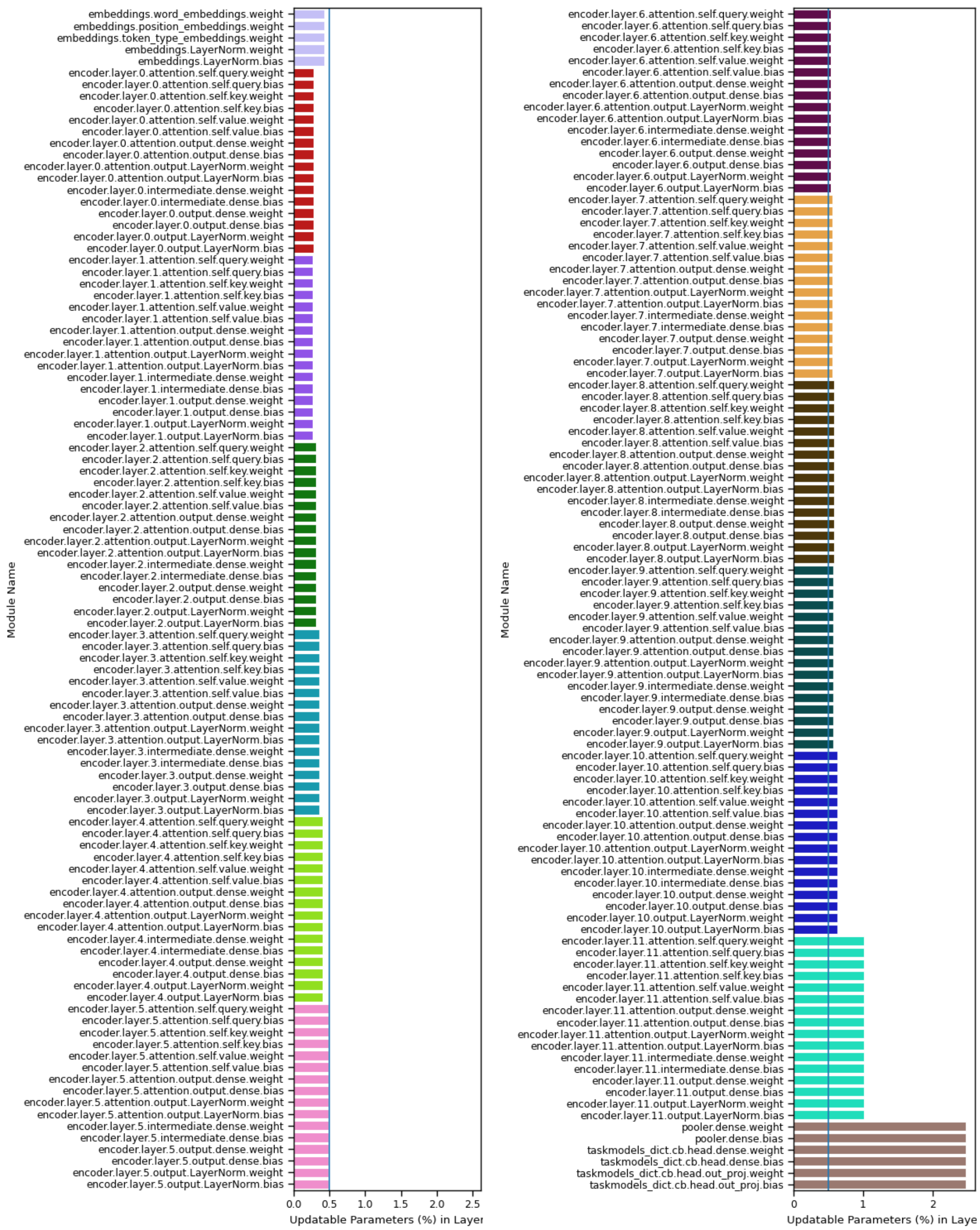


Figure 15: Sparsity pattern of the layers in RoBERTa-base on cb. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

copa (**Module-wise** sparsity pattern)

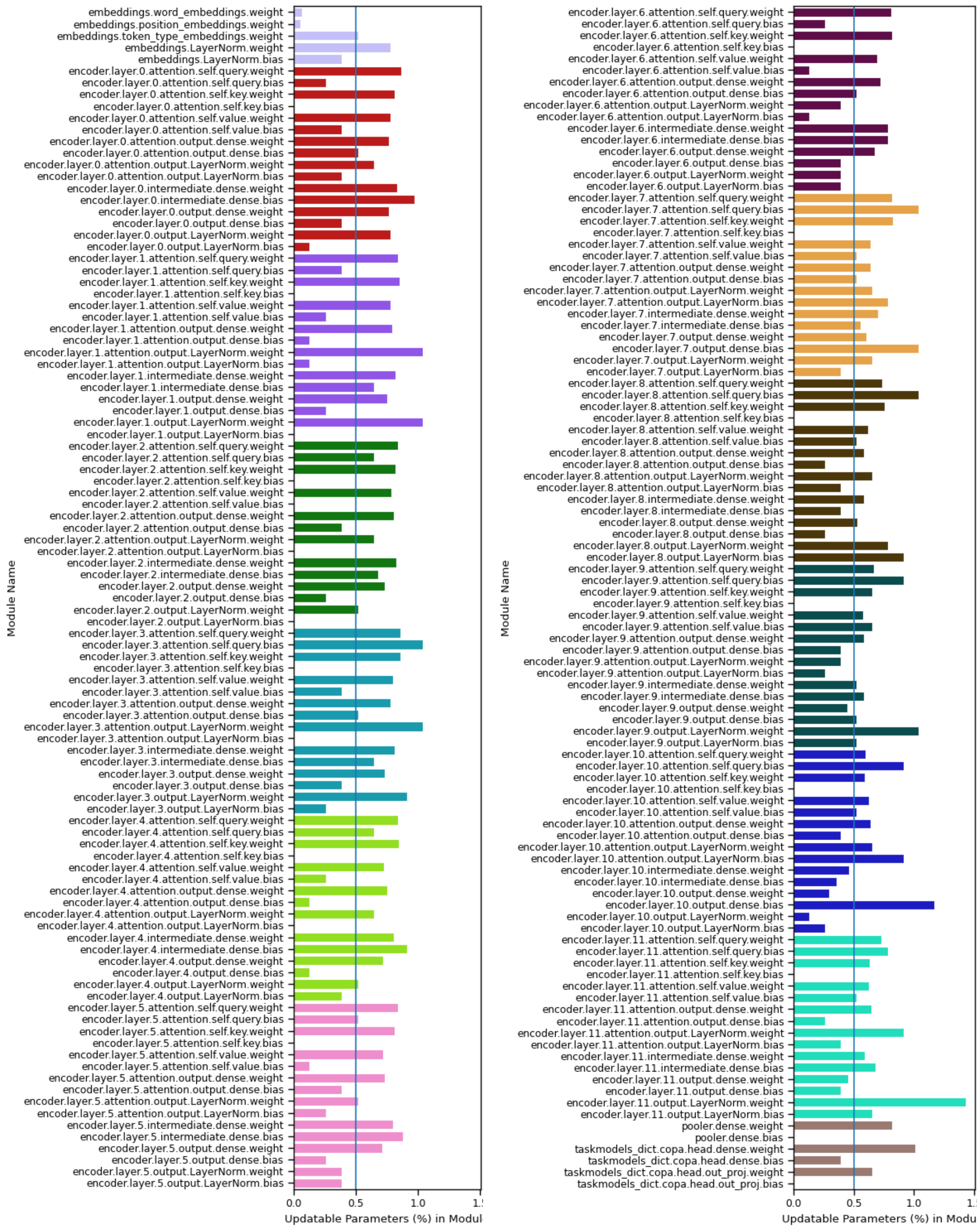


Figure 16: Sparsity pattern of the modules in RoBERTa-base on copa. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

copa (**Layer-wise** sparsity pattern)

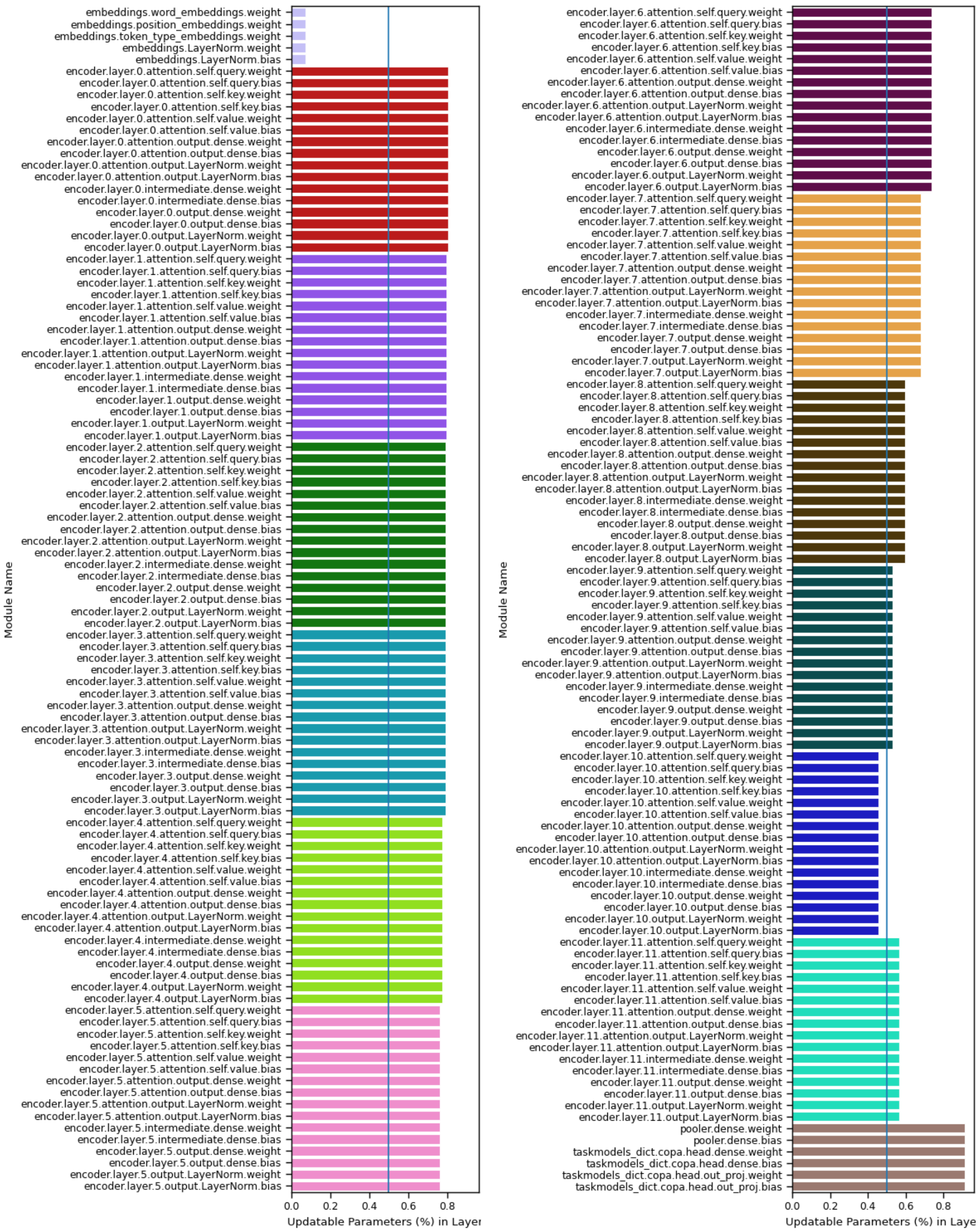


Figure 17: Sparsity pattern of the layers in RoBERTa-base on copa. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

wsc (**Module-wise** sparsity pattern)

Figure 18: Sparsity pattern of the modules in RoBERTa-base on wsc. Module-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

wsc (**Layer-wise** sparsity pattern)

Figure 19: Sparsity pattern of the layers in RoBERTa-base on wsc. Layer-wise updatable parameters (%). (Left) The first half of the network and (Right) the second half. The default sparsity level $p^* = 0.5\%$ is shown as vertical line.

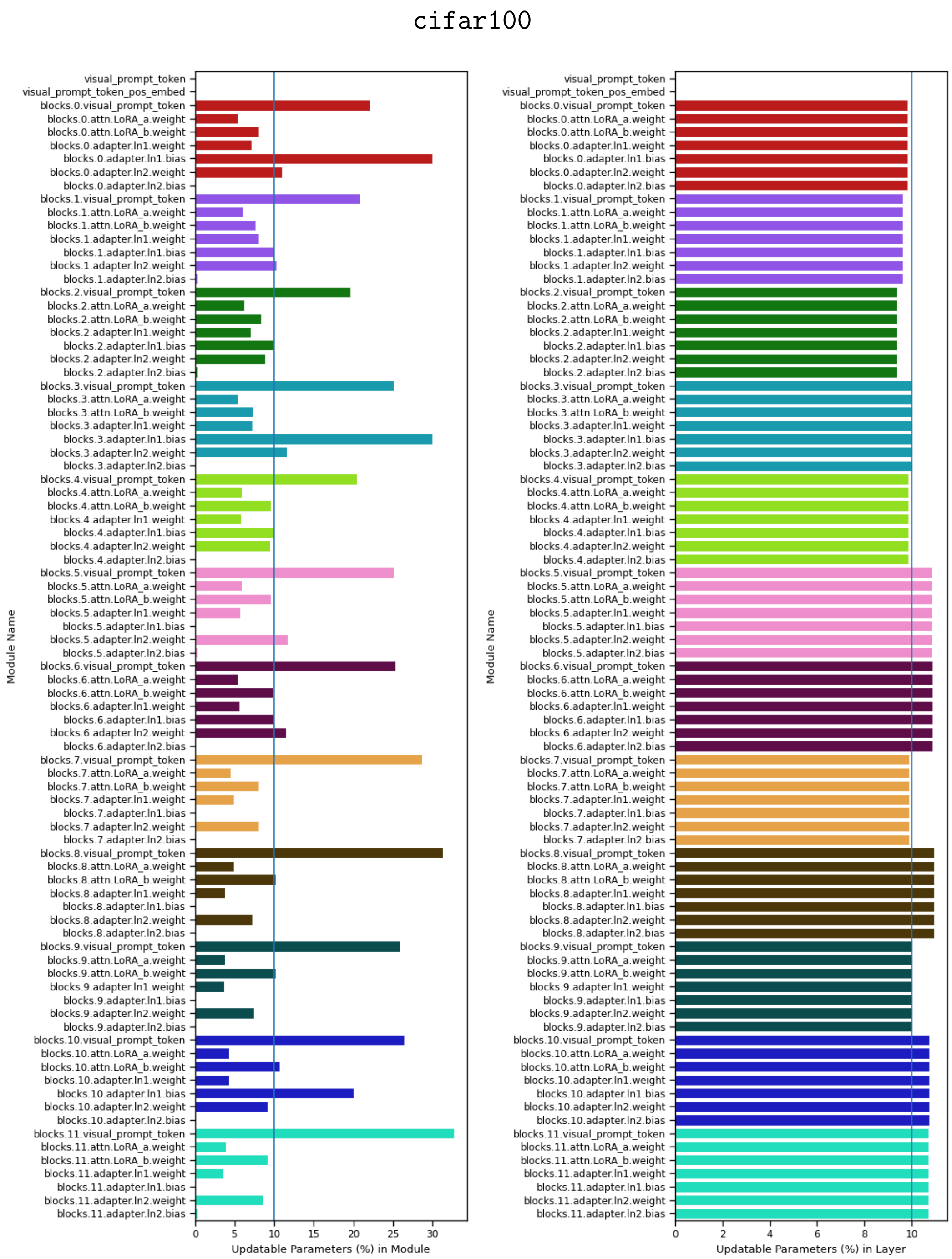


Figure 20: Sparsity pattern of attached modules to ViT-B/16 on `cifar100`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 10\%$ is shown as vertical line.

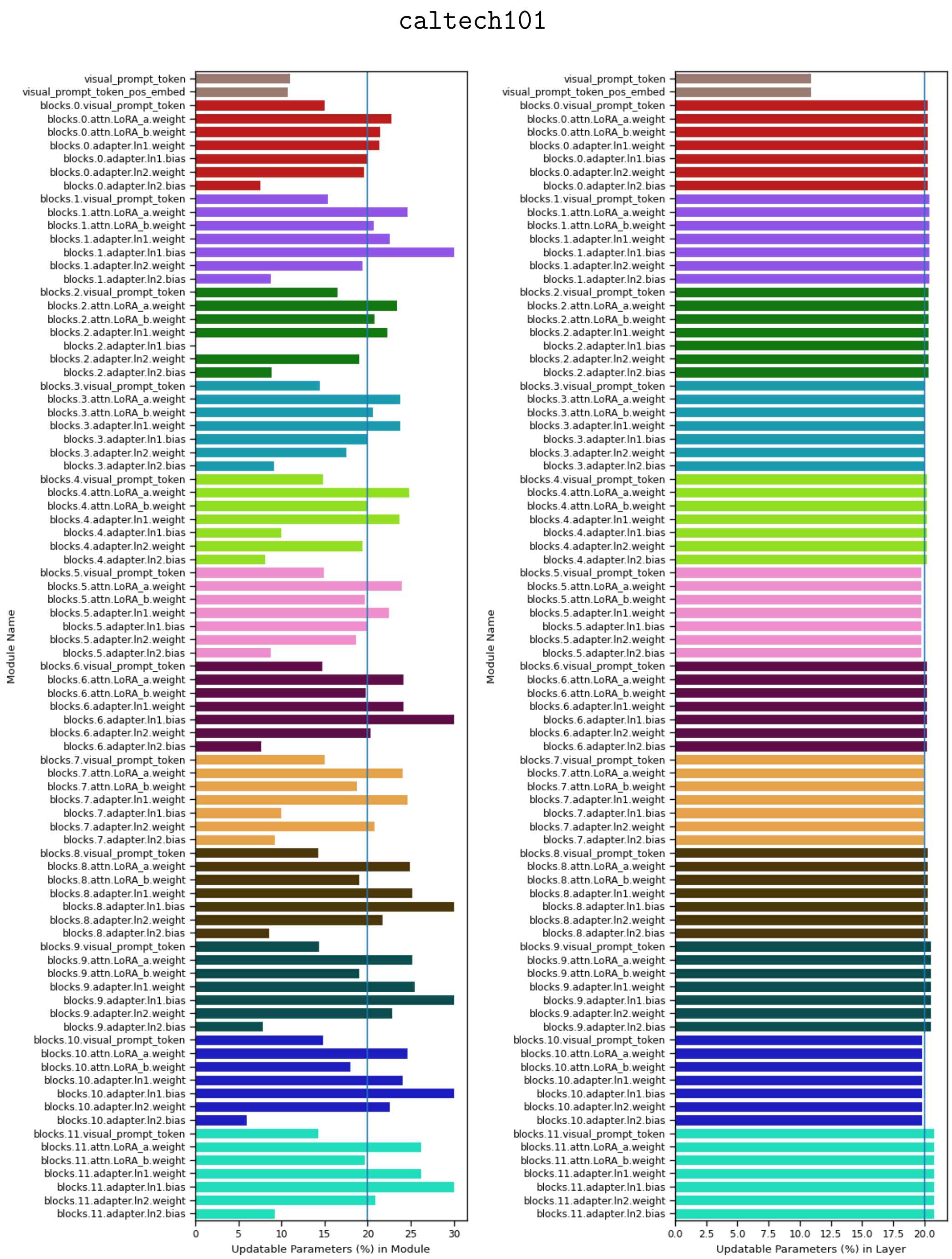


Figure 21: Sparsity pattern of attached modules to ViT-B/16 on `caltech101`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 20\%$ is shown as vertical line.

dtd

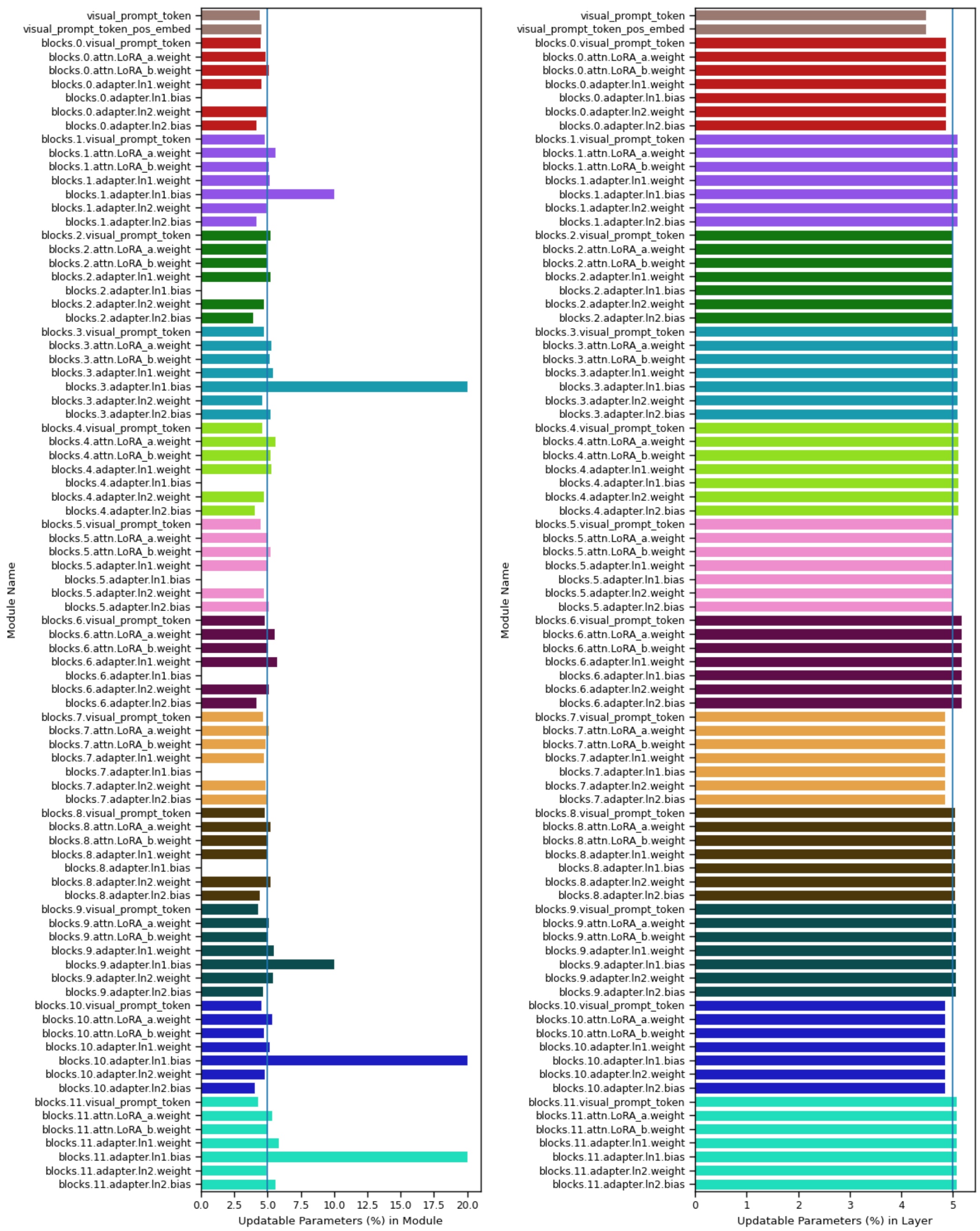


Figure 22: Sparsity pattern of attached modules to ViT-B/16 on `dtd`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 5\%$ is shown as vertical line.

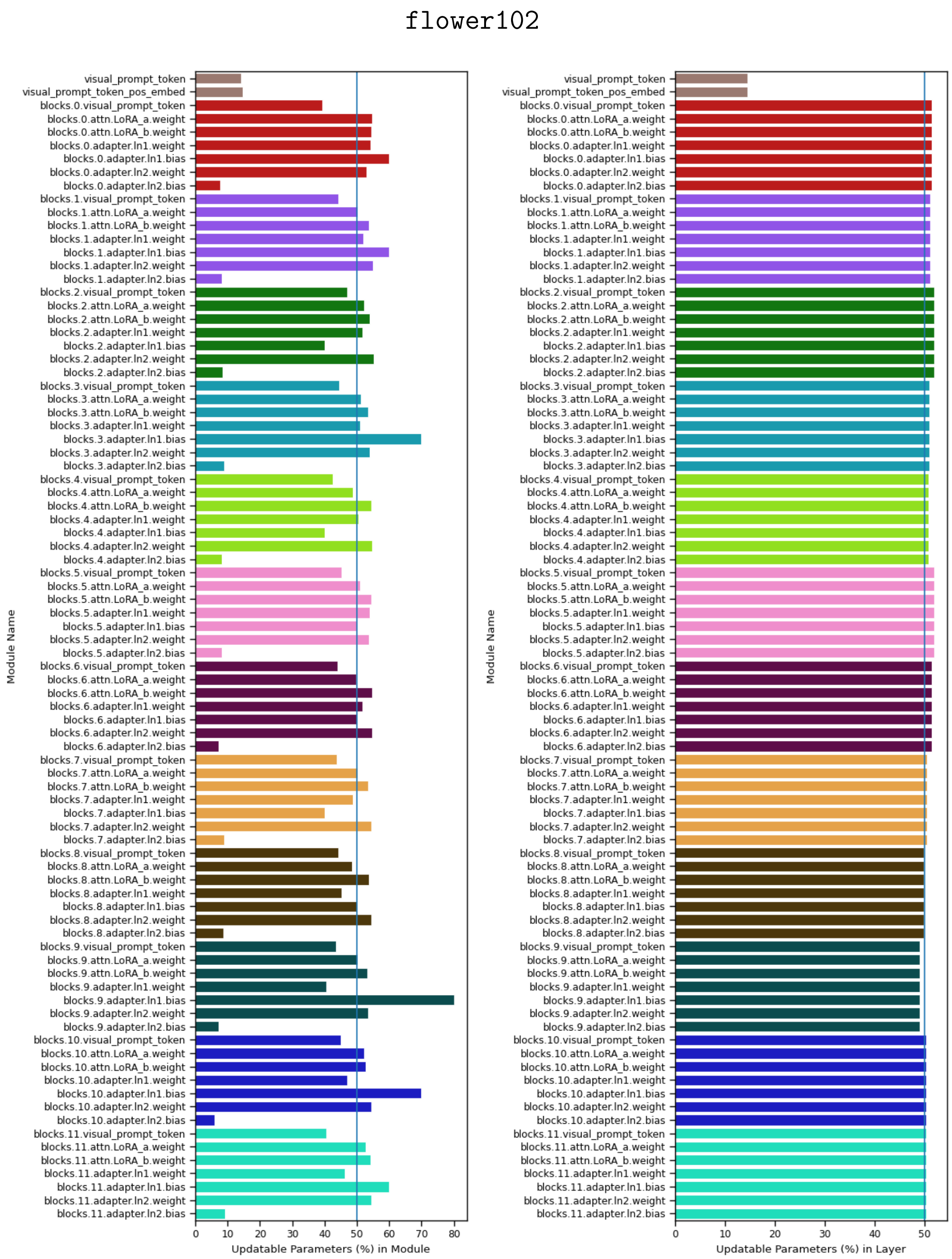


Figure 23: Sparsity pattern of attached modules to ViT-B/16 on `flower102`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 50\%$ is shown as vertical line.

pets

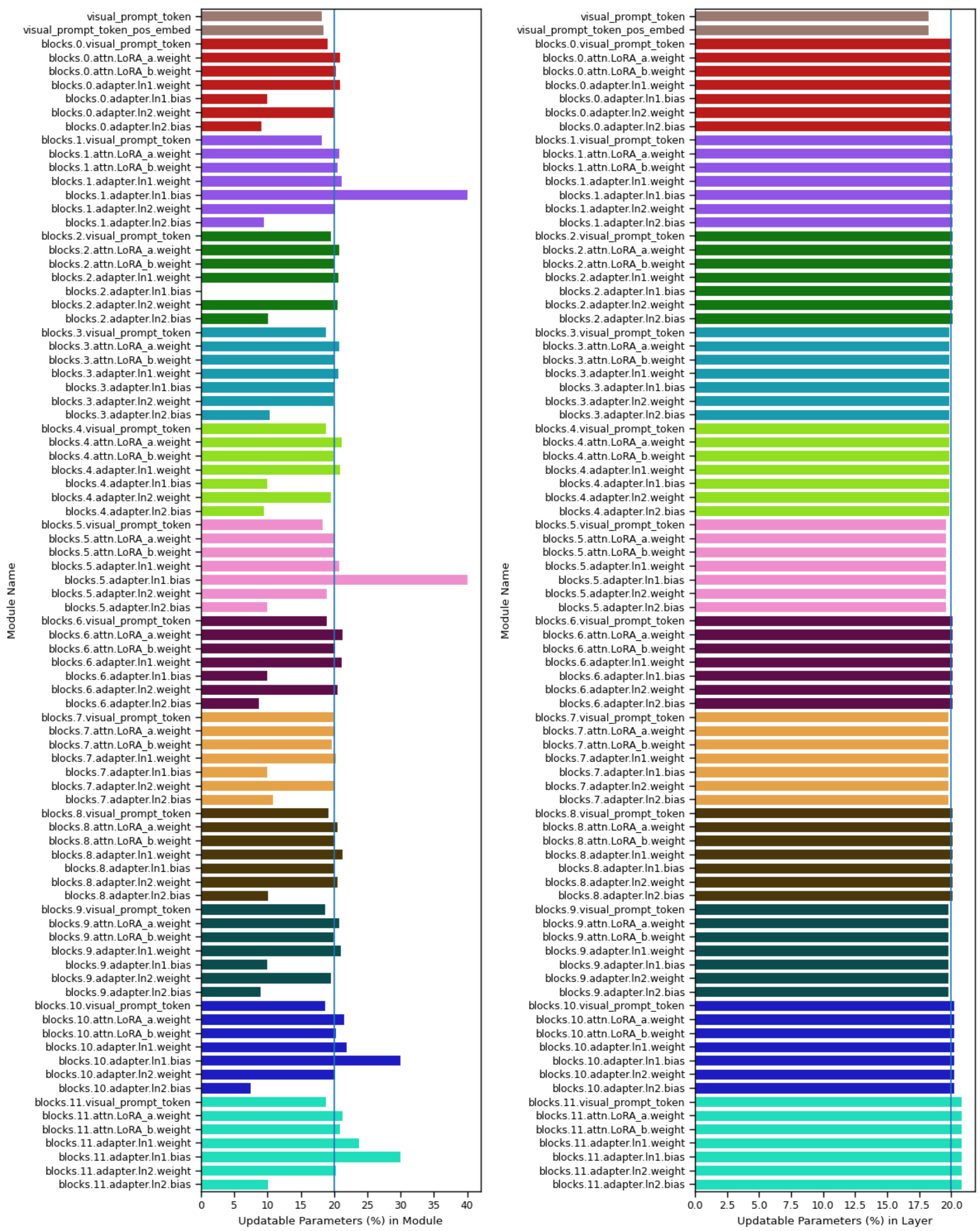


Figure 24: Sparsity pattern of attached modules to ViT-B/16 on pets. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 20\%$ is shown as vertical line.

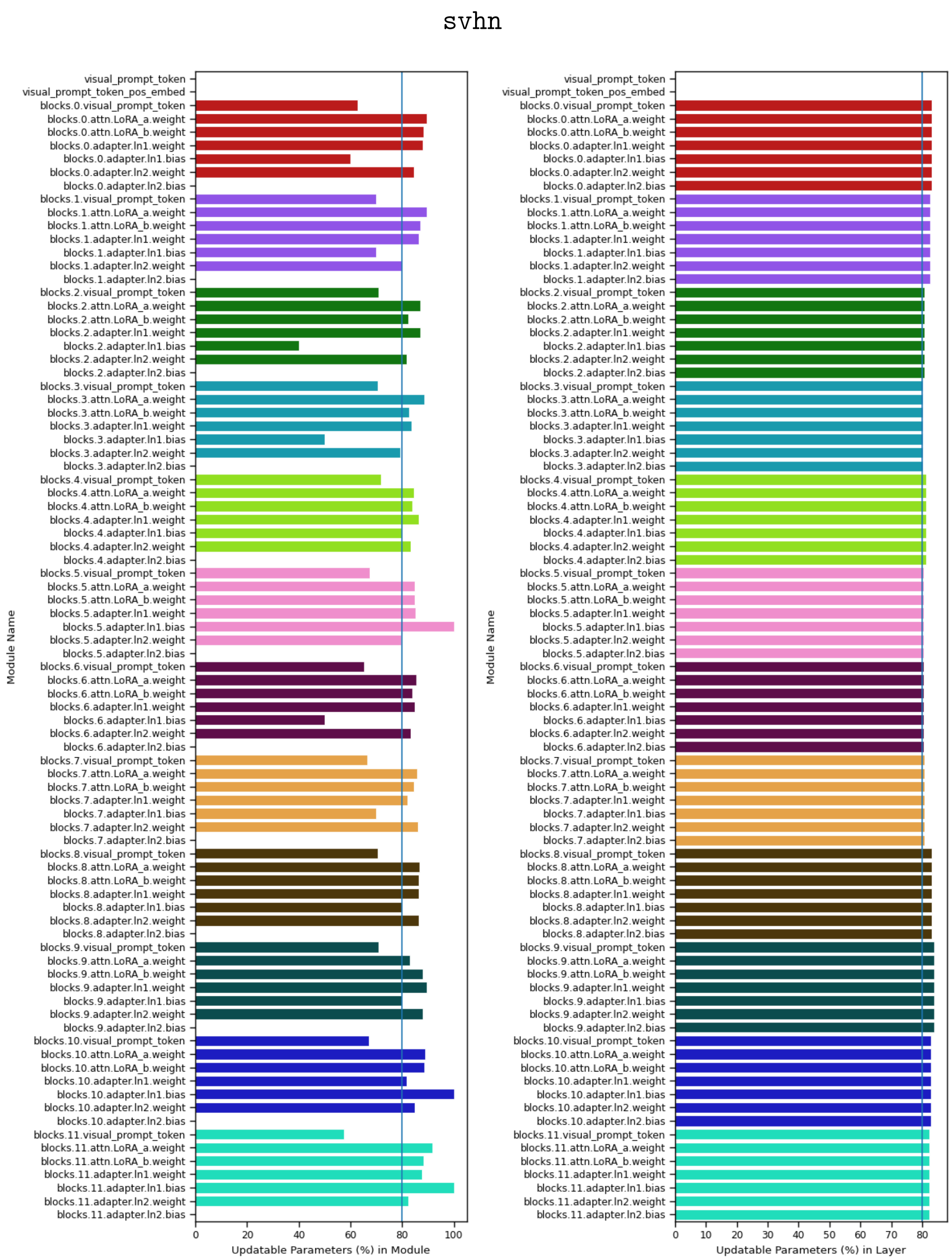


Figure 25: Sparsity pattern of attached modules to ViT-B/16 on `svhn`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 80\%$ is shown as vertical line.

sun397

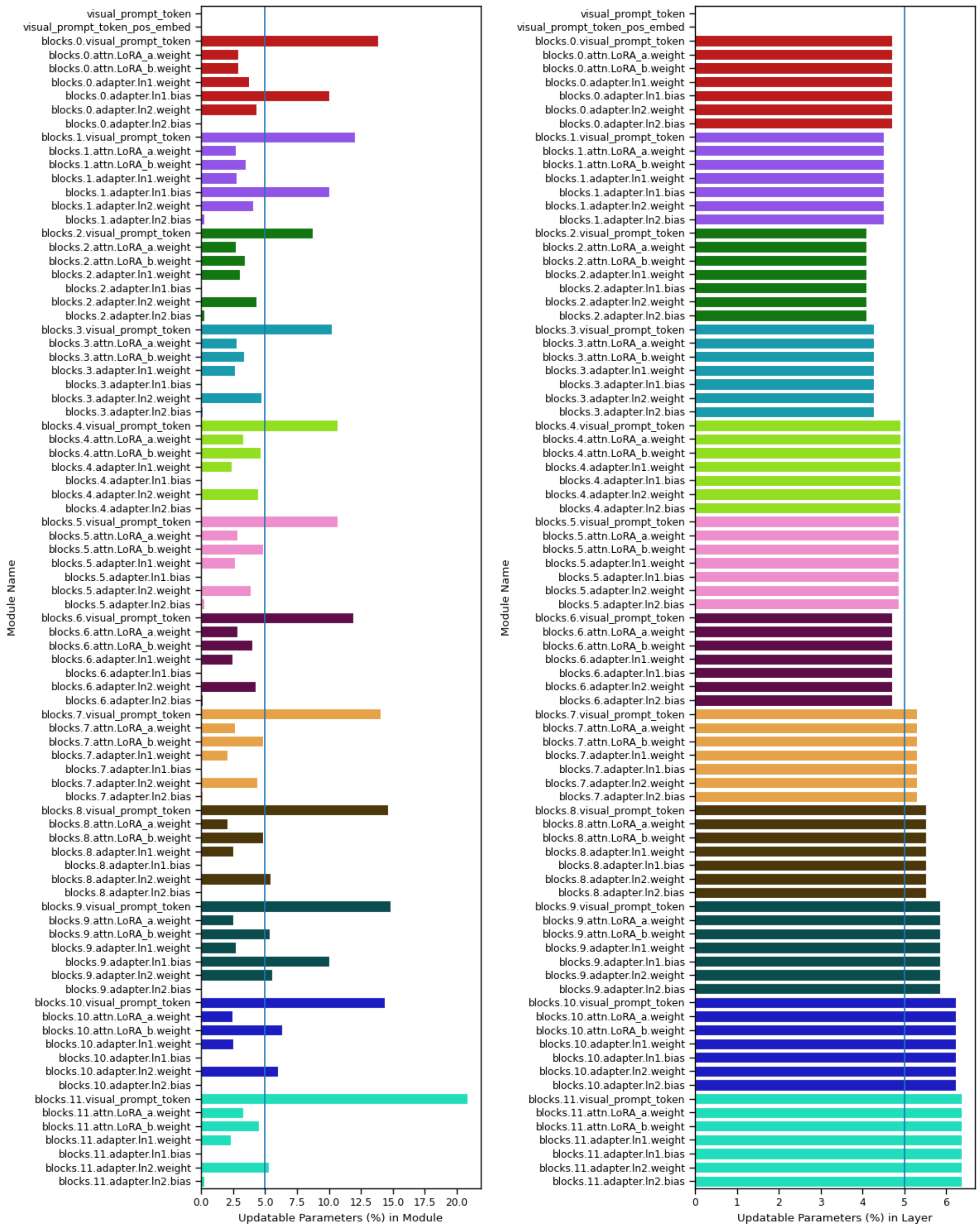


Figure 26: Sparsity pattern of attached modules to ViT-B/16 on sun397. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 5\%$ is shown as vertical line.

camelyon

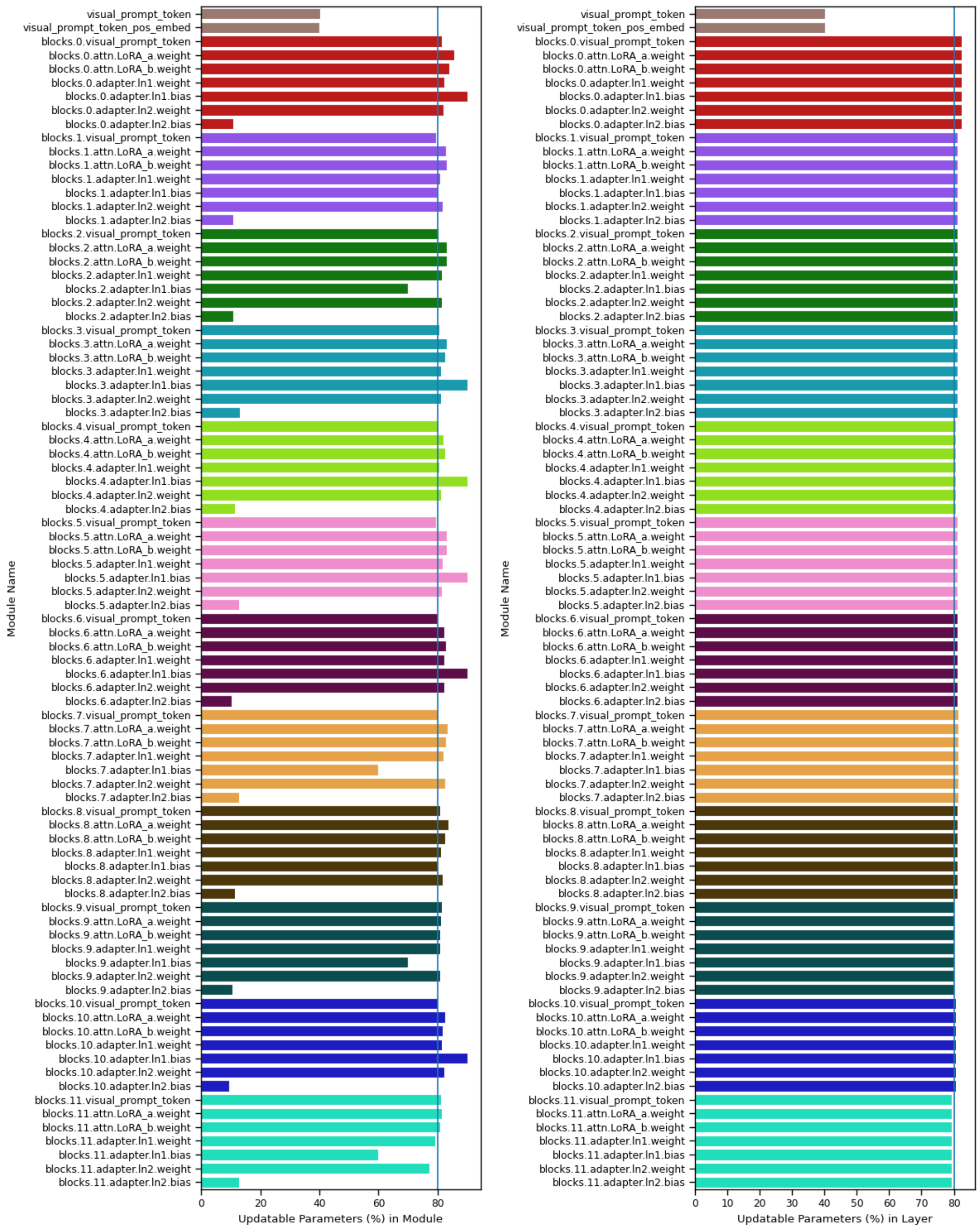


Figure 27: Sparsity pattern of attached modules to ViT-B/16 on `camelyon`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 80\%$ is shown as vertical line.

eurosat

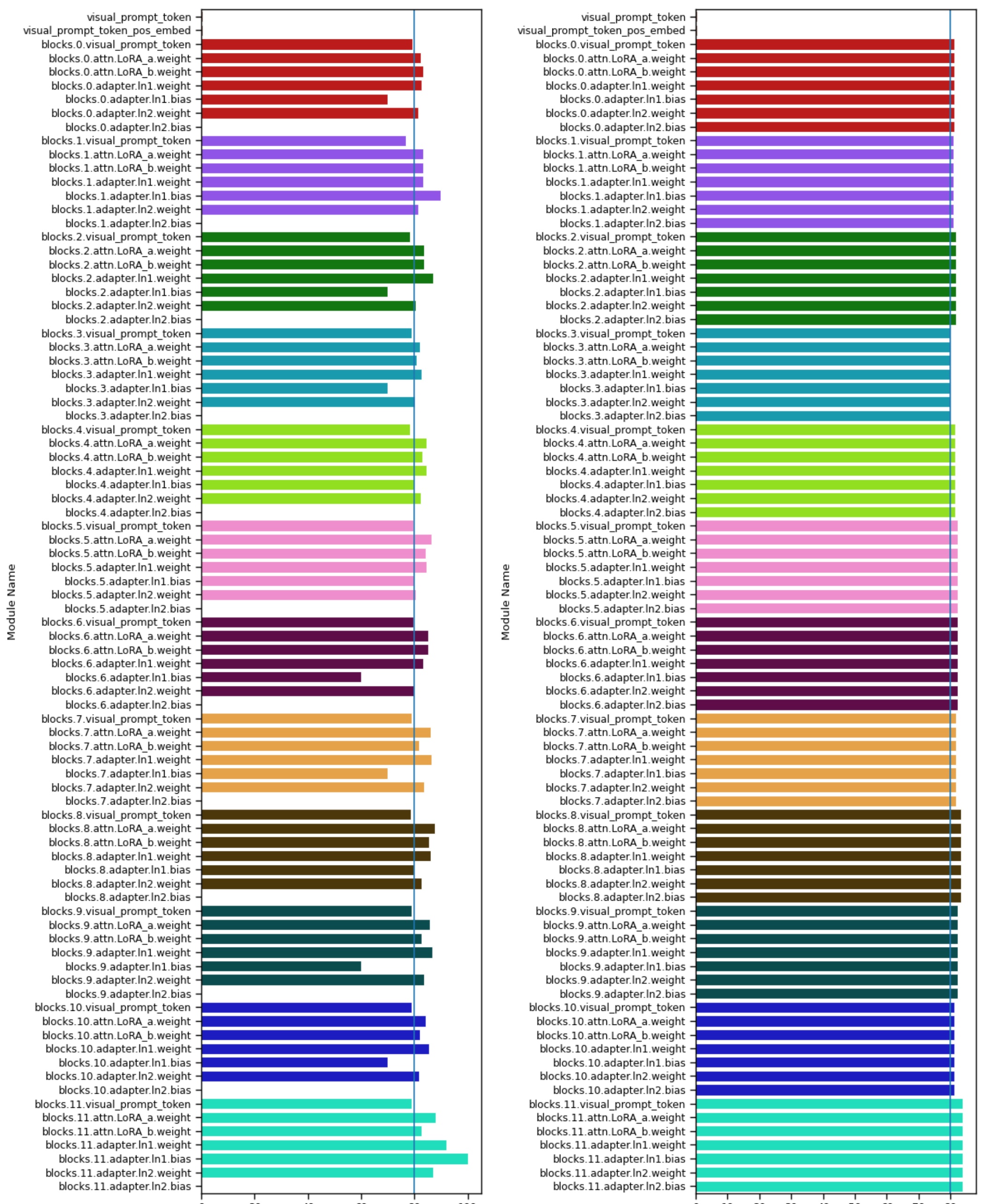


Figure 28: Sparsity pattern of attached modules to ViT-B/16 on eurosat. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 80\%$ is shown as vertical line.

resisc45

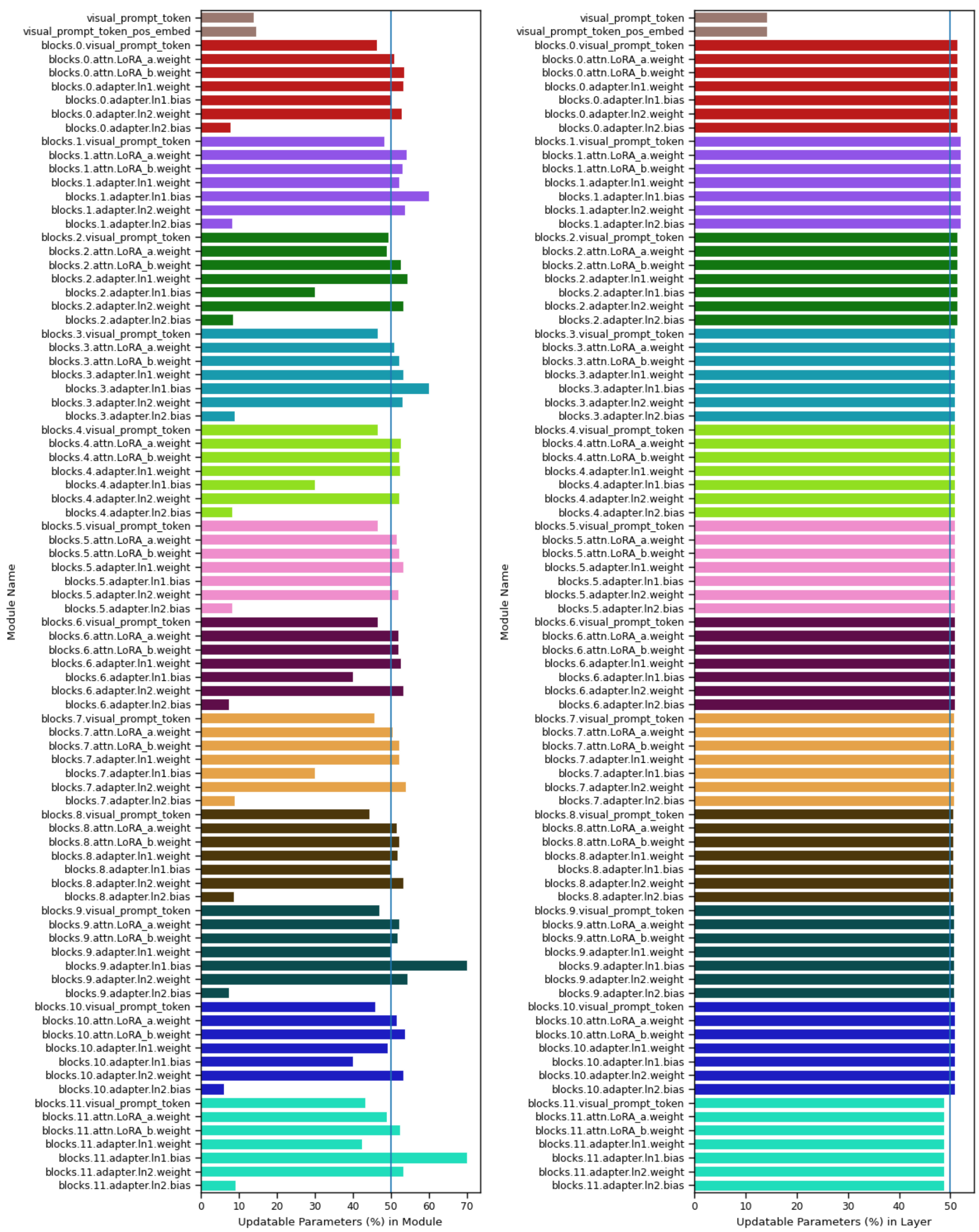


Figure 29: Sparsity pattern of attached modules to ViT-B/16 on `resisc45`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 50\%$ is shown as vertical line.

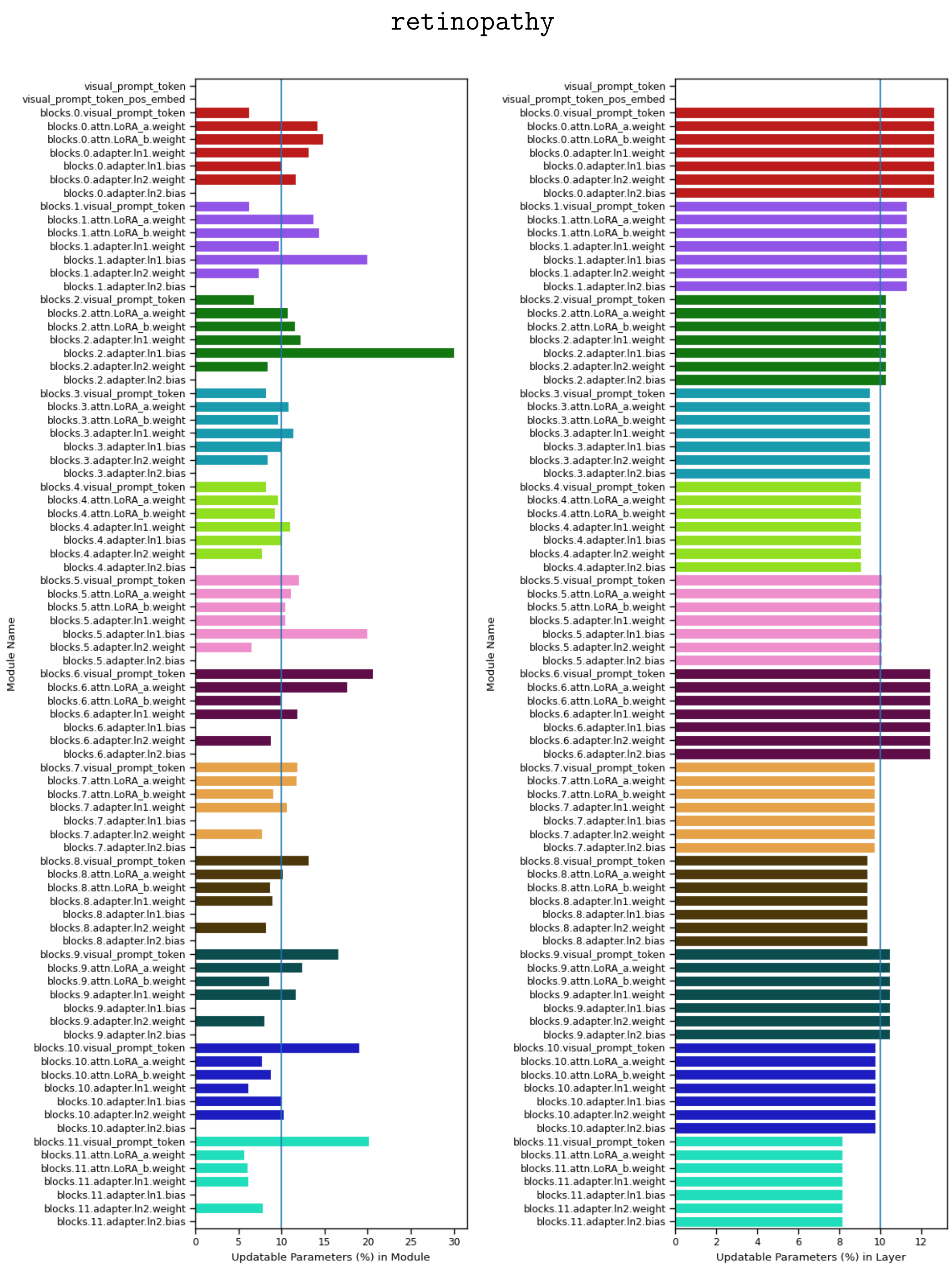


Figure 30: Sparsity pattern of attached modules to ViT-B/16 on `retinopathy`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 10\%$ is shown as vertical line.

`clevr-count`

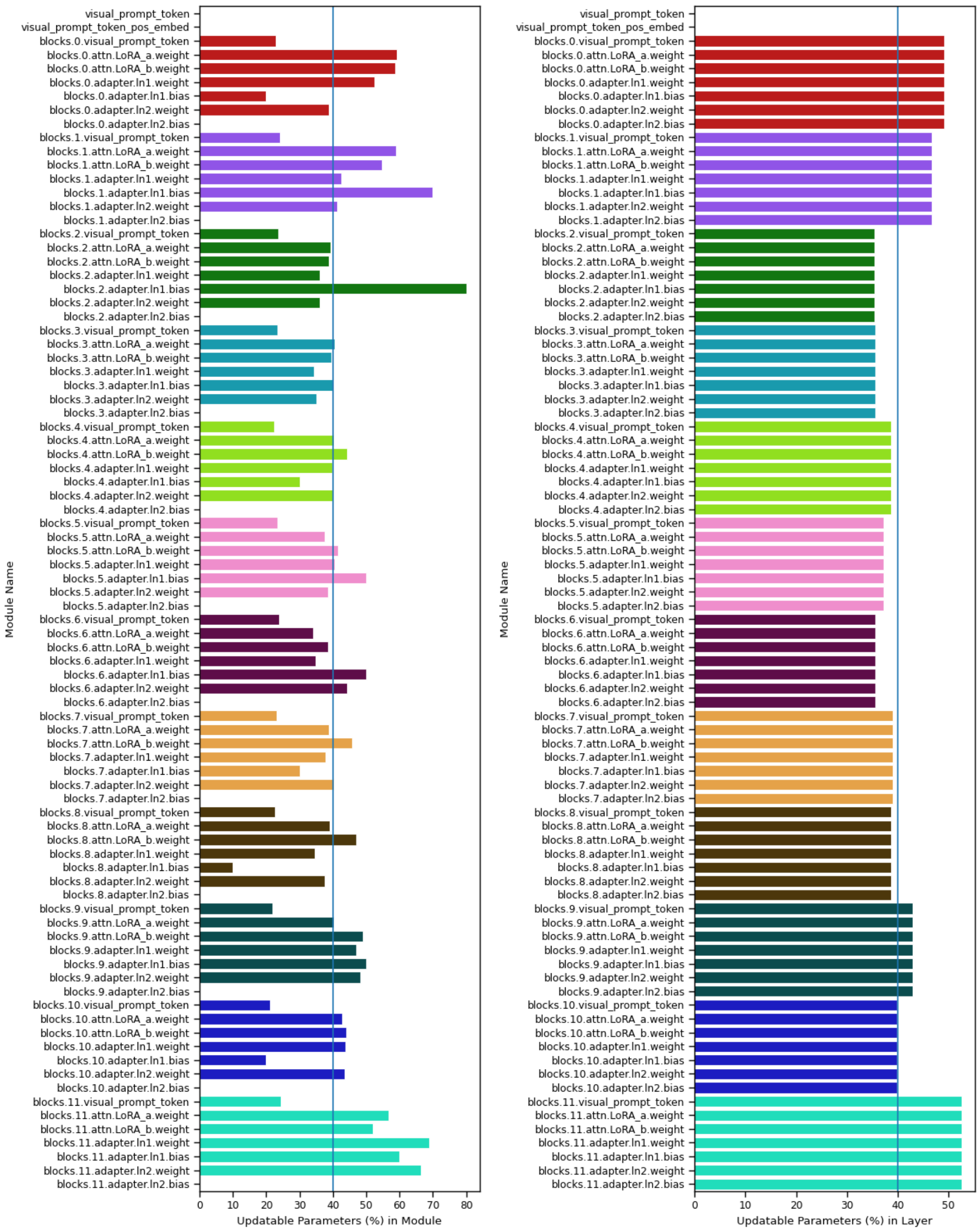


Figure 31: Sparsity pattern of attached modules to ViT-B/16 on `clevr-count`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 40\%$ is shown as vertical line.

clevr-dist

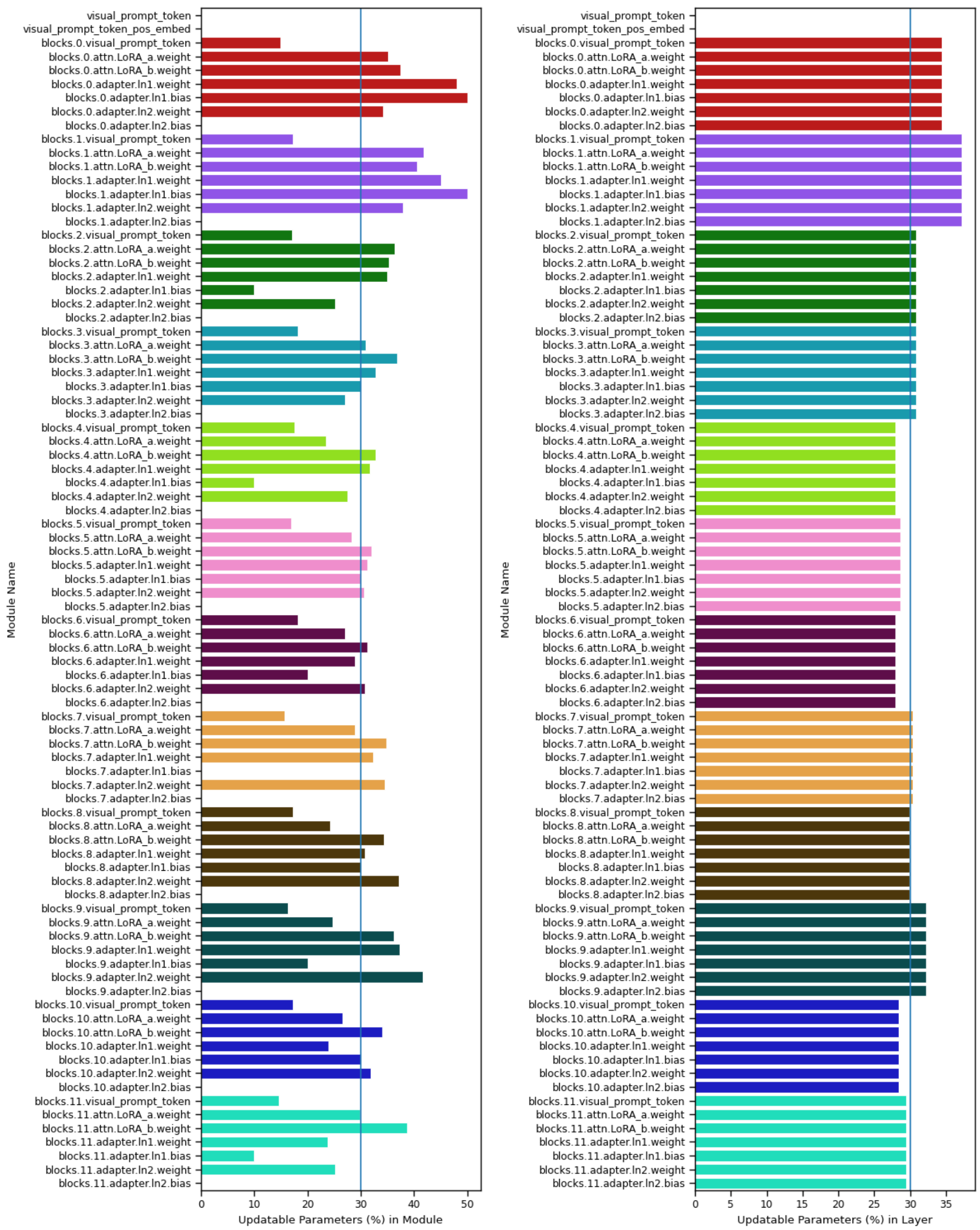


Figure 32: Sparsity pattern of attached modules to ViT-B/16 on `clevr-dist`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 30\%$ is shown as vertical line.

dmlab

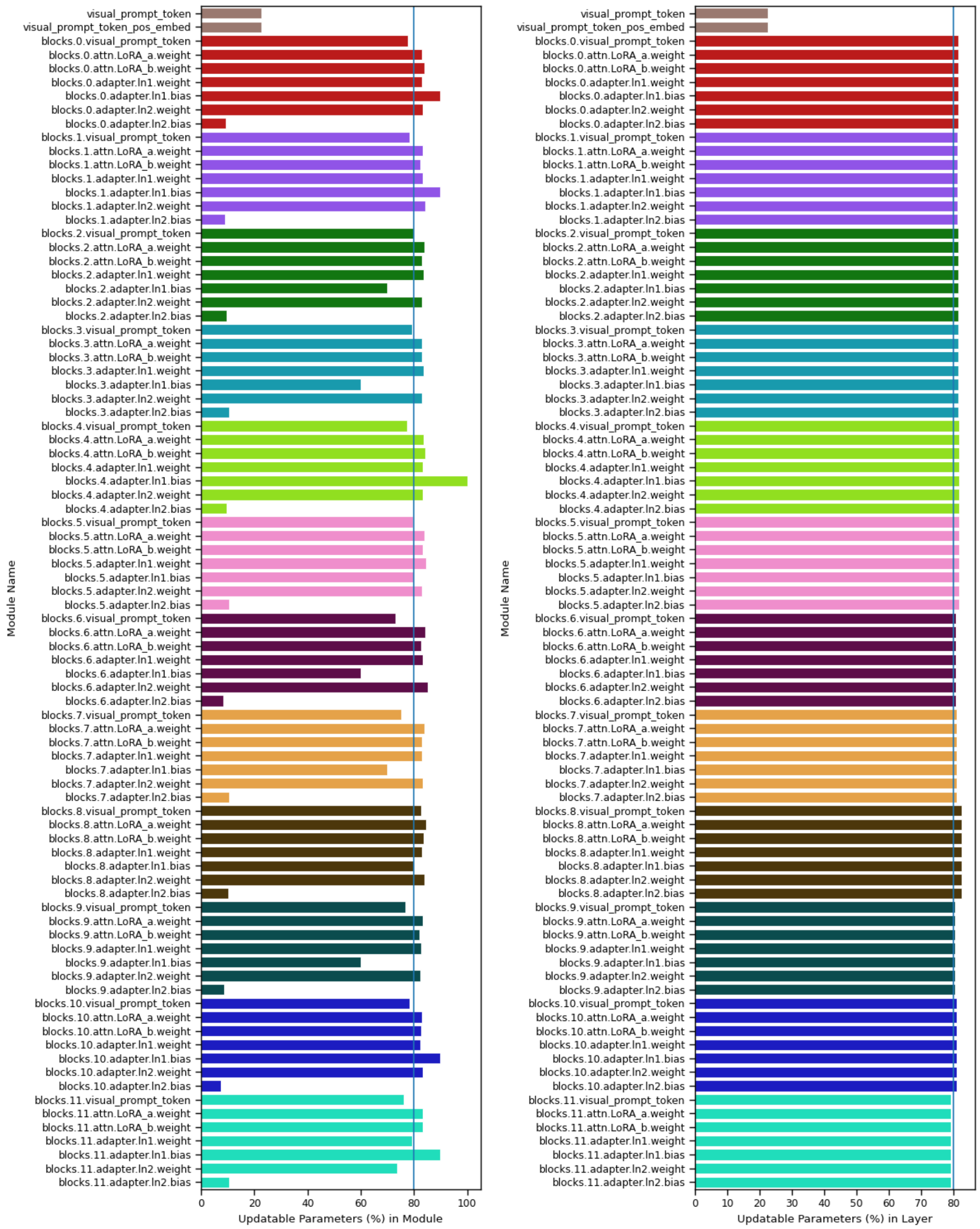


Figure 33: Sparsity pattern of attached modules to ViT-B/16 on `dmlab`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 80\%$ is shown as vertical line.

kitti

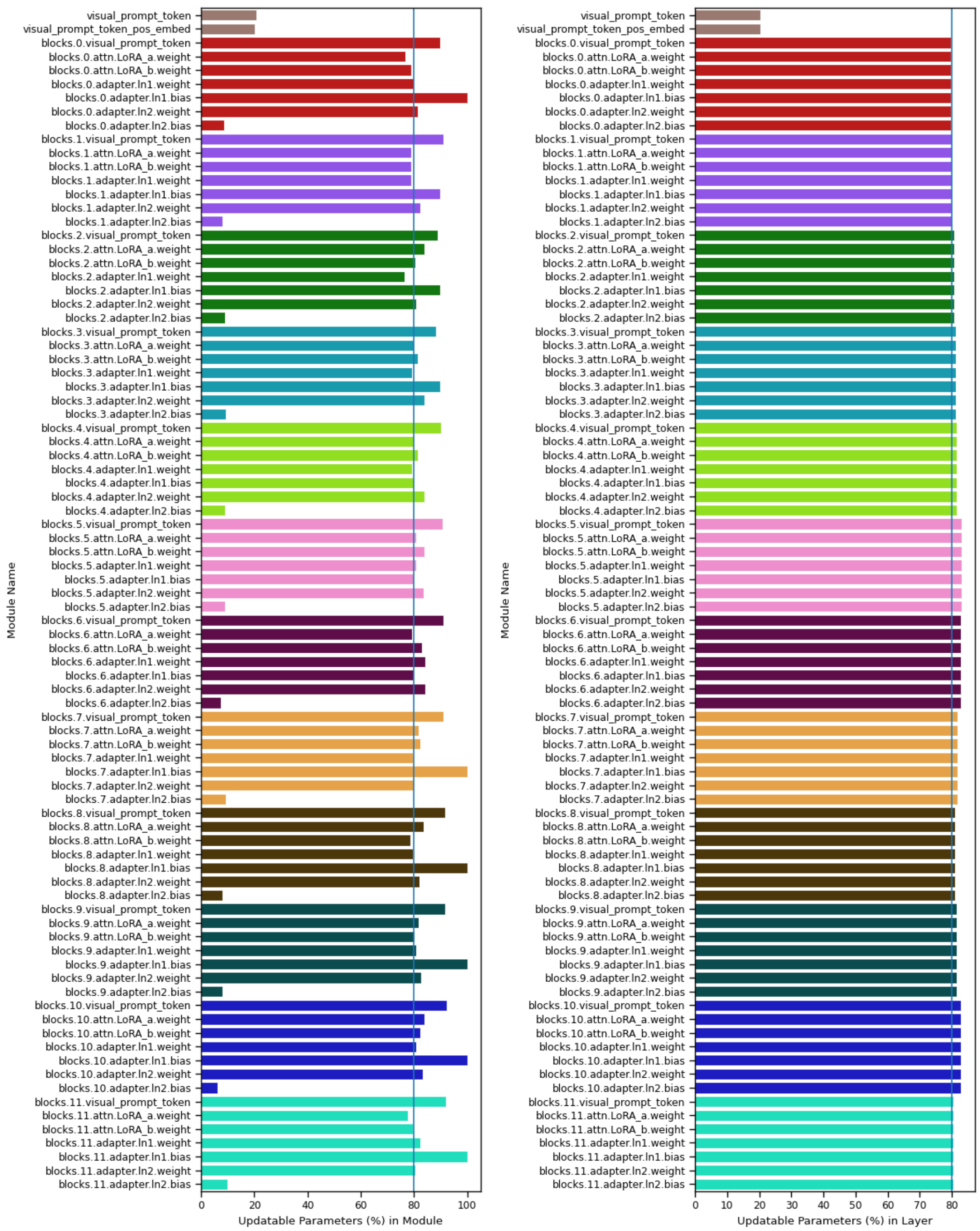


Figure 34: Sparsity pattern of attached modules to ViT-B/16 on `kitti`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 80\%$ is shown as vertical line.

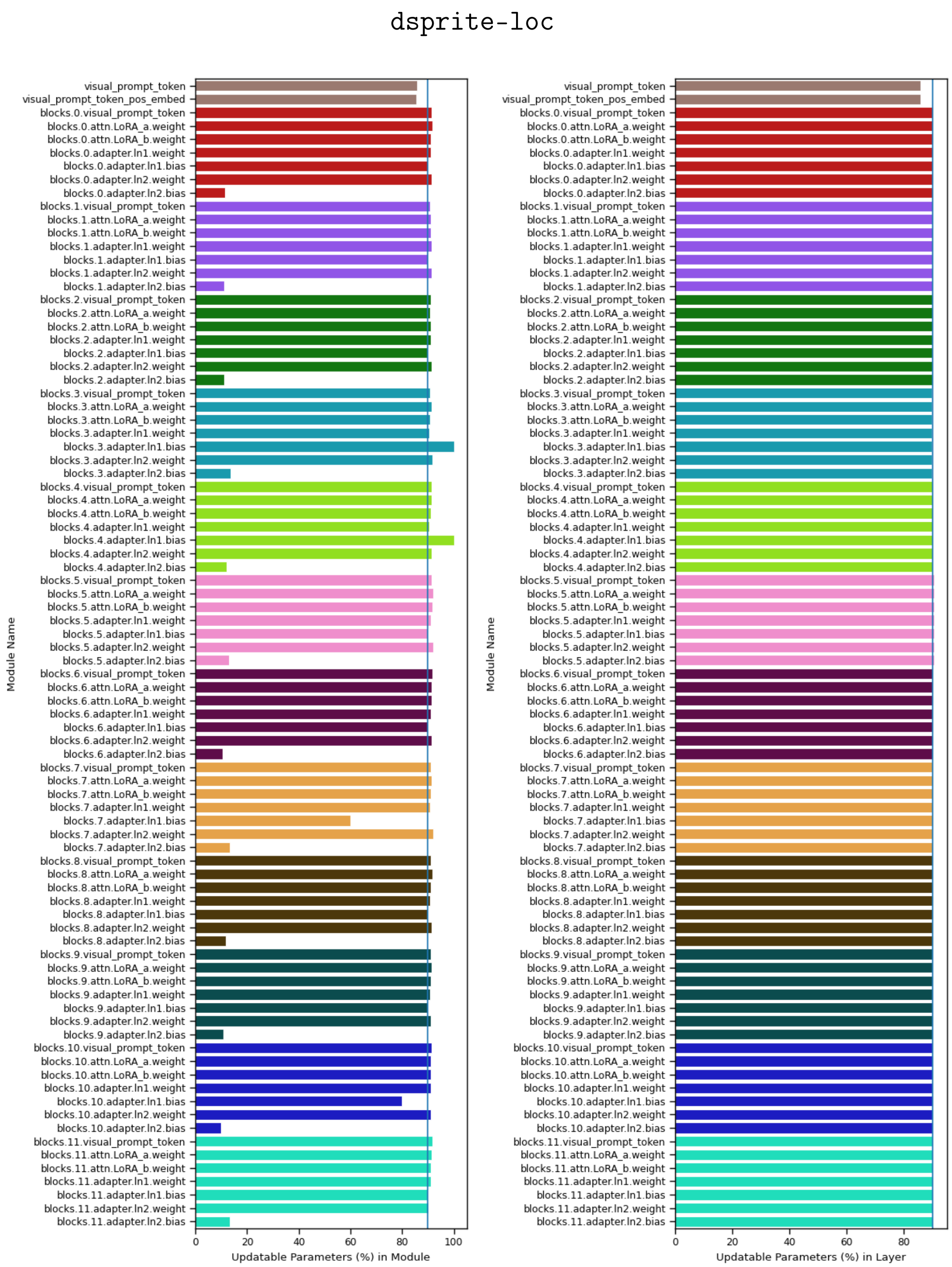


Figure 35: Sparsity pattern of attached modules to ViT-B/16 on `dsprite-loc`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 90\%$ is shown as vertical line.

dsprite-ori

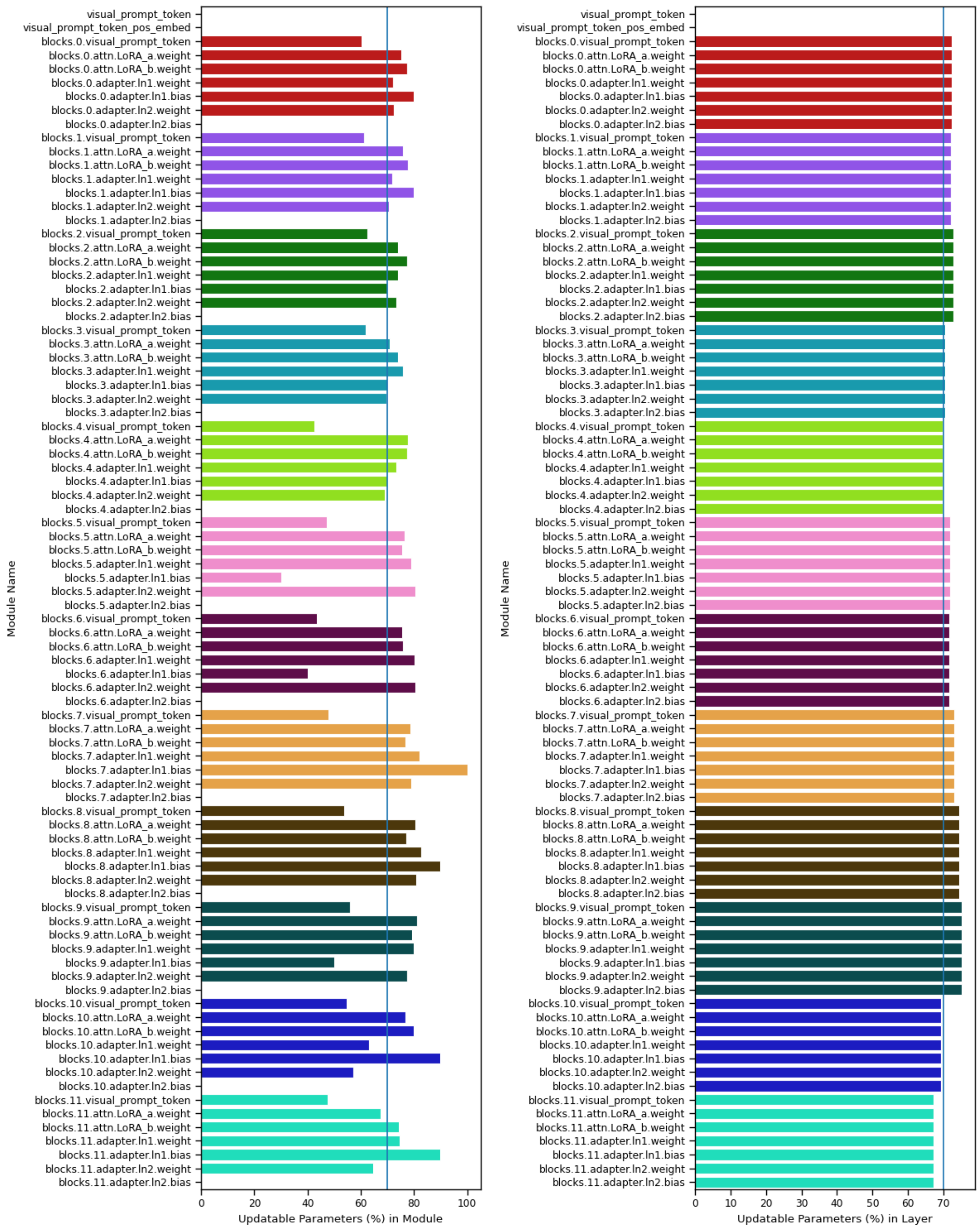


Figure 36: Sparsity pattern of attached modules to ViT-B/16 on `dsprite-ori`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 70\%$ is shown as vertical line.

snorb-azim

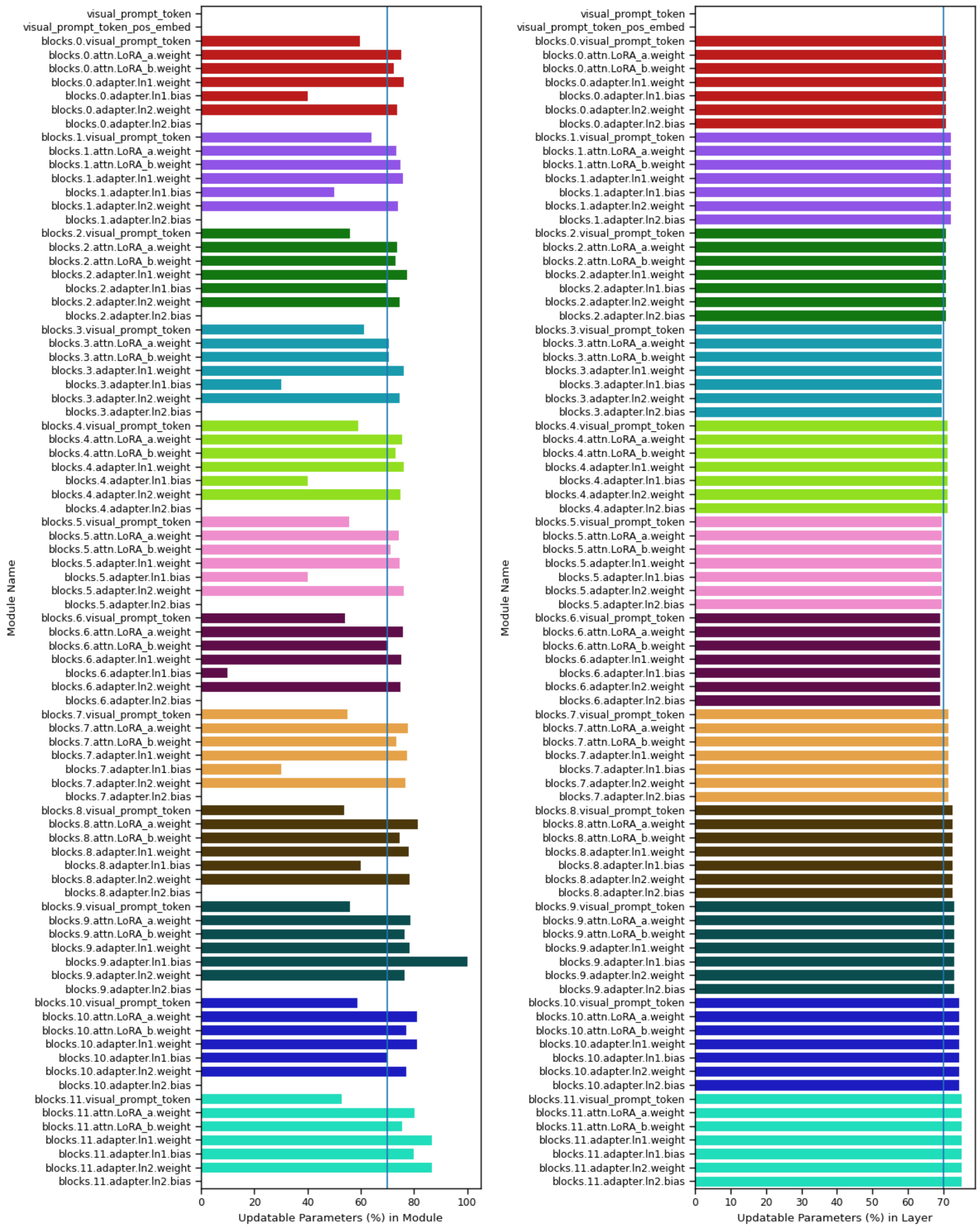


Figure 37: Sparsity pattern of attached modules to ViT-B/16 on `snorb-azim`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 70\%$ is shown as vertical line.

snorb-ele

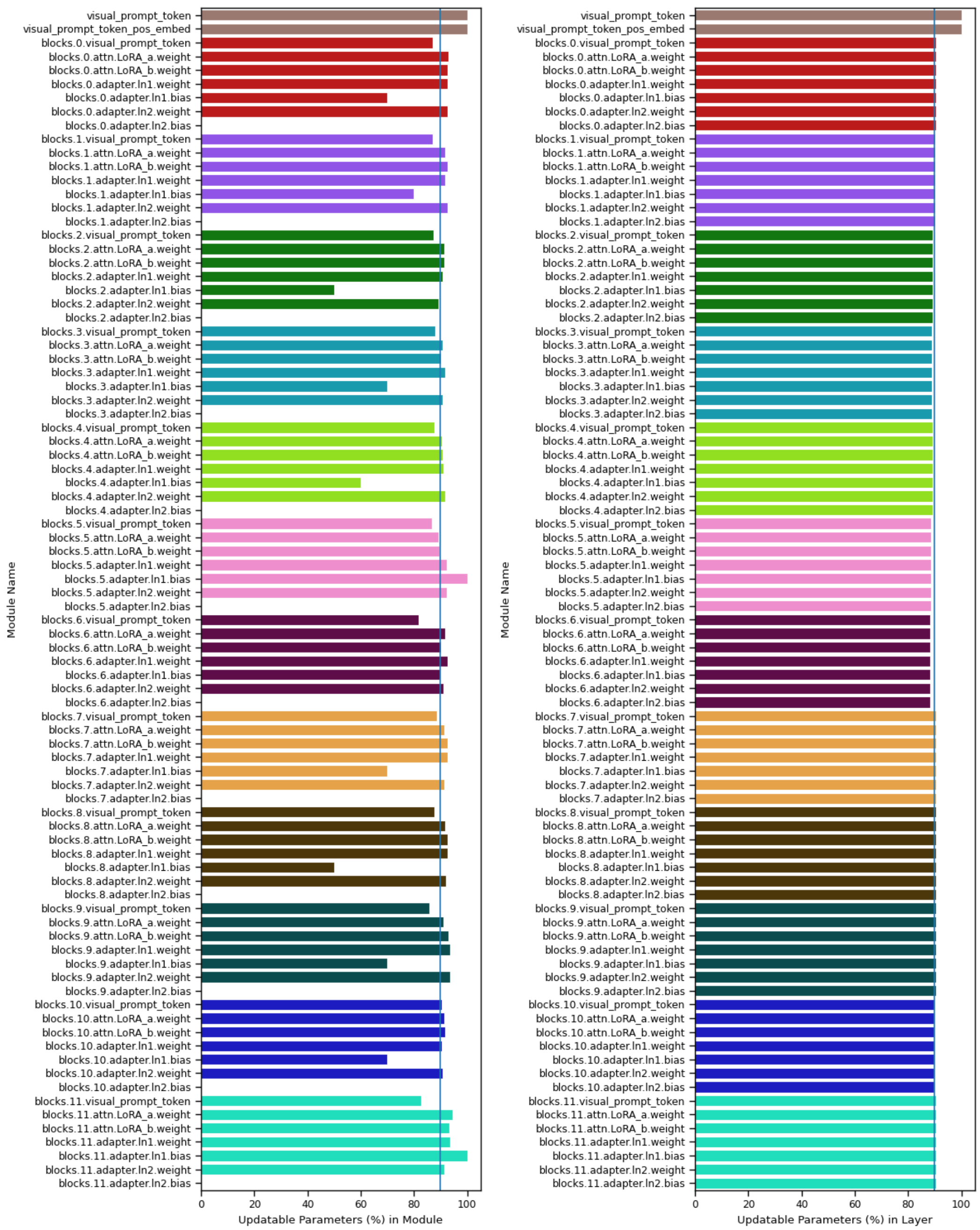


Figure 38: Sparsity pattern of attached modules to ViT-B/16 on `snorb-ele`. (Left) Module-wise updatable parameters (%) and (Right) Layer-wise updatable parameters (%). The BayesTune's optimal sparsity level $p^* = 90\%$ is shown as vertical line.