# A COMPUTE RESOURCES USED

All models were trained either on 48GB A6000s or 24GB A5000s. Each experiment on synthetic image reconstruction took about 5 hrs to train. Each experiment on T5 with Adapters for GLUE (T5-GLUE) took about 16 hours to train. Each experiment on ResNet with Adapters for DomainNet (Res-Dom) took about 11 hours to train.

# B EXPERIMENT DETAILS

## B.1 DAE-SYN

All of the experiments use learning rate of $1e^{-4}$ with batch size of 128 and were trained for $200k$ steps with the warmup ratio of $0.1$. For ST-Gumbel estimator, we use $\tau$ value of 10 and anneal rate of $1e^{-6}$ in the equation 3. For REINFORCE, we use values of 1, $1e^{-6}$, and $1e^{-2}$ for the corresponding $\alpha$, $\beta$, and $\gamma$ hyperparameters in the equation 2. The values for $\alpha$ and $\gamma$ are tuned over $\{1, 0.1\}$ and $\{1e^{-1}, 1e^{-2}, 1e^{-3}\}$ by fixing $\beta$ to 0 and then the value of $\beta$ is tuned over $\{0, 1e^{-6}, 1e^{-5}, 1e^{-4}\}$ by taking the best values of $\alpha$ and $\gamma$ found in the previous setup. The value of $\delta$ in the equation 1 is set to standard value of $0.5$ for all the REINFORCE experiments in all settings. The loss weight for the supervising the router is taken as 10 for all of the estimators after tuning over $\{10, 1, 0.1, 0.01\}$.

## B.2 T5-GLUE

All T5 models are trained for $2^{18}$ steps with learning rate of $1e^{-3}$, with $2k$ warmup steps, and batch size of 128. We use $\tau$ value of 10 and anneal rate of $1e^{-6}$ for the ST-Gumbel estimator. The values of $\alpha$, $\beta$, and $\gamma$ for the REINFORCE estimators are $1e^{-2}$, $5e^{-4}$, and $1e^{-2}$ following (Clark et al., 2022). The weight of the loss for supervising the router is taken as 1 for all of the estimators after tuning over $\{10, 1, 0.1, 0.01\}$.

The adapters used in this settings are simple bottleneck architectures with $swish$ non-linearity in between. The inputs values are added back to the output of the bottleneck block and then layer normalization is applied for calculating the final output of the adapter.

## B.3 RES-DOM

All ResNet models are trained for $100k$ steps with batch size of 128 and learning rate of $1e^{-3}$ and no warm up. We use $\tau$ value of 10 and anneal rate of $1e^{-4}$ for the ST-Gumbel estimator. The values of $\alpha$, $\beta$, and $\gamma$ for the REINFORCE estimators are $1e^{-2}$, $5e^{-4}$, and $1e^{-2}$ similar to T5-GLUE experiments. The supervised loss weight is taken as $0.1$ after tuning over $\{1, 0.1, 0.01\}$.

The adapters used are same bottleneck architectures with the same non-linearity as in T5-GLUE. The inputs are first batch normalized and then passed through the bottleneck architectures. The final output of the adapter is the sum of input and the output of the bottleneck block.

# C Focus AND Cover SCORES

In tables 3 and 4 we list the Focus and Cover scores for all estimators in all scenarios we consider, with and without varying levels of tag annotation supervision.

# D CONTINUE TRAINING FROM SUPERVISION CHECKPOINT

In table 5 we show the performance of continuing training the model after removing routing supervision loss.

# E FULL RESULTS ON T5-GLUE AND RES-DOM

We show full results of T5-GLUE in table 6 and Res-Dom in table 7.

| Routing | DAE-Syn | T5-GLUE | Res-Dom |
|---|---|---|---|
| Tag | 1.00 | 1.00 | 1.00 |
| Monolithic | 0.03 | 0.09 | 0.08 |
| Hash | 0.15 | 0.21 | 0.21 |
| Top-k | 0.49 | 0.35 | 0.36 |
| w/ 1% supervision | 0.52 (+0.03) | 0.98 (+0.63) | 0.75 (+0.39) |
| w/ 10% supervision | 1.00 (+0.51) | 0.99 (+0.64) | 0.77 (+0.41) |
| w/ 30% supervision | 1.00 (+0.51) | 0.99 (+0.64) | 0.77 (+0.41) |
| w/ 100% supervision | 1.00 (+0.51) | 0.99 (+0.64) | 0.78 (+0.42) |
| ST-Gumbel | 0.49 | 0.32 | 0.13 |
| w/ 1% supervision | 0.55 (+0.06) | 0.98 (+0.66) | 0.75 (+0.62) |
| w/ 10% supervision | 0.90 (+0.41) | 0.99 (+0.67) | 0.77 (+0.64) |
| w/ 30% supervision | 0.99 (+0.50) | 0.99 (+0.67) | 0.77 (+0.64) |
| w/ 100% supervision | 1.00 (+0.51) | 0.99 (+0.67) | 0.78 (+0.65) |
| REINFORCE | 0.57 | 0.35 | 0.29 |
| w/ 1% supervision | 0.59 (+0.02) | 0.98 (+0.63) | 0.74 (+0.45) |
| w/ 10% supervision | 1.00 (+0.43) | 0.99 (+0.64) | 0.77 (+0.48) |
| w/ 30% supervision | 1.00 (+0.43) | 0.99 (+0.64) | 0.78 (+0.49) |
| w/ 100% supervision | 1.00 (+0.43) | 0.99 (+0.64) | 0.77 (+0.48) |

Table 3: Focus scores for routing schemes learned by different estimators. X% supervision corresponds to including a tag annotation and training against it for X% of the training data.

| Routing | DAE-Syn | T5-GLUE | Res-Dom |
|---|---|---|---|
| Tag | 1.00 | 1.00 | 1.00 |
| Monolithic | 0.22 | 0.18 | 0.28 |
| Hash | 0.15 | 0.08 | 0.16 |
| Top-k | 0.45 | 0.70 | 0.64 |
| w/ 1% supervision | 0.47 (+0.02) | 1.00 (+0.30) | 0.80 (+0.16) |
| w/ 10% supervision | 1.00 (+0.55) | 1.00 (+0.30) | 0.82 (+0.18) |
| w/ 30% supervision | 1.00 (+0.55) | 1.00 (+0.30) | 0.82 (+0.18) |
| w/ 100% supervision | 1.00 (+0.55) | 1.00 (+0.30) | 0.83 (+0.19) |
| ST-Gumbel | 0.44 | 0.57 | 0.44 |
| w/ 1% supervision | 0.53 (+0.09) | 0.99 (+0.42) | 0.79 (+0.35) |
| w/ 10% supervision | 0.90 (+0.46) | 1.00 (+0.43) | 0.82 (+0.38) |
| w/ 30% supervision | 0.99 (+0.55) | 1.00 (+0.43) | 0.82 (+0.38) |
| w/ 100% supervision | 1.00 (+0.56) | 1.00 (+0.43) | 0.82 (+0.38) |
| REINFORCE | 0.54 | 0.75 | 0.55 |
| w/ 1% supervision | 0.56 (+0.02) | 1.00 (+0.25) | 0.78 (+0.23) |
| w/ 10% supervision | 1.00 (+0.46) | 0.99 (+0.24) | 0.82 (+0.27) |
| w/ 30% supervision | 1.00 (+0.46) | 1.00 (+0.25) | 0.82 (+0.27) |
| w/ 100% supervision | 1.00 (+0.46) | 1.00 (+0.25) | 0.82 (+0.27) |

Table 4: Cover scores for routing schemes learned by different estimators. X% supervision corresponds to including a tag annotation and training against it for X% of the training data.

| Routing | DAE-Syn | T5-GLUE | Res-Dom |
|---|---|---|---|
| Top-k | 0.1 | 81.9 | 61.8 |
| from supervision checkpoint | 0.1 (+0.0) | 81.0 (-0.9) | 62.2 (+0.4) |
| ST-Gumbel | 0.4 | 81.0 | 61.8 |
| from supervision checkpoint | 0.2 (-0.2) | 81.0 (+0.0) | 61.5 (-0.3) |
| REINFORCE | 0.1 | 81.6 | 61.8 |
| from supervision checkpoint | 0.1 (+0.0) | 81.7 (+0.1) | 62.1 (+0.3) |

Table 5: Performance of estimators from further training of tag supervised models by removing the supervision. Tag supervision considered for DAE-Syn, T5-GLUE, and Res-Dom are 30%, 1%, and 10% of total tags.

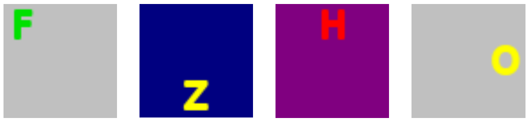

Figure 3: Synthetic images used for DAE-Syn. The (alphabet identity, alphabet location, foreground color, background color) tags for each image are (F, top-left, green, grey), (Z, bottom-mid, yellow, violet), (H, top-mid, red, purple) and (O, mid-right, yellow, grey) in the order of images.

| Routing | RTE<br>acc | SST-2<br>acc | MRPC<br>f1 | MRPC<br>acc | STS-B<br>pearson | STS-B<br>spearman | QQP<br>f1 | QQP<br>acc | MNLI<br>acc | QNLI<br>acc | CoLA<br>mcc | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tag | 58.9 | 90.9 | 90.7 | 87.2 | 88.1 | 87.7 | 85.9 | 89.5 | 81.7 | 90.2 | 43.5 | 81.3 |
| Hash | 51.7 | 88.5 | 84.4 | 78.3 | 84.6 | 84.4 | 82.8 | 87.2 | 79.1 | 88.6 | 6.5 | 74.0 |
| Monolithic | 60.1 | 91.3 | 88.3 | 84.1 | 88.0 | 87.8 | 84.7 | 88.6 | 81.1 | 90.0 | 13.0 | 77.9 |
| Tag+ | 65.5 | 90.7 | 90.7 | 87.2 | 88.2 | 87.8 | 85.9 | 89.5 | 81.8 | 90.2 | 44.2 | 82.0 |
| Top-$k$ | 58.7 | 90.5 | 91.1 | 88.0 | 88.0 | 87.7 | 85.3 | 89.0 | 81.3 | 89.9 | 21.2 | 79.2 |
| + subnet init | 61.4 | 90.5 | 90.8 | 87.5 | 88.4 | 88.1 | 85.2 | 88.9 | 81.6 | 89.9 | 21.1 | 79.4 |
| w/ 1% supervision | 65.7 | 90.4 | 91.2 | 87.8 | 88.2 | 87.9 | 85.9 | 89.4 | 81.6 | 90.1 | 42.8 | 81.9 |
| w/ 10% supervision | 60.9 | 90.2 | 90.1 | 86.4 | 88.6 | 88.2 | 85.9 | 89.4 | 81.8 | 90.2 | 42.3 | 81.2 |
| w/ 30% supervision | 59.7 | 90.6 | 90.7 | 87.4 | 88.2 | 87.9 | 86.0 | 89.5 | 81.7 | 90.1 | 41.2 | 81.2 |
| w/ 100% supervision | 64.0 | 90.9 | 91.0 | 87.7 | 88.0 | 87.7 | 86.0 | 89.5 | 81.6 | 90.2 | 39.2 | 81.4 |
| from supervision checkpoint | 60.1 | 90.0 | 90.1 | 86.7 | 88.4 | 88.1 | 86.2 | 89.6 | 82.0 | 89.8 | 39.7 | 81.0 |
| ST-Gumbel | 61.6 | 91.2 | 90.0 | 86.4 | 87.9 | 87.6 | 85.2 | 88.8 | 81.1 | 90.1 | 19.9 | 79.1 |
| + subnet init | 59.7 | 90.5 | 89.9 | 86.5 | 88.3 | 88.1 | 85.2 | 88.9 | 81.0 | 90.0 | 12.5 | 78.2 |
| w/ 1% supervision | 60.6 | 90.7 | 90.4 | 86.9 | 88.5 | 88.0 | 86.0 | 89.5 | 81.6 | 90.3 | 38.6 | 81.0 |
| w/ 10% supervision | 58.7 | 90.9 | 90.0 | 86.4 | 88.1 | 87.8 | 85.9 | 89.5 | 81.7 | 90.2 | 41.2 | 80.9 |
| w/ 30% supervision | 59.9 | 90.6 | 90.8 | 87.4 | 88.0 | 87.6 | 86.0 | 89.5 | 81.6 | 90.3 | 44.2 | 81.4 |
| w/ 100% supervision | 61.4 | 90.7 | 90.4 | 86.9 | 88.2 | 88.0 | 85.9 | 89.5 | 81.8 | 90.1 | 40.6 | 81.2 |
| from supervision checkpoint | 62.3 | 90.4 | 91.5 | 88.3 | 87.9 | 87.7 | 86.1 | 89.6 | 81.8 | 90.1 | 35.7 | 81.0 |
| REINFORCE | 64.7 | 90.5 | 90.6 | 87.5 | 88.3 | 88.1 | 85.6 | 89.2 | 81.6 | 90.3 | 28.8 | 80.5 |
| + subnet init | 61.1 | 90.4 | 90.5 | 87.2 | 88.3 | 88.2 | 85.8 | 89.4 | 81.8 | 90.5 | 27.5 | 80.0 |
| w/ 1% supervision | 64.0 | 90.9 | 90.7 | 87.2 | 88.4 | 88.0 | 85.8 | 89.4 | 81.5 | 90.1 | 42.2 | 81.6 |
| w/ 10% supervision | 62.6 | 91.3 | 91.9 | 89.0 | 88.3 | 88.0 | 85.9 | 89.4 | 81.8 | 90.3 | 42.2 | 81.9 |
| w/ 30% supervision | 62.8 | 91.1 | 91.6 | 88.7 | 88.0 | 87.6 | 85.9 | 89.4 | 81.8 | 90.3 | 42.9 | 81.8 |
| w/ 100% supervision | 59.7 | 91.3 | 91.6 | 88.7 | 88.3 | 87.9 | 85.9 | 89.4 | 81.8 | 90.3 | 37.3 | 81.1 |
| from supervision checkpoint | 65.9 | 90.6 | 91.2 | 88.0 | 87.6 | 87.1 | 86.1 | 89.6 | 82.1 | 89.9 | 40.0 | 81.7 |

Table 6: Full T5-GLUE results.

| Routing | Clipart | Infograph | Painting | Quickdraw | Real | Sketch | Average |
|---|---|---|---|---|---|---|---|
| Tag | 56.3 | 24.6 | 51.9 | 51.0 | 70.1 | 48.4 | 62.5 |
| Hash | 64.2 | 31.5 | 59.4 | 62.7 | 74.7 | 57.0 | 54.5 |
| Monolithic | 62.7 | 29.3 | 56.5 | 60.6 | 73.3 | 54.4 | 60.5 |
| Top-$k$ | 63.7 | 30.8 | 57.5 | 61.6 | 74.0 | 55.6 | 61.5 |
| + subnet init | 63.5 | 30.6 | 56.9 | 63.3 | 74.3 | 56.2 | 62.0 |
| w/ 1% supervision | 63.6 | 30.6 | 57.2 | 62.9 | 74.2 | 55.7 | 61.6 |
| w/ 10% supervision | 63.3 | 30.8 | 56.9 | 62.8 | 74.3 | 55.6 | 61.8 |
| w/ 30% supervision | 63.5 | 30.5 | 57.4 | 62.9 | 74.2 | 55.3 | 61.9 |
| w/ 100% supervision | 63.6 | 30.6 | 57.2 | 62.9 | 74.2 | 55.7 | 61.9 |
| from supervision checkpoint | 63.5 | 30.6 | 57.0 | 63.8 | 74.3 | 56.1 | 62.2 |
| ST-Gumbel | 62.4 | 29.2 | 56.4 | 60.2 | 73.2 | 54.3 | 60.3 |
| + subnet init | 61.7 | 29.3 | 56.8 | 61.3 | 73.9 | 53.7 | 60.8 |
| w/ 1% supervision | 63.9 | 31.3 | 57.5 | 62.6 | 74.4 | 55.8 | 61.4 |
| w/ 10% supervision | 63.3 | 31.3 | 57.5 | 62.4 | 74.3 | 55.6 | 61.8 |
| w/ 30% supervision | 63.5 | 31.4 | 57.6 | 62.7 | 74.2 | 55.7 | 61.9 |
| w/ 100% supervision | 63.9 | 31.3 | 57.5 | 62.6 | 74.4 | 55.8 | 62.0 |
| from supervision checkpoint | 63.2 | 30.6 | 57.3 | 61.8 | 74.3 | 55.4 | 61.5 |
| REINFORCE | 62.8 | 30.5 | 57.4 | 62.3 | 73.8 | 55.1 | 61.5 |
| + subnet init | 63.3 | 30.7 | 57.6 | 63.2 | 74.0 | 55.6 | 61.9 |
| w/ 1% supervision | 63.6 | 31.4 | 57.7 | 62.6 | 74.4 | 55.9 | 61.5 |
| w/ 10% supervision | 63.3 | 31.1 | 57.6 | 62.5 | 74.2 | 55.7 | 61.8 |
| w/ 30% supervision | 63.6 | 31.2 | 57.7 | 62.7 | 74.5 | 55.9 | 62.0 |
| w/ 100% supervision | 63.6 | 31.4 | 57.7 | 62.6 | 74.4 | 55.9 | 62.0 |
| from supervision checkpoint | 63.8 | 31.1 | 57.1 | 63.5 | 74.2 | 56.0 | 62.1 |

Table 7: Full Res-Dom results.