A Appendix

A.1 Dataset Specification

Domain/Attribute	# of Cause Example	# of IsolateExample	# of Entity
City	34899/7016	49500/9930	3552/3374
Country	7925/1544	8250/1655	3528/2411
Language	6207/1252	8250/1655	3471/2221
Continent	8254/1658	8250/1655	3543/2567
Timezone	5371/1144	8250/1655	3414/1900
Latitude	3813/743	8250/1655	3107/1519
Longitude	3329/675	8250/1655	2989/1357
Nobel Laureate	39771/6754	44628/7600	928/928
Country of Birth	7218/1356	8908/1520	928/909
Award Year	11037/1904	8930/1520	928/926
Gender	854/96	8930/1520	592/149
Field	9518/1558	8930/1520	928/922
Birth Year	11144/1840	8930/1520	928/927
Occupation	54444/1582	29052/864	799/785
Work Location	24216/724	9684/288	799/708
Duty	12090/371	9684/288	785/522
Industry	18138/487	9684/288	799/600
Physical Object	49114/4659	35285/3636	563/563
Color	14707/1518	8825/909	563/563
Category	13540/1273	8820/909	563/562
Texture	14666/1265	8821/909	563/561
Size	6201/603	8819/909	563/528
Verb	70003/3806	14396/782	986/984
Past Tense	34043/1848	7188/391	986/975
Singular	35960/1958	7208/391	986/978

Table 1: The details of the dataset used for the experiment, in the format of train/test splits. For every model in each setting. Methods are trained on the full dataset of that setting with 5 epochs. The prompts used by the train/test splits are completely disjoint.

A.2 Information Masking

Figure 10 shows how the HyperDAS may learn a trivial solution to the RAVEL benchmark if the relevant information (base prompt attribute) is not properly masked.

A.3 Subspace Dimension

We experiment with different feature subspace dimension, as shown in Figure 11.

A.4 Intervention Patterns

Here we include a few demonstrations of the intervention pattern that HyperDAS generates on RAVEL, as shown in Figure 12.

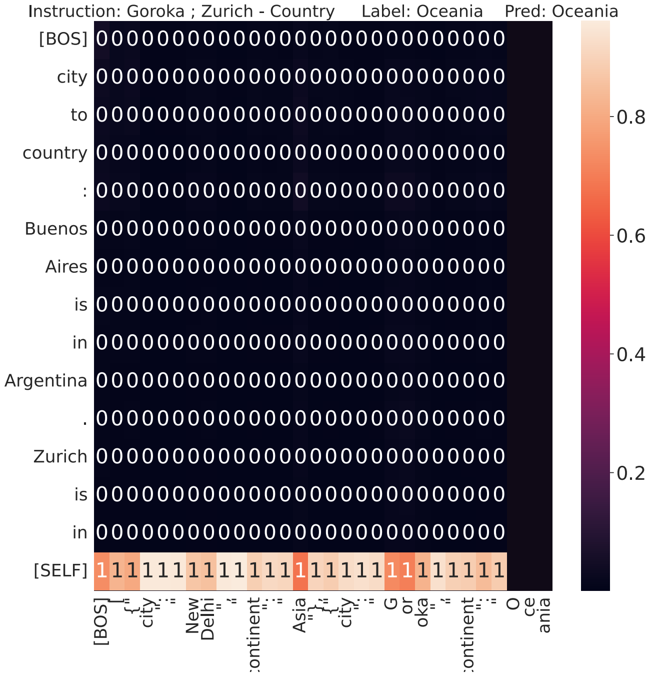


Figure 10: The trivial solution learnt by the HyperDAS on isolate examples when no mask is applied on the attribute token in the prompt. HyperDAS learns to do no intervention at all if it sees the base prompt attribute to be different than the attribute in the instruction.

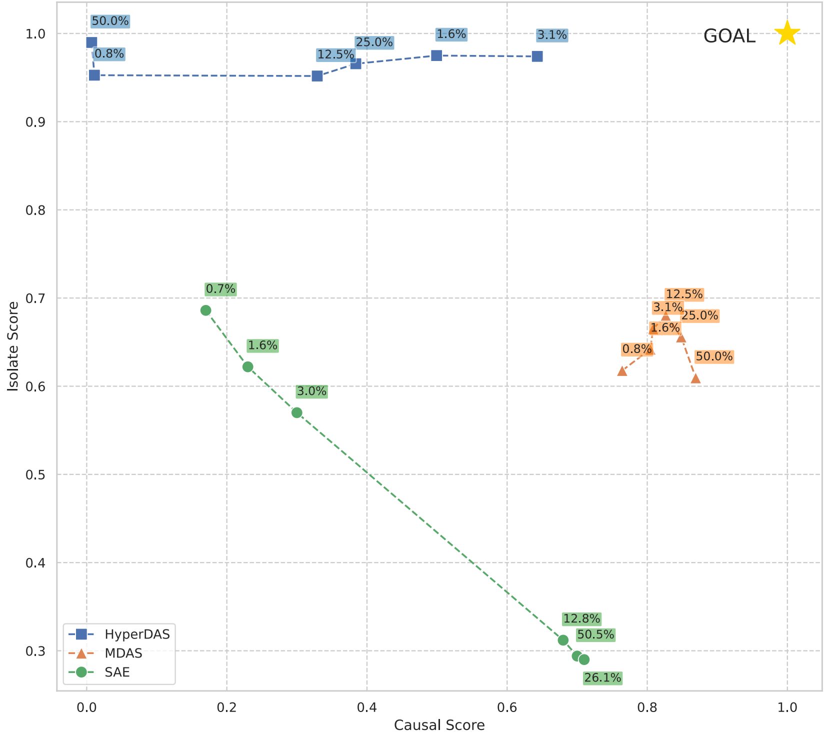
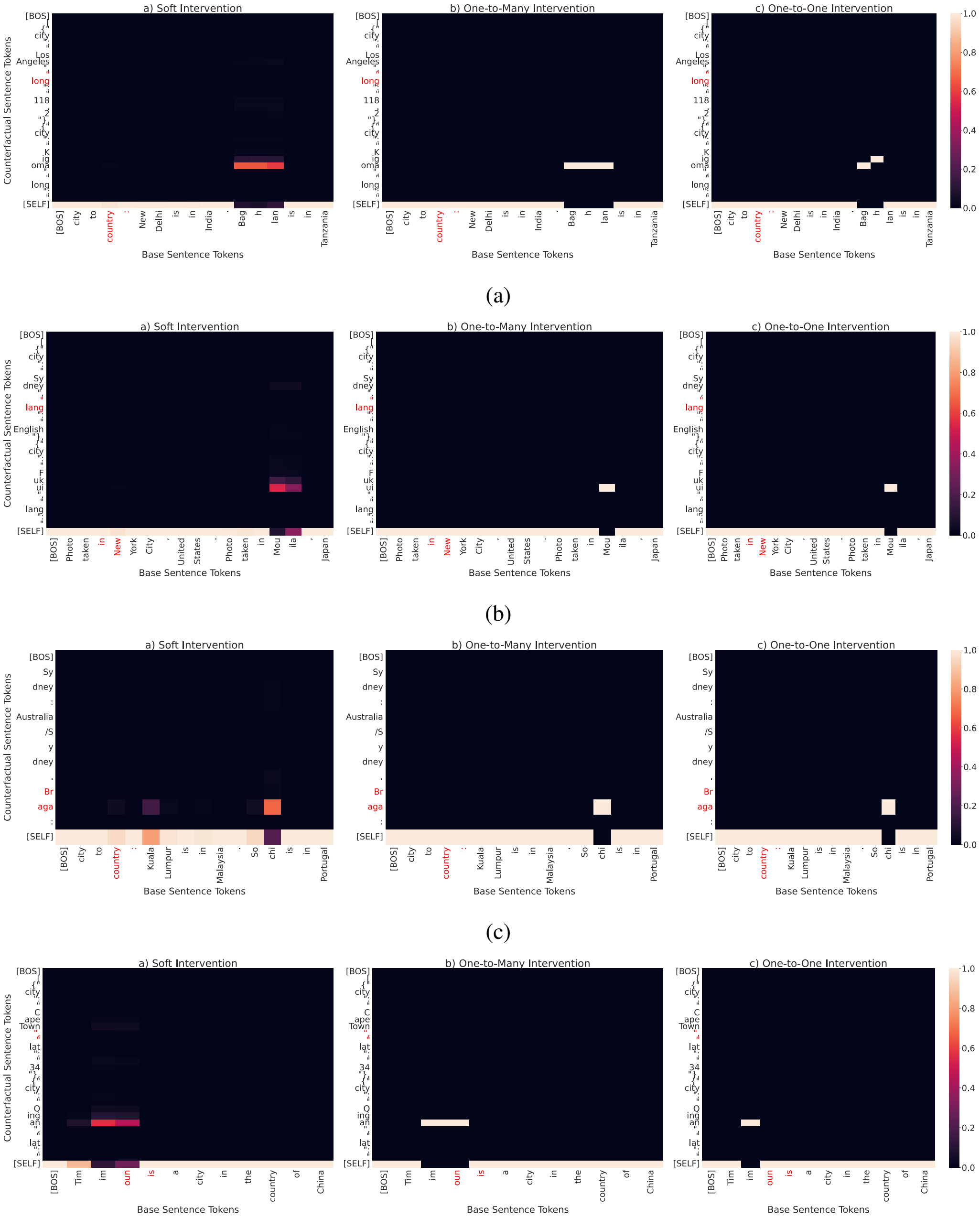


Figure 11: Cause and Iso scores for HyperDAS, MDAS, and SEA when using different feature size shown as the ratio $\%$.



(a)

(b)

(c)

(d)

Figure 12: Four types of intervention patterns.