	NMI (↑)	ARI (↑)	ACC (↑)	LP (↑)	DP (↑)	LHD (↓)
Lowe et al (2024)	0.753	0.499	0.616	0.664	0.445	0.303
L2H-TEMI	0.778	0.565	0.682	0.701	0.502	0.298
L2H-TURTLE	0.917	0.831	0.896	0.897	0.803	0.235

Table: Performance comparison of the approach proposed by Lowe et al (2024) on the CIFAR-100 dataset with our proposed L2H approach with TURTLE and TEMI as backbone models. Hierarchical and flat clustering metrics are reported. Note that given Lowe et al. 2024 is based on agglomerative clustering *it does not allow for inference on a hold-out test set*. Hence while we train our approach on the training set, to then report results on the test set, for Lowe et al. we necessarily both train and evaluate on the test set. Note this inevitably results in an unfair comparison as the method for Lowe et al. sees the test data for training. In spite of this, our approach (L2H-TURTLE, L2H-TEMI) still shows superior performance across all metrics, which further validates its effectiveness.