1 Additional Experiments

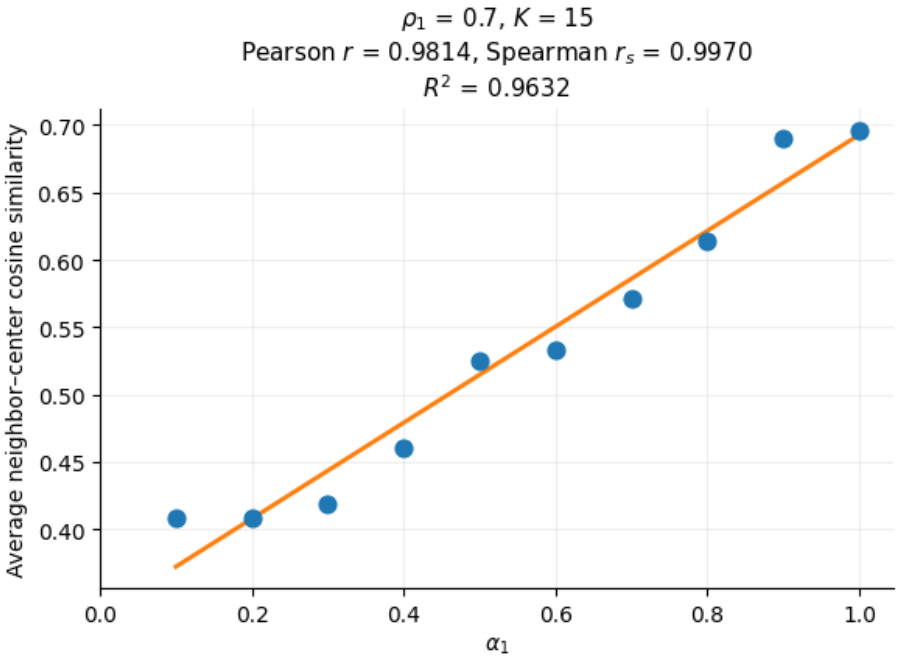


Figure 1: Correlation between the raw cosine similarity vs α_1

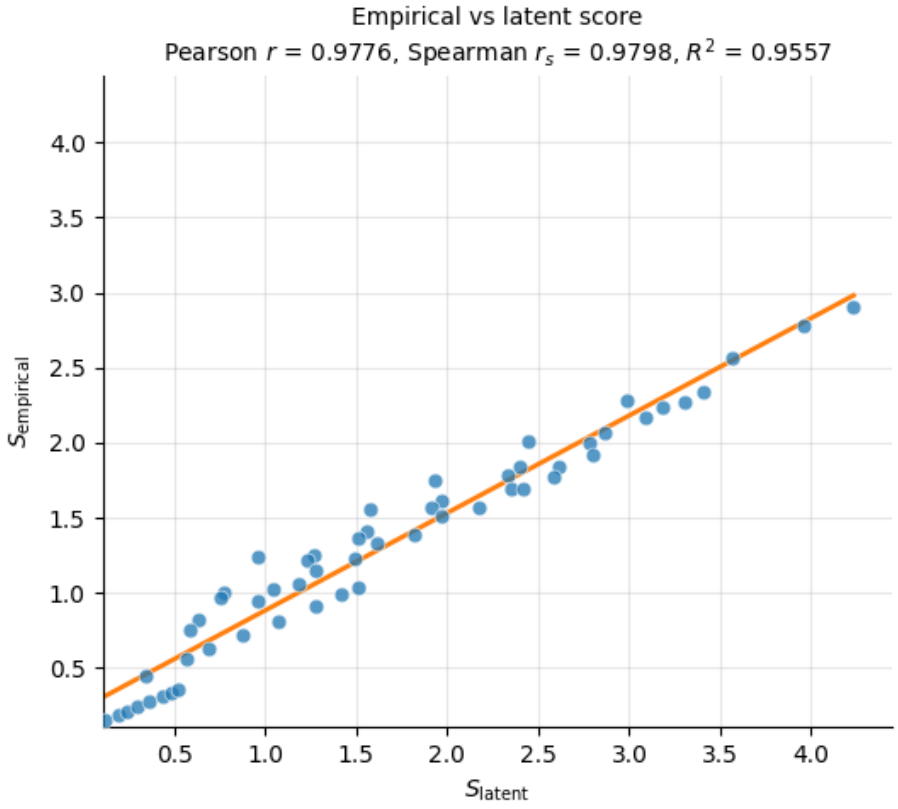


Figure 2: $S_{empirical} = \hat{\alpha}\hat{\rho^2}(\hat{K}+1)$ vs $S_{latent} = \alpha_1\rho_1^2(K+1)$

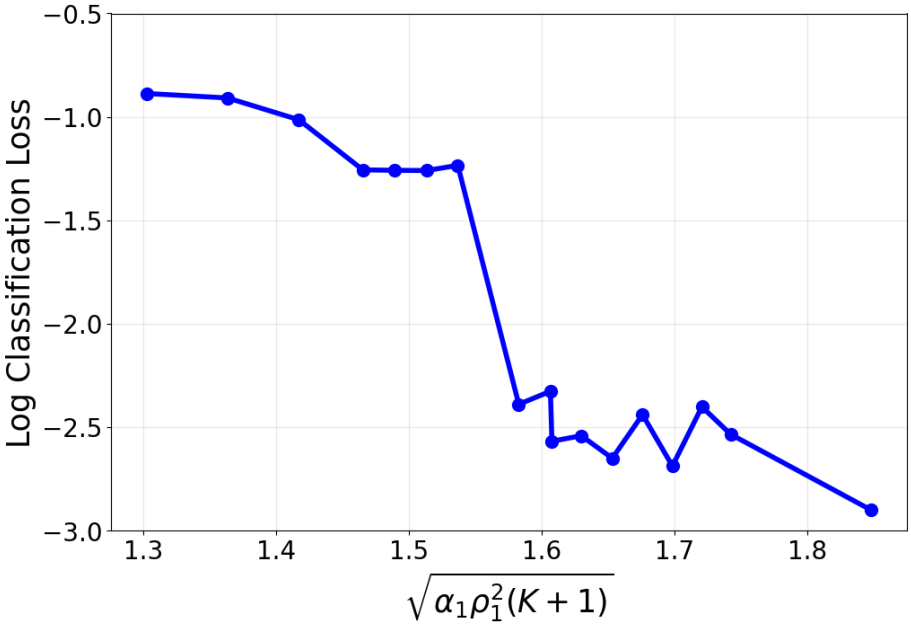


Figure 3: Downstream loss