

Reproducibility Checklist

Question 1: Are all datasets used in the study publicly available, and do they involve any access restrictions, licensing requirements, or special procedures for obtaining them?

All datasets used are publicly available and can be accessed without restrictions.

Question 2: Does the study ensure fully reproducible train/validation/test splits, and how are these splits managed across different datasets?

All datasets provide official train, validation and test splits, which we follow directly to ensure reproducibility and facilitate result verification.

Question 3: Is the codebase and software environment fully released, including all dependencies and configurations needed for reproduction?

The complete codebase is publicly available, specifying all dependencies.

Question 4: Are all training hyperparameters fully documented, including those for optimization, scheduling, and number of epochs?

The hyperparameters, including batch size, learning rate, optimizer, scheduler, and number of epochs, are documented in the supplementary materials.

Question 5: Are all baseline comparisons conducted fairly and consistently, using the same data splits, training protocols, and evaluation metrics?

All baselines are reproduced following official implementations or original papers with consistent splits, training, and evaluation settings.

Question 6: Are the experimental results stable and reproducible across multiple runs with different random seeds?

All experiments are repeated 3 times with different random seeds. Final results are reported as the mean \pm standard deviation.

Question 7: Are the hardware and computational resources fully specified, including GPUs, memory, and environment details?

We provide the experimental setup, memory details, and computational environment used for training and evaluation to ensure reproducibility.

Question 8: Are all evaluation metrics, calculation procedures, and experimental protocols fully specified to ensure fair and reproducible comparisons across baselines and models?

The evaluation metrics, calculation methods, and protocols are explicitly described to allow fair comparison with baselines.