PQPP: A Joint Benchmark for Text-to-Image Prompt and Query Performance Prediction*

Eduard Poesina[1,*,†], Adriana Costache[1,†], Josiane Mothe[2,†], Adrian-Gabriel Chifu[3,†] and Radu Tudor Ionescu[1,†]

[1]Department of Computer Science, University of Bucharest, 14 Academiei Street, Bucharest, Romania
[2]INSPE, IRIT UMR5505 CNRS, Université Toulouse Jean Jaurès, 5 allées Antonio Machado, 31058 Toulouse, France
[3]Aix-Marseille Université, Université de Toulon, CNRS, LIS UMR 7020, Marseille, France

Abstract

Query Performance Prediction (QPP) has been widely studied in information retrieval (IR), but its role in text-to-image generation remains largely unexplored. Unlike retrieval, which aims to return pre-existing images, generative models create new images from text prompts, requiring novel approaches to assess prompt difficulty.

We introduce PQPP, the first large-scale benchmark for Prompt and Query Performance Prediction across both text-to-image retrieval and generation. PQPP provides 10,200 prompts/queries manually annotated with human relevance judgments. The dataset combines DrawBench prompts and MS COCO captions, evaluated using two diffusion models (Stable Diffusion XL, GLIDE) and two retrieval models (CLIP, BLIP-2).

Our findings reveal a weak correlation between retrieval and generation difficulty, highlighting the need for distinct QPP models. We evaluate pre- and post-generation/retrieval predictors, establishing competitive baselines. Key results include:

- Fine-tuned CLIP and BERT predictors achieve the best performance, with CLIP excelling post-retrieval/generation and BERT performing well pre-retrieval/generation.

- Simple linguistic heuristics fail to predict query/prompt difficulty, emphasizing the need for learned models.

- Cross-model generalization remains challenging, as predictors trained on one model often underperform on others.

Beyond benchmarking, PQPP has significant implications for real-world applications. In generative AI, accurately predicting prompt difficulty can enable adaptive user guidance, where systems suggest prompt modifications to improve generation outcomes. In retrieval, it can enhance ranking and search efficiency by identifying ambiguous or overly complex queries. Additionally, PQPP lays the foundation for multimodal QPP research, bridging the gap between IR and generative AI by providing a common evaluation framework for both tasks.

PQPP provides a standardized evaluation framework for QPP in multimodal retrieval and generation, advancing research in prompt optimization and query difficulty estimation. We publicly release our dataset and code to encourage further exploration.

Our benchmark and code are publicly available at https://github.com/Eduard6421/PQPP.

Published in CVPR 2025 [1]

Keywords

Information Systems, Information Retrieval, Query Performance Prediction, Image Retrieval, Prompt

References

[1] E. Poesina, A. Costache, A.-G. Chifu, J. Mothe, R. T. Ionescu, PQPP: A joint benchmark for text-to-image prompt and query performance prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

QPP++ at ECIR

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.
†These authors contributed equally.
eduardgabriel.poe@gmail.com (E. Poesina); adriana16costache@gmail.com (A. Costache); josiane.mothe@irit.fr (J. Mothe); adrian.chifu@univ-amu.fr (A. Chifu); 0000-0002-9301-1950 (R. T. Ionescu)
https://cs.unibuc.ro/ (E. Poesina); https://www.irit.fr/~Josiane.Mothe/ (J. Mothe); https://adrianchifu.com (A. Chifu)
0009-0003-1116-0851 (E. Poesina); 0000-0001-9273-2193 (J. Mothe); 0000-0003-4680-5528 (A. Chifu); 0000-0002-9301-1950 (R. T. Ionescu)