Github

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

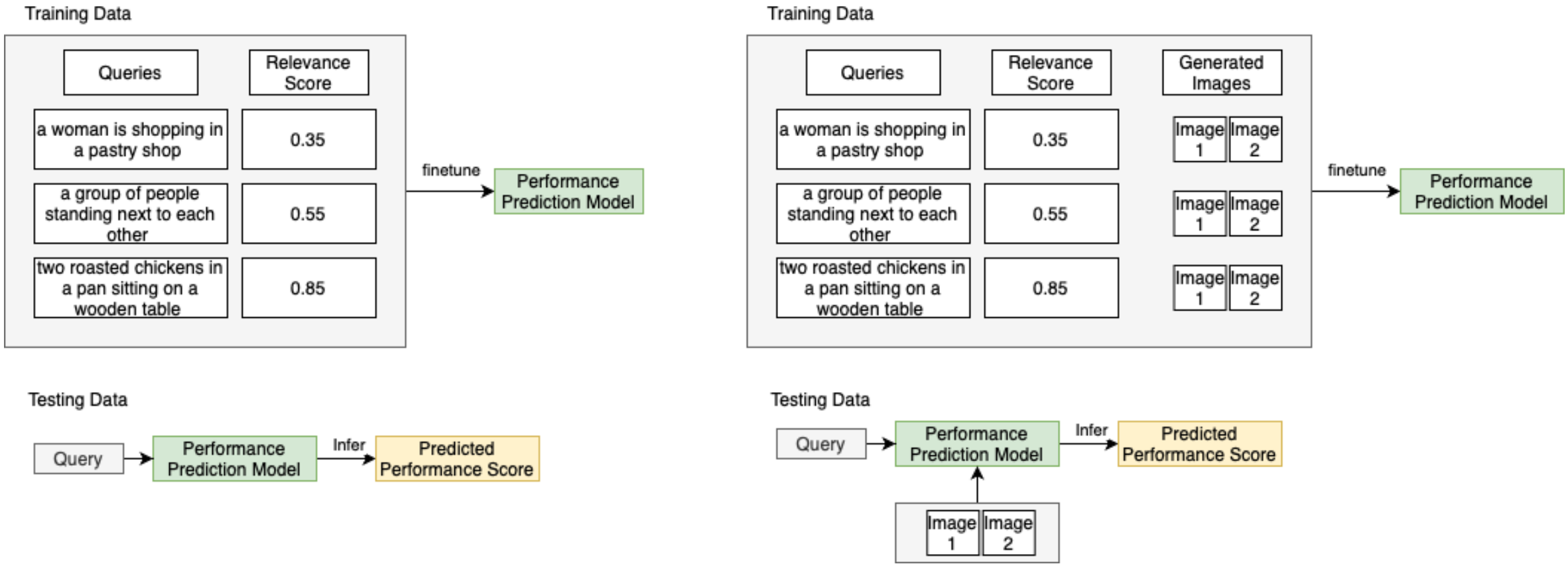
Eduard Poesina[1], Adriana-Valentina Costache[1], Adrian-Gabriel Chifu[2], Josiane Mothe[3], Radu Tudor Ionescu[1]

[1]University of Bucharest
[2] Aix-Marseille Université, LIS
[3] Univ. de Toulouse, Institut de Recherche en Informatique de Toulouse

Github

Introduction

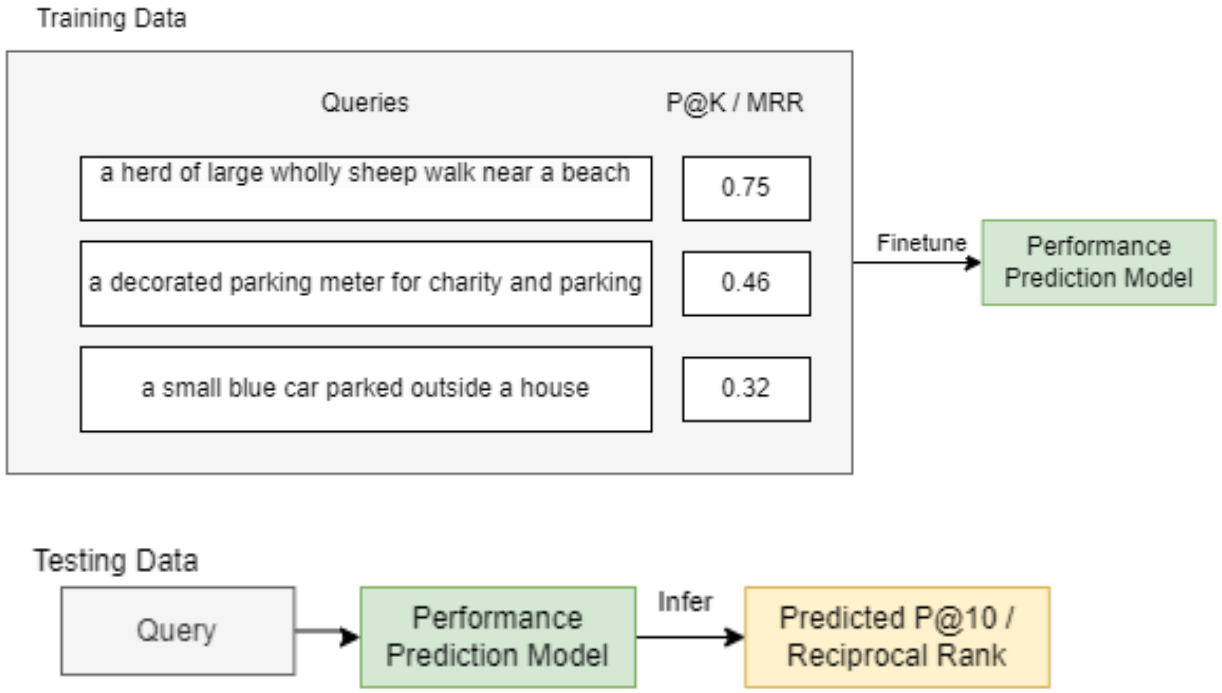
Prompt Performance Prediction – Pre/Post-Generative

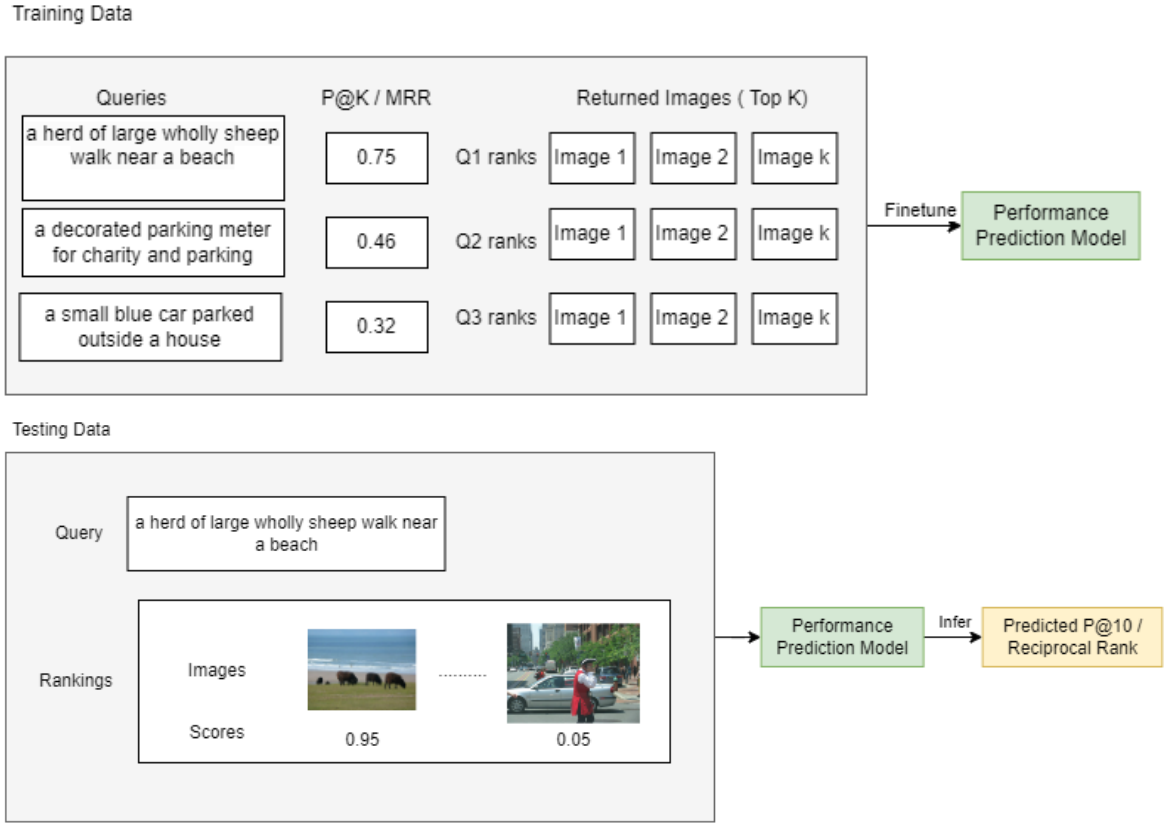


Github

Introduction

Query Performance Prediction – Pre/Post-Retrieval





Github

Introduction

Context & Challenges:

No prior work explores the relationship between
Query Performance Prediction (QPP) and Prompt Performance Prediction (PPP)

Research questions:

RQ1: Are the two tasks related, or they require specifically trained models?
RQ2: Could a QPP model also be used for PPP and vice-versa?

Our contributions:

1. The first joint benchmark for query/prompt performance prediction
2. A database of ~ 10K queries/prompts, with > 1.5M total annotations
3. Baselines: multiple pre-generative/retrieval and post-generative/retrieval

Github

Related Work

Prompt Performance Prediction

Bizozerro et al,2024 - Prompt Performance Prediction and training models on automated relevance scoring

Pavlichenko et al., 2023, SIGIR – Method to adapt a given prompt to improve text to image model performance

Kirstain et al., 2023, NeurIPS - Preference annotations by asking humans to choose an image from a pair of generated images

Query Performance Prediction

Xing et al,2010,ECIR - Groundwork of exploring query difficulty prediction in image retrieval

Liu et al., 2012, Neurocomputing – Focus on estimating query difficulty with unigram language models and visual word verification

Github

Proposed Benchmark

- Source datasets for prompts:
  - MS COCO: 10K captions (sampled via k-means for diversity)
  - DrawBench: all 200 prompts (more suitable for generation)
- Generative models:
  - GLIDE: two image samples per prompt
  - SDXL: two image samples per prompt
- Retrieval Models:
  - CLIP
  - BLIP2

Github

Prompt Performance Assessment

Men sit in the outfield in white uniforms waiting for the pitch.

- ● high relevance
- ○ low relevance
- ○ no relevance
- ○ unrealistic

- ● high relevance
- ○ low relevance
- ○ no relevance
- ○ unrealistic

- ● high relevance
- ○ low relevance
- ○ no relevance
- ○ unrealistic

- ○ high relevance
- ○ low relevance
- ● no relevance
- ○ unrealistic

- ○ high relevance
- ○ low relevance
- ● no relevance
- ○ unrealistic

Label definitions:

- high relevance - over half of the concepts mentioned in the prompt
- low relevance - at least one concept, but fewer than half
- no relevance - unrelated, yet realistic
- Unrealistic - notable artifacts

Github

Prompt Performance Assessment

Target Annotations: 10,200 prompts × 5 images × 3 annotations = 153,000 annotations

Criteria for exclusion:
Cohen's K agreement with control prompts < 0.4

Annotators:
Recruited: 173
Validated: **147 (247,050 annotations)**

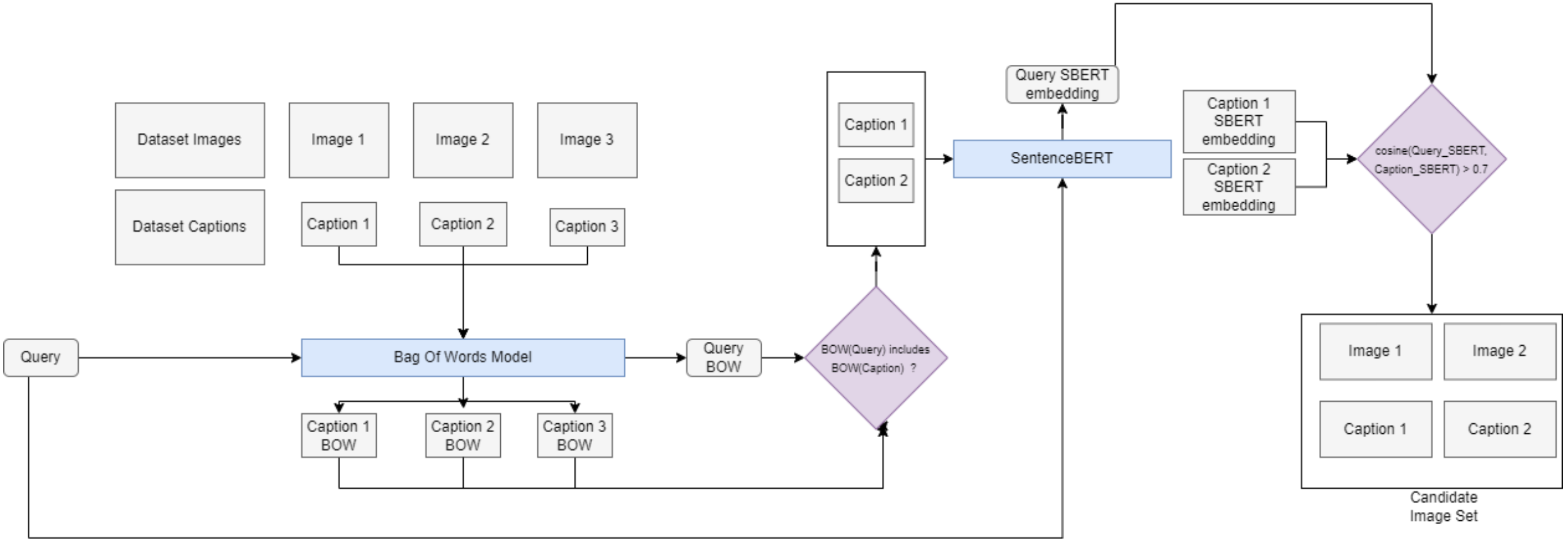
Statistic	Min	Mean	Max
#annotations per person	30	1,681	15,845
Fleiss' κ	0.41	0.54	1.00

Table 1. Statistics about the annotators enrolled in the annotation process for generated images.

https://www.researchgate.net/figure/Fleiss-Kappa-and-Inter-rater-agreement-interpretation-24_tbl3_281652142

Github

Query Performance Assessment



Github

Query Performance Assessment

Target Annotations: 10,200 prompts × Max 2000 images × 3 annotations
Annotated Images: 1,393,363

Criteria for exclusion:
F1 score < 0.4 with control prompts

Annotators:
Recruited: 100
Validated: 93

Min F_1	Mean F_1
0.477	0.727

Github

Predictors

Pre-generation/retrieval Predictors

Basic text predictors
diversity of concepts
lexical density
morphological complexity
frequency of grammatical structures

Fine-tuned BERT
Base architecture, based on cased inputs

Post-generation/retrieval Predictors

Fine-tuned CLIP
Long-CLIP with a Vit-B/32 backbone

Correlation-based CNN
Model trained over the correlation matrix between image pairs

Github

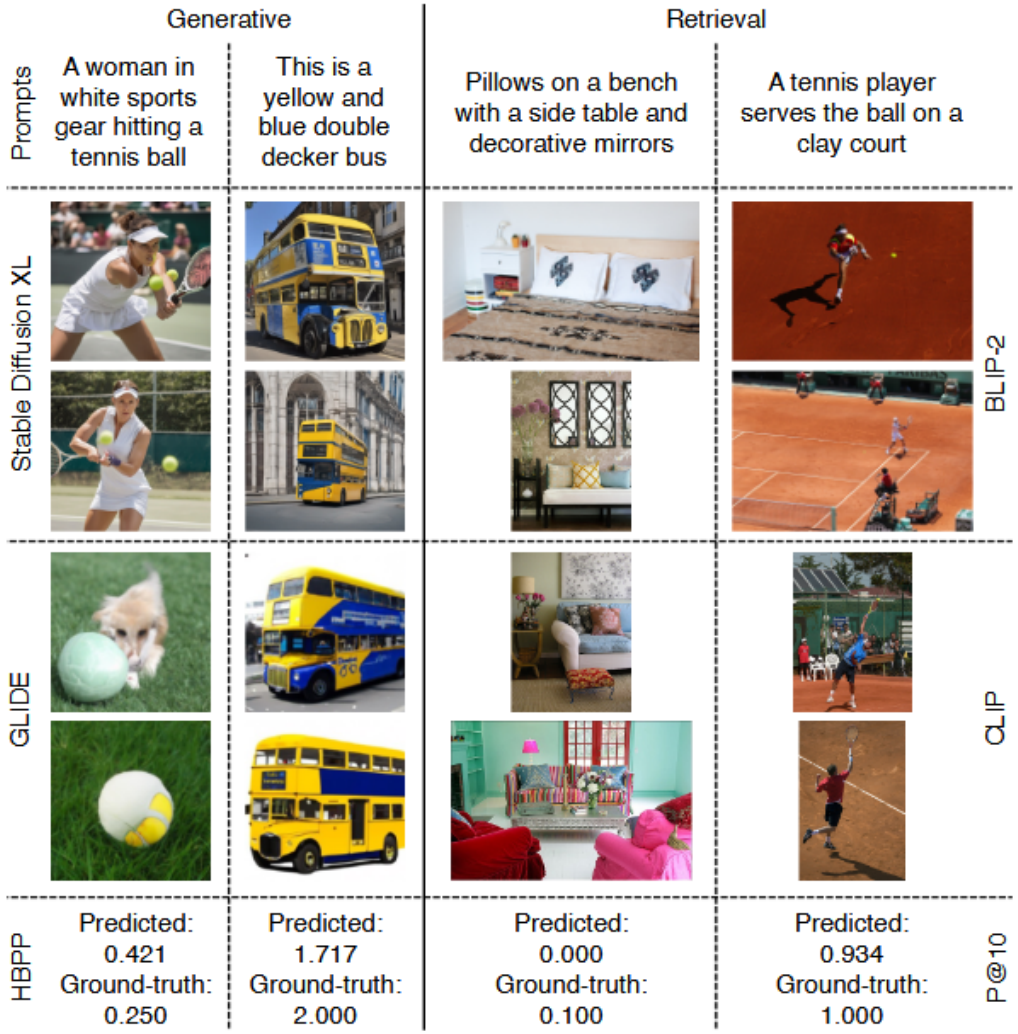
Experiments and Results

Predictor Type	Predictor Name	Generative Task				Retrieval Task							
		GLIDE		SDXL		CLIP				BLIP-2			
		HBPP		HBPP		P@10		RR		P@10		RR	
		Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
Pre-	#synsets	$-0.112^{\ddagger}$	$-0.076^{\ddagger}$	$-0.087^{\ddagger}$	$-0.080^{\ddagger}$	$-0.110^{\dagger}$	$-0.058^{\ddagger}$	-0.034	-0.012	$-0.115^{\ddagger}$	$-0.070^{\ddagger}$	-0.038	-0.010
	#words	$-0.090^{\dagger}$	$-0.084^{\ddagger}$	$-0.105^{\ddagger}$	$-0.109^{\ddagger}$	$-0.133^{\ddagger}$	$-0.104^{\ddagger}$	-0.035	-0.026	$-0.175^{\ddagger}$	$-0.136^{\ddagger}$	-0.038	-0.015
	Fine-tuned BERT	$0.566^{\ddagger}$	$0.406^{\ddagger}$	$0.281^{\ddagger}$	$0.232^{\ddagger}$	$0.451^{\ddagger}$	$0.277^{\ddagger}$	$\mathbf{0.221}^{\ddagger}$	$\mathbf{0.176}^{\ddagger}$	$\mathbf{0.511}^{\ddagger}$	$0.328^{\ddagger}$	$0.168^{\ddagger}$	$0.139^{\ddagger}$
Post-	Fine-tuned CLIP	$\mathbf{0.649}^{\ddagger}$	$\mathbf{0.474}^{\ddagger}$	$\mathbf{0.380}^{\ddagger}$	$\mathbf{0.246}^{\ddagger}$	$\mathbf{0.473}^{\ddagger}$	$\mathbf{0.299}^{\ddagger}$	$0.200^{\ddagger}$	$0.149^{\ddagger}$	$0.498^{\ddagger}$	$\mathbf{0.358}^{\ddagger}$	$0.166^{\ddagger}$	$0.150^{\ddagger}$
	Correlation CNN	$0.548^{\ddagger}$	$0.393^{\ddagger}$	$0.159^{\ddagger}$	$0.107^{\ddagger}$	$0.270^{\ddagger}$	$0.186^{\ddagger}$	$0.189^{\ddagger}$	$0.162^{\ddagger}$	$0.159^{\ddagger}$	$0.133^{\ddagger}$	$\mathbf{0.206}^{\ddagger}$	$\mathbf{0.158}^{\ddagger}$

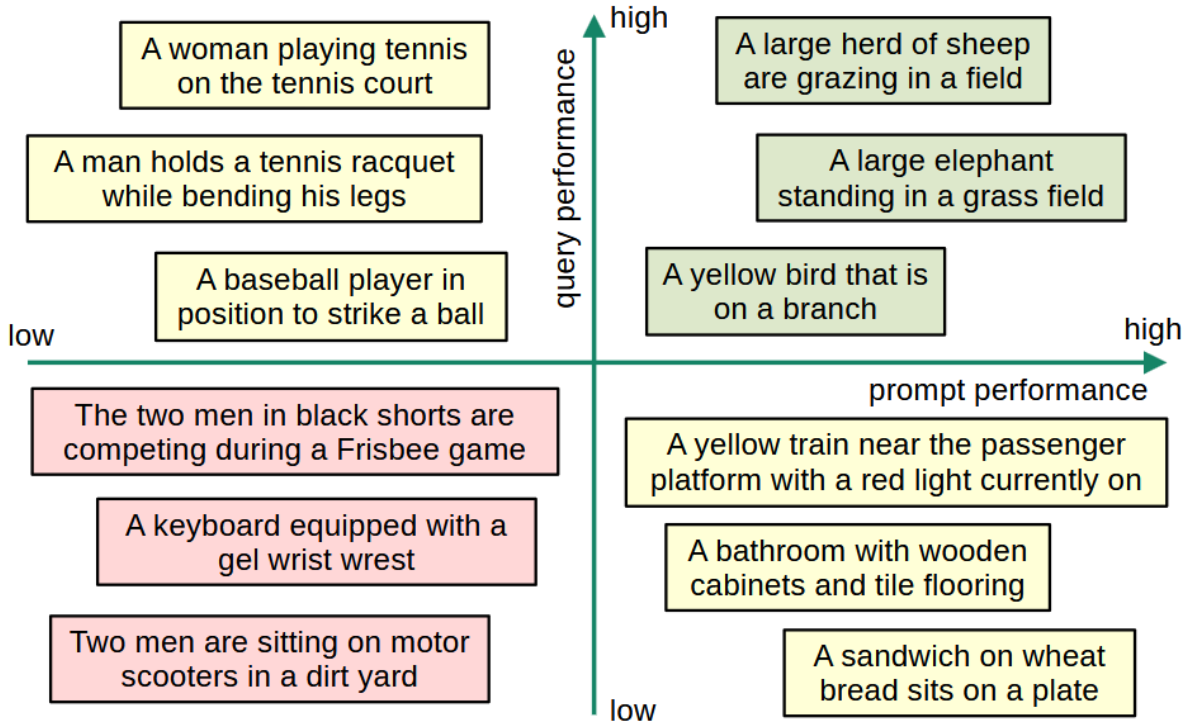
Predictor Results

Github

Experiments and Results



Metric	Pearson	Kendall
HBPP vs. P@10	0.135‡	0.093‡
HBPP vs. RR	0.072‡	0.048†
P@10 vs. RR	0.560‡	0.512‡



UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

amU Aix Marseille Université

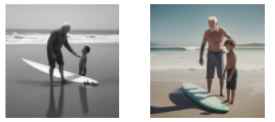
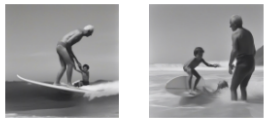
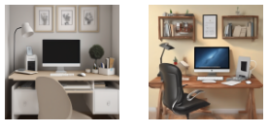
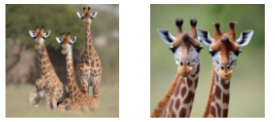
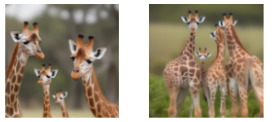
IRIT Institut de Recherche en Informatique de Toulouse
CNRS - Toulouse INP - UT - UT Capitole - UT2

Github

ECIR 2025

Experiments and Results

Prompt	GLIDE	Predicted Score	SDXL	Predicted Score
A cat rests on a dogs back as he lies on the sidewalk		0.67		1.71
A young man riding a wave on his surf board.		0.49		2.00
A man jumps to catch a frisbee on the beach.		-0.02		1.87
A man in a village has pots full of food and other food on the stove.		0.01		1.95
Two ultimate Frisbee players jumping to contest a Frisbee.		-0.23		1.60

Original Captions	Rephrased Captions		
An older man is holding a surfboard while a young boy stands on it.	A young boy standing on a surfboard, held steady by an elderly man.	An older man grips a surfboard, supporting a young boy standing on it.	An elderly man steadying a surfboard as a young boy balances on top.
1.88	1.95	1.88	1.83
			
A study table with computer, mouse and keyboard. Photo frame are also kept.	A neatly arranged study table with a computer, mouse, keyboard, and a photo frame nearby.	A study desk featuring a computer setup with a mouse and keyboard, alongside a photo frame.	A computer, mouse, and keyboard on a study table, with a photo frame placed beside them.
1.70	1.90	1.79	1.74
			
Two little giraffes standing between two slightly bigger ones.	Two small giraffes standing between two slightly taller giraffes.	Two young giraffes nestled between two slightly bigger ones.	A pair of little giraffes positioned between two larger giraffes.
0.30	1.48	1.08	1.40
			

Github

Thank you!

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

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

[1]University of Bucharest
[2] Aix-Marseille Université, LIS
[3] Univ. de Toulouse, Institut de Recherche en Informatique de Toulouse

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

amU Aix Marseille Université

IRIT Institut de Recherche en Informatique de Toulouse
CNRS - Toulouse INP - UT - UT Capitole - UT2

Github

ECIR 2025