

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

## Supplementary Material


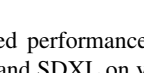
Prompt	GLIDE	Predicted Score	SDXL	Predicted Score
A cat rests on a dogs back as he lies on the sidewalk		0.67		1.71
A young man riding a wave on his surf board.		0.49		2.00
A man jumps to catch a frisbee on the beach.		-0.02		1.87
A man in a village has pots full of food and other food on the stove.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

Figure 5. Predicted performance scores by a fine-tuned BERT model for GLIDE and SDXL on various test prompts. The scores are presented alongside the images generated by each model. Best viewed in color.

## 8. Task Usefulness

In the context of text-to-image generation, if a prompt is predicted as difficult, the system could initiate a conversation to refine the prompt in order to overcome the difficulty and improve the final output. Moreover, the system could indicate to the user its inability to provide a satisfactory image, or it could give positive feedback to the user when a query is predicted as easy. For image generation and image retrieval, when a system is predicted to perform poorly on a prompt/query, additional processes can be activated to improve performance, such as:

- Automatic query reformulation: If a query is predicted to perform poorly, it can be automatically reformulated to improve retrieval effectiveness.
- Automatic query expansion: For queries expected to perform poorly, QPP can trigger automatic query expansion, adding terms that might improve search performance.
- Model selection: Search engines can allocate more computational resources to queries predicted to perform poorly. QPP helps in choosing the most appropriate retrieval algorithm based on the predicted performance for a specific type of query.
- Query proposals: Users can be provided with alternative query suggestions if their original query is predicted to perform poorly, improving user satisfaction.
- Adapted filtering: In content-based filtering systems, QPP can adapt filtering strategies based on the predicted performance of the query, leading to better results.

We further harness the PQPP benchmark and the trained

prompt performance predictors to exemplify their utility in two of the applications listed above.

**Use case 1: Generative model selection.** Next, we illustrate the application in generative model selection via a series of prompt examples that are predicted to perform poorly for GLIDE, but are well-handled by SDXL. Therefore, one could use the more efficient GLIDE model to generate images for easy prompts, and turn to the less efficient (but more capable) SDXL for difficult prompts. The prompt examples, along with their predicted performance, and images generated by GLIDE and SDXL are shown in Figure 5. The illustrated samples, selected from our test set, indicate that a pre-generation model such as BERT can guide the selection of a more suitable generator, when required.

**Use case 2: Automatic query reformulation.** Another important use case of our benchmark is automatic query reformulation. In Figure 6, we illustrate examples of rephrased captions by ChatGPT 4.0 using the following prompt:

You will need to behave as a large language model made to assist with query reformulation for the application of prompt performance prediction. You will need to reformulate the query in order to increase a text-to-image model to its maximum performance. Here are a few examples of texts and their estimated performance scores:

<Caption>  
Black and white of windsurfers on a lake.  
<Score>  
0.15

<Caption>  
A black and white picture of several buses driving down a street.  
<Score>  
0.1

<Caption>  
A bed made up with linens, is outside on a street corner.  
<Score>  
0.15



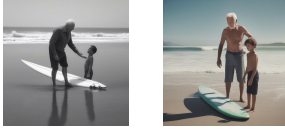
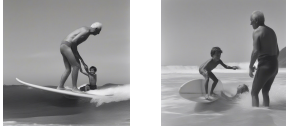



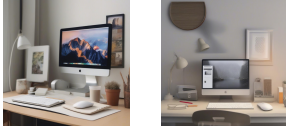




Original Captions	Rephrased Captions		
<p>An older man is holding a surfboard while a young boy stands on it.</p> <p>1.88</p> 	<p>A young boy standing on a surfboard, held steady by an elderly man.</p> <p>1.95</p> 	<p>An older man grips a surfboard, supporting a young boy standing on it.</p> <p>1.88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

Figure 6. Examples of prompts reformulated by ChatGPT 4.0 and associated scores predicted by the fine-tuned BERT performance predictor. The images generated by SDXL for reformulated prompts with higher scores are typically more relevant than those generated for the original prompts. Best viewed in color.

You will receive a caption and you must offer 3 rephrases of the given caption, considering the best performance for obtaining the desired image with a text-to-image model.

<Caption>  
[target caption placeholder]

Figure 6 shows that the rephrased captions lead to higher scores predicted by the fine-tuned BERT pre-generation predictor, as well as images that are better aligned with the original/rephrased prompt.

## 9. Related Work on QPP in Text Retrieval

Pre-retrieval features are based on information available prior to the execution of the query. Some are independent to the document collection, such as query length, part-of-speech features (including the number of words of some grammatical categories), query ambiguity [15], and query complexity [32]. Other pre-retrieval features depend on the document collection statistics, such as the inverse doc-

ument frequency [45], the query scope [23] (which measures the coverage of a query within the context of a document collection and estimates the proportion of documents that are relevant to the query), and the SCQ [60] (which is a similarity score between the query and the collection). Pre-retrieval predictors have the huge advantage of being determined before running the search, but they have been found to be less effective than post-retrieval ones on textual ad hoc retrieval [22, 32, 38, 43]. Unlike pre-retrieval features, post-retrieval features require conducting document retrieval with the query. Most of these features are calculated based on the scores of the retrieved documents, quantifying the robustness of the document list, or considering the distribution of the document scores [8, 11, 12, 38, 59]. In textual IR, the Clarity Score estimates the specificity of a query considering the language distribution of the document collection and that of the top-retrieved documents [11]. The Normalized Query Commitment (NQC), also known as the query drift [43], measures how much the retrieved documents deviate from the central topic of the query. The Weighted Information Gain (WIG) calculates the difference in information content between the documents retrieved for

a specific query and a baseline distribution of information in the collection or corpus, based on the scores of the top-retrieved documents [61].

The main conclusions from the earlier studies on QPP for textual ad hoc retrieval are that post-retrieval predictors outperform pre-retrieval ones [38], and combinations of predictors using supervised approaches are the most effective [16].

Some recent studies investigated QPP on neural IR (NIR) systems [3, 13, 20, 31, 44, 57]. Datta et al. [13] employed convolutional neural layers for their Deep-QPP predictor. This architecture has further been combined with LETOR post-retrieval predictors with some success [14]. According to Faggioli et al. [20], QPP models which have been developed for sparse IR methods perform worse when applied to NIR systems. However, the authors did not consider linguistic-based predictors in their work. On the other hand, supervised BERT-based QPP models seem to work better. Arabzadeh et al. [4] used BERT to predict the performance of search queries in terms of their ability to retrieve relevant documents from a corpus. Such predictors may better capture the semantic aspects of the query-document matching.

Other recent studies focused on the transition from ad hoc search to conversational search [31] or question answering [21, 42]. In conversational search, the experiments showed that supervised QPP methods outperform unsupervised ones when a large amount of training data is available, but unsupervised methods are effective in conversational dense retrieval method assessment.

## 10. Details on Generated Image Annotation

To annotate generated images in terms of relevance, the human annotators are essentially asked to count concepts (objects, attributes, actions) that are both mentioned in the input prompt and present in the generated image. Depending on the number of concepts that are present in the image, the annotators are instructed to label images as follows: high relevance (more than half of the concepts are present), low relevance (less than half of the concepts are present), no relevance (no concept is present), unrealistic (the image contains visible generative artifacts, regardless of the number of concepts). The users are informed that a *concept* can be an object, a property of an object or an activity. For example, the caption “a white dog catches a Frisbee in its mouth” contains 5 concepts: the adjective “white”, the noun “dog”, the verb “catch”, the noun “Frisbee”, and the noun “mouth”. The users are also given a list of potential generation artifacts: objects with inconsistent appearance (wrong shape, wrong color), counting artifacts (too many / too few object parts of a certain kind), perspective artifacts (different parts of the same object are jointly depicted from visibly different perspectives), structural artifacts (objects have wrong, missing or added parts), etc.

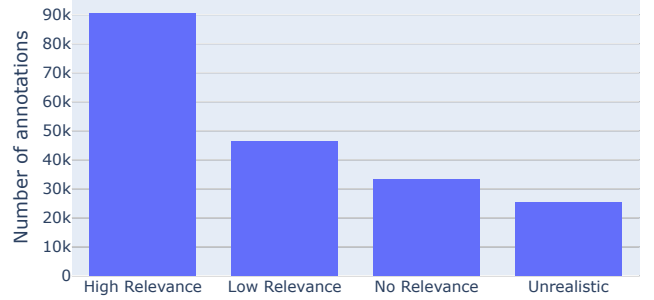


Figure 7. A histogram showing the number of annotations per category for images generated by Stable Diffusion XL and GLIDE.

In Figure 7, we show the number of annotations per category label computed for the four images generated for each prompt. Although the annotations corresponding to the ground-truth images are excluded, it is clear that most images are voted as highly relevant, confirming that Stable Diffusion and GLIDE generally produce relevant results.

The annotators providing the relevance judgments are adults having at least a bachelor or college degree. The recruited annotators willingly agreed to engage in the annotation process, after reading our terms and conditions. Annotators are allowed to opt out at any time during the annotation process. To reduce bias or uncertainty, annotators are permitted to update previously made annotations or skip specific prompts altogether. Annotators are informed about the inclusion of control prompts within their tasks, but are not given specifics on the frequency of such prompts. A fair compensation (proportional to the number of annotated prompts) is given to each annotator with a Cohen’s  $\kappa$  coefficient higher than 0.4 on the control prompts.

To compute HBPP, we first group the annotations into relevant (combining *high relevance* and *low relevance* labels) and irrelevant (combining *no relevance* and *unrealistic* labels). We hereby acknowledge that the distinction between *high relevance* and *low relevance* is more difficult to determine, involving a fine assessment of how many of the prompt elements are depicted in the image. This requires evaluators to consider not just the presence of these elements, but also their significance and portrayal within the image, making the distinction between high and low relevance inherently more subjective and challenging. In contrast, the distinction between the high-level categories (*relevant* and *irrelevant*) can be easily assessed.

## 11. Predictor Implementation Details

**Fine-tuned BERT.** The regression head consists of a dropout layer and two fully connected layers. The dropout rate is set to 0.3 to prevent overfitting. The first dense layer is based on ReLU activation functions, and it takes the [CLS] token returned by BERT and transforms it into

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 <sup>‡</sup>	-0.076 <sup>‡</sup>	-0.087 <sup>‡</sup>	-0.080 <sup>‡</sup>	-0.110 <sup>†</sup>	-0.058 <sup>‡</sup>	-0.034	-0.012	-0.115 <sup>‡</sup>	-0.070 <sup>‡</sup>	-0.038	-0.010
	#words	-0.090 <sup>†</sup>	-0.084 <sup>‡</sup>	-0.105 <sup>‡</sup>	-0.109 <sup>‡</sup>	-0.133 <sup>‡</sup>	-0.104 <sup>‡</sup>	-0.035	-0.026	-0.175 <sup>‡</sup>	-0.136 <sup>‡</sup>	-0.038	-0.015
	Average word length	0.039	0.041 <sup>†</sup>	-0.067	-0.011	-0.090 <sup>†</sup>	-0.066 <sup>‡</sup>	-0.064 <sup>†</sup>	-0.035	-0.150 <sup>‡</sup>	-0.104 <sup>‡</sup>	-0.116 <sup>‡</sup>	-0.079 <sup>‡</sup>
	Ratio of proper nouns	0.002	-0.027	-0.007	-0.034	-0.053	-0.053 <sup>†</sup>	-0.012	0.001	-0.106 <sup>‡</sup>	-0.102 <sup>‡</sup>	-0.063 <sup>†</sup>	-0.040
	Ratio of acronyms	0.001	0.008	0.007	-0.031	0.012	-0.000	0.014	0.018	-0.008	-0.028	0.017	-0.002
	Ratio of numerals	-0.028	-0.026	-0.074 <sup>‡</sup>	-0.072 <sup>‡</sup>	-0.049	-0.046	0.007	0.006	-0.065 <sup>†</sup>	-0.070 <sup>‡</sup>	-0.032	-0.025
	Ratio of conjunctions	0.054	0.044 <sup>†</sup>	0.037	-0.008	-0.079 <sup>‡</sup>	-0.062 <sup>‡</sup>	-0.024	-0.018	-0.121 <sup>‡</sup>	-0.097 <sup>†</sup>	-0.032	-0.030
	Ratio of prepositions	0.043	0.031	0.033	0.003	0.020	0.020	0.035	0.030	0.014	0.007	0.050	0.038
	Edge Count	0.058 <sup>†</sup>	0.084 <sup>‡</sup>	0.020	0.020	0.033	0.047 <sup>†</sup>	0.031	0.011	0.048	0.057	0.030	0.007
	Edge Weight Sum	0.054	0.083 <sup>‡</sup>	0.019	0.026	0.033	0.048 <sup>†</sup>	0.030	0.011	0.033	0.057 <sup>†</sup>	0.028	0.010
	Inverse Edge Frequency	0.119 <sup>‡</sup>	0.062 <sup>‡</sup>	0.018	0.039	0.069 <sup>†</sup>	0.039	0.019	0.012	0.046	0.025	0.008	0.011
	Degree Centrality	0.073 <sup>‡</sup>	0.071 <sup>‡</sup>	0.022	0.021	0.059 <sup>†</sup>	0.030	0.029	0.010	0.066 <sup>†</sup>	0.038	0.034	0.020
	Closeness Centrality	0.032	0.039 <sup>†</sup>	0.133 <sup>‡</sup>	0.048 <sup>‡</sup>	0.077 <sup>†</sup>	0.035	0.036	0.013	0.048	0.027	0.042	0.010
	Betweenness Centrality	0.025	0.019	0.062 <sup>†</sup>	0.047 <sup>†</sup>	0.054	0.038	0.026	0.018	0.040	0.035 <sup>†</sup>	0.034	0.027
	PageRank	0.064 <sup>†</sup>	0.038	0.088 <sup>†</sup>	0.022	0.022	0.021	0.014	0.012	0.049	0.013	0.058 <sup>†</sup>	0.019
	Fine-tuned BERT	0.566 <sup>‡</sup>	0.406 <sup>‡</sup>	0.281 <sup>‡</sup>	0.232 <sup>‡</sup>	0.451 <sup>‡</sup>	0.277 <sup>‡</sup>	<b>0.221<sup>†</sup></b>	<b>0.176<sup>†</sup></b>	<b>0.511<sup>†</sup></b>	0.328 <sup>‡</sup>	0.168 <sup>‡</sup>	0.139 <sup>‡</sup>
Post-	Fine-tuned CLIP	<b>0.649<sup>‡</sup></b>	<b>0.474<sup>‡</sup></b>	<b>0.380<sup>‡</sup></b>	<b>0.246<sup>‡</sup></b>	<b>0.473<sup>†</sup></b>	<b>0.299<sup>‡</sup></b>	0.200 <sup>‡</sup>	0.149 <sup>‡</sup>	0.498 <sup>‡</sup>	<b>0.358<sup>‡</sup></b>	0.166 <sup>‡</sup>	0.150 <sup>‡</sup>
	Correlation CNN	0.548 <sup>‡</sup>	0.393 <sup>‡</sup>	0.159 <sup>‡</sup>	0.107 <sup>†</sup>	0.270 <sup>‡</sup>	0.186 <sup>‡</sup>	0.189 <sup>‡</sup>	0.162 <sup>‡</sup>	0.159 <sup>‡</sup>	0.133 <sup>‡</sup>	<b>0.206<sup>†</sup></b>	<b>0.158<sup>†</sup></b>
	HPSv2	0.482 <sup>‡</sup>	0.352 <sup>‡</sup>	0.026	0.033	-	-	-	-	-	-	-	-

Table 5. Results of the prompt/query performance predictors for the generative and retrieval settings on the PQPP test set. On the generative task, we report the correlation of the predicted value with the HBPP performance of SDXL and GLIDE, respectively. On the retrieval task, the correlation is computed for the P@10 and RR scores of CLIP and BLIP-2, respectively. For each task and model, the highest correlation is highlighted in bold. According to a Student’s t-test, the results marked with <sup>†</sup> and <sup>‡</sup> are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

a 512-dimensional hidden representation. The second layer contains a single neuron (activated by sigmoid) that predicts prompt/query performance. Before training, the ground-truth performance values are normalized to  $[0, 1]$ . We employ grid search on the validation set to establish the optimal hyperparameter configuration. More specifically, we consider learning rates between  $10^{-3}$  and  $10^{-6}$ , and weight decays in the set  $\{0, 0.1, 0.01\}$ . All versions are trained for 15 epochs with early stopping, on mini-batches of 256 samples. We employ AdamW and optimize the mean squared error (MSE) loss. The fine-tuning is independently carried out for each generation and retrieval model.

**Fine-tuned CLIP.** For the generative task, the model uses all four images generated by SDXL and GLIDE. For the retrieval task, we limit the training data to the first 25 images returned by each retrieval model. Although 10 images would be enough for the P@10 metric, estimating the RR measure can require more images. A statistical analysis of the training queries indicates that more than 95% of the queries have the first relevant image at a rank higher than 25, which motivates our choice for limiting the training data to 25 images per model. The regression/classification head is composed of a two-layer neural network of 512 and 256 neurons, respectively. Both layers are based on ReLU activation. A dropout layer with a drop rate of 0.5 is added

after each dense layer. For the generative task, another layer comprising a single neuron is added to predict prompt performance. The objective of the model is to minimize the MSE loss. For the retrieval task, the last neuron has to determine if an input (query, image) pair is relevant or not. This is a binary classification task, so the model is trained via binary cross-entropy. We perform a grid search to find the best hyperparameters, considering learning rates between  $10^{-3}$  and  $10^{-6}$ , and weight decays in the set  $\{0, 0.1, 0.01\}$ . We employ the AdamW optimizer for 25 epochs with early stopping, using a batch size of 256.

**Correlation-based CNN.** For the generative task, the size of the input correlation matrix is  $4 \times 4$ , comprising images generated by both SDXL and GLIDE. For the retrieval task, we apply the same limit to the number of retrieved images per query as for the fine-tuned CLIP predictor. Hence, the size of the correlation matrix for one retrieval model is  $25 \times 25$ . We concatenate the correlation matrices for CLIP and BLIP-2 models in the channel dimension, which results in a tensor of  $25 \times 25 \times 2$  components that is given as input to the CNN.

The CNN architecture is composed of four convolutional-pooling blocks, followed by two linear layers. This is a custom architecture that comprises  $3 \times 3$  convolutional filters applied at a stride of 1, using a padding



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-	Fine-tuned BERT	0.237 <sup>†</sup>	0.304 <sup>†</sup>	0.167	0.137	−0.064	−0.069	<b>−0.020</b>	−0.175	−0.101	−0.071	−0.063	−0.102
Post-	Fine-tuned CLIP	<b>0.417<sup>†</sup></b>	<b>0.317<sup>‡</sup></b>	<b>0.412<sup>†</sup></b>	<b>0.318<sup>†</sup></b>	<b>−0.021</b>	<b>−0.030</b>	−0.198	−0.019	<b>0.161</b>	<b>0.129</b>	0.083	<b>0.138</b>
	Correlation CNN	0.387 <sup>†</sup>	0.276 <sup>‡</sup>	0.157	0.058	−0.185	−0.098	−0.177	<b>0.117</b>	0.130	0.092	<b>0.099</b>	0.018

Table 6. Cross-dataset results of the prompt/query performance predictors for the generative and retrieval settings, using MS COCO for training and DrawBench for testing. On the generative task, we report the correlation of the predicted value with the HBPP performance of SDXL and GLIDE, respectively. On the retrieval task, the correlation is computed for the P@10 and RR scores of CLIP and BLIP-2, respectively. For each task and model, the highest correlation is highlighted in bold. According to a Student’s t-test, the results marked with <sup>†</sup> and <sup>‡</sup> are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

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-	Fine-tuned BERT	0.219 <sup>‡</sup>	0.129 <sup>‡</sup>	0.050	<b>0.086<sup>‡</sup></b>	<b>0.032</b>	0.028	−0.011	−0.012	−0.026	−0.009	−0.043	−0.034 <sup>†</sup>
Post-	Fine-tuned CLIP	<b>0.287<sup>‡</sup></b>	<b>0.188<sup>‡</sup></b>	<b>0.078<sup>†</sup></b>	0.034 <sup>†</sup>	0.030	<b>0.040<sup>†</sup></b>	<b>0.086<sup>‡</sup></b>	<b>0.100<sup>‡</sup></b>	−0.041	−0.022	<b>0.070<sup>†</sup></b>	<b>0.054<sup>†</sup></b>
	Correlation CNN	0.194 <sup>‡</sup>	0.130 <sup>‡</sup>	0.047	0.085 <sup>‡</sup>	−0.094 <sup>†</sup>	−0.081 <sup>‡</sup>	0.045 <sup>†</sup>	0.032 <sup>†</sup>	<b>0.065<sup>†</sup></b>	<b>0.050<sup>†</sup></b>	0.018	0.026

Table 7. Cross-dataset results of the prompt/query performance predictors for the generative and retrieval settings, using DrawBench for training and MS COCO for testing. On the generative task, we report the correlation of the predicted value with the HBPP performance of SDXL and GLIDE, respectively. On the retrieval task, the correlation is computed for the P@10 and RR scores of CLIP and BLIP-2, respectively. For each task and model, the highest correlation is highlighted in bold. According to a Student’s t-test, the results marked with <sup>†</sup> and <sup>‡</sup> are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

of 1. The number of filters in each of the four convolutional layers is 64, 128, 256 and 512, respectively. A max-pooling is applied after each convolutional layer. The pooling operation uses  $2 \times 2$  filters applied at a stride of 2. The first fully connected layer comprises 1024 units. Each hidden neuron is followed by a ReLU activation. The final layer comprises a single neuron that is trained in a regression setting via the MSE loss. The hyperparameter tuning is identical to the one employed for the fine-tuned CLIP model. The correlation-based CNN is trained for 25 epochs using AdamW with early stopping, on mini-batches of 256 samples.

## 12. More Quantitative Results

**Results with more predictors.** In Table 5, we present the results of all the considered predictors, while Table 3 only shows the most interesting ones. We consider that it is important to also report failed attempts with specific predictors. The additional predictors are generally based on basic features extracted from queries. The tested predictors are the following: the diversity of concepts (number of WordNet synsets per prompt/query), the lexical density (number of words per prompt/query), the morphological complexity

(average word length measured in characters), and the frequency of specific grammatical structures (ratio of proper nouns, ratio of acronyms, ratio of numerals, ratio of conjunctions and ratio of prepositions).

Following the work of Arabzadeh et al. [2], we implement a suite of predictors based on neural embeddings. In their work, the authors use an ego network to represent each query as a graph. The ego network construction relies on a pre-trained embedding model, such as *word2vec*, which is guided by two hyperparameters:  $\alpha$ , controlling network depth, and  $\beta$ , specifying the minimum similarity threshold for node connections. To build the network, terms directly connected to the root term (ego) must have a similarity of at least  $\beta$ . For subsequent levels, the similarity threshold is dynamically adjusted as  $\beta$  is multiplied with the connecting term’s similarity from the previous level. Each child node identifies and connects to its most similar terms meeting this criterion, creating a hierarchical structure. Graph-based metrics, including Edge Count, Edge Weight Sum, Inverse Edge Frequency, Degree Centrality, Closeness Centrality, Betweenness Centrality, and PageRank, are computed over these networks and aggregated to predict query performance.

Predictor Type	Predictor Name	Generative Task				Retrieval Task							
		CLIP→GLIDE		BLIP-2→SDXL		GLIDE→CLIP				SDXL→BLIP-2			
		P@10→HBPP		P@10→HBPP		HBPP→P@10		HBPP→RR		HBPP→P@10		HBPP→RR	
		Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
Pre-	Fine-tuned BERT	<b>0.108<sup>‡</sup></b>	<b>0.071<sup>‡</sup></b>	<b>0.103<sup>‡</sup></b>	<b>0.157<sup>‡</sup></b>	<b>0.165<sup>‡</sup></b>	<b>0.174<sup>‡</sup></b>	<b>0.118<sup>‡</sup></b>	<b>0.109<sup>‡</sup></b>	0.155 <sup>‡</sup>	<b>0.167<sup>‡</sup></b>	0.094 <sup>‡</sup>	0.087 <sup>‡</sup>
Post-	Fine-tuned CLIP	0.075 <sup>‡</sup>	0.039	0.092 <sup>‡</sup>	0.121 <sup>‡</sup>	0.134 <sup>‡</sup>	0.103 <sup>‡</sup>	0.090 <sup>‡</sup>	0.071 <sup>‡</sup>	<b>0.174<sup>‡</sup></b>	0.155 <sup>‡</sup>	<b>0.135<sup>‡</sup></b>	<b>0.114<sup>‡</sup></b>
	Correlation CNN	0.111 <sup>‡</sup>	0.066 <sup>‡</sup>	0.080 <sup>‡</sup>	0.037	0.053	0.037	0.030	0.024	0.026	0.022	0.030	0.022

Table 8. Cross-task results of the prompt/query performance predictors on the PQPP benchmark. We report the correlation results for two cross-task model pairs: (GLIDE, CLIP) and (SDXL, BLIP-2). This pairing generates the following evaluation cases: CLIP→GLIDE, BLIP-2→SDXL, GLIDE→CLIP and SDXL→BLIP-2. For each case, the highest correlation is highlighted in bold. According to a Student’s t-test, the results marked with † and ‡ are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

Dataset	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
MS COCO	Pre-	Fine-tuned BERT	0.550 <sup>‡</sup>	0.400 <sup>‡</sup>	0.254 <sup>‡</sup>	0.244 <sup>‡</sup>	0.454 <sup>‡</sup>	0.271 <sup>‡</sup>	0.257 <sup>‡</sup>	0.197 <sup>‡</sup>	0.489 <sup>‡</sup>	0.320 <sup>‡</sup>	0.149 <sup>‡</sup>	0.112 <sup>‡</sup>
	Post-	Fine-tuned CLIP	0.657 <sup>‡</sup>	0.479 <sup>‡</sup>	0.360 <sup>‡</sup>	0.245 <sup>‡</sup>	0.435 <sup>‡</sup>	0.315 <sup>‡</sup>	0.127 <sup>‡</sup>	0.105 <sup>‡</sup>	0.488 <sup>‡</sup>	0.399 <sup>‡</sup>	0.058	0.097 <sup>‡</sup>
DrawBench	Pre-	Fine-tuned BERT	0.358 <sup>‡</sup>	0.274 <sup>‡</sup>	0.511 <sup>‡</sup>	0.216 <sup>†</sup>	−0.050	−0.046	−0.152	−0.149	−0.162	−0.102	0.016	0.024
	Post-	Fine-tuned CLIP	0.456 <sup>‡</sup>	0.335 <sup>‡</sup>	0.462 <sup>‡</sup>	0.205	−0.118	−0.109	−0.161	−0.128	−0.060	−0.039	−0.048	−0.348 <sup>†</sup>

Table 9. Results of prompt/query performance predictors on MS COCO vs. DrawBench. On the generative task, we report the correlation of the predicted value with the HBPP performance of SDXL and GLIDE, respectively. On the retrieval task, the correlation is computed for the P@10 and RR scores of CLIP and BLIP-2, respectively. According to a Student’s t-test, the results marked with † and ‡ are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

In general, we find that predictors based on simple heuristics are not capable of capturing prompt/query performance, showcasing typically low correlations, under 0.1. The predictors based on ego networks [2] do not seem to be any better. We perform an additional experiment with the pre-trained HPSv2 [51] model, employing it to predict the HBPP scores. This model is not as good as the fine-tuned predictors, failing to predict HBPP for SDXL. In general, we find that the only predictors able to consistently predict performance across all models and tasks are the supervised ones, namely the fine-tuned BERT, the fine-tuned CLIP and the correlation CNN.

**Cross-dataset results.** In Table 6, we present results of supervised predictors trained on prompts/queries from MS COCO and tested on prompts/queries from DrawBench. Conversely, in Table 7, we show the results of the same predictors trained on DrawBench and evaluated on MS COCO. We first observe that the cross-dataset results are generally higher for the image generation task than for the image retrieval task. This observation can be attributed to the fact that many of the DrawBench queries (around 50%) have no relevant results in the MS COCO database (as per the collected ground-truth annotations), which places the respective queries in the “very difficult” zone. This exacerbates

the distribution gap between MS COCO and DrawBench in the retrieval setting. Therefore, it is very challenging for predictors to generalize across datasets. Comparing the two scenarios, MS COCO→DrawBench vs. DrawBench→MS COCO, in the image generation context, we find that training on MS COCO leads to better results. This can be attributed to the fact that the number of prompts from MS COCO (10K) is much higher than the number of prompts from DrawBench (200), even after applying our filtering based on k-means to select captions from MS COCO. Nevertheless, both cross-dataset settings are difficult, opening a new avenue for future research: proposing prompt/query performance predictors able to generalize across different data distributions.

**Cross-task results.** Although the correlations between the ground-truth scores for image generation and image retrieval are moderate (see Table 2), we also aim to assess how well predictors perform across tasks. To this end, we present cross-task results for two model pairs, namely (GLIDE, CLIP) and (SDXL, BLIP-2), in Table 8. As expected, the correlation coefficients are typically low, indicating that predictors are not able to generalize across tasks. However, this apparent inability of the predictors should be attributed to the low correlations between the image gener-

Predictor Type	Predictor Name	Generative Task				Retrieval Task							
		GLIDE		SDXL		CLIP				BLIP-2			
		HPSv2		HPSv2		CLIP-P@10		CLIP-RR		CLIP-P@10		CLIP-RR	
		Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
Pre-	Fine-tuned BERT	0.806 <sup>‡</sup>	0.608 <sup>‡</sup>	0.696 <sup>‡</sup>	0.505 <sup>‡</sup>	0.437 <sup>‡</sup>	0.255 <sup>‡</sup>	0.207 <sup>‡</sup>	0.167 <sup>‡</sup>	0.495 <sup>‡</sup>	0.329 <sup>‡</sup>	0.144 <sup>‡</sup>	0.110 <sup>‡</sup>
Post-	Fine-tuned CLIP	0.257 <sup>‡</sup>	0.169 <sup>‡</sup>	0.729 <sup>‡</sup>	0.530 <sup>‡</sup>	0.463 <sup>‡</sup>	0.305 <sup>‡</sup>	0.160 <sup>‡</sup>	0.122 <sup>‡</sup>	0.484 <sup>‡</sup>	0.358 <sup>‡</sup>	0.159 <sup>‡</sup>	0.144 <sup>‡</sup>

Table 10. Results of performance predictors for automatic relevance judgments. On the generative task, we report the correlation of the predicted value with the HPSv2 performance of SDXL and GLIDE, respectively. On the retrieval task, the correlation is computed for the CLIP-based P@10 and CLIP-based RR scores of CLIP and BLIP-2, respectively. According to a Student’s t-test, the results marked with † and ‡ are significantly better than the random chance baseline at p-values 0.01 and 0.001, respectively.

ation and retrieval tasks reported in Table 2, which clearly indicate that the two tasks are not very well aligned.

**MS COCO vs. DrawBench.** To assess the disparity between MS COCO and DrawBench, we train and test the fine-tuned BERT and fine-tuned CLIP predictors on the individual subsets (see Table 9). On the generative task, predictors obtain comparable results across the two datasets. Since DrawBench is specifically designed for text-to-image generation, its queries are too difficult for the retrieval setup, so predictors fail in this case. In contrast, MS COCO queries have about the same difficulty (on average) in generation and retrieval. This supports our decision to include more captions from MS COCO than DrawBench into PQPP.

**Results for automatic metrics.** We conduct additional experiments with automatic evaluation metrics instead of the proposed metrics based on human relevance judgments. More specifically, we rely on HPSv2 [51] for generated images and CLIP for retrieved images. We report the corresponding results in Table 10. Predictors seem to have higher correlation with HPSv2 than with human labels (in image generation), indicating that automatic labels are easier to predict.

### 13. More Qualitative Results

In Figure 8, we present a t-SNE visualization of the test queries embedded in the latent space of the BERT predictor fine-tuned on image retrieval with BLIP-2. We observe that the learned latent space correlates well with the ground-truth P@10 values, explaining the high accuracy of the fine-tuned BERT predictor on the retrieval task. The separation between easy and difficult queries is evident in the retrieval setting, which is consistent with the quantitative results reported in Table 3, where the fine-tuned BERT exhibits generally higher Pearson and Kendall  $\tau$  correlation coefficients than other predictors.

We showcase examples of easy and difficult prompts/queries for the generation and retrieval tasks in Figure 9. The generative models exhibit a clear proficiency with prompts referring to inanimate objects,

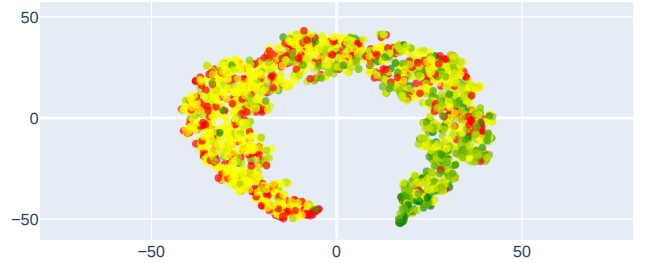


Figure 8. t-SNE visualization of the test queries embedded in the latent space of the fine-tuned BERT on image retrieval with BLIP-2. The ground-truth P@10 performance is encoded via a color map from green (high) to red (low). The visualization confirms that the fine-tuned BERT predictor learns a meaningful representation of the queries. Best viewed in color.

generating images with high relevance. However, their capability falls short when faced with more intricate prompts involving complex actions based on human-object interactions, leading to inaccuracies in object composition. Such cases exhibit artifacts, such as duplicate or missing body parts and misplaced objects, showcasing the lack of deep understanding in both generative models. The retrieval systems are capable of fetching images for prompts centered around single, loosely-defined objects. However, they struggle when the prompts require images containing multiple, specific elements, retrieving results that only partially match the query. This limitation highlights a gap in the ability of retrieval systems to interpret and respond to the multifaceted nature of some queries. This exploration into both generative and retrieval tasks underscores the nuanced challenges faced by systems in accurately capturing and responding to the inherent complexity of certain prompts. It also reinforces the importance of the prompt/query performance prediction task, setting realistic expectations for the outcomes of both generative and retrieval models, based on the detailed content and structure of the prompts/queries.

	Generative		Retrieval	
Prompts	A woman in white sports gear hitting a tennis ball	This is a yellow and blue double decker bus	Pillows on a bench with a side table and decorative mirrors	A tennis player serves the ball on a clay court
Stable Diffusion XL				
				
GLIDE				
				
HBPP	Predicted: 0.421 Ground-truth: 0.250	Predicted: 1.717 Ground-truth: 2.000	Predicted: 0.000 Ground-truth: 0.100	Predicted: 0.934 Ground-truth: 1.000
				P@10

Figure 9. Examples of generative (first two columns) and retrieval (last two columns) results for difficult (first and third columns) and easy (second and fourth columns) queries. For the retrieval systems, we show only the top two results. The prompt/query performance values, namely HBPP and P@10, are predicted by the correlation-based CNN. Best viewed in color.

## 14. Limitations

We recognize specific limitations in the annotation processes for both generative and retrieval tasks. In the generative setting, prompts can be subjectively interpreted by users, which may introduce variability in the results. To mitigate this, we incorporated a control set, excluded annotations from annotators who failed on the control set, and computed an average score from multiple user annotations for each image/prompt pair, enhancing the robustness of the evaluations.

In the retrieval setting, the initial ground-truth image bank was generated using a method that combines Sentence-BERT and the bag-of-words model. We acknowledge that some images may contain content that is not fully or accurately captured by their paired captions, which could introduce occasional false negatives or false positives in the ground-truth collection. To minimize the number of false negatives, we set a relatively low similarity threshold for the inclusion of candidate images. Then, the false positives were curated by the enrolled annotators.

## 15. Potential Negative Societal Impact

The development and deployment of text-to-image generation and retrieval systems come with several societal im-

plications that warrant careful consideration. Here, we outline key areas of concern. Our enhanced dataset is built upon pre-existing datasets, which may inherit and perpetuate biases present in the original data. Additionally, user annotations might have been influenced by their own cultural backgrounds, potentially introducing subjective biases into the final decisions. The pre-trained generative models employed in our study could also exhibit inherent biases, affecting the generated outputs. These combined factors could lead to unfair representations or reinforce existing stereotypes. We acknowledge the necessity of ongoing efforts to identify, measure, and mitigate these biases to ensure the fairness and inclusiveness of our models. In our work, we address these concerns by promoting transparency in our methodology and unifying multiple user annotations to mitigate possible individual biases.

The computational resources required for data collection, filtering, and model training contribute to energy consumption and carbon emissions. We recognize the environmental impact of our work and emphasize the importance of optimizing computational processes and exploring sustainable practices to reduce the ecological footprint of AI research.

Moreover, our benchmark can be used to develop and improve generative models. Such models can further be used in unethical scenarios, *e.g.* to generate deep fakes. In recent years, an increase in deep fake materials flooded the web, either to spread false information or to steal sensitive information by impersonating trustworthy individuals. While we strongly believe in the benefits of very capable generative models, we are aware of the potential risks. However, we can see that governments are working very closely with academia and industry on safely developing artificial intelligence, and thus observe and support the increasing focus on models that detect AI-generated content to mitigate the aforementioned risks.

## 16. Ethical Considerations on Data Annotation

Data annotation by students is a common practice in our host institutions and we followed the standard protocols to get approvals from the corresponding ethics committees. The enrolled students were compensated with bonus points. We would like to emphasize that the students understood that the annotation task is optional, and they could also get the extra bonus points by performing alternative tasks (which did not involve data annotation). Moreover, all students were given the opportunity to obtain a full grade without the optional annotation task. Hence, there was no obligation for any of the students to perform the annotations. The students were also able to opt out, at any time during the annotation, without any penalties.



## 17. Computational Resources

We have employed two types of machines to perform our experiments:

- **Local Hardware:**
  - **GPU:** NVIDIA RTX 3090 with 24GB VRAM
  - **CPU:** Intel i9-10920X @ 3.50GHz
  - **Memory:** 64GB RAM
  - **Storage:** 1TB SSD, 5TB HDD
- **Cloud VM:**
  - **GPU:** NVIDIA A100 with 40GB VRAM
  - **CPU:** 12 vCPUs
  - **Memory:** 85GB RAM
  - **Storage:** 100GB HDD

Our annotation platform was hosted using Google Cloud Provider, with authentication developed with Google Firebase Authentication, and image hosting facilitated by Google Cloud Storage. By detailing the utilized computational resources, we aim to provide transparency and reproducibility for our research.

## 18. Computational Time Estimation

We present the following estimation of the compute time (in hours) required to fully replicate the experiments detailed in this paper:

- **Pre-processing of the MS COCO dataset:** The extraction of Sentence-BERT embeddings and the subsequent application of the k-means clustering algorithm across the entire corpus of MS COCO captions require approximately 48 hours.
- **Generative processes:** The generative processes employing both the SDXL and GLIDE methods demand a total time of approximately 120 hours.
- **Preliminary relevance judgments:** The creation of initial relevance judgments for the retrieval task takes 72 hours.
- **Model fine-tuning:** The cumulative time spent on fine-tuning all predictors involved in our study amounts to 50 compute hours.

These estimates are based on the computational resources and configurations described in Section 17.

## 19. Dataset Documentation

### 19.1. Documentation Framework

The dataset is documented using the Data Card framework, which provides a comprehensive overview of its content, collection methods, and intended uses. The structure is as follows:

- **Dataset Overview:** General information about the dataset, including size, number of instances, and collected human labels.

- **Content Description:** Detailed description of the data points, including relevant features and formats.
- **Typical Data Point:** Example of a typical data entry.
- **Dataset Structure:** Explanation of the dataset’s organization, including file and folder descriptions.
- **Provenance:** Information on data collection methods and maintenance status.
- **Licensing:** Details about the dataset’s license and usage terms.

### 19.2. Dataset Overview

The dataset does not contain sensitive data about people and includes original images from the MS COCO dataset. The dataset snapshot is as follows:

- **Size:** 34 GB
- **Query/Prompt Instances:** 10,200
- **Generated Image Instances:** 40,800
- **Human Labels:** 1,589,055

### 19.3. Dataset Format and Preservation

The dataset utilizes widely recognized open data formats. Annotations are provided in CSV format, while images are in standard image formats (PNG). Detailed instructions on reading and using the dataset are provided in the repository.

### 19.4. Structured Metadata

To enhance the discoverability and organization of our dataset, structured metadata is included using Web standards (schema.org). This metadata is encapsulated in a dataset.json file within our repository.

### 19.5. Content Description

Each data point includes the following features:

- **id:** Number, ID of the query in MS COCO / DrawBench.
- **image\_id:** Number, ID of the image in MS COCO.
- **best\_caption:** String, text containing selected prompt.
- **blip2\_rr:** Float, reciprocal rank for query using BLIP-2 retrieval method.
- **clip\_rr:** Float, reciprocal rank for query using CLIP retrieval method.
- **blip2\_pk:** Float, precision@10 for the query using BLIP-2 retrieval method.
- **clip\_pk:** Float, precision@10 for the query using CLIP retrieval method.
- **glide\_score:** Human annotated generative score for the GLIDE model.
- **sdxl\_score:** Human annotated generative score for the SDXL model.

### 19.6. Typical Data Point

A typical data point is shown in Table 11.

Column Name	Value
id	319365
image_id	363951
best_caption	Black and white of windsurfers on a lake.
blip2_rr	1.0
clip_rr	1.0
blip2_pk	0.1
clip_pk	0.1
glide_score	0.5
sdxl_score	2.0

Table 11. Example of a typical data point.

## 19.7. Dataset Structure

The dataset folder structure can be viewed in the official repository:

- **Dataset Files:** CSV files for training, validation, and test splits containing MS COCO image IDs, P@10/RR scores for retrieval, and HBPP scores for the generative setting.
- **Image Folder:** Contains the SDXL/GLIDE generated images alongside the original MS COCO images.

The folder structure is:

- **Dataset Files:**

```
\dataset
  \ train.csv
  \ validation.csv
  \ test.csv
```

- **Image Folder:**

```
\images
  \{ IMG_ID}
    \image_4.png
    \image_5.png
    \image_6.png
    \image_7.png
    \image_8.png
```

The structure of the additional resources is explained in extenso in the official repository.

## 20. Maintenance and Support

### 20.1. Maintenance

Although there is no plan to make new versions available in the future, this dataset will be actively maintained by the authors, including but not limited to updates to the data.

### 20.2. Support

We commit to maintaining the dataset and providing support through the following channels:

- Official Github repository ticketing system.
- Direct contact via email at: [eduardgabriel.poe@gmail.com](mailto:eduardgabriel.poe@gmail.com).

## 21. Licensing and Responsibility Statement

We release our dataset, which includes annotations alongside images created with generative models, under the CC BY 4.0 license. We also acknowledge the license offered by the original authors of the MS COCO dataset annotations (CC BY 4.0) and the Flickr Terms of Use for the images, as detailed at <https://cocodataset.org/#termsofuse> and <https://www.flickr.com/creativecommons/>.

In the event that it is determined that we have violated any rights or licenses associated with the used resources, we take full responsibility and guarantee our cooperation in resolving any such issues with any affected third parties. Potential resolutions will include, as appropriate, the modification, substitution, or deletion of data or code that infringe on copyrights or licenses.

## 22. Intended Uses

This dataset is intended for use in either commercial or research and development within the domains of machine learning, computer vision, query performance prediction, and prompt performance prediction. It is designed to facilitate the training, validation, and testing of models for these applications.