# Generation Through Search with Image Aesthetic Assessment


Aditya Akula, Justin Blalock, Alex Hobby, David Munechika, Rithvik Rajavelu
Georgia Institute of Technology, Atlanta, GA
{akula, jblalock30, alex.hobby, dmunechika3, rrajavelu3}@gatech.edu


## Summary


As **text-to-image generative models** have exploded in popularity, search engines for generated images have emerged allowing users to find images without needing to personally generate them. However, these **"generation through search"** systems do not incorporate the aesthetics of images, making high-quality exploration difficult. We propose a system for computing **image aesthetic scores** on generated images and provide a **scalable, interactive interface** for users to search for **relevant, high-quality images.**


## Generation Through Search

Using diffusion models to generate images from prompts can be a **computationally expensive** process.

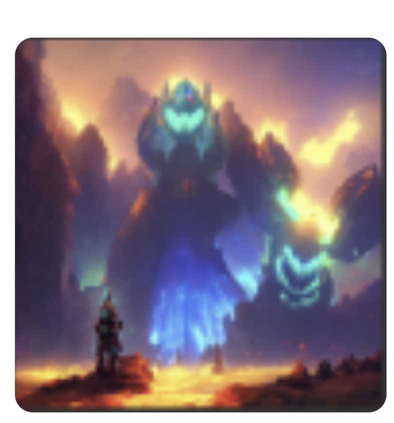
The problem of "generation through search" aims to allow users to find **relevant, pre-generated images** that fit their needs so they do not need to perform the generation themselves.

## Image Aesthetic Assessment

Our novel approach extends generation through search to incorporate information about **image aesthetics.**

We compute **image aesthetic assessment (IAA)** scores for each image using an ML model called **MUSIQ**.

IAA = 10.2

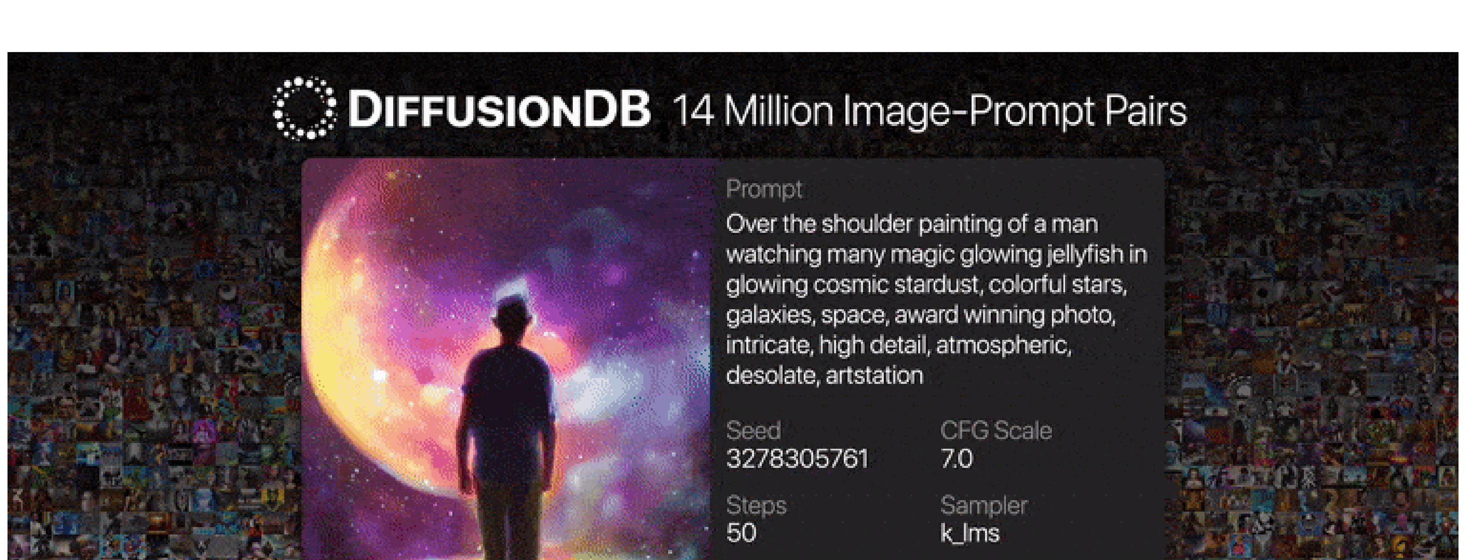


IAA = 57.6

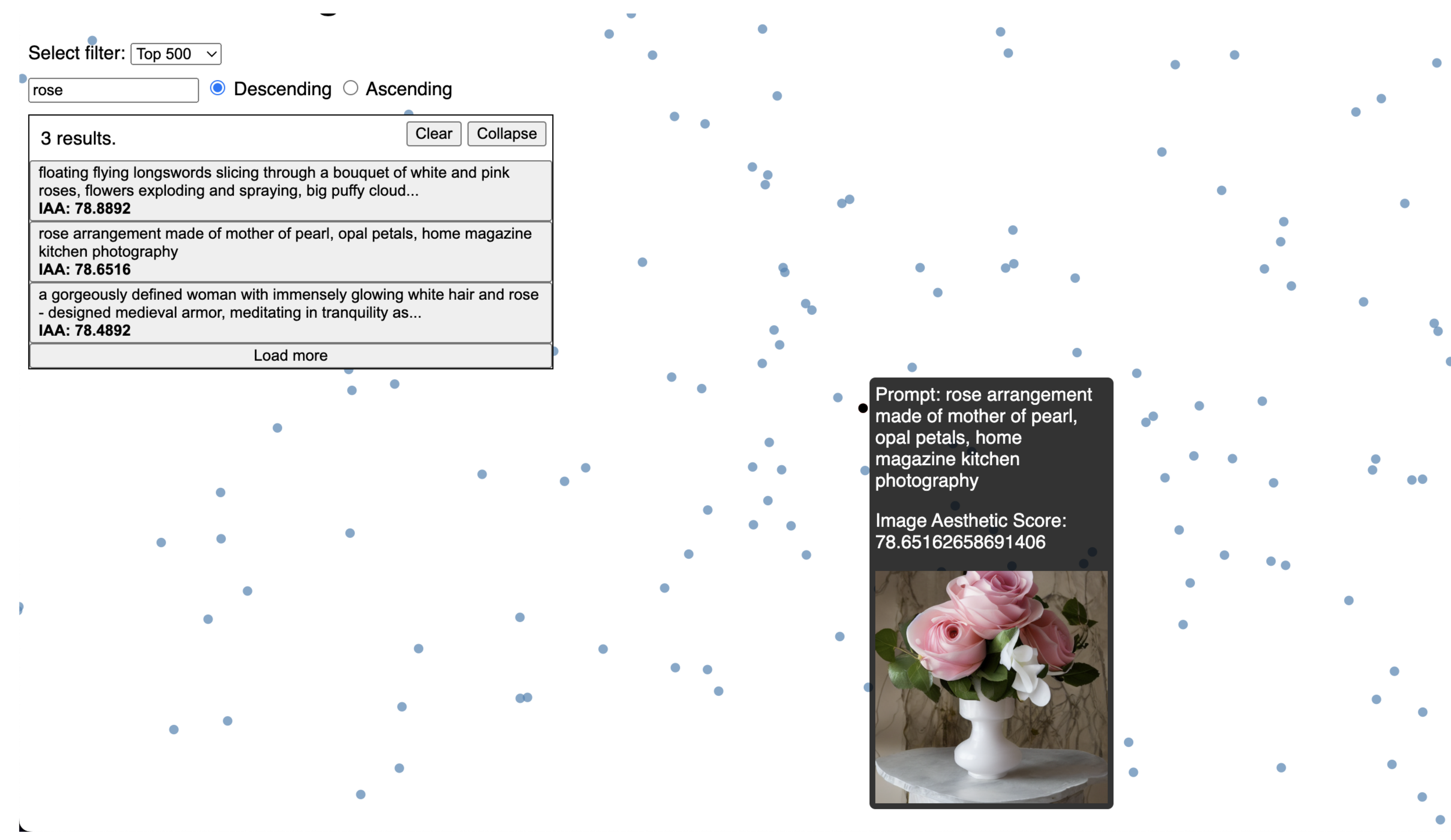
## DiffusionDB Dataset

DiffusionDB is the first **large-scale prompt gallery dataset**. We use a subset of 100,000 prompt-image pairs for our system. We wrote scripts to scrape the dataset and we host the 56 GB of images in our Cloud Storage bucket.

For each sample, we compute the **IAA score** and **CLIP image embedding** which allows us to cluster similar images together in our visualization.



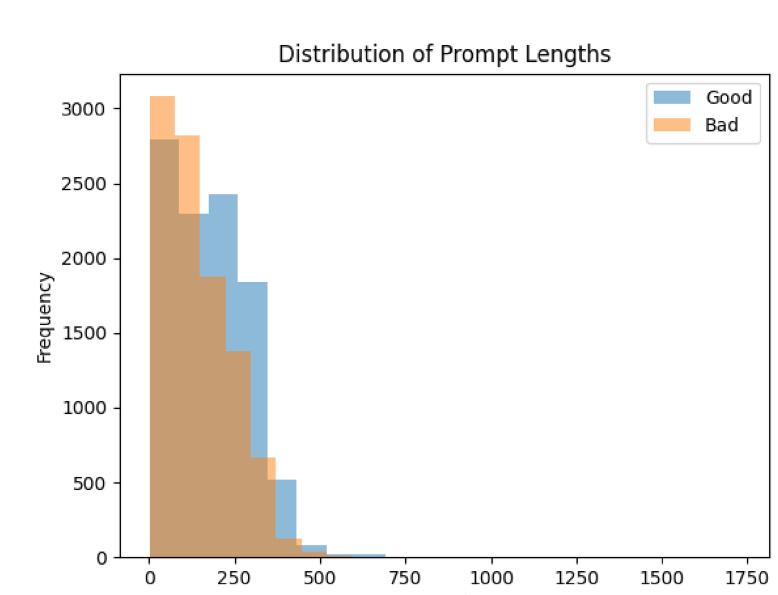

## Interactive Visualization




Our interactive interface clusters similar images together using **CLIP image embeddings** and **UMAP dimensionality reduction**. It supports **searching** for images by prompt, but differs from other methods by also **filtering** by IAA score.

## Experimental Results

### Prompt Length Distribution

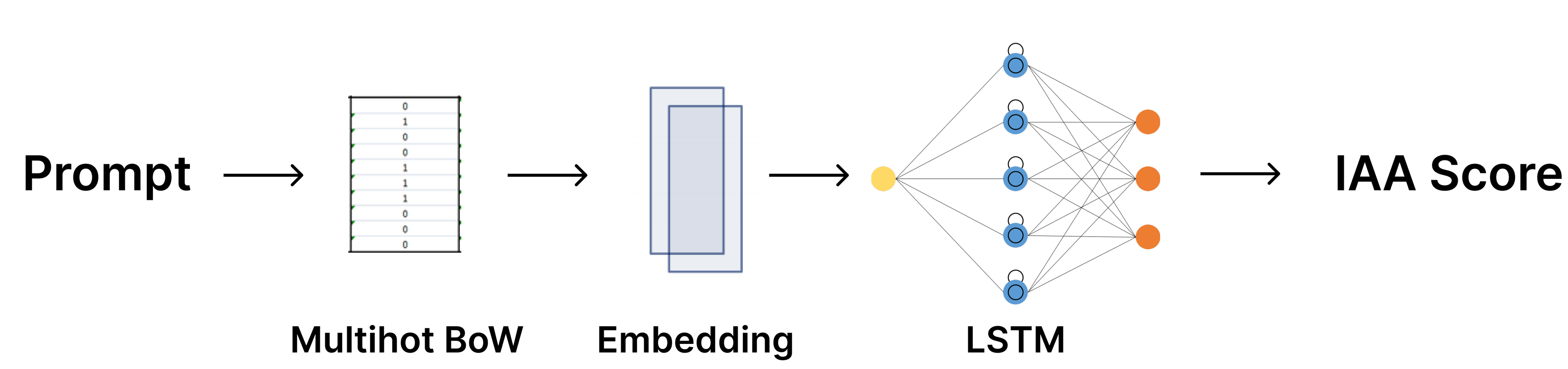



### Best vs. Worst Prompt TF-IDF Scores

| Bottom 10% | Top 10% |
|---|---|
| black: 0.012 | render: 0.017 |
| lighting: 0.015 | focus: 0.018 |
| light: 0.015 | sharp: 0.018 |
| portrait: 0.015 | intricate: 0.019 |
| still: 0.015 | concept: 0.019 |
| artstation: 0.016 | illustration: 0.020 |
| movie: 0.017 | highly: 0.022 |
| detailed: 0.018 | digital: 0.022 |
| film: 0.019 | with: 0.023 |
| from: 0.019 | painting: 0.023 |
| with: 0.019 | the: 0.024 |
| cinematic: 0.022 | on: 0.024 |
| on: 0.023 | artstation: 0.026 |

We analyzed the best and worst prompts (by IAA score) to understand the differences:

- Good prompts are on average **20% longer** than bad prompts.
- The **TF-IDF** scores show certain **keywords** are more important for good prompts vs. bad prompts.




We train a **neural network** to predict IAA score (prompt strength) given an input prompt, resulting in a **mean absolute error of 7.5**.

## User Study

For the user study, we recorded user responses to survey questions based on a Likert scale with 1 being strongly disagree and 5 being strongly agree. Below are the average scores: