Prompt Expansion for Adaptive Text-to-Image Generation

Siddhartha Datta[1,2 †]**, Alexander Ku**[1,3]**, Deepak Ramachandran**[1]**, Peter Anderson**[1 ‡]

[1]Google DeepMind, [2]University of Oxford, [3]Princeton University

Figure 1: Prompt Expansion is an alternative paradigm for interaction between users and text-to-image models. **Top**: Outputs from Straight-Query Generation may result in images that are less visually compelling and diverse. **Bottom**: Prompt Expansion samples uncommitted aspects of the image in text space, improving visual quality and diversity while enabling interaction modes beyond prompt engineering / iteration.

Abstract

Text-to-image generation models are powerful but difficult to use. Users craft specific prompts to get better images, though the images can be repetitive. This paper proposes the Prompt Expansion framework that helps users generate high-quality, diverse images with less effort. The Prompt Expansion model takes a text query as input and outputs a set of expanded text prompts that are optimized such that when passed to a text-to-image model, they generate a wider variety of appealing images. We conduct a human evaluation study that shows that images generated through Prompt Expansion are more aesthetically pleasing and diverse than those generated by baseline methods. Overall, this paper presents a novel and effective approach to improving the text-to-image generation experience.

† Work done while at Google. Correspondence with author at siddhartha.datta@cs.ox.ac.uk.
‡ Now at Balyasny Asset Management.
‡ https://github.com/google-deepmind/t2i-prompt-expansion

1 Introduction

Text-to-image generation models (Ramesh et al., 2022; Saharia et al., 2022a; Yu et al., 2022b) are capable of rendering a stunning variety of high-quality images, from highly-realistic professional-looking photos to fanciful dreamworlds in almost any visual style. However, interaction with these models frequently exposes two significant usability problems. First, achieving high-quality outputs often requires users to include arcane lighting and camera jargon ('35mm', 'DSLR', 'backlit' 'hyper detailed'), idiosyncratic descriptors ('audacious and whimsical'), and social media tags ('trending on artstation'). The prompts that produce the best images are not necessarily stable across different models, or even across different versions of the same model, leading to a focus on prompt-sharing and 'prompt engineering', which is the process of iterating over prompts to craft the optimal textual input for a desired image.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3449–3476
August 11-16, 2024

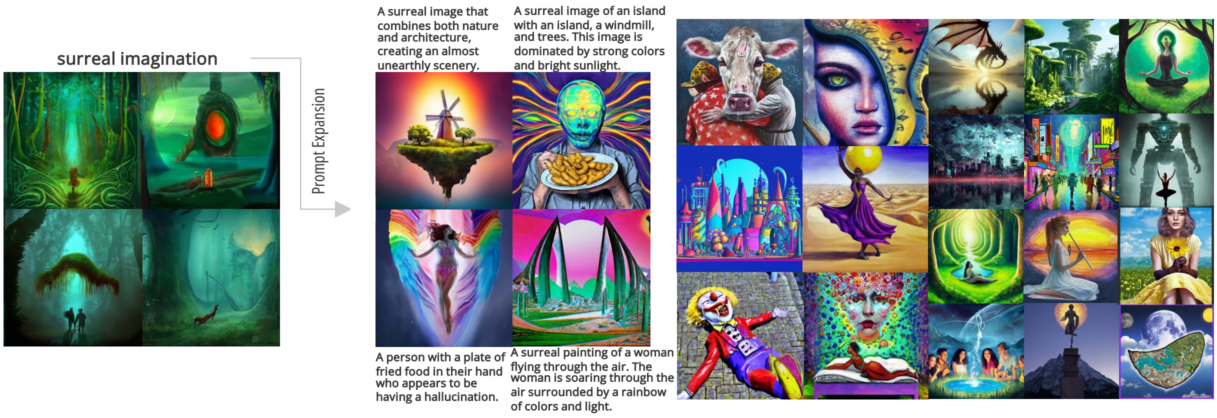


Figure 2: **Abstract queries**: While Straight-Query Generation returns a set of similar images (Left), Prompt Expansion returns a diverse range of images, varying in topic, subject, location, colours, etc. This range is reflected in the expanded prompts for the first four images generated (Middle), and continues in the next 18 images (Right).

Second, when sampling sets of images from these models, while the outputs are *different*, they are not necessarily *diverse*. For example, randomly sampling four images for the prompt 'jack o lantern designs' from a diffusion model similar to Imagen (Saharia et al., 2022a) produces highly similar outputs (Figure 1 Top). Even though the prompt does not specify the composition, style, viewpoint or lighting of the image, whether it is night or day, or whether there are other features in the scene, samples from the model do not vary in any of these aspects. This lack of diversity can amplify harmful social biases when generating people (e.g. generating all doctors as male) (Naik and Nushi, 2023), and also fails to lead the user towards more interesting imagery. Just as people are non-committal about basic aspects of their mental images (Bigelow et al., 2023), text-to-image models should also exhibit non-commitment by generating diverse outputs for any image characteristics that are unspecified (Hutchinson et al., 2022).

We reduce the burden of prompt engineering and improve diversity in text-to-image generation by proposing a *Prompt Expansion* (PE) framework. A Prompt Expansion model takes a text prompt as input, which we refer to as a query, and outputs a set of N expanded text prompts that include specialized keywords (to improve image quality) and interesting additional details (to add diversity to the generated images). To do this, we construct a Prompt Expansion dataset by inverting a dataset of high aesthetic images to text, and few-shot prompting the inverted text to queries. We then train a PALM 2 (Anil et al., 2023) text-to-text model on the query:prompt pairs, and iteratively re-fine-tune based on the generated output. As illustrated in Figure 1 Bottom, incorporating Prompt Expansion into text-to-image generation produces a greater variety of appealing, high-quality images and also opens up additional modes of user interaction beyond iterating over prompts. Prompt Expansion achieves improved diversity, aesthetics, and text-image alignment compared to Straight-Query Generation (i.e., sampling multiple images for the same query) in automatic and human rater evaluation.

Contributions. Motivated by the frustrations of prompt-sharing and prompt-engineering, we propose a new framework, Prompt Expansion, to improve image quality and diversity while opening up new modes of interaction. We construct a Prompt Expansion dataset by reverse engineering prompts and high-level queries from aesthetic images, and train a variety of Prompt Expansion models. We undertake a detailed human evaluation demonstrating that Prompt Expansion can improve diversity, aesthetics, and text-image alignment.

2 Related Work

To better control and introduce diversity into image output, most prior work focuses on techniques to allow end-users to edit the generated images, such as text-guided image editing (Hertz et al., 2022), or to fine-tune to create personalized images (Collell and Moens, 2016; Ruiz et al., 2022). These methods focus on local edits and iterating with respect to a given image. Our goal is to introduce diversity in the generated images set. The standard approaches to introducing variation are increasing guidance (Ho and Salimans, 2021) in diffusion models (e.g. Imagen (Saharia et al., 2022a)), or

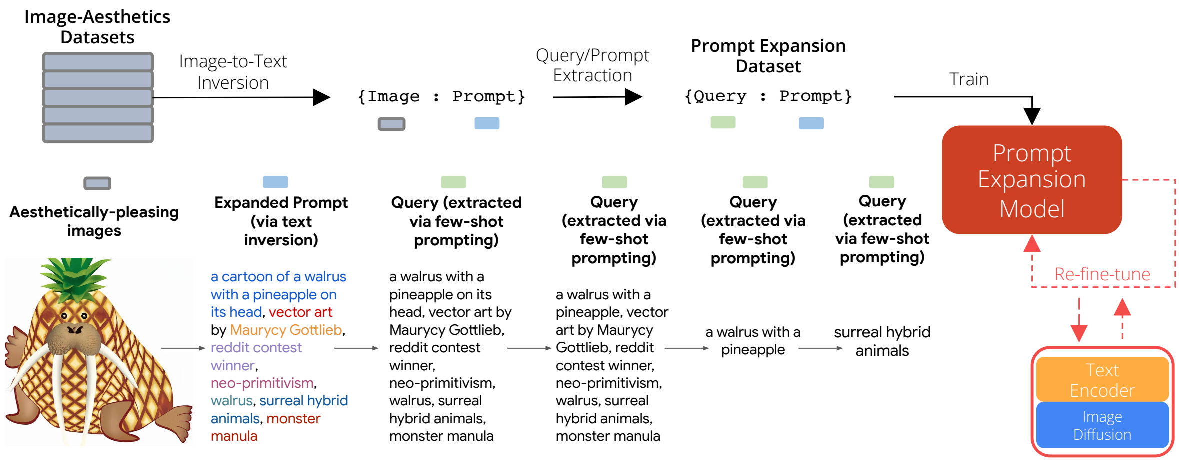


Figure 3: *(Outer)* Overview of Prompt Expansion dataset construction and model training. Beginning with the Image-Aesthetic datasets (Sec 3.1), we generate expanded prompts through Image-to-Text Inversion (Sec 3.2), then we extract queries for each prompt through Query/Prompt Extraction (Sec 3.3). The resulting Prompt Expansion dataset is used to train a base `Prompt Expansion` model. We then align with our downstream text-to-image model to create the `PE: Re-fine-tuned` model (Sec 4.2). *(Inner)* An example of the Image-Text inversion (using COCA-Interrogator) generating an expanded prompt containing caption, art form, artist, medium, style, and other flavours / objects / descriptors. Subsequently, the prompt is mapped by successive few-shot prompting inference calls into shorter and more abstract queries that are each paired with the prompt to construct the dataset.

increasing temperature in autoregressive models (e.g. Parti (Yu et al., 2022b)). Changing the sampler/decoder hyperparameters, however, may not return meaningfully-diverse images. Setting the temperature/guidance too high will insert excessive noise and yield images of low quality. The noise inserted is random, thus it may not return variations along targeted dimensions (e.g. style, lighting, attributes of people / objects / places). To interpretably sample images, we shift our attention to sampling within text space.

Prompt Optimization techniques such as Promptist (Hao et al., 2022) or Promptify (Brade et al., 2023) investigate automatic and human-in-the-loop approaches to improving the aesthetic quality of a prompt's generated image. The task of Prompt Optimization presumes that an ideal prompt may exist for a given query that returns the "best" image. Instead, the task of Prompt Expansion acknowledges that the ideal generated image depends on more than the query alone, and that the query may have different intents or context (e.g. varying by user preferences). It addresses the uncommitted aspects of the image (e.g. query ambiguity) and that images are evaluated in sets. For example, compared to Straight-Query Generation, Prompt Optimization (e.g. Promptist) can return a prompt that maximizes aesthetics, but it would still return a similar set of black-and-white dancers (Figure 11).

3 Prompt Expansion Dataset

The *Prompt Expansion* (PE) framework requires a model to take a user text *query* as input and return N text *prompts* as output, such that the N text prompts through text-to-image generation will return a set of N diverse, aesthetic images aligned to the query. To train a PE model, we require a dataset mapping queries to prompts, which we construct in reverse. First, we collect aesthetically high-quality images, including both model-generated and natural images (refer Sec 3.1). Second, we invert the images to a closely corresponding prompt that includes alt-text jargon (which we refer to as *flavors*, refer Sec 3.2). Finally, we map the inverted text to a range of high-level queries that more closely correspond to user input (refer Sec 3.3). These queries are paired with the prompts from the second step to form the `{query:prompt}` dataset.

3.1 Image-Aesthetic Datasets

We curate two image datasets. The first, Webli-Align, is composed of images from the Webli (Chen et al., 2023) and Align (Jia et al., 2021) datasets, filtered to retain only images with high MUSIQ (Ke et al., 2021) aesthetic scores. The second, CrowdSourced, is obtained by crowd-sourcing output from a text-to-image model. We provide an interface for text-to-image generation (Figure 15), allowing users from a large organization to enter prompts that generate images. Users also have the option to upvote images that they like. We use this signal to retain only the most appealing images. We retain 80k Webli-Align (natural) and 40k CrowdSourced (generated) images.

3.2 Image-to-Text Inversion

The second step is to invert the images in the Image-Aesthetic datasets to prompt text. While the user *query* is the input a user provides, the *prompt* is the text that generates a specific image. We use the *Interrogator* (CLIP-Interrogator) approach to image-to-text inversion. The computed prompt text is generated by concatenating: (i) a caption, and (ii) a set of 'flavors'. The caption is a description of the content of the image (e.g. who, what, where, when). To generate the caption, we use COCA (Yu et al., 2022a) fine-tuned for the captioning task. A *flavor* refers to a descriptive word/phrase that alters the style of the image, without intending to add/change the content of the image, like "impressionism" or "dslr". We generate the lists of flavors from words and phrases used in a large number of collected prompts of generated images (details in Section J).

3.3 Query/Prompt Extraction

The final step in dataset preparation is to compute a range of potential user queries that are suitable to map to the inverted text (prompt). We use few-shot prompting with FLAN-PaLMChilla 62B (Chung et al., 2022) to generate successively shorter queries and longer prompts. The model receives few-shot prompting sets of long prompts mapped to short queries as examples. The few-shot prompts are in the format `{prompt:query}`, and examples of these pairs can be seen in Figure 3 and Table 1. For each prompt from Image-to-Text inversion, the few-shot prompt examples are prepended before the prompt as context, and a corresponding query is generated by the text-to-text model.

We extract a range of different queries that can be mapped to the expanded prompt, and use few-shot prompting to generate queries that are abstract, concrete, short-length, medium-length, or long-length. Appendix A has further details on how the query types are generated (e.g. grounded queries, eliciting specificity). This results in a Prompt Expansion Dataset of 600k `{query:prompt}` pairs. We perform a 70-20-10 train-val-test split, and split the train set 50-50 for base and re-fine-tuning.

4 Prompt Expansion Model

We describe the two stages to train the Prompt Expansion model: (i) we train a base Prompt Expansion model on the Prompt Expansion Dataset; then (ii) we re-fine-tune the base model with respect to the downstream text-to-image model.

4.1 Base Prompt Expansion Model

Our Prompt Expansion model is a text-to-text generation model trained to map query text to expanded prompt text with an architecture based on the PaLM 2 language model family (Anil et al., 2023). PaLM 2 is a decoder-only transformer-based architecture trained with the UL2 objective (Tay et al., 2023). We train a PaLM 2 1B parameter model with prompt-tuning (Lester et al., 2021), after evaluating different model configurations, as described in Table 6. We chose this relatively small size for the base architecture, as it needs to serve as a front-end to a complex high-latency text-to-image model such as Imagen (Saharia et al., 2022a), and thus needs to be low-resource/latency to make the entire pipeline usable. For the base dataset, we use a 50% split of the Prompt Expansion Dataset described in Section 3, consisting of 300k `{query:prompt}` examples.

4.2 Re-fine-tuning for Aligned Prompt Expansion

After training the base model, we observe that it may generate prompts that the text-to-image model cannot generate good images for. The main reason for this is that the expanded prompts generated by the base Prompt Expansion model are based on the alignment between text and images favored by the COCA image-to-text inversion model. Therefore, we propose a general *re-fine-tuning* procedure: given a target behavior for the model, re-fine-tuning filters for expanded prompts generated from the base model that align with the target behaviour. To align with the text-to-image model, we construct a dataset where the expanded prompts are aligned closer to the behavior of the downstream text-to-image model.

For queries in the remaining 50% split of the Prompt Expansion dataset, we generate expanded prompts from our base model, which are then input to the downstream text-to-image model (Imagen (Saharia et al., 2022a) in our experiments). We score these images using a weighted average of the query-image embedding distance and prompt-

Table 1: Examples of prefixed {query:prompt} pairs, where prepending the prefix controls the output generated.

Augmentation	Query	Prompt
Abstract	*ABST hope*	*a tunnel with a light at the far end*
Grounded	*GRD colossal dragon*	*a colossal dragon eating humans*
	GRD monster bike	*a monster bike being ridden by a daredevil*
Specificity	*SPCT animal drawings*	*animal drawings, for example a children's drawing of a dog*
	SPCT groups of animals	*groups of animals, specifically a herd of sheep being gathered by a farmer*
Flavor	*FLV a drawing of a dog swimming in a river*	*pointillism*
	FLV a giant robot fighting a dinosaur	*pixel art*
Multi-step expansion	*MSTP a brain on a wall*	*a brain on a wall, poster art by Robert Beatty*
	MSTP a brain on a wall, poster art by Robert Beatty	*a brain on a wall, poster art by Robert Beatty, featured on behance*
	MSTP a brain on a wall, poster art by Robert Beatty, featured on behance	*a brain on a wall, poster art by Robert Beatty, featured on behance, psychedelic art, psychedelic artwork, brains, 8k archival print*
	MSTP a brain on a wall, poster art by Robert Beatty, featured on behance, psychedelic art, psychedelic artwork, brains, 8k archival print	*a picture of a brain on a wall, poster art by Robert Beatty, featured on behance, psychedelic art, psychedelic artwork, brains, 8k archival print*

image embedding distance (See Appendix D for details) and filter out {query:prompt} pairs whose scores are below a fixed threshold. We then continue re-fine-tuning from the base model checkpoint using only these filtered {query:prompt} pairs, thus producing a `PE: Re-fine-tuned` model which is optimized to return expansions of the query and flavors that the text-to-image model can faithfully generate high quality images for.

5 Controllable Generation

5.1 Prefixes for Controlled Prompt Expansion

Till now, we have presented our approach for building a model for generic use cases of Prompt Expansion. However, it is often the case that the user or application designer would like to control the direction of the Prompt Expansion strategy towards e.g. adding more flavors or adding specific kinds of diverse details. To support these use cases, we implement a controllable version of our Prompt Expansion model that can be directed to produce specific kinds of expansions by prepending the query with one of 8 supported *Prefixes*. For example, we can direct the model to produce flavors only using the `FLV` prefix, or to iteratively expand the original query for interactive Multi-Step Prompt Expansion scenarios with the `MSTP` prefix. A few examples of controlled generation are shown in Table 1 and the full list of supported flavors are in Table 4.

To train the `PE: Multi-Prefix` model, we begin with the Prompt Expansion dataset of Section 3. Each {query:prompt} pair is assigned with an appropriate prefix. During Query/Prompt Extraction, human annotators designed {query:prompt} pairs for prefix types (examples in Table 1). The {query:prompt} pairs of each prefix type are then used to few-shot prompt the model to generate {query:prompt} of a certain type (e.g. `ABST`, `DTL`). Some {query:prompt} pairs need to be classified to its prefix, for example `HAST` prefixes are assigned to prompts whose images return good aesthetics. Some prefixes (e.g. `RFT`, `MSTP`) are known as their {query:prompt} pairs are synthesized. Prefix assignment resulted in a new version of the Prompt Expansion dataset with every query prepended with a prefix; and this was used to fine-tune and train the `PE: Multi-Prefix` model.

5.2 Prefix Dropout for Generic Prompt Expansion

With the Multi-Prefix dataset in hand, we explored the possibility of using controllable generation hints to improve the performance on the generic Prompt Expansion task. The idea is to initialize the training of the model using controlled generation, and then gradually shift its behavior over the course of the training to guess an appropriate prefix for a given query and generate the matching expansions e.g. for highly abstract queries such as *"Undying Love"*, the model's behavior should match that of the `ABST` prefix (See Table 1). This is achieved through a novel curriculum learning technique `Prefix Dropout`: we start with the prefix-annotated dataset described above, but over the course of training steadily increase the percentage of examples where the prefix is randomly removed or dropped-out from the query starting from a 0.4 dropout rate to 1.0. This yields the `PE: Prefix Dropout` model which can be compared to our base and re-fine-tuned models as a candidate for generic Prompt Expansion.

5.3 Multi-Step Prompt Expansion

Exploration can be a multi-step process. After the user's query returns a set of expanded prompts, the user can select amongst the prompts, and this prompt is fed back into the Prompt Expansion model. This allows users to iterate on the expanded prompts without the need for manually prompt engineering the text. Using `PE: Re-fine-tuned`, we generate expanded prompts on held-out queries, and iteratively generate prompts upon the previous step's prompts. This results in multi-step training data of expanded prompts to next-step expanded prompts. We re-fine-tune the Prompt Expansion model with this data prepended with prefix `MSTP`.

6 Experiments

We evaluate different flavors of PE described above, focusing mainly on their ability to generate diverse and aesthetically pleasing images, without significant semantic drift in the prompt. We describe here the experimental setup, task design, and metrics.

6.1 Evaluation Setup

Evaluation set. For both automatic and human evaluation experiments, we sample queries from 2 sources: (1) n=200 prompts from the PartiPrompts (PP) (Yu et al., 2022b) dataset of prompts that is designed to represent different domains and features of language (such as counting, negation, etc); (2) n=500 prompts from a novel test set of potential queries constructed by applying the PE dataset generation process of Section 3 (Text-Inversion + Query/Prompt Extraction) to Webli-Align (WA) (Chen et al., 2023; Jia et al., 2021) images. Queries are categorized as abstract, concrete, short (<4 words), medium length (4-7 words), or long length (>7 words). For WA, we obtain prompts of different lengths by selecting prompts from different steps of the prompt shortening process (Sec 3).

Models. We evaluate three variations of Prompt Expansion: (1) A `Few-shot Prompting` baseline, where we few-shot prompt a FLAN-PaLMChilla 62B (Chowdhery et al., 2022) model to generate expanded prompts given examples of `{query:prompt}` pairs. This is similar to the setup in Section 3, except we swap the input and output to generate prompts from queries. (2) The base `Prompt Expansion` model, which is constructed following all the steps in Section 4.1 except for re-fine-tuning. (3) A model trained using the full `Prompt Expansion` pipeline, which we call `PE: Re-fine-tuned`. We compare the results of these to `Straight-Query Generation`, where the text-to-image generation model is given the original query as input. All our prompts are given as input to Imagen (Saharia et al., 2022b) and the resulting images are evaluated using the metrics below.

6.2 Metrics for Automatic Evaluation

We use 3 metrics to evaluate the images generated from the prompts returned by the models described above. For **aesthetics**, we use MUSIQ (pre-trained on the AVA dataset) (Ke et al., 2021). A higher score is better, and is primarily used for evaluating distortions in photorealistic images (cartoon images, pixel art, and paintings tend to receive lower scores). For **diversity**, we use the variance (σ_p) of the COCA (Yu et al., 2022a) image embeddings of the generated images. For **text-image alignment**, we use the cosine similarity between the COCA text embeddings of the query text against the COCA image embeddings of the generated image, similar to CLIP-score (Hessel et al., 2021). For a query q used to generate an expanded prompt p, which is then used to generate its corresponding image $I(p)$, to measure the COCA embedding distance $\texttt{COCA}(\cdot,\cdot)$ we denote the COCA score between query text and expanded prompt image as $\texttt{COCA}(q, I(p))$, and between prompt text and expanded prompt image as $\texttt{COCA}(p, I(p))$.

6.3 Task Design for Human Evaluation

We perform a side-by-side (SxS) evaluation task: we show raters a pair of images side by side for a given text query, where each image is generated by a prompt from one PE model or the other, and ask the rater to pick the best one. We split our query set into two subsets of 350 queries each to be used in the following settings:

1. **Random vs. Random (1x1)**: A random image generated from each prompt is compared SxS.
2. **Best vs. Best (4x4)**: For each prompt, we generate 4 images using the text-to-image model, ask raters to select the best one for each prompt in separate tasks, and then ask a fresh rater to compare the selected images SxS.

Each SxS evaluation is conducted for 2 criteria: aesthetic preference and text-image alignment (details of the task design in Appendix H). To evaluate aesthetics, the rater is asked *"Which image do you personally prefer?"* They are instructed to select images based on personal preference (e.g. aesthetically, stylistically, compositionally). To evaluate text-image alignment, the rater is asked *"Which image is more consistent with the text description?"* To avoid image rejection unrelated to the metrics, raters are also instructed to ignore other aspects (e.g. resolution/quality, poorly-rendered text, out-of-place objects).

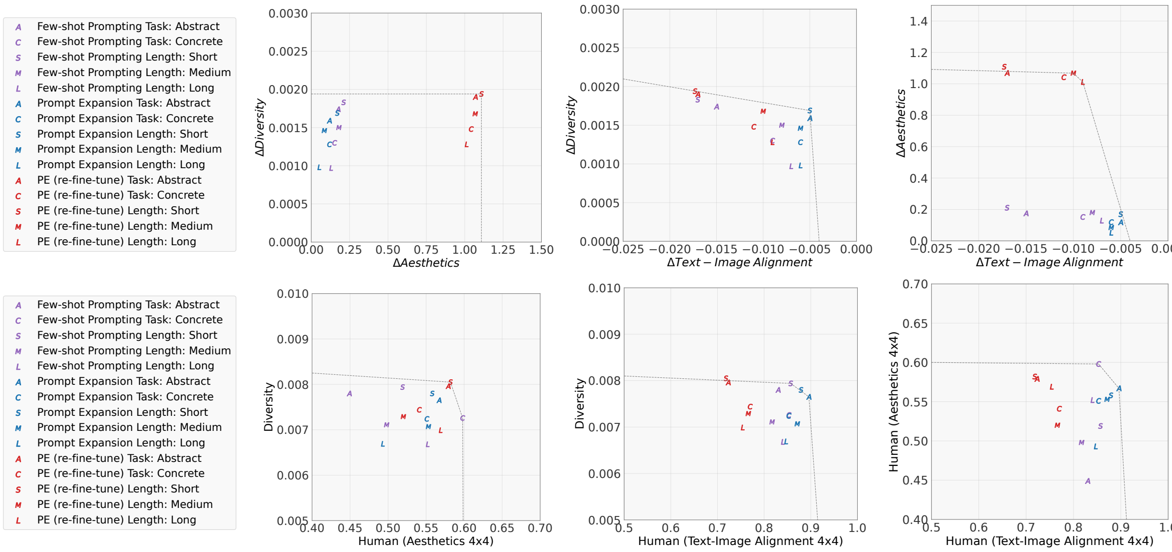


Figure 4: We evaluate each Prompt Expansion method relative to Straight-Query Generation. For automatic evaluation metrics, we plot the difference $\Delta = PE - SQG$. For human rater evaluation metrics, the rater study already compares relative to Straight-Query Generation. For Human (Text-Image Alignment) values, we plot Prompt+Equivalent scores. For completeness, we provide all numbers in Table 3 and Figures 5-6.

If two images contain all the attributes and details specified by the query, there are no further distinctions from which a rater may rate one image as being more aligned to the query than the other. Thus, for text-image alignment we allow the rater to judge the SxS rating task as "Equivalent" when they find no significant difference between the images, but for Aesthetics they are required to select one of the images as better. We analyze the results using Prompt-win rates vs Query-win rates (the percentage of time each rater chose the prompt-generated image over the query-generated image), along with reporting Equivalent-rate for text-image alignment. Between the 3 raters, the inter-rater agreement is on average 46.1% for 3/3, and 52.0% for 2/3 (Table 9).

7 Results and Discussion

From the results of the automatic (Table 2) and human evaluation (Figures 4, 5, and 6), we derive some insights relevant to the motivations and research questions raised in Section 1.

—**Does Prompt Expansion increase diversity of the generated images?** Table 2 shows a small but consistent increase in diversity scores between our various Prompt Expansion models and the Straight-Query baseline. Though small compared to the confidence interval, the gains are similar across all systems and hold consistently across query types (see Table 3), increasing our confidence in the positive impact of PE on diversity. We observe `PE: Re-fine-tuned` has marginally higher diversity than other PE models, both in aggregate (Table 2) and per query type (Table 3).

We also found it possible to further increase overall diversity by using techniques such as temperature sampling and post-hoc filtering (See Table 7 for results). Qualitatively, this increase in diversity manifests in topical, semantic, and expressive diversity, and in fairness considerations such as age, gender, and culture (See the examples in Table 10).

We also observe that the diversity is consistent over multiple generations. With Multi-Step Prompt Expansion, Figure 7 evaluates the N generated images at expansion step t, and shows consistent diversity over the number of expansion steps. Figure 8 evaluates consistency over the number of expanded prompts generated for a single expansion step, and also finds similar consistent diversity as the number of expanded prompts increase.

—**Does prompt optimization lead to better aesthetics?** The `PE: Re-fine-tuned` model shows a significant increase in the aesthetic score vs. Straight-Query Generation, while the gain for the other PE models are more modest. We conclude, therefore, that the text-to-image model-relative parts of our pipeline are critical to improving aesthetic values. This conclusion is further strengthened by examining the rater preference results (Figure 5), where we see that re-fine-tuning consistently achieves higher win-rates over straight-query base-

Table 2: Aggregate Prompt Expansion performance.

Method	Aesthetics (MUSIQ-AVA) ↑	Text-Image Alignment `COCA(q,I(p))` ↑	Diversity (σ_p) ↑
Straight-Query Gen.	5.121 ± 0.519	0.125 ± 0.0147	0.00582 ± 0.00275
Few-shot Prompting	5.295 ± 0.549	0.114 ± 0.0199	0.00726 ± 0.00339
Prompt Expansion	5.225 ± 0.585	0.120 ± 0.0175	0.00720 ± 0.00341
PE: Re-fine-tuned	6.185 ± 0.474	0.113 ± 0.0199	0.00746 ± 0.00354
PE: Multi-Prefix	5.712 ± 0.616	0.125 ± 0.0157	0.00624 ± 0.00297
PE: Prefix Dropout	5.410 ± 0.622	0.121 ± 0.0156	0.00634 ± 0.00304

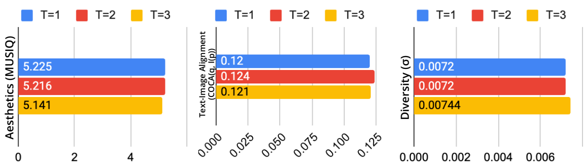


Figure 7: Evaluating consistency in Multi-Step Prompt Expansion: We find that the diversity of Prompt Expansion is consistent over number of steps.

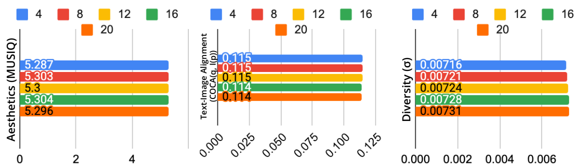


Figure 8: Evaluating consistency with scale: We find that the diversity of Prompt Expansion is consistent over number of expanded prompts per step.

lines (scoring > 0.52 across all query types) than the other models. For the 1x1 setting, this can be entirely attributed to the prompt being optimized during re-fine-tuning for higher aesthetics. A significant source of the gains in aesthetics is due to the Prompt Expansion model learning which flavors the downstream text-to-image model is most responsive to (examples shown in Table 5).

—**Does increased diversity of samples lead to better aesthetics?** We hypothesize that in addition to the benefits of tuning the expanded prompt to generate higher quality images, there is an additional gain of quality achieved when images from a single prompt are more diverse. With higher diversity, the likelihood of a highly aesthetically pleasing image being generated and selected by the user goes up. To demonstrate this effect, we compare the difference in Prompt-win rates between the 4x4 (where the best images from each system are compared) and 1x1 settings (Figure 5). The Prompt-win rates are consistently larger across all systems in the 4x4 setting, showing the aesthetic benefits of increased diversity at both the prompt and image level.

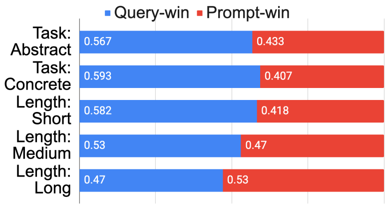
—**Does Prompt Expansion balance between aesthetics and text-image alignment?** An important observation is that while an expanded prompt can be designed to preserve all the semantics of the original query (by ensuring no details are lost etc.), it can never be expected to increase the degree of alignment of the image to the original intent of the user as expressed by the query. Thus, the main goal of Prompt Expansion with respect to text-image alignment is to minimize the decrease of alignment scores while optimizing for aesthetics and diversity. Thus, we expect to see a trade-off between alignment and aesthetic/diversity scores.

Indeed, in Table 2, we see that there is a minimal drop in the alignment scores of our PE models, inversely correlated with the other 2 metrics of diversity and aesthetics. The alignment score drop, however, is well within the confidence intervals, showing that the semantics of the expanded prompt do not stray too far from the meaning of the original query. As further evidence of this, we turn to the human rater evaluation for text-image alignment (Figure 6), with large Equivalent-rates in the comparisons, and comparable win-rates for prompt and query. For example, in the 1x1 setting for the re-fine-tuning model, we see 70% equivalent-rate and prompt/query win-rates of 15% each.

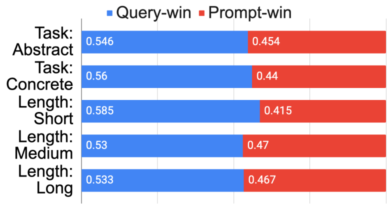
We further explore the trade-offs between diversity, aesthetics, and text-image alignment by presenting Pareto-optimality curves between the various automatic and human metrics for all the different query types in Figure 4. While `PE: Re-fine-tuned` dominates `Prompt Expansion` in terms of Diversity-Aesthetics, we notice that the converse is true for Diversity-Alignment. If text-image alignment is more important than diversity and aesthetics collectively, then it may be beneficial to omit the re-fine-tuning step.

—**Can controllable generation improve performance?** In Table 2, we see that `PE: Multi-Prefix` obtains better aesthetics than our base PE model, and remarkably without any loss in text-image alignment. Given the expansion is

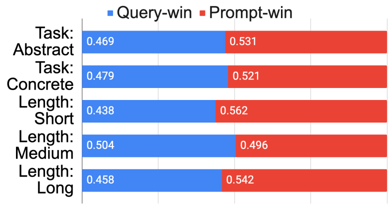
Figure 5: **Aesthetics**: Human Evaluation results for 1x1 and 4x4 settings.



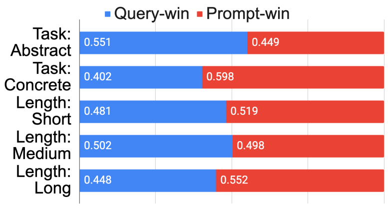
(a) 1x1 Aesthetics, Few-Shot Prompting



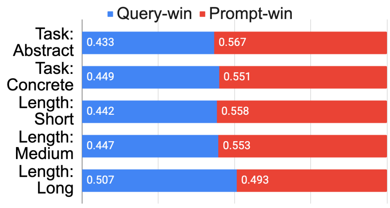
(b) 1x1 Aesthetics, Prompt Expansion



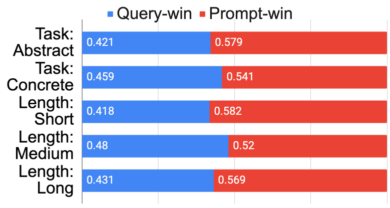
(c) 1x1 Aesthetics, PE: re-fine-tuned



(d) 4x4 Aesthetics, Few-Shot Prompting

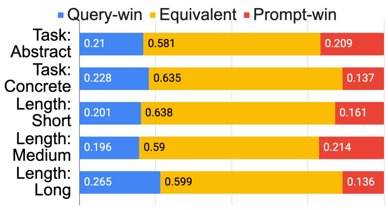


(e) 4x4 Aesthetics, Prompt Expansion

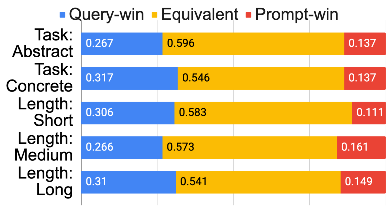


(f) 4x4 Aesthetics, PE: re-fine-tuned

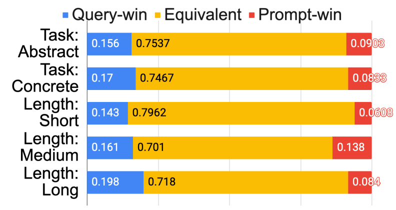
Figure 6: **Text-Image Alignment**: Human Evaluation results for 1x1 and 4x4 settings.



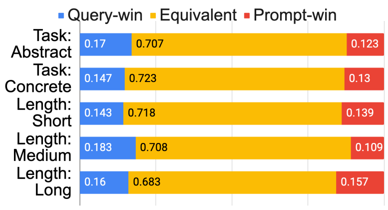
(a) 1x1 Text-Image Alignment, Few-Shot Prompting



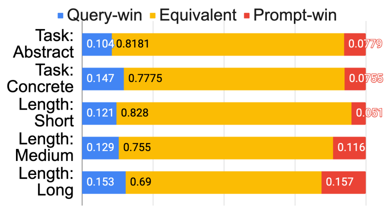
(b) 1x1 Text-Image Alignment, Prompt Expansion



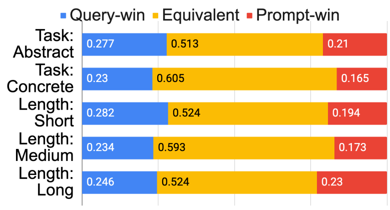
(c) 1x1 Text-Image Alignment, PE: re-fine-tuned



(d) 4x4 Text-Image Alignment, Few-Shot Prompting



(e) 4x4 Text-Image Alignment, Prompt Expansion



(f) 4x4 Text-Image Alignment, PE: re-fine-tuned

determined by query type, these results should be interpreted as an upper-bound on how much controlled generation can benefit the generic Prompt Expansion task. However, we also see that the `PE: Prefix Dropout` model retains much of the advantage of `PE: Multi-Prefix`, without matching the aesthetic performance of the `PE: Re-fine-tuned` model. It has the advantage, however, of not being optimized for a particular downstream text-to-image model, and is thus more generally applicable.

8 Conclusion

Text-to-image generation models are capable of producing high-quality images from text prompts, but they require specialized skills and knowledge to use effectively. Prompt Expansion reduces the need for users to iteratively prompt-engineer and over-specify their text prompt. Human raters find that images generated through Prompt Expansion are more aesthetically-pleasing and diverse than baselines. Empirical metrics for aesthetic and diversity also show that Prompt Expansion outperforms comparable baselines. In addition, the paper discusses the use of Prompt Expansion as a building block for other use cases such as multi-step adaptation and controllable generation.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Eric J. Bigelow, John P. McCoy, and Tomer D. Ullman. 2023. Non-commitment in mental imagery. *Cognition*, 238:105498.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. In *arXiv:2001.08361*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

CLIP-Interrogator. Clip-interrogator. https://github.com/pharmapsychotic/clip-interrogator.

Guillem Collell and Marie-Francine Moens. 2016. Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *COLING*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. In *arXiv preprint arXiv:2208.01626*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1172–1184, Online only. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. ssociation for Computational Linguistics.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022a. Photorealistic text-to-image diffusion models with deep language understanding.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022a. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022b. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. In *arXiv:2206.10789*.

Table 3: Comparison of PE variations: The base Prompt Expansion (PE) model maps varying query lengths to a fixed expanded prompt; we evaluate the ablation where each query maps to the subsequent slightly-longer query, resulting in varying output prompt length. We also ablate the use of all prefixes and all mixtures in the dataset. We ablate this further by comparing the use of prefixes to perform query-specific inference (i.e. the model knows what kind of query it is handling) against query-agnostic inference (through the use of Prefix Dropout).

Method	**Prompt Category**	**Aesthetics (MUSIQ-AVA)** $\uparrow$	**Text-Image Alignment** `COCA(q,I(p))` $\uparrow$	**Diversity** (σ_p) $\uparrow$	**Output prompt length (mean+std)**
Straight-Query Generation	Task: Abstract	5.114 ± 0.528	0.117 ± 0.0106	0.00606 ± 0.00286	2.927 ± 2.249
	Task: Concrete	5.141 ± 0.505	0.126 ± 0.0137	0.00596 ± 0.00281	7.590 ± 2.893
	Length: Short	5.0632 ± 0.516	0.117 ± 0.0106	0.00611 ± 0.00289	2.196 ± 0.686
	Length: Medium	5.142 ± 0.523	0.125 ± 0.0140	0.00561 ± 0.00265	5.417 ± 1.114
	Length: Long	5.164 ± 0.513	0.134 ± 0.0180	0.00571 ± 0.00268	11.525 ± 5.773
	Dataset: WAA	5.128 ± 0.517	0.122 ± 0.0130	0.00601 ± 0.00283	5.263 ± 3.486
	Dataset: PartiPrompts	5.105 ± 0.523	0.134 ± 0.0189	0.00535 ± 0.00253	9.156 ± 7.483
Few-shot Prompting	Task: Abstract	5.289 ± 0.560	0.102 ± 0.0161	0.00780 ± 0.00372	11.218 ± 3.0391
	Task: Concrete	5.293 ± 0.530	0.117 ± 0.0176	0.00726 ± 0.00346	17.0727 ± 3.0695
	Length: Short	5.274 ± 0.568	0.100 ± 0.0164	0.00794 ± 0.00379	9.911 ± 3.0202
	Length: Medium	5.322 ± 0.522	0.117 ± 0.0181	0.00711 ± 0.00338	12.842 ± 2.620
	Length: Long	5.293 ± 0.552	0.127 ± 0.0198	0.00668 ± 0.00318	11.420 ± 3.502
	Dataset: WAA	5.291 ± 0.545	0.109 ± 0.0186	0.00753 ± 0.00359	14.169 ± 3.401
	Dataset: PartiPrompts	5.305 ± 0.559	0.127 ± 0.0233	0.00658 ± 0.00312	16.138 ± 4.120
Prompt Expansion (PE)	Task: Abstract	5.232 ± 0.596	0.112 ± 0.0118	0.00765 ± 0.00362	34.736 ± 9.0216
	Task: Concrete	5.260 ± 0.567	0.120 ± 0.0154	0.00724 ± 0.00342	39.362 ± 8.167
	Length: Short	5.233 ± 0.593	0.112 ± 0.0122	0.00780 ± 0.00369	34.287 ± 8.850
	Length: Medium	5.227 ± 0.577	0.119 ± 0.0153	0.00707 ± 0.00334	37.165 ± 8.169
	Length: Long	5.216 ± 0.584	0.128 ± 0.0200	0.00669 ± 0.00316	42.645 ± 9.129
	Dataset: WAA	5.246 ± 0.582	0.116 ± 0.0143	0.00744 ± 0.00352	37.054 ± 8.909
	Dataset: PartiPrompts	5.173 ± 0.590	0.129 ± 0.0208	0.00660 ± 0.00311	40.467 ± 10.276
PE: Re-fine-tuned	Task: Abstract	6.183 ± 0.505	0.100 ± 0.0162	0.00796 ± 0.00380	42.0169 ± 7.986
	Task: Concrete	6.183 ± 0.451	0.115 ± 0.0184	0.00744 ± 0.00353	46.143 ± 8.088
	Length: Short	6.172 ± 0.510	0.0997 ± 0.0167	0.00805 ± 0.00384	41.985 ± 7.978
	Length: Medium	6.210 ± 0.443	0.115 ± 0.0187	0.00729 ± 0.00345	43.126 ± 7.551
	Length: Long	6.176 ± 0.461	0.125 ± 0.0201	0.00699 ± 0.00332	49.490 ± 8.861
	Dataset: WAA	6.183 ± 0.479	0.108 ± 0.0189	0.00770 ± 0.00366	44.0845 ± 8.298
	Dataset: PartiPrompts	6.189 ± 0.464	0.125 ± 0.0227	0.00686 ± 0.00325	47.0 ± 9.754
PE: step-by-step expansion	Task: Abstract	5.191 ± 0.539	0.114 ± 0.0120	0.00756 ± 0.00360	11.161 ± 3.941
	Task: Concrete	5.179 ± 0.505	0.124 ± 0.0141	0.00670 ± 0.00318	19.344 ± 6.125
	Length: Short	5.173 ± 0.520	0.114 ± 0.0123	0.00765 ± 0.00365	9.888 ± 2.719
	Length: Medium	5.187 ± 0.527	0.122 ± 0.0149	0.00663 ± 0.00316	15.405 ± 3.787
	Length: Long	5.161 ± 0.526	0.131 ± 0.0184	0.00617 ± 0.00290	25.926 ± 9.401
	Dataset: WAA	5.185 ± 0.522	0.119 ± 0.0139	0.00713 ± 0.00339	15.261 ± 6.579
	Dataset: PartiPrompts	5.144 ± 0.529	0.132 ± 0.0198	0.00610 ± 0.00289	21.582 ± 12.526
PE: Multi-Prefix	Task: Abstract	5.653 ± 0.610	0.116 ± 0.0153	0.00658 ± 0.00315	20.198 ± 8.617
	Task: Concrete	5.718 ± 0.617	0.123 ± 0.0162	0.00630 ± 0.00301	21.569 ± 7.280
	Length: Short	5.663 ± 0.619	0.116 ± 0.0156	0.00652 ± 0.00312	20.196 ± 8.0729
	Length: Medium	5.744 ± 0.603	0.125 ± 0.0165	0.00627 ± 0.00298	22.471 ± 7.849
	Length: Long	5.709 ± 0.639	0.126 ± 0.0163	0.00616 ± 0.00293	21.976 ± 6.117
	Dataset: WAA	5.705 ± 0.609	0.122 ± 0.0156	0.00638 ± 0.00305	20.994 ± 6.833
	Dataset: PartiPrompts	5.729 ± 0.635	0.131 ± 0.0161	0.00588 ± 0.00278	23.692 ± 6.501
PE: Prefix Dropout	Task: Abstract	5.394 ± 0.618	0.116 ± 0.0147	0.00667 ± 0.00320	18.217 ± 8.298
	Task: Concrete	5.457 ± 0.631	0.120 ± 0.0149	0.00640 ± 0.00308	19.264 ± 8.672
	Length: Short	5.384 ± 0.637	0.117 ± 0.0152	0.00660 ± 0.00317	18.377 ± 8.472
	Length: Medium	5.401 ± 0.619	0.122 ± 0.0167	0.00630 ± 0.00302	20.212 ± 8.729
	Length: Long	5.445 ± 0.608	0.122 ± 0.0158	0.00624 ± 0.00300	19.903 ± 7.868
	Dataset: WAA	5.425 ± 0.625	0.118 ± 0.0149	0.00654 ± 0.00314	18.741 ± 8.504
	Dataset: PartiPrompts	5.374 ± 0.614	0.129 ± 0.0174	0.00583 ± 0.00280	22.148 ± 7.359

A Types of queries

Queries may have different structures or user intents. We extract a range of potential user queries that would be mapped to the inverted text (prompt). One intent for this is that a range of user queries may map to the same prompt text. For example, the user query may be abstract and underspecified, and a Prompt Expansion model is needed to introduce ideas and potential content directions. Or the user query may be very concrete and detailed, but needs refinement in fine-grained details or phrasing to improve image aesthetics, while remaining faithful to the original query.

Focused on adapting Prompt Expansion to the level of image specification provided by the user, there are two general settings that we consider: *semantic expansion*, and *detailed expansion*. Semantic expansion is a Prompt Expansion setting where the user queries are abstract or underspecified, the user intent may be towards ideation or exploration, and the output prompts intend to add details and expand on the query conceptually. Detailed expansion is a Prompt Expansion setting where the queries may be quite detailed and concrete, and the expanded prompt needs added details while still remaining faithful to the original query. This consequently results in the construction of additional sub-datasets (mixtures) to be added to the base Prompt Expansion dataset. These augmentative mixtures include abstract queries (to be used in evaluating semantic expansion), detailed queries (such as grounded / specificity mixtures, to be used in evaluating detailed expansion), etc.

To further support these wider range of query types (abstract, detailed, varying sentence length), we few-shot prompt to generate more training data (Table 1). For example, grounded `{query:prompt}` pairs require the sub-phrases in the input query to reside in the output prompt. Specificity `{query:prompt}` pairs require the output prompt to begin with the query, and the rest of the prompt is an extended example of the query. Constructing these augmentations need to be mindful of grammatical errors, which may reduce human readability in the expanded prompts. We also need to be mindful of the token distribution of the training dataset, so as to avoid skewing the model towards generating prompts on very specific topics, or repeating tokens (e.g. in Figure 9b).

B Adapting to query specification

B.1 Controllable Generation with `Multi-Prefix`

Not all queries are the same, and different queries may have different requirements on what their expanded prompts should look like. We assign different query types with a corresponding prefix (Table 4), and train a Prompt Expansion model. Having prepended the queries with prefixes during training, at test-time we prepend the queries with query-specific prefixes to determine the format of the expanded prompts. In Table 3, assuming the query type is known and the query-specific prefix is used during inference, we find that `PE: Multi-Prefix` maintains higher text-image alignment against other Prompt Expansion models.

Other than handling varying levels of image specification, prefixes have also been used for indicating Multi-Step Prompt Expansion. The use of prefixes sets a clear task boundary between different query sets/mixtures, such that the style of expanded prompt for one query type does not interfere with others. This also means we can continue to add new query types, their data, and prefixes over time, without risking interference with respect to other query types. For a given dataset, if we assign a prefix for each data point, the trade-off is that assigning (more) prefixes (further) dilutes the number of samples per prefix. Encountering fewer samples per prefix may result in underfitting per prefix's data, in contrast to fitting on the whole prefix-less dataset. To balance this trade-off, the Prompt Expansion models in the main results only make use of the `MSTP` prefix in the training dataset and during multi-step inference, and does not assign any other prefixes. Models trained on other prefixes are used in `PE: Multi-Prefix` (query-specific) and `PE: Prefix Dropout` (query-agnostic) models in Table 3. `PE: Multi-Prefix` makes use of all augmentative mixtures, assigns prefixes to queries during training (though replacing `GRD` and `SPCT` with `DTL`), and relies on the type-specific prefix during inference.

B.2 Query Ambiguity with `Prefix Dropout`

At test-time, we may not be able to assume we know the query type or which prefix to use. Thus, it is necessary for the Prompt Expansion model to infer the query type such that it can return appropriate expanded prompts. The Prompt Expansion model will be required to (implicitly) infer the

Table 4: Types of prefixes used for controllable generation. The models used in the paper make use of `ABST`, `DTL` (* indicates we replace the non-parenthesed prefix with the paranthesed prefix), `MSTP`.

Prefix	Full name	Description
ABST	ABSTract	Returns output sentences conditioned on the input sentence being an abstract query (output sentence structure is in-line with the abstract augmentation/mixture in training).
DTL	DeTaiLed	Adding this prefix returns output sentences conditioned on the input sentence being a detailed query and requiring a detailed expansion (output sentence structure is in-line with the detailed augmentation/mixture in training).
GRD (DTL*)	GRounDed	Adding this prefix returns output sentences conditioned on the input sentence being a query requiring a grounded expansion (output sentence structure is in-line with the grounded augmentation/mixture in training).
SPCT (DTL*)	SPeCificiTy	Adding this prefix returns output sentences conditioned on the input sentence being a query requiring an expansion that elicits specificity (output sentence structure is in-line with the specificity augmentation/mixture in training).
FLV	FLaVor	Adding this prefix returns only the flavor alone.
HAST	HighAeSThetics	Adding this prefix returns an output sentence that should return images with good aesthetics (MUSIQ > 6).
RFT	Re-Fine-Tuned	Adding this prefix in front of the the input sentence will return images based on the re-fine-tuning objectives (e.g. based on aesthetics and renderable flavors).
MSTP	MultiSTeP	Adding this prefix returns an output sentence specifically for Multi-Step Prompt Expansion.

query type correctly, then perform type-specific expansion. To tackle this, we develop Prefix Dropout, a curriculum learning approach where we assign a prefix to each query type, and train the Prompt Expansion model on the data with prefixes, but with a linearly increasing dropout rate (starting from 0.4 dropout rate to gradually 1.0 dropout rate), where the prefixes are randomly dropped out. This allows the model to first develop the representations for each query type, being activated by the use of prefixes. As dropout progresses, the model learns that the prefixes are not a reliable indicator. The model then develops representations to perform task inference to infer the query type, and uses this to access type-specific representations. We observed during training that the distribution of sampled expanded prompts would initially be either one query type or the other for a given query (implicit classification of the query type, i.e. type inference), but as the model begins to encounter a higher dropout rate such that no prefixes are available, it begins to diversify the generated samples such that there is a mix of sentences from all query types. During training, `PE: Prefix Dropout` makes use of all the augmentative mixtures, uses all the original prefixes, and during inference does not require the prefix. As seen in Table 3's results for `PE: Prefix Dropout`, Prefix Dropout can resolve query type ambiguity at test-time with output samples from multiple types, while also retaining text-image (query-prompt) alignment.

C Multi-Step Prompt Expansion

Rather than continuously increasing expanded prompt sequence length, we construct a dataset that maps an expanded prompt from step $t-1$ to an expanded prompt for step t. Using a re-fine-tuned model, we generate expanded prompts on held-out queries, and iteratively generate prompts upon the previous step's prompts. We re-fine-tune the model on this multi-step dataset.

As demonstrated in Figure 3, few-shot prompting of prompts to queries return multiple steps of expansion based on input length and output length. This enables Multi-Step Prompt Expansion for shorter queries. In the initial model, there is a maximum observed output length in our training dataset, and thus an expanded prompt may tend to stop expanding in length at this limit. Additionally, the text-to-text model also has a token length limit.

To enable expanded prompts past such limits, we first construct a paired training dataset of expanded prompts to next-step expanded prompts. The expanded prompt passed to the Prompt Expansion model returns expanded prompts (of similar output length but replaced details). We reuse the Prompt Expansion model by passing the expanded prompt and receiving a set of variant expanded prompts of alternative details. We then append sentences with additional details to the input prompt. We repeat this until the input prompt length reaches the token limit. In practice, if the query exceeds the token limit, it will need to be truncated.

We re-fine-tune the Prompt Expansion model with the multi-step data, and add the `MSTP` prefix. As the user is reusing the previous step's prompt, we know that the query type is that of "multi-step", and know from Table 3 (`PE: Multi-Prefix`) that a prefix can return better metric performance if the query type is known and its prefix is used. We also may wish to avoid the performance drop that may arise from having output sequences of varying lengths, as seen in Table 3 (`PE: step-by-step expansion`). `PE: step-by-step expansion` is a supporting ablation that mapping queries to the full expanded prompt (as opposed to the subsequent query) is not ideal.

After generating $N = 4$ expanded prompts from the initial query at step $t = 0$, at step $t = 1$ we then subsequently generate N expanded prompt for each expanded prompt, and so on, resulting in N^{t+1} expanded prompts in a set to evaluate. In Figure 7, we observe consistent diversity across number of expansion steps. We evaluate against N^{t+1} expanded prompts at the t-th step. We observe that diversity is consistent in a multi-step setting over several iterations, where the expanded prompt from a previous iteration is passed back into the model to return another set of expanded prompts.

While Figure 7 evaluates the consistency of Prompt Expansion over the number of steps, Figure 8 evaluates its consistency over the number of expanded prompts generated. We observe that the diversity of the generated images continue to increase, and aesthetics and text-image alignment remains consistent.

D Probing the alignment between text-to-text and text-to-image models

Using the base model, we probe the ability of the Prompt Expansion model in rendering aesthetically-pleasing and diverse images. We observe that the text-to-image model has difficulty in rendering certain prompts (e.g. human hands, faces). This can become an issue when the Prompt Expansion model generates prompts such that the text-to-image model cannot render its image.

We begin by evaluating how well flavors can be rendered. Compared to specific details in captions, flavors have a higher frequency of overlap / re-use between different expanded prompts. As such, we would be able to evaluate how well a flavor is rendered across multiple expanded prompts. We evaluate the COCA similarity between the query text embedding and generated image embedding, as well as the COCA similarity between the prompt text embeddings and generated image embedding.

	flavor	`COCA(q,I(p))` ↑	`COCA(p,I(p))` ↑
Top-5 flavors	art deco	0.163	0.169
	vorticism	0.162	0.168
	classical realism	0.161	0.165
	fine art	0.155	0.166
	figurative art	0.159	0.161
Worst-5 flavors	academic art	0.096	0.110
	international gothic	0.107	0.098
	pixel art	0.085	0.114
	arte povera	0.093	0.102
	generative art	0.102	0.088

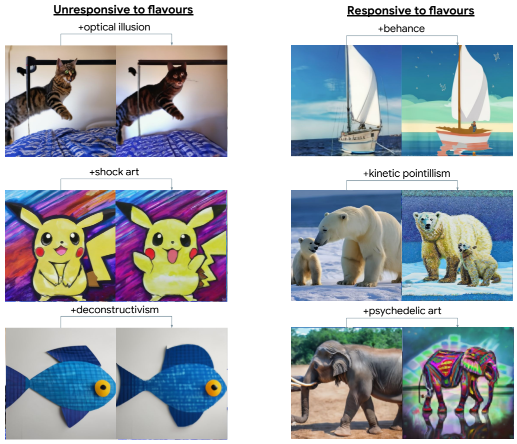
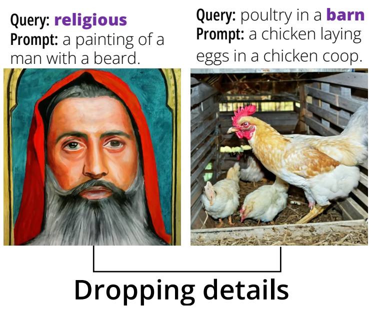


Table 5: **Top:** The average query-image embedding distance and prompt-image embedding distance for flavors are ranked. **Bottom:** For the same seed, image generation does not respond to the insertion of some flavors, while responding to others.

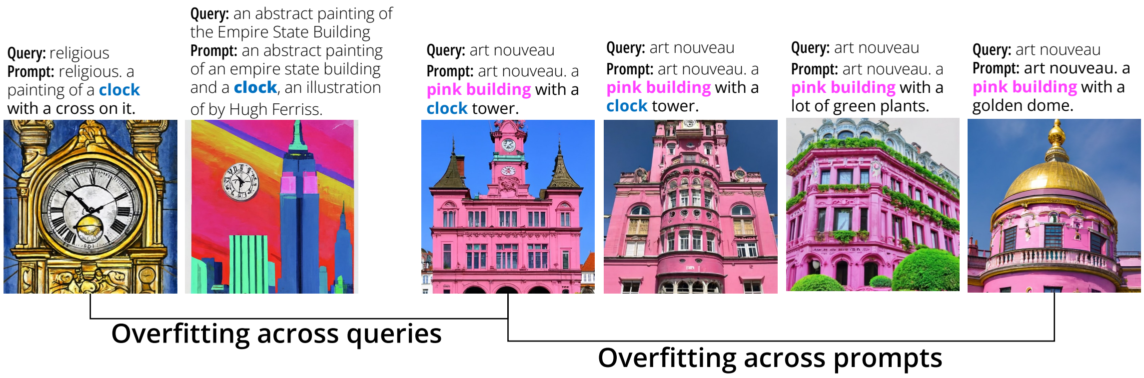
As shown in Table 5 (Top), sorted by the average of the two COCA scores, the similarity scores between the Top-5 highest flavors diverges greatly from that of the Worst-5 flavors. Referring to Table 5 (Bottom), where we generate two images (without and with flavor) with the same seed, we can visually observe that flavors in the left column are unresponsive (e.g. the seeded image does not change much), while the flavors in the right column visually change the image. Probing the difficulty in rendering certain prompts motivates the need to re-fine-tune the Prompt Expansion model to align it with the target text-to-image model (Section 4.1).

We also evaluate how re-fine-tuning against one text-to-image model performs against other models. In Table 8, we evaluate Stable Diffusion v1.4 against expanded prompts generated by the Base PE model and the PE model re-fine-tuned for Imagen. Even without re-fine-tuning, comparing PE vs Straight-Query Generation, we observe that Prompt Expansion can already improve the aesthetics/diversity regardless of text-to-image archi-

Figure 9: Examples of challenges mitigated in the current Prompt Expansion Model.



(a) Dropping details from step-by-step expansion: Qualitatively supporting the findings from Table 3, training on variable output sequence length and mapping step-by-step (`PE: step-by-step expansion`) encounters challenges that others do not. We observe that a drop of details may occur (e.g. barn is not mentioned in the prompt for "poultry in a barn"), thus resulting in lower text-image alignment, especially for concrete / long queries.



(b) Overfitting from fine-tuning: Qualitatively supporting the findings from Table 6, prompt-tuning alleviates certain challenges that fine-tuning may present. We also observe overfitting to the training distribution, where tokens (e.g. clock, pink building) are repeated, regardless of same or different query.

tecture. With re-fine-tuning to align with Imagen, we observe that Stable Diffusion does not have as big a jump in aesthetics as Imagen. Both these observations confirm that Prompt Expansion can in general improve performance regardless of text-to-image architectures, and that re-fine-tuning can further enhance performance for a target text-to-image architecture.

E Skewed token distribution

In general, we find that prompt-tuning avoids a skewed token distribution. With Table 6, we establish the training setup of the PaLM 2 model. We evaluate fine-tuning against prompt-tuning, and a smaller 1B against a larger 24B parameter model. We show that prompt-tuning a 1B PaLM 2 model is the optimal setup, achieving pareto-optimality in parameter count, aesthetics, text-image alignment, and diversity scores compared to other configurations. In-line with Welleck et al. (2020)'s evaluation of repetition, we also evaluate the repetitiveness of expanded prompt generation. We observe that fine-tuned models returned a higher repetition rate, indicating a potentially skewed token distribution, likely towards tokens seen during training. This manifests in expanded prompts as recurring terms appearing regardless of varying queries, and even recurring terms for the same query (examples in Figure 9b).

F Prompt length

In general, we find that fixed/long length is the optimal prompt length compared to variable/short length. In Table 3, we ablate the sequence length of the expected output prompt. We do this by comparing two query-prompt mapping settings. When we few-shot prompt a prompt p to a (shorter) query q_n, where n is the nth-order query, we then few-shot prompt the query q_n to a (shorter) query q_{n-1}. We repeat this till query q_1 , until we have a set of n queries mapped to each p. While `Prompt Expansion` maps each $q_i \mapsto p$ for every $i \in n$, `PE: step-by-step expansion` maps each (shorter) query q_i to the next (longer) query q_{i+1} such that $q_i \mapsto q_{i+1}$ (Figure 10).

Though the variance of output sequence length

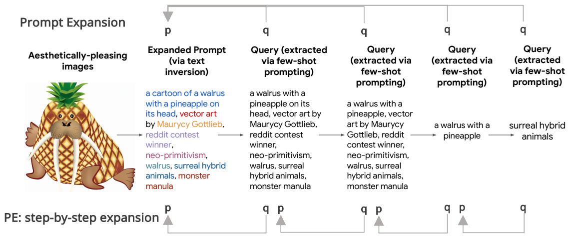
Table 6: Evaluating different model configurations, namely model size and parameter tuning method.

Method	Size	Aesthetics (MUSIQ-AVA) ↑	Text-Image Alignment `COCA(q,I(p))` ↑	Diversity (σ_p) ↑	Repetition Rate (r)) ↓
Fine-Tune	1B	5.282 $\pm$ 0.570	0.122 $\pm$ 0.0170	0.00621 $\pm$ 0.00297	0.0101
Fine-Tune	24B	5.411 $\pm$ 0.622	0.1216 $\pm$ 0.0166	0.00657 $\pm$ 0.00312	0.112
Prompt-Tune	1B	5.225 $\pm$ 0.585	0.120 $\pm$ 0.0175	0.00720 $\pm$ 0.00341	0.00285
Prompt-Tune	24B	5.335 $\pm$ 0.625	0.119 $\pm$ 0.0177	0.00702 $\pm$ 0.00331	0.00715

Table 7: Ablations for diversity in Prompt Expansion and its subsequent text-to-image generation.

Method	Hyperparameters	Aesthetics (MUSIQ-AVA) ↑	Text-Image Alignment `COCA(q,I(p))` ↑	Diversity (σ_p) ↑
Straight-Query Gen.	4 random seeds	5.121 $\pm$ 0.519	0.125 $\pm$ 0.0147	0.00582 $\pm$ 0.00275
Prompt Expansion	Temperature = 0.1	5.267 $\pm$ 0.534	0.119 $\pm$ 0.0177	0.00702 $\pm$ 0.00331
	Temperature = 0.5	5.283 $\pm$ 0.5266	0.116 $\pm$ 0.0215	0.00683 $\pm$ 0.00326
	Temperature = 1.0	5.225 $\pm$ 0.585	0.120 $\pm$ 0.0175	0.00720 $\pm$ 0.00341
Prompt Expansion	decoder=greedy	5.294 $\pm$ 0.538	0.115 $\pm$ 0.0213	0.0 $\pm$ 0.0
	decoder=beam search	5.0839 $\pm$ 0.550	0.118 $\pm$ 0.0211	0.00426 $\pm$ 0.00332
Prompt Expansion	+ text post-hoc filtering	5.318 $\pm$ 0.542	0.114 $\pm$0.0216	0.00726 $\pm$0.00346
	+ image post-hoc filtering	5.316 $\pm$0.562	0.113 $\pm$0.0215	0.00739 $\pm$0.00350

Figure 10: An example of queries are mapped to prompts for `PE: step-by-step expansion`, compared to `Prompt Expansion`.



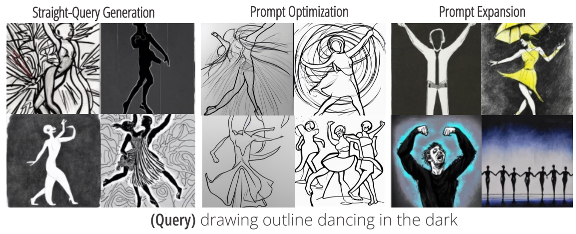


Figure 11: We demonstrate the difference between prompt optimization techniques and prompt exploration techniques. While prompt optimization (Promptist (Hao et al., 2022)) can return the optimized prompt *"drawing outline dancing in the dark, trending on artstation"*, it aims to return one optimized prompt, and thus 4 images generated from this optimized prompt may return less visual diversity and explores less of the prompt space than Prompt Expansion. We evaluate Promptist's optimized prompt against Stable Diffusion v1.4, which is the model it is aligned with during its training.

between `PE: step-by-step expansion` and `Prompt Expansion` is similar, the mean length is much higher for `Prompt Expansion`. We find that `PE: step-by-step expansion` consistently underperforms `Prompt Expansion` across varying query types across all metrics. Qualitatively, we find that `step-by-step expansion` also results in drop of query details, as shown in Figure 9a.

G Diversity ablations

We further evaluate diversity in the output images compared to other ablation methods in Table 7. The baseline is Straight-Query Generation, which generates 4 images with 4 random seeds.

Decoding. Prompt Expansion uses temperature-based decoding, with default temperature being 1.0. The first set of ablations is varying temperature between 0 to 1. A higher temperature tends to result in more diverse output for text models using temperature-based decoding. The next set of ablations is changing the decoding method, e.g. greedy decoding and beam search decoding. Beam search (with beam size 4) returns 4 expanded prompts per query. Greedy decoding returns 1 expanded prompt per query. The embedding variance of 1 image generated per expanded prompt σ_p is 0.0.

Post-hoc filtering. Our final set of ablations is to maximize diversity after text/image generation. Specifically, post-hoc filtering aims to generate a large number (N=20) of prompts/images, and filter out 4 that return the highest combinatorial diversity score. For post-hoc text filtering, we return 20 prompts through Prompt Expansion, compute the COCA text embeddings for each prompt, then enumerate through each combination of 4 prompts, and the selected 4 prompts is the combination that has the highest variance in text embeddings. For post-hoc image filtering, we return 20 prompts through Prompt Expansion and generate 1 image per prompt, compute the COCA image embeddings

Table 8: Evaluating different generated prompts against text-to-image models.

Input Prompt	Text-to-Image Model	Aesthetics (MUSIQ-AVA) ↑	Text-Image Alignment `COCA(q,I(p))` ↑	Diversity (σ_p) ↑
Straight-Query Gen.	StableDiffusion v1.4	4.617 ± 0.634	0.117 ± 0.0139	0.00521 ± 0.00166
Straight-Query Gen.	Imagen	5.121 ± 0.519	0.125 ± 0.0147	0.00582 ± 0.00275
Prompt Expansion	StableDiffusion v1.4	4.878 ± 0.646	0.112 ± 0.0213	0.00689 ± 0.00249
Prompt Expansion	Imagen	5.225 ± 0.585	0.120 ± 0.0175	0.00720 ± 0.00341
PE: Re-fine-tuned for Imagen	StableDiffusion v1.4	4.836 ± 0.638	0.108 ± 0.0236	0.00707 ± 0.00302
PE: Re-fine-tuned for Imagen	Imagen	6.185 ± 0.474	0.113 ± 0.0199	0.00746 ± 0.00354

for each prompt, then enumerate through each combination of 4 images, and the selected 4 images is the combination that has the highest variance in image embeddings.

H Human Rater Task

We retain the queries (n=700) and expanded prompts used in the automatic evaluation, and evaluate them with human raters. We perform a side-by-side (SxS) evaluation task, where we show a rater a set of images side by side for a given text query, and ask the rater to pick the best image. We split the 700 queries into two sets of 350 for 1x1 and 4x4 settings. Each instance was evaluated by 3 human raters, with the strong consensus (Table 9) indicating consistency in preferences between raters. Figure 13 shows the annotation interface for the rater experiment.

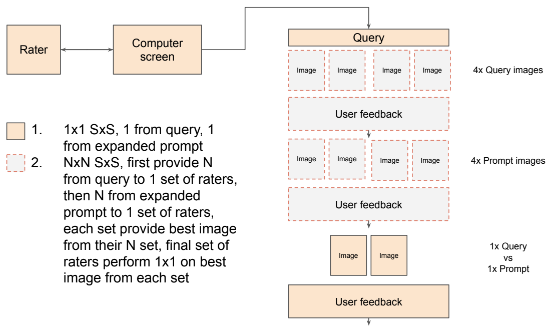
Aesthetics. In this setting, raters are first provided instructions on how to evaluate Aesthetics. We describe it as follows:

> In this task, please select the image that you personally prefer. You may find this image more appealing due to aesthetic, stylistic or compositional qualities. Please only evaluate this image for aesthetic appeal, and ignore all other aspects (overall image quality).

We provide examples of one image of high aesthetic appeal side-by-side with another image of low aesthetic appeal. In addition to positive examples, we provide examples of images not to reject due to flaws in image quality.

> Please only evaluate this image for aesthetic appeal, and ignore all other aspects (e.g. overall image quality). Image quality measures the visual flaws of the image. Examples of visual flaws include, watermarks, illegible text, hallucinated content, warped features or objects, blurriness, and out of place objects. If these flaws appear, this does not mean the image must be rejected. Only reject the image if the other image is one that you personally prefer (e.g. more appealing due to aesthetic, stylistic or compositional, but not image quality).

Figure 12: An overview of the 1x1 and NxN rater experiments. A Query vs Prompt evaluation always takes place, but an NxN experiment requires a set of 4 images to be evaluated against each other before the 1x1 experiment.



Once the rater is acclimated with the metric, we proceed with the task and can ask them:

> Q: Which image do you personally prefer? (aesthetically, stylistically, or compositionally).

Text-Image Alignment. We provide positive examples of images more aligned to the text description (query) compared to the other side-by-side image. We similarly provide negative examples of images not pertaining to text-image alignment. We describe the task to the rater as follows:

> Please only evaluate this question for consistency, and ignore all other aspects of the image (image quality, or overall image preference).

Furthermore, we add an option `Unsure` when the rater perceives all images as equally likely to be picked. Specifically, we describe to the raters to select this option as follows:

Table 9: Consensus between Raters: This details how many of the raters agree on an each Prompt-win image (i.e. number of raters that all picked the same Prompt-win image).

Prompt-win Consensus	Consensus Agreement (3/3)	Consensus Agreement (2/3)	Consensus Agreement (0/3)
1x1 Aesthetics	0.450 ± 0.0103	0.549 ± 0.0103	0.0 ± 0.0
4x4 Aesthetics	0.396 ± 0.0167	0.604 ± 0.0167	0.0 ± 0.0
1x1 Text-Image Alignment	0.513 ± 0.146	0.446 ± 0.117	0.0407 ± 0.0288
4x4 Text-Image Alignment	0.486 ± 0.150	0.481 ± 0.128	0.0333 ± 0.0259

Select `Unsure` if:

- You do not understand the text prompt / it is unclear
- Both images are equally consistent
- Both images are equally inconsistent

Once the rater is acclimated with the metric, we proceed with the task and can ask them:

Q: Which image is more consistent with the text description?

Random vs. Random (1x1). Can we attribute aesthetic preference of prompt expanded images over straight-query images to improved prompt construction / engineering? To answer this question, we perform a one-to-one side-by-side (1x1 SxS) comparison, where we pick the *first* image generated from straight-query generation and the *first* image generated from Prompt Expansion, then ask the raters to evaluate which one is better. This tests the ability of Prompt Expansion to reduce the need for manual prompt engineering by the user. For both Aesthetics and Consistency, raters have 700 discrete rating decisions (we split to n=350, going through 1,400 images). The raters can pick between the options: Left, Right, (Unsure). For the consistency task, we provide the Unsure option in case two images are equally good (or bad), since text-image alignment can be objectively evaluated. For aesthetics, we do not provide the Unsure option to make their aesthetic preference.

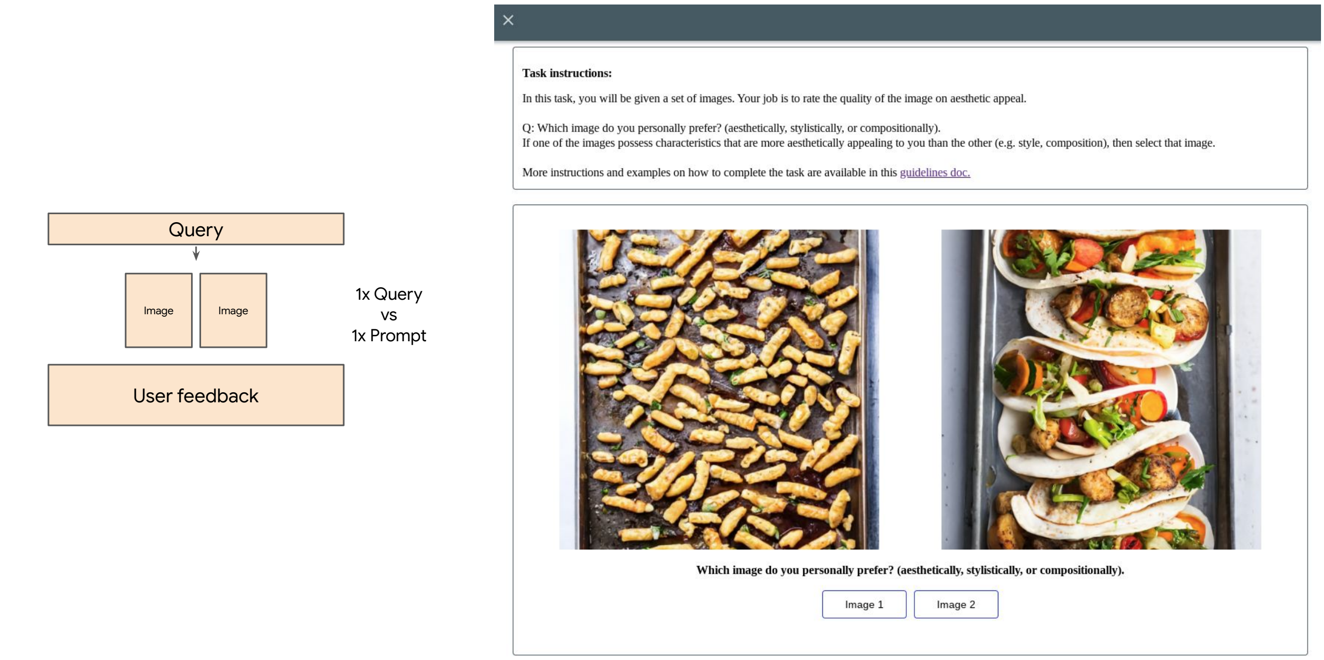
Best vs. Best (4x4). Can we attribute aesthetic preference of prompt expanded images over straight-query images to increased diversity of samples? To answer this question, we perform a N-to-N side-by-side (NxN SxS) comparison followed by another 1x1 SxS comparison. We begin by showing the rater N=4 images from straight-query generation. We also show the rater N=4 images from Prompt Expansion. The raters are not shown straight-query generation and Prompt Expansion images in separate stages; the 4x4 SxS is in one stage, and the rating decisions are shuffled. The raters pick the best image out of N=4. The raters can pick between the options: 1, 2, 3, 4, (Unsure). Then the raters enter a second stage, where they are shown the best straight-query image against the best Prompt Expansion image for the same query, and asked to select the best. The raters can pick between the options: Left, Right, (Unsure). For both Aesthetics and Consistency, raters have 2,100 discrete rating decisions (we split to a distinctly-different N=350, going through 7,000 images).

Discussion: Human evaluation of open-ended generation. Though outside our scope, we share our human rater designs (Figure 14) on evaluating Multi-Step Prompt Expansion. Evaluating this setting requires comparing user behaviour with iterative image selection against standard prompt engineering.

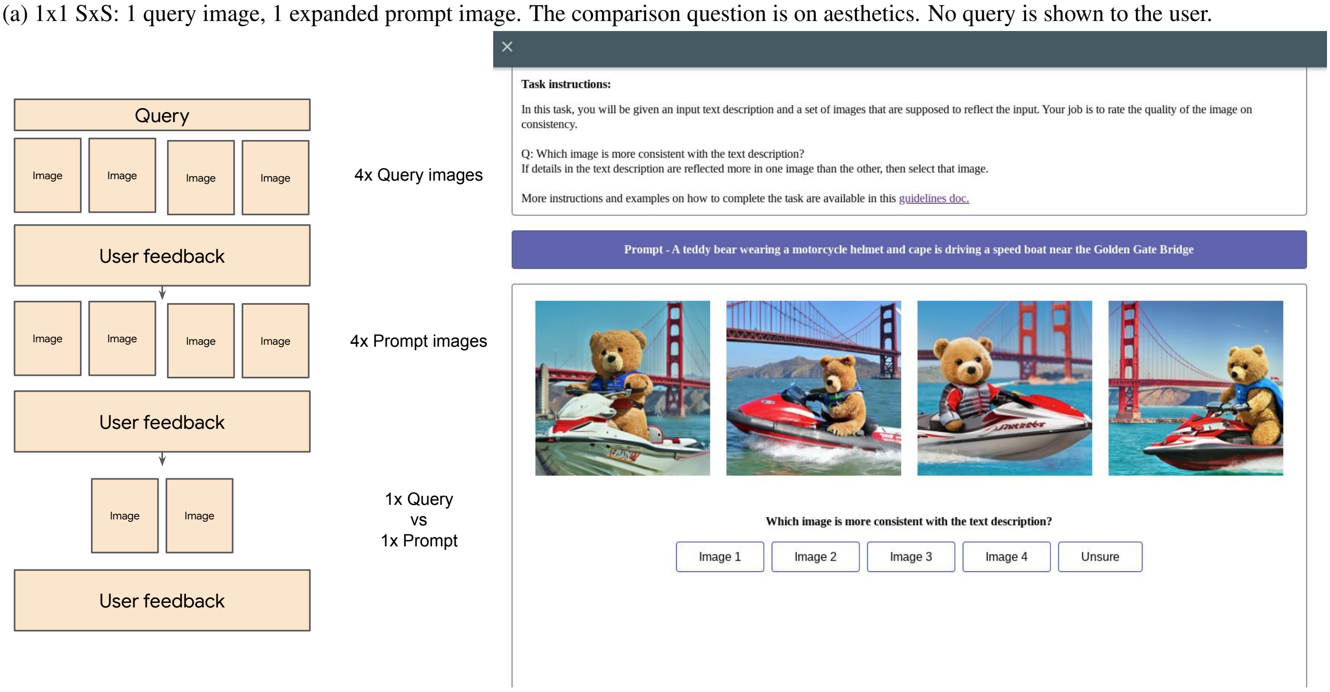
First we collect a set of reference images, and raters will attempt to generate images that are as equivalent or superior to the reference image (e.g. w.r.t. aesthetics or text-image alignment). For each reference image, raters iterate on their query T times, and choose the best of 4 images per prompt hacking iteration t. After a maximum of T prompt iterations, the rater is shown the best image selected per iteration step against each other, and pick the best image out of the T images.

Other than measuring the improvement in aesthetics, text-image alignment or diversity, this setup can measure changes to the user experience in text-to-image generation. For example, we can measure the cognitive load/effort reduced for the user. We can track the reduction in prompt iteration steps needed to obtain the best image, the reduction in number of text edits, etc. In addition to evaluating Multi-Step Prompt Expansion with Image selection, there are 2 other prompt iteration baselines:

- **Manual prompt engineering:** Raters manually prompt engineer the entire process on their own, irrespective of ability.
- **Suggestions:** During prompt engineering / free-text entry, we show raters recommended prompts



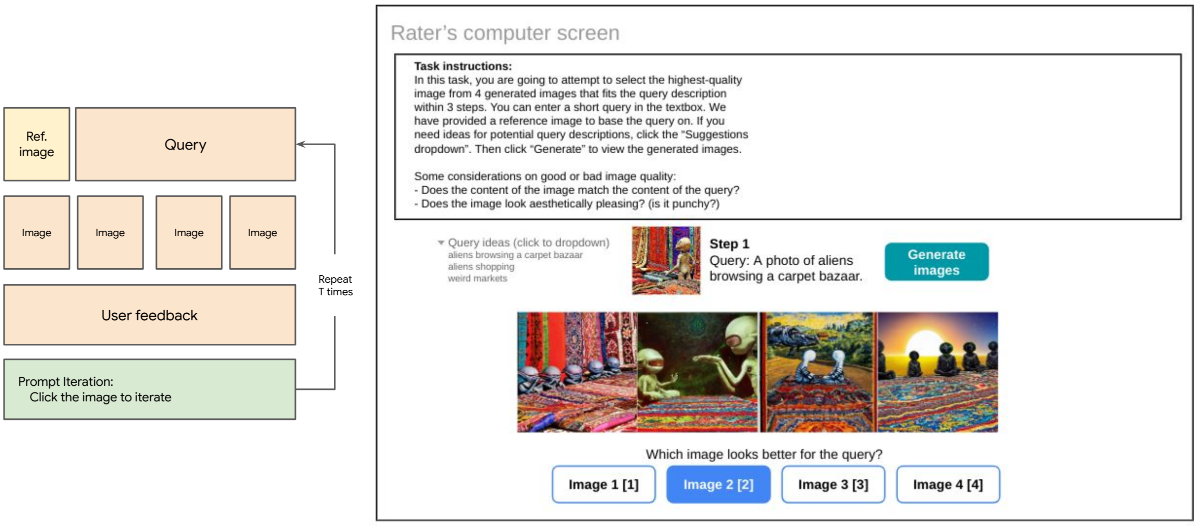
(a) 1x1 SxS: 1 query image, 1 expanded prompt image. The comparison question is on aesthetics. No query is shown to the user.



(b) 4x4 SxS: best image from 4 query images, best image from 4 expanded prompt images, raters perform 1x1 on best image from each set. The comparison question is on text-image alignment, thus the query is shown to the user.

Figure 13: The user interface for our human rater experiments.

Figure 14: UI design for evaluating Multi-Step Prompt Expansion with human raters.



generated from Prompt Expansion. They can use it directly, use it for inspiration, etc.

- **Image selection:** There is no free text entry other than for inserting the initial query, and they iterate by just clicking the image to iterate on. Images per iteration are generated through Multi-Step Prompt Expansion.

I Datasets

Several datasets varying by modality, training-evaluation use, captioned/non-captioned, etc are used throughout the paper. We provide further details on all of them here.

Webli-Align (Jia et al., 2021; Chen et al., 2023). Webli-Align is composed of two existing datasets Webli (Chen et al., 2023) and Align (Jia et al., 2021). We filter out non-aesthetic images in Webli-Align with the MUSIQ (Ke et al., 2021) aesthetics metric (refer to Sec 6). The filtered dataset contains 80k images. We also perform a 70-20-10 train-val-test split. Within the train set, we split 50-50 for base and re-fine-tune training.

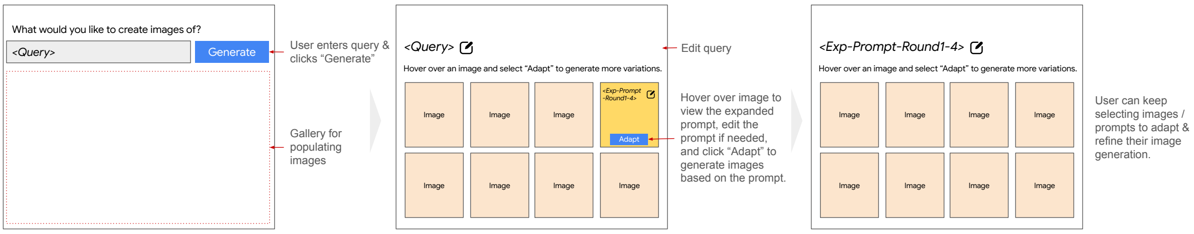
CrowdSourced. CrowdSourced is obtained through crowd-sourcing, where we provide an interface for text-to-image generation (Figure 15), and users enter queries to generate images. Users also have the option to upvote images that they liked, and we use this to filter for aesthetically-appealing images. An example of an upvoted, aesthetically-appealing image is shown in Figure 3. The filtered dataset contains 40k images. We also perform a 70-20-10 train-val-test split. Within the train set, we split 50-50 for base and re-fine-tune training.

Base Prompt Expansion Dataset. The procedure for generating the base dataset is detailed in Section 3. We also generate variations of this dataset, e.g. Multi-Prefix dataset (Section 5.1), Multi-Step dataset (Section 5.3).

Re-fine-tune Prompt Expansion Dataset. From the queries of the held-out set from the base dataset, we compute the expanded prompts with the base Prompt Expansion model (4 prompts per query). We discard the original prompt mapped to the query, and replace it with the new expanded prompts. For each expanded prompt, we compute the generated image. From this probing set, we score the images in order to rank the images based on renderability with respect to its prompt text. Our scoring metric is a weighted average of the query-image embedding distance and prompt-image embedding distance: $0.6 \times \texttt{COCA(q,I(p))} + 0.4 \times \texttt{COCA(p,I(p))}$. We filter out query-prompt-image pairs that are below a threshold. This filtered `{query:prompt}` dataset is used for re-fine-tuning.

While other scoring tools such as using MUSIQ scores or measuring embedding distance with respect to common image defects (e.g. distorted hands/faces) were also tested, we qualitatively find that this weighted average as a general heuristic can sufficiently filter out images that are non-renderable, unaligned to neither query nor prompt, and non-aesthetic. We also avoid completely filtering out every prompt/flavor that do not work. If the user intentionally requests for specific flavors / objects, we would need to generate expanded prompts that still fit the query.

Figure 15: UI design for the image generation tool used in crowdsourcing images. Users enter queries to generate expanded prompt images. Users can hover over images to directly pass the expanded prompt of the image through the Prompt Expansion model to generate a second round of expanded prompt images. The user may optionally edit the expanded prompt. In this way, users can engage in Multi-Step Prompt Expansion.



(Evaluation) Webli-Align (WA) (Jia et al., 2021; Chen et al., 2023). We sample n=500 queries from a test set of potential queries constructed by applying the PE dataset generation process of Section 3 (Image-to-Text Inversion + Query/Prompt Extraction) to Webli-Align (WA) (Chen et al., 2023; Jia et al., 2021) images. Queries are categorized based on whether they are abstract (WA=249), concrete (WA=251), short (<4 words) (WA=224), medium length (4-7 words) (WA=134), or long length (>7 words) (WA=143).

(Evaluation) PartiPrompts (PP) (Yu et al., 2022b). We sample n=200 prompts from the PartiPrompts (PP) dataset of prompts that is designed to represent different domains and features of language (such as counting, negation, etc). Queries are categorized based on whether they were short (<4 words) (PP=35), medium length (4-7 words) (PP=66), or long length (>7 words) (PP=66).

J COCA-Interrogator

The *Interrogator* (CLIP-Interrogator) approach to image-to-text inversion requires an image-text model to probe what set of words and phrases are most similar to a given image, and concatenate the words/phrases together as the prompt text. The prompt text computed through Interrogators are composed of (i) a caption, and (ii) a set of flavors. The caption is a description of the content of the image (e.g. who, what, where, when).

We detail the implementation steps for COCA-Interrogator. First, it requires an image captioning model (COCA fine-tuned on captioning (Yu et al., 2022a)) to compute the image captions. Then, it computes flavors pertaining to the image, and appends them to the caption. The distance between the text embedding of the flavor and image embedding of the image is used to measure how likely the image manifests the given flavor. To curate the lists of flavors, we first aggregate a large number of prompts written by users to generate images for text-to-image models. We split the prompts into words and phrases, measure its COCA distance w.r.t. the corresponding generated image, and filter out infrequent flavors. Flavors are categorized by art form (e.g. vector art), artist (e.g. Maurycy Gottlieb), medium (e.g. reddit contest winner), style (e.g. neo-primitivism, photorealistic), and other commonly-used phrases to prompt engineer images (e.g. monster manula, glamor pose, matte finish). From each potential flavor, we compute the text embeddings, and measure the cosine similarity with respect to the image embeddings of the image. Both text and image embeddings are computed with COCA (Yu et al., 2022a). We enumerate through each flavor across the categories to identify the set of flavors that maximize similarity. Flavors are organized by categories, and we ensure that at least one flavor from each category is included. Though COCA-Interrogator is used here, other image-to-text inversion methods can be used in-place, such as CLIP-Interrogator (CLIP-Interrogator) or PEZ Dispenser (Wen et al., 2023). We pursue the interrogator approach to inversion, given the high interpretability / human readability of the output, and because it enables us to explicitly include prompt-engineering flavors that govern the style and tone of the generated image.

Aspect	Query	Prompt
Religion	religious light	—An iconic statue of Buddha in a buddhist monastery. The statue is majestic and dominating in its surroundings. The statue represents peace and harmony, and the monastery is filled with monks and monks. art deco. —A depiction of a church with many people worshipping in it during evening service. The church is full of sunlight and the people are all dressed up in black. The church is ornate and the people are singing and dancing.
Culture	A woman wearing a long, flowing orange and white dress.	—A beautiful portrait of a woman in a flowing orange and white dress. The dress is long and flowing, with a deep v-neck and flowing sleeves. The background is a bright sunrise, with a lone cloud in the sky. The overall mood of the work is nostalgic and romantic. futurism. —A portrait of a young woman wearing a traditional Indian wedding dress. The woman is smiling and looking over at her husband. The dress is flowing, flowing, flowing, and the lighting is soft and romantic. futurism.
Age & Gender	angular eyebrows	—A black-and-white photograph of a man's eyes with thick, angular eyebrows. The photograph is at least two decades old, and the man is most likely still alive. hyperrealism. —A young woman with wide-open eyes and an angular, sagginess to her eyebrows. The woman is looking into the mirror, her face is radiant with confidence. trending on cg society. —A young boy has angular eyebrows with dark hair. The boy is smiling, and he is wearing a dark shirt and dark pants. The boy is standing in the grass, and he is looking at the sky with a dark grin on his face. post-impressionism.

Table 10: **Fairness from diversity**: Prompt Expansion generates prompts that introduce variations across religion, culture, and demographics at the textual level. We highlight the sentences that are indicative of the variation, and highlight an example. For exanple, despite no explicit training for specific topical diversity, Prompt Expansion responds to a query on religion by showing imagery from different religious backgrounds from Buddhism, Christianity, and Hinduism. It responds to another underspecified query "emotional conflict people" by showing people in different situations of emotional conflict and of varying age demographics. For another example, Prompt Expansion still finds small ways to introduce cultural variations that remain faithful to a long and detailed query, such as describing dresses from different global contexts.

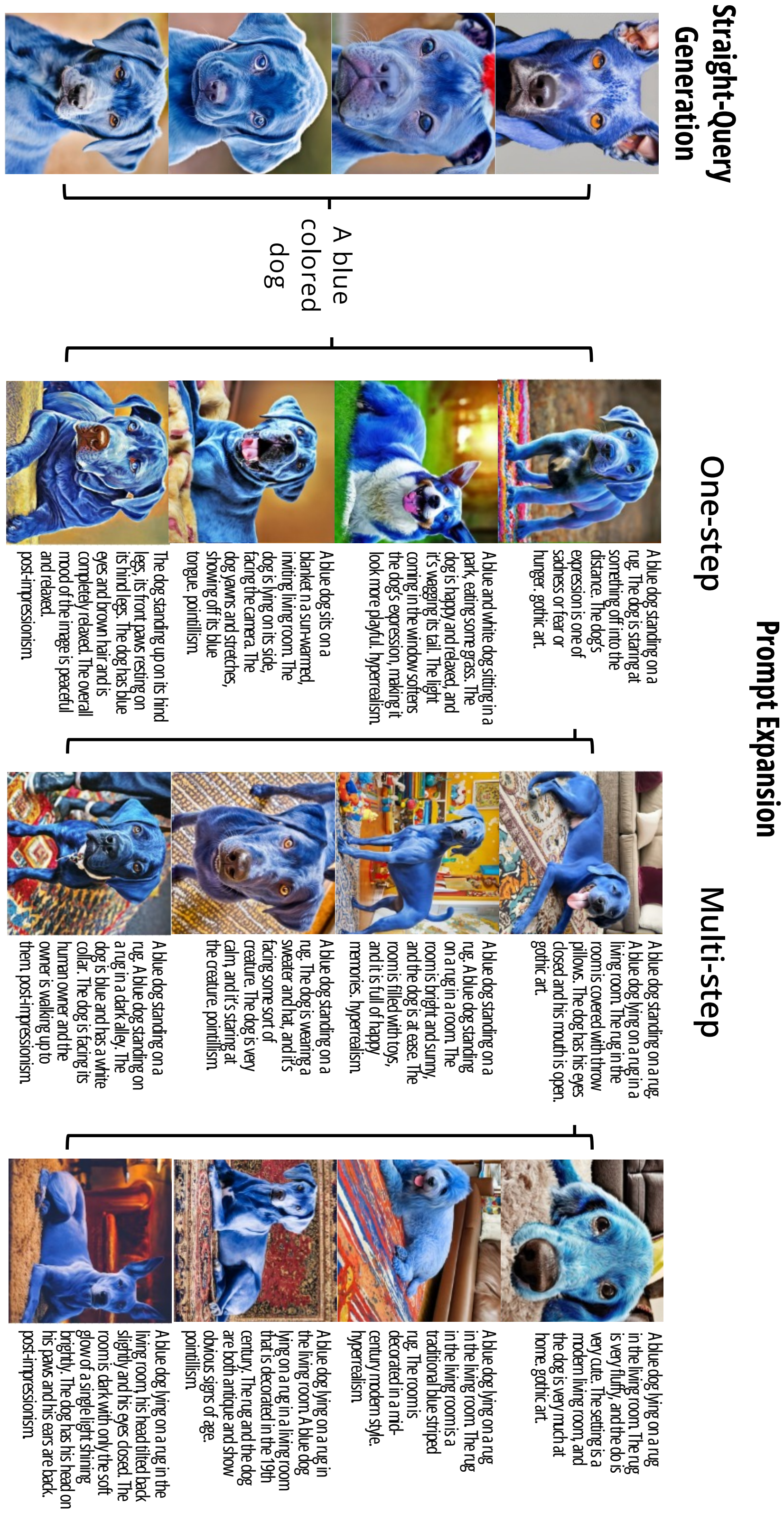


Figure 16: Tree of Multi-Step Prompt Expansion: A comparison of images generated of straight-query generation *(left)* and Prompt Expansion *(right)*. The first image from each step is further expanded upon, in-line with the implementation described in Sec C. We observe diversity in pose, action, point-of-view, background, colours.

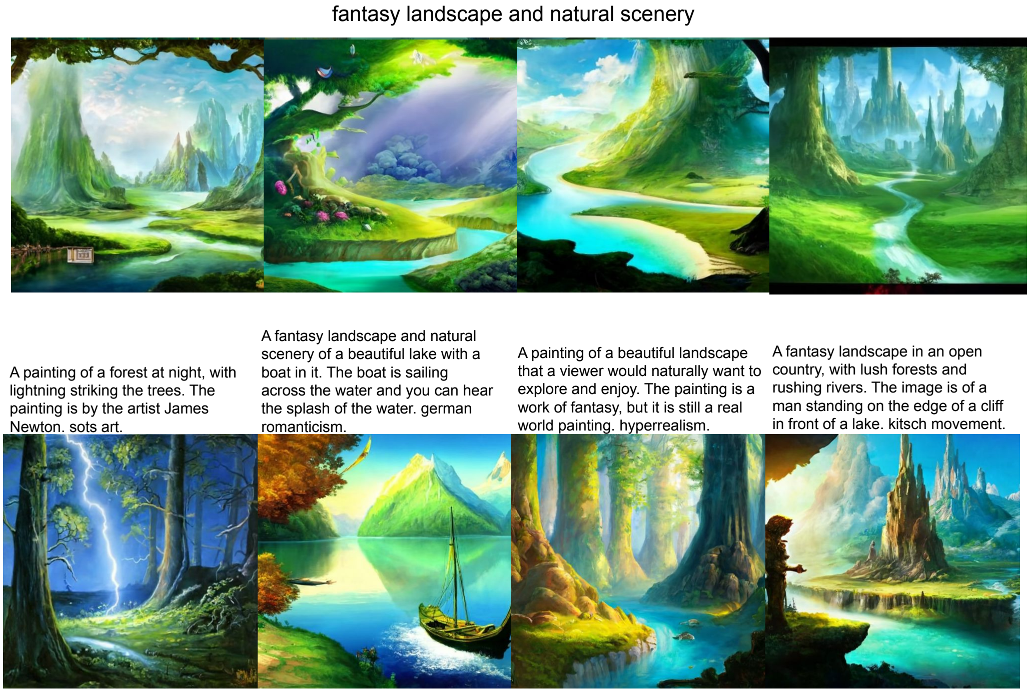


Figure 17: **Abstract-Concrete queries**: For a query that is broad in topic but contains specific details that should not be left out, straight-query generation still returns similar-looking images. Prompt Expansion returns images that remain entailed and faithful to the query, but adds variations in lighting, fine-grained details, etc.

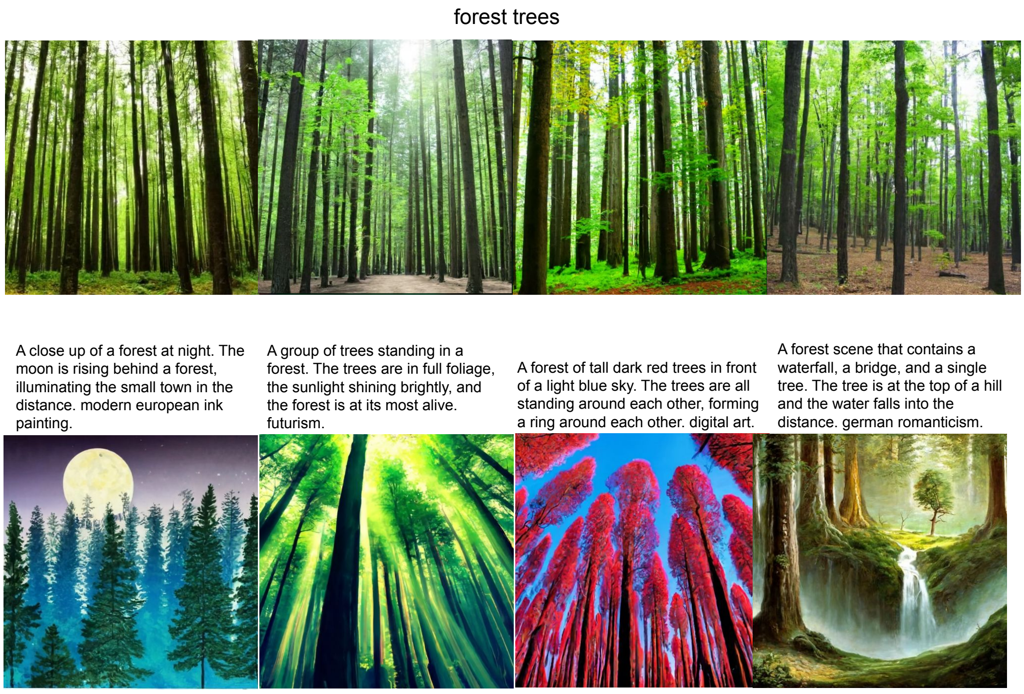


Figure 18: **Concrete queries**: For a query that is very specific in nature, Straight-Query Generation (top) returns images of the same forest setting with trees from the same angle, same species, ground vegetation, etc. Prompt Expansion (bottom) explores colours, styles of imagery, angles of view, time of day, scenery, etc.

Figure 19: **Navigational / ideation queries**: These types of abstract queries are intended to be exploratory, and the user does not have a specific image in mind necessarily. They are looking for "designs", so showing as many variations would improve the user experience.

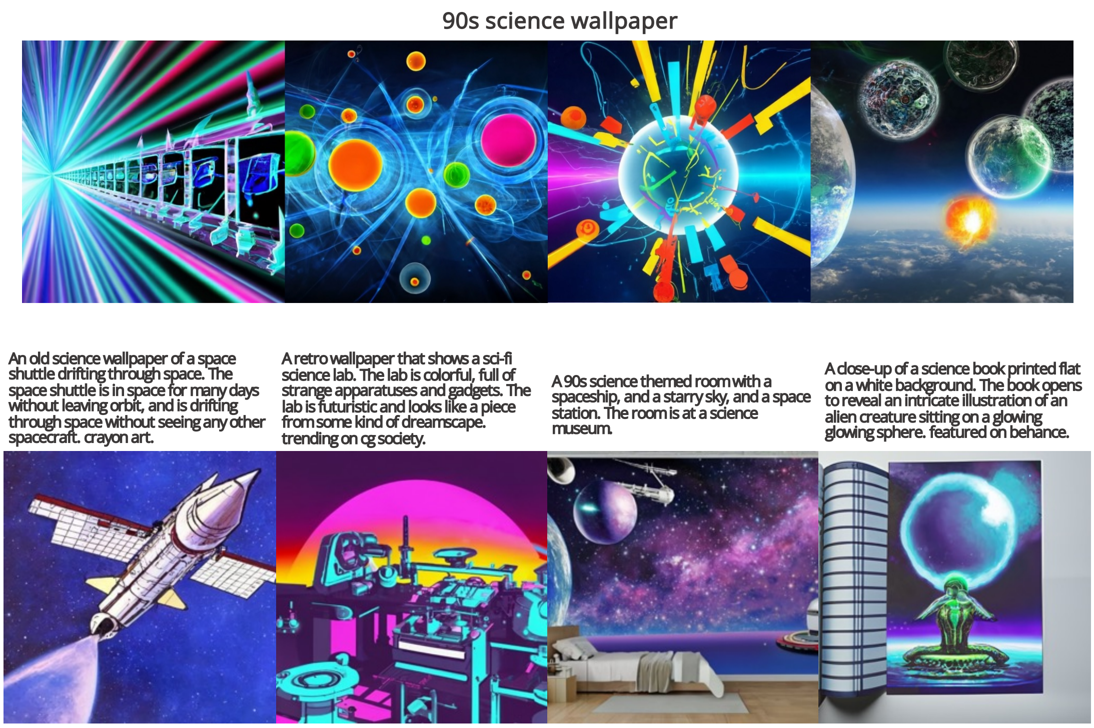


Figure 20: **Navigational / ideation queries**: The interpretation of *wallpaper* can vary by user intent. As a result, Prompt Expansion returns images across different mediums (e.g. poster wallpapers, room wallpapers). It also shows a variation of aesthetic styles referencing the 90s.

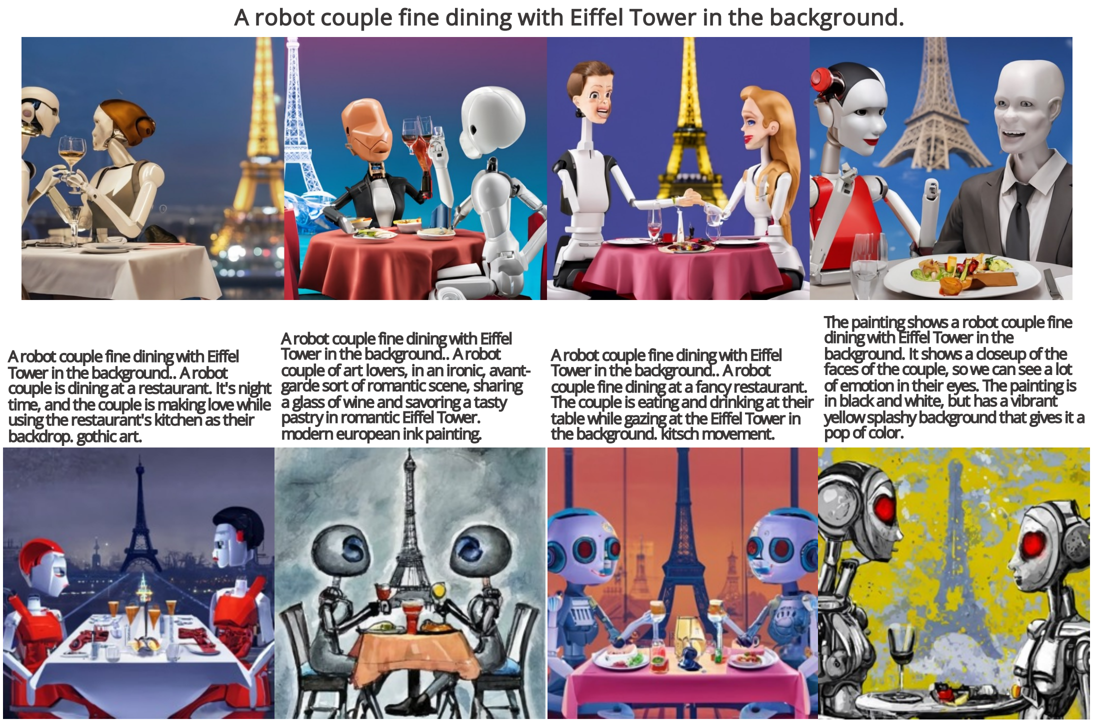


Figure 21: **Concrete / Long queries**: Using a query from PartiPrompts, we find both retainment and faithfulness to the original query, as well as stylistic variations in the subjects and environment.

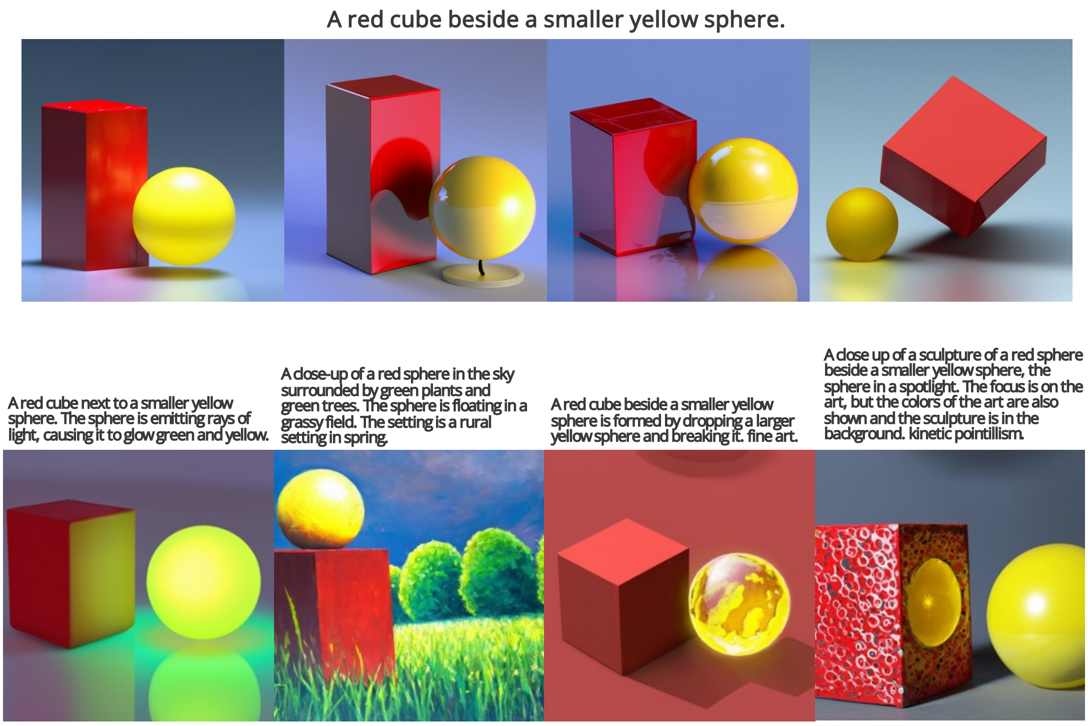
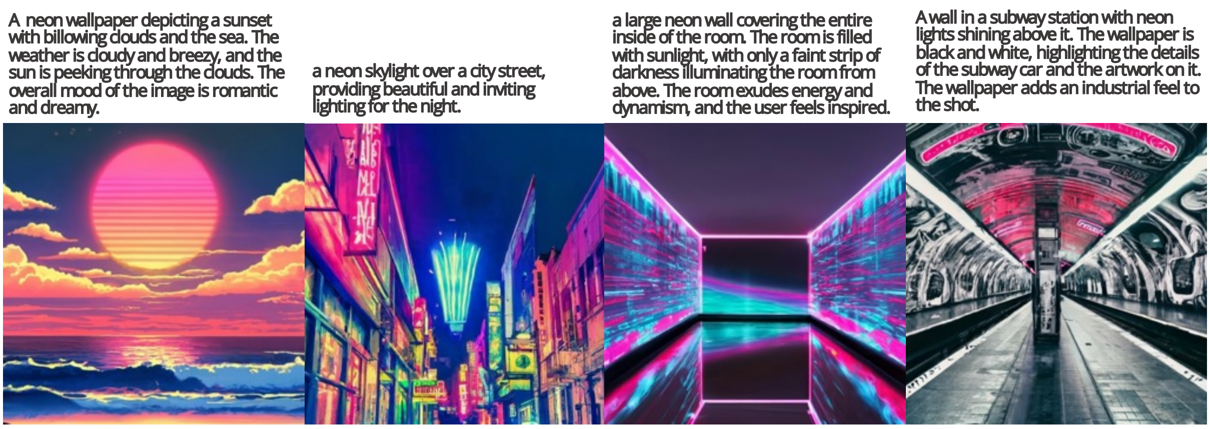


Figure 22: **Concrete queries**: Using another query from PartiPrompts, we find both retainment and faithfulness to the original query (e.g. size and colors of the objects), as well as variations in the characteristics of the objects and environment.

(a) Straight-Query Generation

(b) Prompt Expansion



(c) Prompt Expansion (re-fine-tuned)

Figure 23: We compare images generated between Prompt Expansion and PE (re-fine-tuned). Both explore different interpretations of the query, e.g. poster, city environments, rooms. The re-fine-tuned model explores more types of locations and types of *neon*.