Prompt diversification for iterating with text-to-image models

Francisco Ibarrola and **Kazjon Grace**
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
[francisco.ibarrola,kazjon.grace]@sydney.edu.au

Abstract

The recent appearance of new generative models has transformed Creative Computing, allowing for the development of striking and original art and design. Nevertheless, achieving creative objectives depends heavily on supplying particular prompts for guiding the generation process. In this work, we use semantic models and affect to develop two methods to help the prompt building process, promoting exploration and subsequent specificity. We show some results obtained with these proposals and discuss the implications to image generation.

Introduction

Creative AI has been revolutionised by the recent emergence of new generative models that can produce visually stunning works of art and design (Rombach et al. 2022; Saharia et al. 2022) from a simple text prompt. However, this process is in practice rarely one-shot, as users iteratively refine their prompt, both to communicate a specific desired outcome to the model as well as (perhaps more importantly given what we know about the creative process) to refine and explore what it is they are after (Liu and Chilton 2021). In creative settings users often struggle to articulate their vision in precise enough terms, obtaining suboptimal results from generative models. This suggests an avenue for a new kind of co-creative interaction: suggesting prompt modifications to aid in this iterative exploration process. In this paper, we propose two novel approaches to address these challenges, intended to enhance the capabilities of creators working with generative models.

The first approach has to do with helping users refine their prompts to more accurately reflect their creative intent. The idea is based on Affect modelling (Osgood et al. 1975), a psychometrically validated approach which establishes three affective dimensions (Valence, Arousal and Dominance), quantifying peoples' feelings about a wide range of stimuli, including both words and images. This can be used to provide users with refinement suggestions that are diverse in terms of affect, and hence convey different impressions, guiding their creative process more accurately. Tapping into how words "feel" as opposed to (or in addition to) their semantic meaning provides an additional vector for prompt diversification.

The second approach is image-based, allowing users to provide a second image possessing certain attribute that they desire to imbue into their generated image but cannot quite grasp the term for. By identifying these underlying key points using image semantic latents (Radford et al. 2021) and presenting them as options, we enable users to guide the generative process towards their intent more precisely. Both of our approaches allow for greater creative control over the behaviour of generative models, but are also tuned towards generating more-diverse images and increasing the potential for serendipitous discoveries and creative pivots.

In the next section, we develop these two approaches in detail, and then provide some practical examples.

Prompt Modification Suggestions

Specificity enhancement

Let us consider a text prompt $\bar{y} \in \mathbb{Y}$ provided by the user as a first draft, on which we want to improve by suggesting some additional characterisation. Additionally, let us consider a set of words $Y = \{y_1, \dots, y_N\} \subset \mathbb{Y}$ that are semantically similar to $\bar{y}$. This set can be constructed by means of the CLIP (Radford et al. 2021) encoder, which is a function $g : \mathbb{Y} \to \mathbb{R}^D$ that maps text prompts into a latent space, where similar vectors account for semantic similarity. In other words, Y can simply be built from a list of words maximising the normalized inner product

$$\langle g(y), g(\bar{y}) \rangle. \tag{1}$$

Given that all the elements in Y are close to $\bar{y}$ in the sense of (1), by transitivity they are all close to each other, and hence have some degree of semantic similarity. In order to propose meaningfully different suggestions to a user, we propose picking a subset of Y whose elements have different affect expressions.

To do this, let us consider the set of affect scores of these words, $A \doteq \{a_1, \dots, a_N\}$, where $a : \mathbb{Y} \to \mathbb{R}^3$ is a function mapping a word to its three-dimensional affect score, and $a_n = a(y_n)$. Note that we are assuming we know the affect scores of these words, which can be extracted from a dataset or estimated using an approach such as the one proposed in (Ibarrola, Lulham, and Grace 2023).

We want to extract the "most diverse" subset $\hat{A} \subset A$ of K words to propose as enhancement options to the user. This

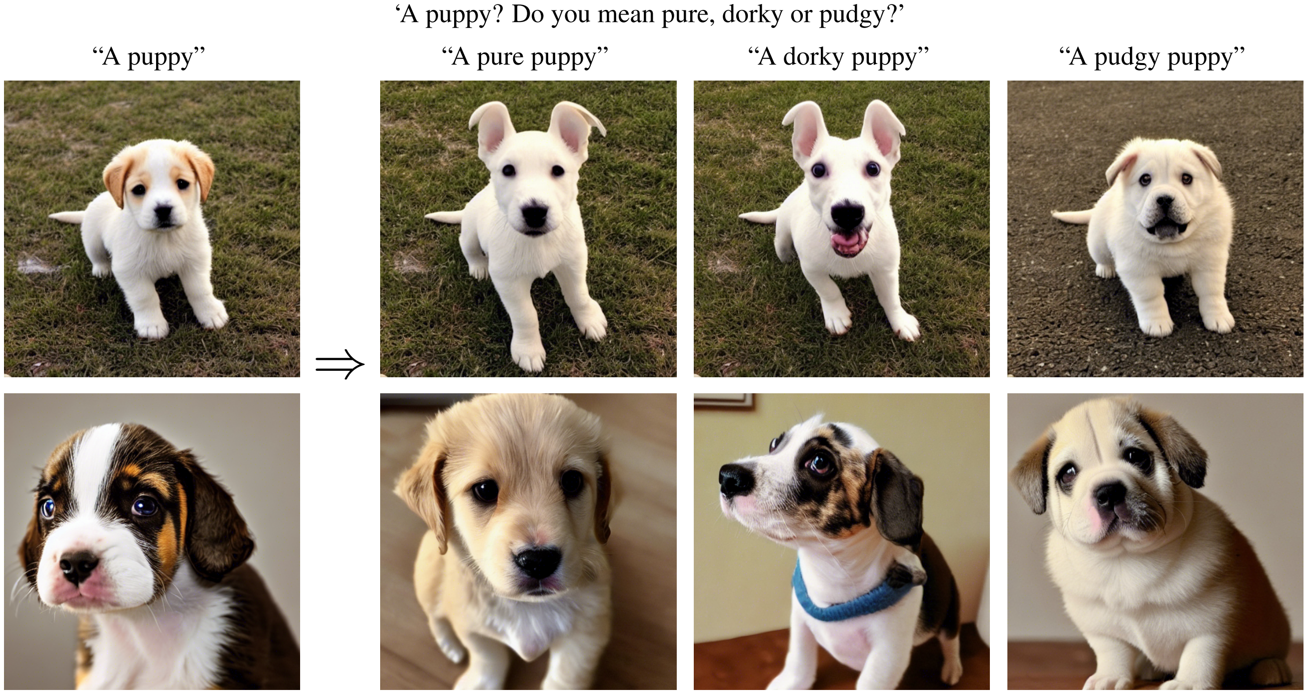

Figure 1: Illustration of two images generated by Stable Diffusion from the prompt “A puppy” (on the left), and those obtained after incorporating the enhancement suggestions made by our first approach (which in this case suggests “pure”, “dorky”, and “pudgy” as possible modifications. Each row was generated from the same random seed.

notion of diversity can be defined in many ways, depending how we choose to quantify it, and in this case we choose the largest minimum distance between the elements of a set. That is

$$\hat{A} \doteq \text{arg max}_{a_1,\ldots,a_K}(\text{min}_{k\neq j}\|a_j - a_k\|).$$

Or, in other words, the subset of words for which the *most* affectively similar pair of words between them is as *dissimilar* as possible. Given that finding $\hat{A}$ according to this definition is intractable for large values of K, we propose to find an approximation using Algorithm 1.

Algorithm 1 Word selection

```
Initialization
  Let $A_0$ be a random subset of $A$, of size $K$
  $\hat{A} \leftarrow A_0$
Search
  for $b \notin A_0$
    $\hat{a} = \text{argmin}_{a\in\hat{A}}\|a - b\|$
    $m_{\hat{a}} = \text{min}_{a\in\hat{A}\setminus\{\hat{a}\}}\|a - \hat{a}\|$
    $m_b = \text{min}_{a\in\hat{A}\setminus\{\hat{a}\}}\|a - b\|$
    if $m_b > m_{\hat{a}}$
      $\hat{A} \leftarrow \hat{A} \cup \{b\}\setminus\{\hat{a}\}$
    end if
```

We can then use the words associated to $\hat{A}$ to present the user with options for modifying the prompt, either through a traditional UI or through a language model.

Image-driven modifiers

We consider the problem of a user who wants their generated image to be more like another target image they have seen, but in a very specific way that may not be obvious to them or easy to put into words.

Let $x_0 \in [0,1]^{3\times M\times M}$ be the (pixel) matrix associated to the current state of the generated image, and let $x_t \in [0,1]^{3\times M\times M}$ be the target image. Then, the problem can be stated as sampling from a distribution

$$\pi(x|x_0, h(x_t)),$$

where h is a feature extraction function that should isolate the aspects of the image on which the user is actually trying to condition the output. In order to discern the aspect the user is seeking to imbue in x, we can use the CLIP image encoder f (as well as the corresponding text encoder g) to figure out which words can be associated with x_t but not with x_0. That is, given a large set of available words Y, we seek a subset maximising

$$\langle g(y), f(x_t)\rangle - \langle g(y), f(x_0)\rangle, \tag{2}$$

w.r.t. y. By presenting the user with a set of words $\hat{Y} \subset Y$ maximizing Equation 2, we can get their choice, and then use it as conditioning input, thus making human decision a component of h.

From here on, we can use $\hat{y} = h(x_t)$ as the conditioning input. Alternatively, if the generative model has joint latent space for images and text, build the conditioning input as the projection of the latents as follows

$$z \doteq f(x_t) \cdot g(\hat{y}) \frac{g(\hat{y})}{\|g(\hat{y})\|}.$$

It is timely to mention that we can combine this proposal with Algorithm 1 by taking A as the set of affect scores associated to $\hat{Y}$. The suggestions presented to the user would thus become an affectively diverse subset of the descriptive words that matched the target image but not the current one. While we have not yet conducted any user studies with these techniques, this could help providing a more varied set of suggestions should the set of available words Y contain too many synonyms.

Results

For the following experiments we used the adjectives from the word dataset developed in (Warriner, Kuperman, and Brysbaert 2013), which contains word classifications into nouns, adjectives or verbs, and their corresponding affect scores.

Specificity enhancement

For the first experiment we tested enhancement suggestions provided by Algorithm 1 with five different prompts. The obtained results are described as follows, in the format that an interface may use to propose the suggestions.

- A puppy? Do you mean pure, dorky or pudgy?
- A meal? Do you mean healthy, appetizing or nutritious?
- A chair? Do you mean random, disabled or quick?
- A dragon? Do you mean righty, gorgeous or beastly?
- A king? Do you mean sensible, solid or powerful?

Some of these suggestions may be considered very good to help narrowing down the user's intentions regarding the output, while some others may be a little strange. Nonetheless, surprise is a good indicator of the creative potential of an interaction, and may lead to explore new possibilities.

In order to illustrate the complete process, we took one of the prompts and suggestions and used Stable Diffusion (Rombach et al. 2022) to generate some samples, shown in Figure 1. It can be seen that adding each suggestion does steer the drawing in a distinctive direction while retaining at least some aspects of the original image. The degree to which this is useful awaits further evaluation, but to the authors the dorky puppies are at least a bit dorky and the pudgy puppies are at least a bit pudgy.

Image-driven modifiers

In order to test the proposal of modifications through suggestions derived from a target image, we picked image pairs of objects in the same category, and produced five suggestions using Equation 2. Some of the results are shown in Figure 2, where there are two observations to be made.

Firstly, the suggestions seem pertinent and do reflect characteristics of the target images not observed on the current one.

Secondly, we compared the results obtained after modifying the prompt using one of the words suggested by our method (under "Image-driven prompt modifier") and those of guiding Stable Diffusion with two images (under "image mixture"). The naïve approach of setting both images as targets has, in the dog example on the left, introduced some unwanted changes (such as the background flowers) along with visual changes such as lightening the dog's fur. In the chair example the direct image mixture appears to have performed better, perhaps due to the absence of background detail. On the other hand, guiding the generation process by introducing a specific characteristic to the prompt, derived from the target image, results in changes much more aligned with that aspect. In the dog example on the left it is clear that "fluffiness" has increased without (significantly) altering the dog, pose, or setting. The chair results are somewhat more mixed, although here perhaps the task was harder, as antique chairs are typically not visually similar to modern moulded plastic ones.

Conclusions

In this work, we focus on the ability of a co-creative system based on generative AI to make diverse suggestions and aid its user in the task of iterative prompt exploration. We have proposed two different methods to do so in the prompt space, one based on affective modelling of words, and one based on extracting target aspects from images. Additionally, early experimental results were presented, highlighting the potential value of co-creative prompt suggestions.

It is worth mentioning that the proposed approaches for prompt improvement are generator-agnostic, meaning that they can be used with any prompt-based generative model. This constitutes a considerable asset given the rate at which new generative models are being developed.

Finally, there is still much work to be done regarding user testing. On one hand, interaction design work will be required to determine effective ways of presenting the options to users. On the other hand, we have only begun exploring the reach and limitations of these approaches.

Acknowledgments

We would like to acknowledge the Australian Research Council for funding this research (ID #DP200101059).

References

Ibarrola, F.; Lulham, R.; and Grace, K. 2023. Affect-conditioned image generation. *arXiv preprint arXiv:2302.09742*.

Liu, V., and Chilton, L. B. 2021. Design guidelines for prompt engineering text-to-image generative models. *arXiv preprint arXiv:2109.06977*.

Osgood, C. E.; May, W. H.; Miron, M. S.; and Miron, M. S. 1975. *Cross-cultural universals of affective meaning*, volume 1. University of Illinois Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from

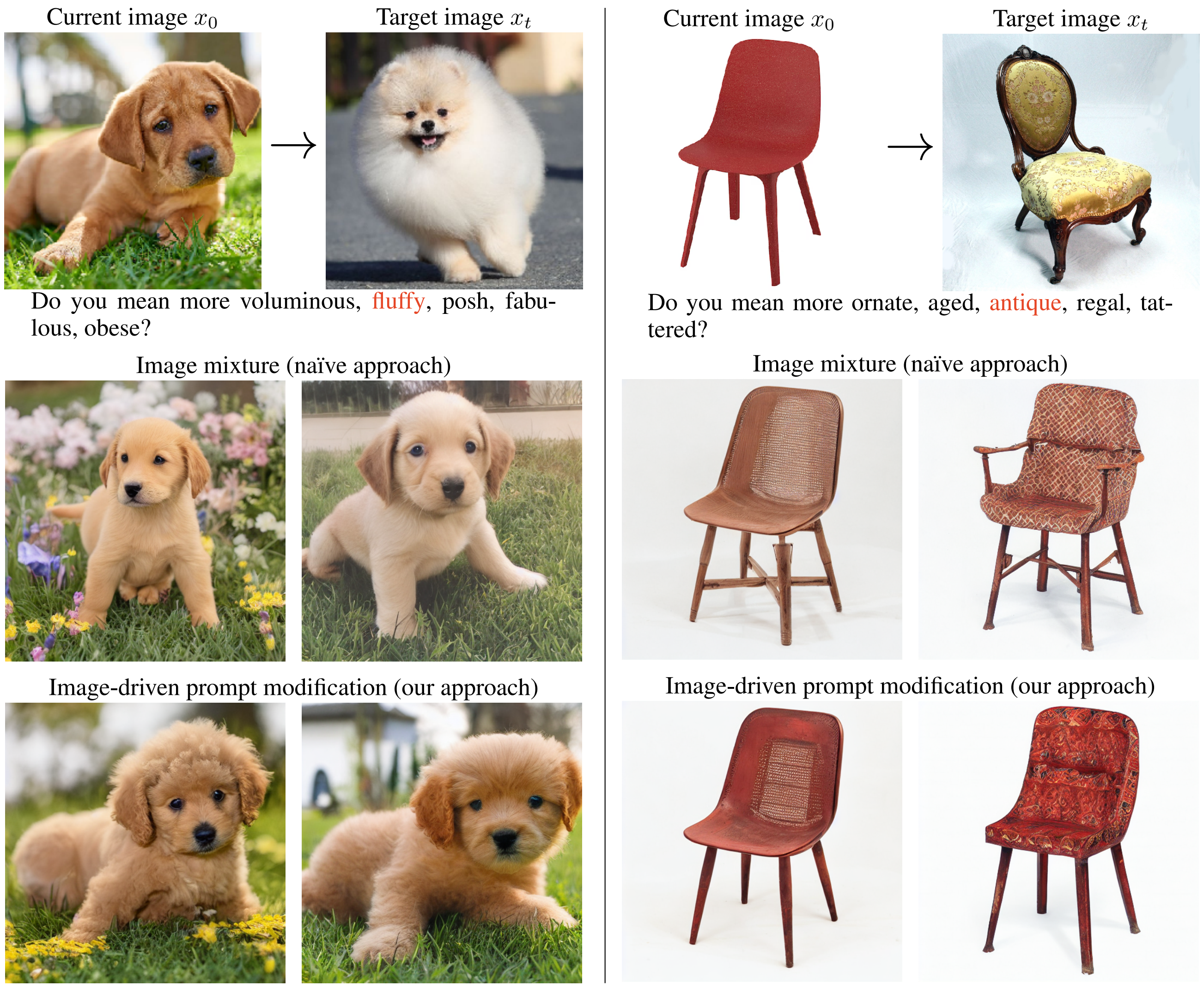
Current image x_0
Target image x_t
Do you mean more voluminous, fluffy, posh, fabulous, obese?
Image mixture (naïve approach)
Image-driven prompt modification (our approach)
Current image x_0
Target image x_t
Do you mean more ornate, aged, antique, regal, tattered?
Image mixture (naïve approach)
Image-driven prompt modification (our approach)

Figure 2: Two examples of the image-driven prompt modification process with Stable Diffusion. The top row shows the current image and a target images, along with the modifiers suggested by the system when presented with this image pairs. Under "Image Mixture", we show the results obtained with the original prompt and the two images as simultaneous targets. Under "Image-driven prompt modification", the results obtained with the current image and the prompt modified according to the highlighted suggestion.

natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35:36479–36494.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* 45:1191–1207.