# Can generative models generate novel objects the same as familiar objects?

Zhaokun Xue*[1], Chang Ye[1], Gamaleldin F. Elsayed[1], Junfeng He[1]

[1]Google

## Abstract

*Text-to-image generative models have demonstrated great performance in generating realistic images. These generations are assumed to reflect a deep understanding of visual scenes. One interesting question is whether these models can possess a zero/few shot generalization capabilities that are known from humans. For example, a human can see an example of a new object and a word associated with this object, use their knowledge in a highly general way to recognize or imagine this novel object in a completely different setting or context. In this work, we are interested in testing whether text-image models can possess this same capability. In this work, we would like to test the hypothesis that text-to-image models may learn familiar objects better than novel objects. We use prompt tuning methods to learn soft token representations for those novel concepts while keeping the text-image models fixed. We prompt tune the model as well to learn familiar concepts, and evaluate the generalization ability for novel objects compared to familiar objects by running generation in different contexts/environments. In addition, instead of initializing the embedding vectors with some similar concepts, we use randomly initialized embedding vectors for both familiar and novel objects. Our human-survey evaluation result demonstrates that in some settings text-image models learn familiar objects better than novel objects.*

## 1. Introduction

Text-to-image generation has rapidly made significant progress with the development of large-scale and pre-trained models [1, 8, 10, 12, 14–17]. These models leverage extensive training data to create high-quality images from text description, making them useful in applications ranging from design to media creation. Despite these achievements, significant challenges still persist, particularly when dealing with novel objects that are hard to describe in concise text. The ability to effectively represent and generate such novel objects is important for improving the capabilities of text-to-image models.

This paper tests the hypothesis that text-to-image models may learn familiar objects better than novel objects by extending the *textual inversion* method introduced by Gal et al. [4]. While the original approach uses a single embedding vector to represent a common object, we propose an innovative extension that leverages multiple embedding vectors that allows us to capture more complex objects' features, especially those novel objects that cannot be simply described in a single token. In addition, instead of initializing the learnable embedding vector with a single-word descriptor like "cat", "mug", etc., we initialize the learnable embedding vector randomly with tokens from the embedding vocabulary, which ensure the results are not retrieved from similar closely related concepts from the initial descriptors.

We evaluate our approach with the Novel Object and Unusual Name (NOUN) Database [6], which provides a diverse set of objects with human-annotated familiarity and name-ability scores. We test our hypothesis on comparing the performance of our method on high familiar objects with less familiar (novel) objects in this dataset. To improve the variation of the objects in the dataset, we apply data augmentation techniques such as scaling, translation, flipping, and rotation, as detailed in Section 4.1.1, which introduces more variations for each object. We use the T5X model [11] as our text encoder. The encoded embedding vectors are fed into the Imagen model [14] for image generation.

The key contributions of the paper include:

1. We introduce an innovative method for representing objects in text-to-image models using multiple randomly initialized embedding vectors.
2. We demonstrate that in some settings the text-to-image model learns generalization of familiar objects better than novel objects via a human-survey evaluation.

## 2. Related Works

**Zero/Few-Shot Text-to-Image New Concept Learning** Research of leveraging the autoregressive models[3, 9, 10, 16] and pre-trained text-to-image diffusion models[2, 5, 7,

*Corresponding author: zhaokun@google.com

8, 12–14] has been widely studied for image synthesis. For example, DALL-E [10] has demonstrated the capability of doing text-to-image generation with zero-short. Textual Inversion[4] and Dreambooth[13] can generate new concepts with a few provided images by finetuning the text-to-image diffusion model.

**Textual Inversion and Prompt Tuning** Previous work[4] using textual inversion has shown that text-image models can learn soft input tokens from a few examples and generalize the generation of concepts in those examples to different settings. However, it is unclear if the learned objects are truly novel or just close to existing objects in the training set. Additionally, the learned tokens were initialized with token embedding similar to the concept being learned (e.g. initialized with 'cat' token while learning 'cat statue'), which could mean the observed generations are retrieved rather than generalized. Besides doing text-based tuning, [18] achieves text-free image generations by tuning the CLIP[9] embedding space to a shared text-image embedding.

**Gap in studying novelty** Previous methods found that it is possible to inject new concepts and object, but few studies have systematically evaluated whether there is a gap on how text-to-image models handle common vs. novel objects. Our work addresses this by proposing a multi-vector embeddings approach for learning newly introduced objects—using randomly initialized embedding rather than reusing existing tokens to more accurately measure genuine learning instead of retrieval from the already learned concepts during training. Moreover, we compare performance on objects with high familiarity against those that are highly novel, studying systematic differences in the model's capacity to learn and compose them in diverse contexts. This investigation provides new insights into how familiarity influences few-shot learning within text-to-image systems.

## 3. Methods

Our approach is inspired by the Text Inversion method in [4], where a single embedding learnable vector is used to represent a given object. We propose a novel extension that leverages multiple-vector embeddings to represent novel objects from the Novel Object and Unusual Name (NOUN) Database [6] shown in Figure 1. Specifically, we utilize five embedding vectors for each object, which balances image quality and computational efficiency. Using multi-vector embeddings allows the model to better capture the diverse attributes of objects, particularly for the less *familiar* and *name-able* objects in the NOUN database.

Unlike [4], where the learnable token embedding is initialized with a single-word descriptor of the input object(e.g., 'mug,' 'cat'), we initialize our learnable embedding vectors by randomly selecting from the text encoder vocabulary. This randomization approach not only addresses the challenge of using a single general descriptor to summarize novel objects, but also ensure the generations are retrieved from similar concepts.

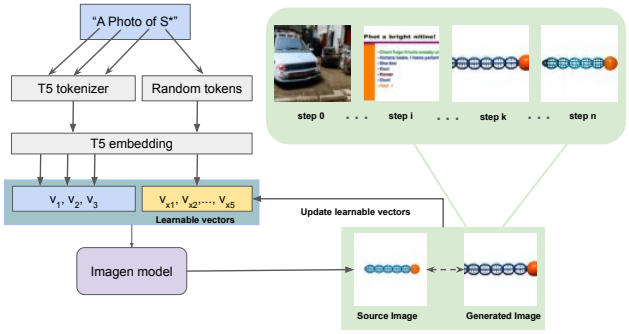


Figure 1. Overview of our approach: A prompt tokenized by the T5 encoder vocabulary. The learnable "words" (S*) in the prompt are initialized with randomly selected token IDs from the T5 vocabulary space. Then, these tokens are transformed into embedding vectors via T5, and fed into the Imagen model to generate the image. We utilize the Imagen loss function to evaluate the generated image against the source image, which is used to guide the updates of the learnable vectors.

We adopt the T5X model [11] as the embedder for the textual encoding step. Like other text encoder models, each input word is transformed into an index that corresponds to a unique dense embedding vector in a pre-defined dictionary. The input prompts are processed using the Transformer-based self-attention mechanism within T5X, generating the text embedding vectors. This embedding space also serves as the basis for finding new embedding vectors to represent objects the NOUN database. The generated embedding vectors are passed to the downstream pre-trained text-to-image model.

We use the Imagen model [14] as the text-to-image model to generate the images. To optimize the quality of the generated images, we minimize the loss function associated with Imagen Model, which evaluates the similarity between the source image and generated image. The loss result guides the updates for the learnable embedding vectors. We use *adafactor* optimizer to optimize the learnable embedding vectors.

## 4. Experiments

### 4.1. Dataset

We utilize the Novel Object and Unusual Name (NOUN) Database [6], which contains 64 principle novel objects. Each object is annotated with two key metrics: a *Familiarity Score* and a *Name-ability Score*, both scores are collected by human surveys. The *Familiarity Score* reflects how familiar the objects are to people, and the *Name-ability Score*

measures how consistently people can name them. For this study, we selected 40 objects from the database, which are divided into two groups. The first group contains the top 20 objects with the highest *Familiarity Score*, representing the most familiar items in the database. The second group is the bottom 20 objects with the lowest *Familiarity Score*, representing the least familiar and more novel items. The selected objects are listed in Supplementary Table 6. Each object's image was augmented using the data augmentation process described in Section 4.1.1 before being passed to the pipeline.

#### 4.1.1 Data Augmentation

To enrich the variation of each object in the selected dataset, we applied various data augmentation techniques:

- **Scaling**: Each object image was randomly resized to a scale ranging from 50% to 100% of its original size.
- **Translation**: The object's position within the image was randomly shifted along both the x and y axes.
- **Horizontal Flipping**: The object's left-right orientation was randomly flipped.
- **Rotation**: Each object was randomly rotated at varying angles.

These transformations produced **12** variations of each object, resulting in a more diverse set of augmented data. More examples are shown in Supplementary 7.

### 4.2. Experimental setup

We began by running experiments to determine the optimal number of embedding vectors to use to represent objects in the dataset. Details of the experiments setup and results can be found in Supplementary 8. We then evaluated whether our method can re-create the objects with the neutral prompt *"A photo of S*"*. Sample outputs are shown in Supplementary 9.1. Next, to test generalization of the new learned objects in new contexts, we generated images with three distinct styles and seven distinct backgrounds using the new learned embedding for each object. The selected styles and backgrounds are shown in Table 1. We used the prompt template, *"A or An {style} of a S* in the center of {background}"* to guide the Imagen model's image generation, where *S** was replaced by the learned representative embedding vectors during the text encoding process. For each combination of the style and the background, 16 images were, provided for evaluating how effectively the model recreates the novel and familiar objects over different contexts.

### 4.3. Qualitative examples

Figure 2 shows some examples that's generated with our method under new scenes. We observed that when generating images with different scenes and styles, the object integrates better with relatively simple backgrounds like beach, rain, river and snow. However, with more complicate backgrounds such as *city, park, and space*, the results are inconsistent. The image generation model fails in some cases for both novel and familiar objects in these complicate backgrounds, where the model may attempt to replace a similar object that can be better fit in these backgrounds. The model successfully generates *DSLR* style images for both familiar and novel objects; however, for the *oil painting* and *surreal digital art* styles, it tends to produce higher-quality results for familiar objects. More successful and failure examples and discussions are included in the Supplementary 9.2.

| | |
|---|---|
| Style | DSLR photo, oil painting, surreal digital art |
| Background | beach, city, park, rain, river, snow, space |

Table 1. 3 styles and 7 backgrounds that are used to compare the ability to recreate images in various contexts for both familiar and novel objects.

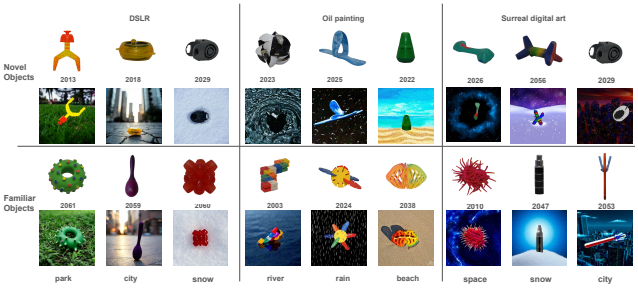


Figure 2. Selected Sample outputs for images generated for different scenes. The images are generated with prompt "A or An {style} photo of S* in the {background}", where the left part shows sample outputs from *DSLR* style, the middle is the sample outputs for the *oil painting* style, and the right part shows the sample outputs for the *surreal digital art* style. The top half is the sample outputs for novel objects and the bottom half shows the examples for familiar objects.

### 4.4. Evaluation

To evaluate and compare the generated images for both familiar and novel objects, we conducted a human rating study. Participants were asked to express their preferences between paired generated images, each containing one object from the familiar set and one from the novel set, generated in the same style and background. The survey form was structured as follows:

For each style listed in Table 1, we randomly shuffled the familiar and novel objects lists to ensure the randomization before creating the pairs. After pairing, the paired list was shuffled again to ensure a randomized presentation. Also, to minimize potential bias from human associations, we ensured that identical objects were not shown consecutively

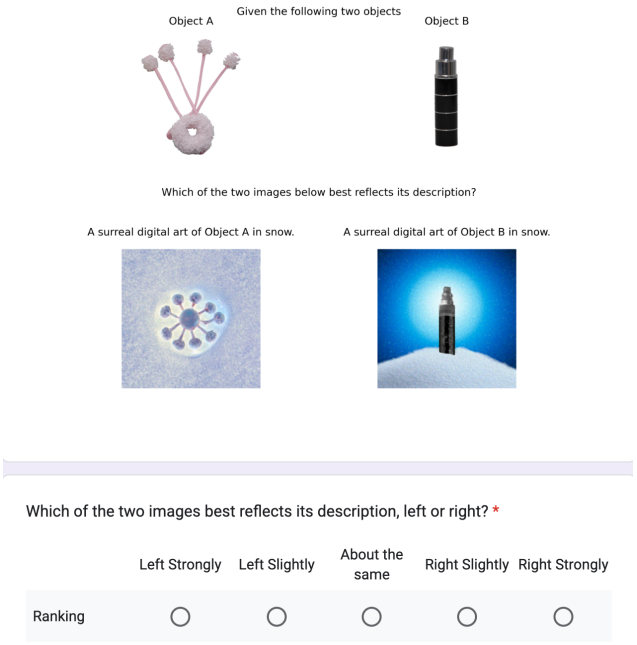


Figure 3. Sample question presented in the survey form, where participants rate two generated images based upon the text prompt and the original objects shown at the beginning.

in the survey. This approach ensured a well-distributed and unbiased presentation order. For each pair of objects, the familiar and novel objects were randomly assigned to either the left or right position, and one pair of style and background was randomly selected for each pair of objects. Based upon the style, background, and object, we randomly chose one generated image. We also ensured that no generated image was repeated in the same form during the study.

Each participant was asked to evaluate 60 pairs of images in total, with 20 pairs for each style. Participants were shown the original objects, along with the text prompt that is used to generate the images for each question. Then, we ask participants to provide their preferences for *"Which image more faithfully reflects its description, A or B?"* as illustrated in Figure 3.

We generated 4 distinct survey forms with different initial random states. We collected responses from 16 participants. In the evaluation, participant responses were scored as follows: "Strongly left/right" was assigned a score of 1, "Slightly left/right" was assigned 0.5, and "About the Same" was scored as 0. The "About the Same" option indicates that the paired images were considered as either equally good or equally bad. Figure 4 presents the distribution of the weighted results of the participants' preferences for the generated images of both novel and familiar objects.

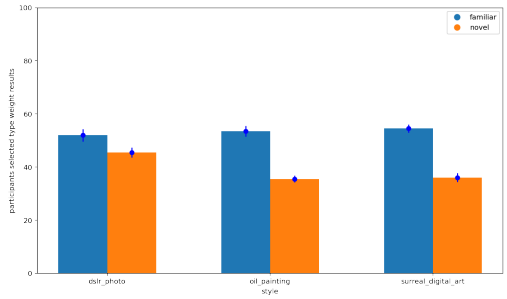


Figure 4. The distribution of weighted results based upon participants' preferences, grouped by the generated images' styles, with standard deviation as error bars, reflecting the for each style that participants agree that the generated images of familiar objects are better than the novel ones.

| Style | Wilcoxon p-value |
|---|---|
| DSLR photo | 0.7209 |
| oil painting | 0.0342 |
| surreal digital art | 0.0587 |

Table 2. P-values from the Wilcoxon signed-rank test comparing participants ratings for generated images of familiar and novel objects grouped by styles.

We performed the Wilcoxon signed-rank test to analyze the statistical differences between the generated images of novel and familiar objects. The results are summarized in Table 2. For "oil painting" style, participants rated that the generated images for familiar objects are significantly better than those of novel objects. For "surreal digital art" style, the generated images of familiar objects tends to be better than those of novel objects. For "DSLR photo" style, the quality of the generated images tends to be the same. These quantitative analysis results match the observations in the qualitative examples in Section 4.3.

Taken together, those results demonstrate that the Text-to-image generative models we studied here learns to generalize concepts and objects that are familiar better than those that are more novel.

## 5. Conclusion

In this work, we extend Textual Inversion[4] by using multiple randomly initialized embedding vectors to represent an object. By testing on the NOUN dataset, we examine the hypothesis that the text-to-image model we used learns generalization of familiar objects better than novel objects. Human rating surveys confirm that these models generalize familiar concepts more effectively than novel ones. In future work, we aim to enhance the model's performance in handling complex backgrounds, such as city, park, etc. and investigate strategies to improve the generation quality of novel objects.

## References


[1] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 1

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 4

[5] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 1

[6] Jessica S Horst and Michael C Hout. The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, 48:1393–1409, 2016. 1, 2

[7] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 1

[8] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[10] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 2

[11] Adam Roberts, Hyung Won Chung, Gaurav Mishra, Anselm Levskaya, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. Scaling up models and data with t5x and seqio. *Journal of Machine Learning Research*, 24(377):1–8, 2023. 1, 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[13] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[15] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023.

[16] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1

[17] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 1

[18] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917, 2022. 2

# Can generative models generate novel objects the same as familiar objects?

## Supplementary Material

| | Selected Categories |
|---|---|
| Familiar | 2003, 2004, 2009, 2010, 2012, 2017, 2024<br>2032, 2034, 2036, 2038, 2042, 2047, 2051, 2052<br>2053, 2059, 2060, 2061, 2062 |
| Novel | 2001 ,2005, 2013, 2018, 2021, 2022<br>2023, 2025, 2026, 2028, 2029, 2030, 2033<br>2035, 2044, 2048, 2049, 2054, 2055, 2056 |

Table 3. 20 familiar and 20 novel objects' categories from the NOUN database that are used in our experiments.

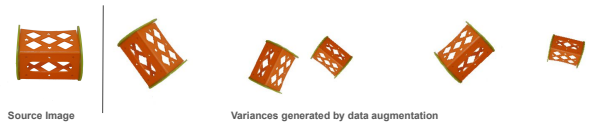


Figure 5. Examples of the source image of Object 2033 and variances generated via data augmentation.

## 6. Selected Familiar and Novel Objects' Categories

Table 3 lists the 20 familiar and 20 novel objects selected to use in our study.

## 7. Data Augmentation Sample Outputs

Figure 5 shows several variations generated via our data augmentation method for Object 2033.

## 8. Determining the Optimal Number of Embedding Vectors for Object Representation

To ensure the number of embedding vectors can effectively capture the variability of objects with different levels of familiarity and name-ability scores, we selected three object categories from the NOUN database: Category 2038, Category 2044, and Category 2053, as shown in Figure 6. These categories were chosen based on their distinct scores. Category 2038 has the highest familiarity score (56%); Category 2053 has the highest name-ability score (79%); Category 2044 has the lowest familiarity score (6%) and the lowest name-ability score (27%).

We trained the model for 5000 steps, repeating this process for 10 times for each configuration of embedding vectors (1, 5 and 10 vectors). At the first step, the embedding vectors were randomly initialized. To constrain the generated images, we randomly sample 27 neutral prompts from

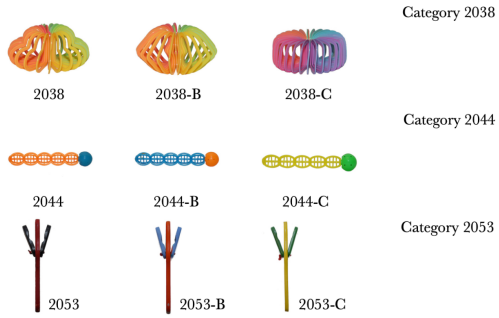


Figure 6. Objects from the NOUN database that are used to determine the optimal number of embedding vectors for representing objects in the dataset.

the CLIP ImageNet templates [9]. Here is the full list of neutral prompts that we used:

- a photo of a S*.
- a rendering of a S*.
- a cropped photo of the S*.
- the photo of a S*.
- a photo of a clean S*.
- a photo of a dirty S*.
- a dark photo of the S*.
- a photo of my S*.
- a photo of the cool S*.
- a close-up photo of a S*.
- a bright photo of the S*.
- a cropped photo of a S*.
- a photo of the S*.
- a good photo of the S*.
- a photo of one S*.
- a close-up photo of the S*.
- a rendition of the S*.
- a photo of the clean S*.
- a rendition of a S*.
- a photo of a nice S*.
- a good photo of a S*.
- a photo of the nice S*.
- a photo of the small S*.
- a photo of the weird S*.
- a photo of the large S*.
- a photo of a cool S*.
- a photo of a small S*.

We generated one 256 x 256 image for each prompt, which then are mapped to the CLIP space using the VIT_B32 model. We evaluated the generated images by computing the pairwise cosine similarity between the generated images

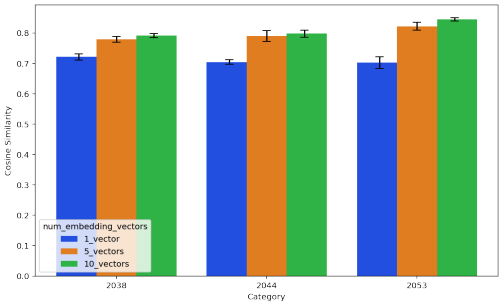


Figure 7. Averaged cosine similarity across 10 repeated experiments for each number of embedding tokens that is used for representing the object, with error bars representing the 95% confidence interval.

and the original source images, and use the averaged score as the final measurement for performance.

This thorough evaluation allows us to assess how well different numbers of embedding vectors performed across objects with varying familiarity and name-ability scores, which ensures our method is robust across a diverse range of objects types. The experimental results are shown in Figure 7.

To statistically evaluate the differences of varying the number of embedding vectors, we performed an one-way ANOVA (Analysis of Variance) on the results obtained using 1, 5, and 10 embedding vectors for Category 2038, 2044, and 2053. The p-values from the ANOVA in Table 4 indicate that there at least one pair is significantly different from the other pairs in each category. As shown in Figure 7, both 5 and 10 embedding vectors yield better results compared to using just 1 vector. To further investigate the differences, we conducted independent t-tests between using 5 and 10 embedding vectors of each category. The t-test results in Table 4 show significant differences in performance for categories 2038 and 2053, which represents more familiar and easily name-able objects. This suggests that increasing the number of embedding vectors improves representation for these categories. However, for Category 2044, which represents more novel and less familiar objects, there was no significant improvement when increasing the number of embedding vectors.

Given the goal of this study is to find a balanced and general approach for both familiar and novel objects, we conclude that using 5 embedding vectors provides a sufficient trade-off between the performance and computational efficiency for both familiar and novel objects. Employing 5 vectors effectively captures objects variability while minimizing the computation time and resource costs. Additionally, using more tokens may affect the model's capabilities of generalization. A larger number of tokens could introduce noisy or redundant information, which may impact

| Category | one-way ANOVA | t-test(5 and 10 vectors) |
|---|---|---|
| 2038 | 2.847e-11 | 0.0402 |
| 2044 | 1.153e-10 | 0.4852 |
| 2058 | 3.026e-14 | 0.0056 |

Table 4. P-values from the one-way ANOVA analysis comparing the use of 1, 5, and 10 embedding vectors across Category 2038, 2044, and 2053, alongside p-values from the t-test comparing the use of 5 and 10 embedding vectors for the same categories.

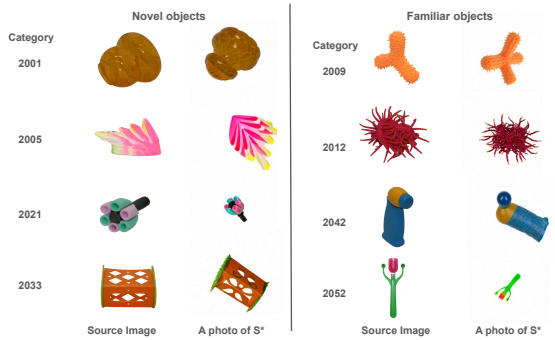


Figure 8. Sample outputs for images generated by neutral prompt "A photo of S*" for both familiar and novel categories. In general, we can recreate high-similar objects for both familiar and novel categories.

the robust generalization across the diverse objects in the dataset. For simpler objects, this could potentially result in overfitting. As part of future work, we plan to conduct a more thorough investigations in comparing the impacts of using longer and shorter token representations.

## 9. Supplementary for Qualitative Examples

### 9.1. Generated Images from Neural Prompts

In Figure 8, we show some sample results generated by prompt "A photo of S*" from both novel and familiar categories, where our methods can successfully recreate similar objects for both familiar and novel categories. The varied orientations result from the diverse orientations in our augmented training data.

### 9.2. More Examples of Generated Images for Different Styles

Figure 9 shows some examples that evaluate our method ability to compose objects into new scenes with prompt "A DSLR photo of S* in the center of {background}". The list of backgrounds is list in Table 1. For relatively simple backgrounds like *beach, rain, river, and snow*, most times the model is able to consistently generate high-quality images of both novel and familiar objects. However, with

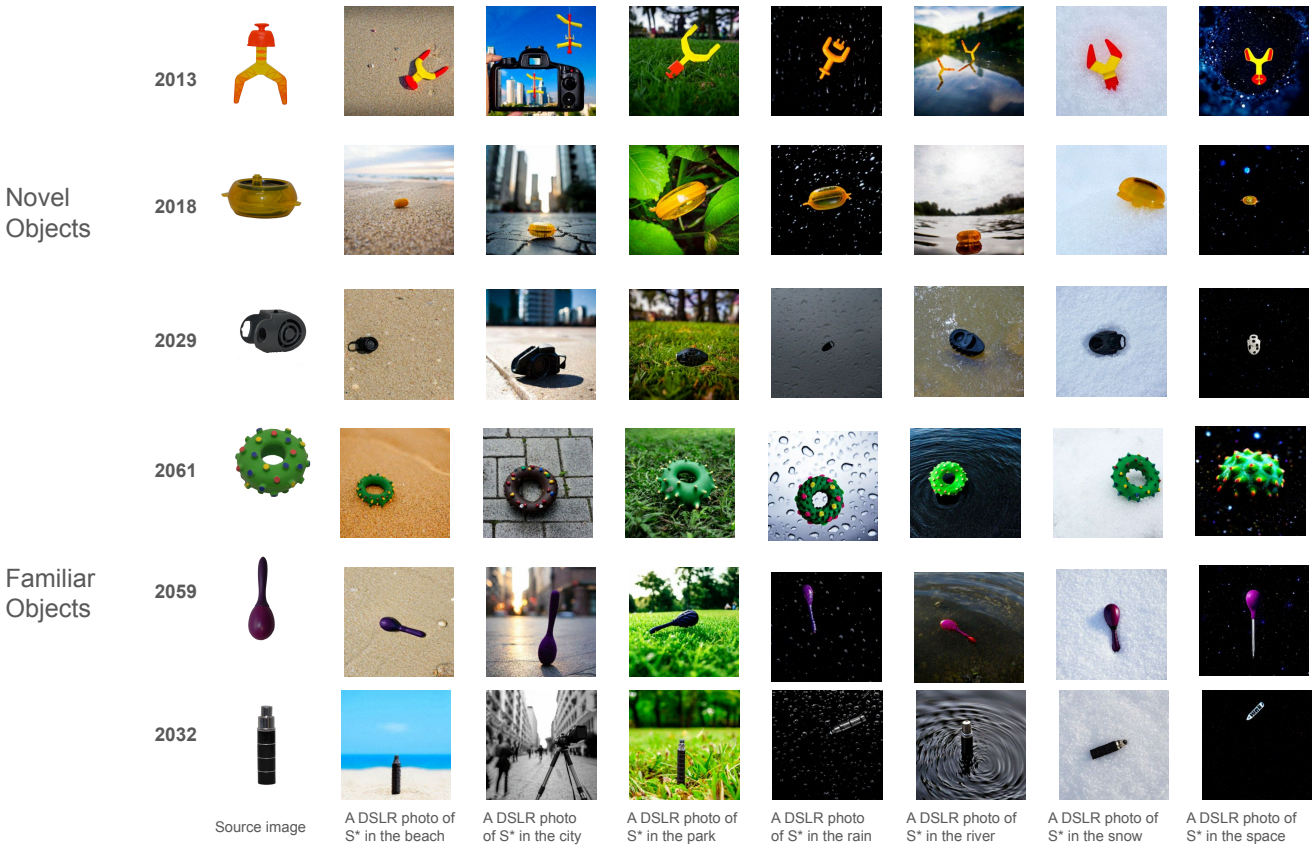


Figure 9. Select sample outputs for images generated for different scenes with prompt "A DSLR photo of S* in the {background}", where Category 2013, 2018 and 2029 are labeled as novel objects and Category 2061, 2059, and 2032 are labeled as familiar objects

more complicate backgrounds such as *city, park, and space*, we observe the results are inconsistent. The image generation model fails in some cases for both novel and familiar objects in these complicate backgrounds, where the model may attempt to replace a similar object that can be better fit in these backgrounds. For example, it replaces Object 2032 with a camera tripod for generating "A DSLR photo of S* in the center of city". Similarly, it re-constructs Object 2061 to a sphere shape instead of the original donuts shape when generating "A DSLR photo of S* in the center of space". We further assess the results generated for the combination of different scenes and different styles using prompt "A or An {style} photo of S* in the center of {background}".

Figure 10 and Figure 11 show more selected samples from the generated images of familiar objects and novel objects with *oil painting* and *surreal digital art* styles. We observed that the model tends to generate higher-quality results for familiar objects in both styles, especially in the *surreal digital art* style, which aligns with the human survey results in Section 4.4. Similar to the observations for the *DSLR* style, we also observed that the model generates inconsistent results for complex backgrounds such as *city, park, and space*. Therefore, we believe improving image generation for complex backgrounds is a promising direction for future work.

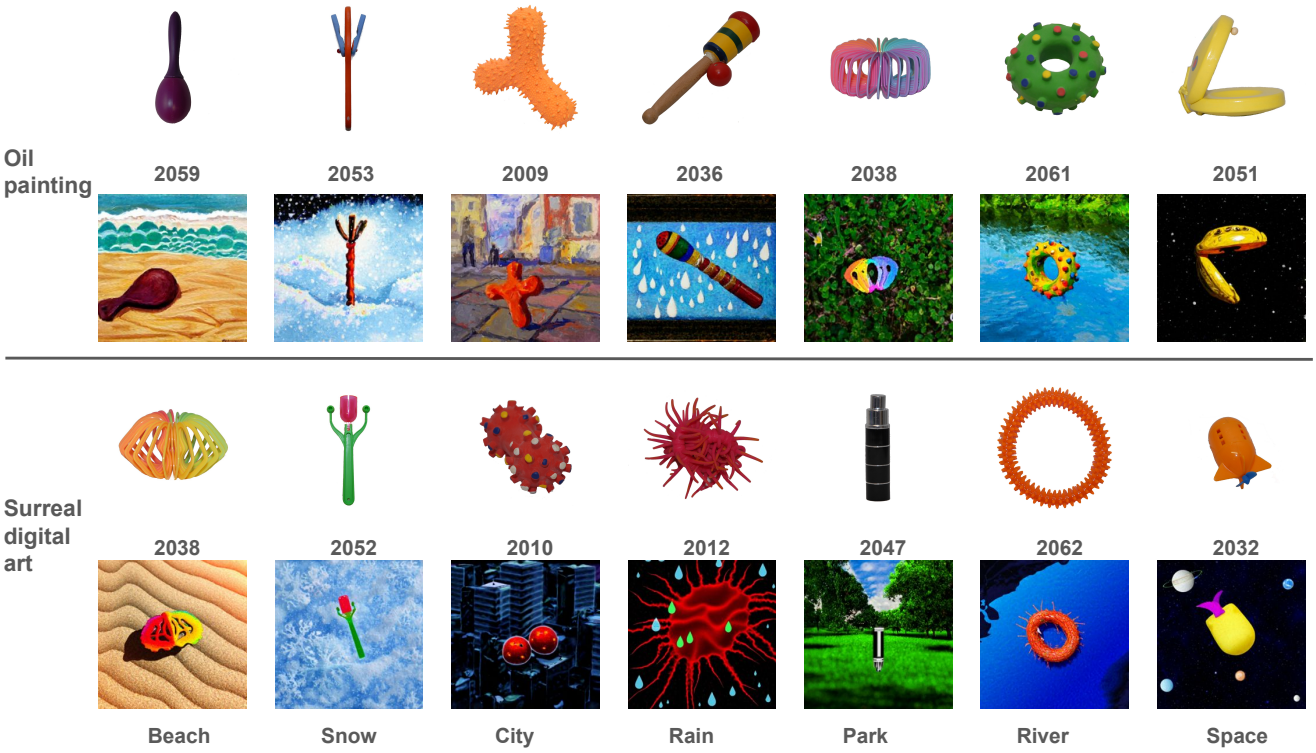


Figure 10. Select sample outputs for familiar objects with ”oil painting” and ”surreal digital art” styles

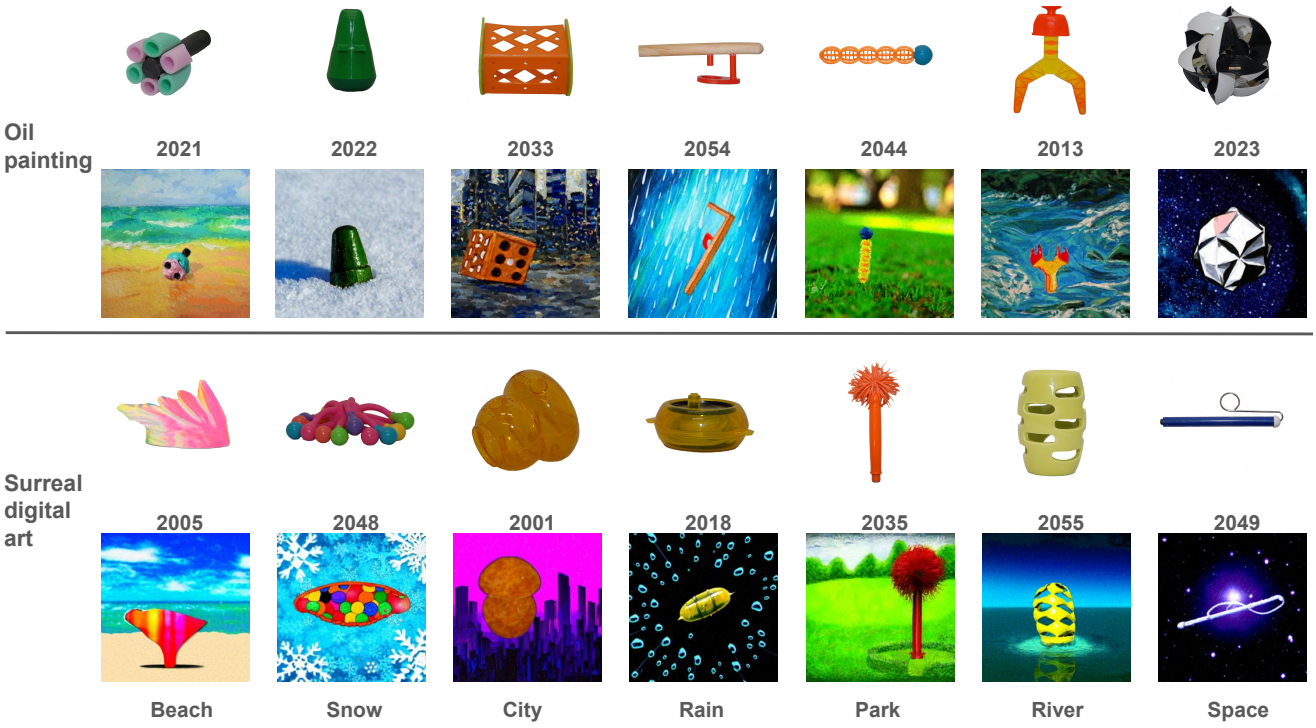


Figure 11. Select sample outputs for novel objects with ”oil painting” and ”surreal digital art” styles