Guided Image Concept Decomposition using Textual Inversion

Stanford CS224N Custom Project

Yvette Lin
Department of Computer Science
Stanford University
yvelin@stanford.edu

Abstract

Textual inversion is a method of introducing new user-defined concepts a text-to-image model by learning new "words" in the model's textual embedding space, allowing for personalized text-to-image generation. However, a limitation of this method is difficulty in learning precise aspects of an image, as the method attempts to incorporate all the semantic essence of a concept into a single learned "word." To address this challenge, drawing inspiration from recent work demonstrating that textual inversion-based methods can be used to decompose image concepts into consituent sub-concepts, we add additional control to textual inversion by isolating the explicitly desired image sub-concept. Given a user-defined prompt capturing the desired relation between sub-concepts of a set of images, our method introduces new "words" representing those sub-concepts, which behave like natural words and thus can be used to generate highly specifically personalized images. We demonstrate that our method is able to successfully isolate desired sub-concepts through a comparison to the naive baseline method using purely single-concept textual inversion.

1 Key Information to include

- Mentor: Kaylee Carissa Burns
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Recent large-scale text-to-image generative models (Ramesh et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) are incredibly powerful at synthesizing shockingly realistic and diverse images given natural-language user prompts. However, out-of-the-box models do not necessarily contain knowledge of personalized concepts that a user may be interested in generating (for example, if a user wants to generate photos of themselves, the model will probably not know what they look like out of the box). Hence, there is interest in modifying such image generation models to introduce new personalized concepts, as retraining the entire model to introduce a new concept with a new dataset is prohibitively expensive.

This modification task poses several challenges, with previous model fine-tuning approaches being subject to forgetting prior knowledge (Kumar et al., 2022). One approach to this problem addressing these challenges involves fixing the generative model and solving the task of *textual inversion* (Gal et al., 2023). Concretely, given a text-to-image model and an image set depicting a visual concept, we wish to find a word embedding representing this concept (that behaves like a natural word, so can be used in prompts like "A photo of S^*").

However, textual inversion is still subject to a key limitation, which is that if a user wishes to generate personalized images focusing on a particular aspect or sub-concept of a visual concept, rather than the entire visual concept (e.g. “A photo of the pattern found on S^*”, rather than “A photo of S^*”), a single word S^* may be insufficient to capture a particular sub-concept. Hence, we propose to draw inspiration from Vinker et al. (2023), which demonstrates that textual inversion can be used to optimize embeddings sub-concepts of an visual concept in addition to the parent concept. Our method adapts Vinker et al. (2023) to the task of isolating specific concepts for the purpose of highly specific and personalized image generation through user-defined prompts that define the desired sub-concept to be extracted. We evaluate our method against the naive baseline of pure single-concept textual inversion by means of quantitative image and text similarity metrics, and a user perceptual study. We demontrate that while our method still exhibits some limitations, we are able to successfully isolate plausible image sub-concepts guided by user prompts as desired, and generate personalized, specific images.

3 Related Work

Large langugage-vision models

There have been many exciting advances in the field of language-vision models, which are now capable of performing extremely sophiscated tasks, including convincing text-to-image generation (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022). A major paradigm for image generation is latent diffusion models (LDMs) (Rombach et al., 2021). An LDM consists of a pretrained image encoder and a diffusion model that can be conditioned on different inputs to guide image generation. For our work, we focus on text-to-image generation, so the conditioning input comes from a text encoder.

Personalization and inversion

Multiple different approaches have been taken to the problem of personalizing image generation (Hu et al., 2022; Gal et al., 2023; Ruiz et al., 2023). In particular, Gal et al. (2023) seeks to invert text-to-image generation by constructing a textual embedding corresponding to a given visual concept (represented by a set of images) by optimizing embeddings in the shared latent space of a given text-to-image model, drawing inspiration from similar GAN inversion methods (Abdal et al., 2019; Gu et al., 2020). We choose an approach most similar to Gal et al. (2023) because unlike some other methods, it does not modify the weights of the underlying network. This is much less memory-intensive, as it does not require a separate model for each personalization (and perhaps also suits the contents of the class a bit better). However, a single word embedding as is optimized in Gal et al. (2023) is may not be sufficient to capture particular desired sub-aspects of an image, which our work seeks to address.

Concept decomposition

Vinker et al. (2023) extend textual inversion (Gal et al., 2023) to the problem of decomposing the parent concept represented by a set of images into constituent sub-concepts. Namely, they seek to optimize not just a single embedding v_p^* as in (Gal et al., 2023) but multiple embeddings v_l^*, v_r^* corresponding to subconcepts S_l and S_r of the parent concept S_p represented by the set of images. Similar to Vinker et al. (2023), we decompose an image set into sub-concepts using textual inversion. However, our work differs in several key ways. Unlike Vinker et al. (2023), we focus on the specific task of personalization in highly specific and guided fashion, and perform evaluations to assess our method on this particular task. Due to this difference, Vinker et al. (2023) does not explicitly specify the relation between sub-concepts on the user end, while we guide the choice of sub-concepts through a user-defined prompt. This warrants several modifications to the concept decomposition method, which are described in Section 4.

4 Approach

Our goal is to improve upon the controllability of the textual inversion method introduced in Gal et al. (2023) by isolating separate visual concepts or objects present in a set of images. Precisely, we wish take a set of images representing multiple visual concepts S_l, S_r (with a given relation) and a text-to-image latent diffusion model (LDM), and obtain personalized word embeddings v_l^*, v_r^* corresponding to the visual sub-concepts that can be used as natural words to generate images via the LDM, e.g. "A car in the style of S_l".

To this end, we build upon Vinker et al. (2023), who demonstrate that a textual inversion can decompose a concept S_p into sub-concepts S_l, S_r by performing textual inversion by conditioning during the training process on prompts of the form "A photograph of S_l S_r". We first modify this approach to suit our specific personalization task by instead conditioning on prompts that capture the explicit, user-specified relation between desired concepts within the images, such a "A photograph of S_l with a S_r pattern". Given

- c_θ a model mapping a conditioning input y (here, text encoding) into a conditioning vector
- the timestep t
- z_t the latent noised to time t
- ϵ the unscaled noise sample
- ϵ_θ the denoising network
- $\mathcal{E}$ an encoder mapping images x to a latent code z

the standard LDM loss is

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\| \epsilon - \epsilon_\theta(z_t, t, c_\theta(y)) \|_2^2 \right] . \tag{1}$$

Our optimization objective is

$$\{v_l, v_r\} = \arg\min_v \mathbb{E}_{x \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\| \epsilon - \epsilon_\theta(z_t, t, c_\theta(y)) \|_2^2 \right] , \tag{2}$$

fixing c_θ and ϵ_θ.

Vinker et al. (2023) additionally performs a consistency test allowing for the selection of the most coherent sub-concepts. The *consistency* between two sets of images I^a, I^b is defined as

$$\mathcal{C}(I^a, I^b) = \underset{I_i^a \in I^a, I_j^b \in I^b, I_i^a \neq I_j^b}{\text{mean}} (\text{sim}(\text{CLIP}(I_i^a), \text{CLIP}(I_j^b))) \tag{3}$$

where sim is the cosine similarity $\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$ from a pretrained CLIP encoder. The *consistency test* is performed by choosing the optimized embeddings v_l^*, v_r^* from a set V_s such that

$$\{v_l^*, v_r^*\} = \underset{\{v_l^i, v_r^i \in V_s\}}{\text{argmax}} \left[(C_l^i + C_r^i) + (\min(C_l^i) - \mathcal{C}(I^{v_l^i}, I^{v_r^i})) \right] \tag{4}$$

where $C_l^i = \mathcal{C}(I^{v_l^i}, I^{v_l^i})$ and $C_r^i = \mathcal{C}(I^{v_r^i}, I^{v_r^i})$. The first term encourages a choice of embeddings v_l^i, v_r^i which maximizes the self-consistency of each concept, and the second term discourages sub-concepts from being too similar to each other.

For our method, we hypothesize that a user-defined prompt can sufficiently specify the relation between S_l and S_r. We modify the consistency test in Eq. 4 to obtain:

$$\{v_l^*, v_r^*\} = \underset{\{v_l^i, v_r^i \in V_s\}}{\text{argmax}} \left[(C_l^i + C_r^i) - \max(\mathcal{C}(I^{v_l^i}, I^{v_p^i}), \mathcal{C}(I^{v_r^i}, I^{v_p^i})) \right] \tag{5}$$

where each v_p^i is an embedding corresponding to the parent concept S_p derived from v_l^i, v_r^i, and the user-defined prompt relating S_l and S_r (e.g. "S_l with a S_r pattern"). The first term, as before, encourages, a choice of embeddings which maximizes the self-consistency of each concept. Unlike Vinker et al. (2023), which does not explicitly specify the relationship between the sub-concepts, our method takes as input a user-defined prompt that specifies the relation between S_l and S_r; hence, we do not need to explicitly discourage the the concepts to be dissimilar, so we remove the second term of Eq. 4. We additionally assume that the user of our method likely wishes to isolate particular sub-concept(s) of S_p, rather than being interested in concepts close to S_p. Hence, S_l and S_r should be relatively dissimilar to S_p, which is encouraged by our second term in Eq. 5.

We build upon and modify the existing released codebases of Vinker et al. (2023). To summarize, we make the following additions/modifications to the code:

- Add functionality to allow specification of user-defined relations between concepts, rather than just the default presumed "S_l S_r" (which assumes something like an adjective-noun relationship between S_l and S_r, which may not apply to the concepts in an image set).
- Modify the consistency score to reflect Eq. 5.
- Implement the CLIP image and text similarity metrics described in Section 5.2.
- Fix existing bugs in the existing codebase which caused the code to crash out-of-the-box at the seed selection step.

5 Experiments

5.1 Data

We currently use the datasets released by Gal et al. (2023), which are found here. To be clear, our task inputs and outputs can be described as

- Input: A set of images representing multiple visual concepts (with a user-defined relation) and a generative LDM, and
- Output: A "word" corresonding to the visual sub-concepts that can be used as a natural word in prompts to generate images from the LDM.

Each dataset is used as an instance of the set of images representing multiple visual concepts in the input.

5.2 Evaluation method

Taking inspiration from Gal et al. (2023), to evaluate the ability of our method to replicate a desired target sub-concept, we measure *image similarity* between a set of images of a learned sub-concept and generated images, using a naive method as a baseline. The naive method of specifying a sub-concept is simply trying to describe the sub-concept through words e.g. "The pattern found on S_p", where S_p is the parent concept. Therefore, concretely, for each dataset, we generate 64 images using the following prompts:

(1) "A photo of S^*", where S^* is a subconcept $S^* \in \{S_l, S_r\}$. This represents our method.

(2) "A photo of S'_p", where S'_p is a string describing the sub-concept S^* of the parent concept S_p through words. For example, if $S_p =$ "S_l with a S_r pattern", $S^* = S_r$, then $S'_p =$ "the pattern found on S_p". This serves as our baseline.

We then compute the semantic CLIP-space distances between the images generated by each prompt with a set of images corresonding to S^*.

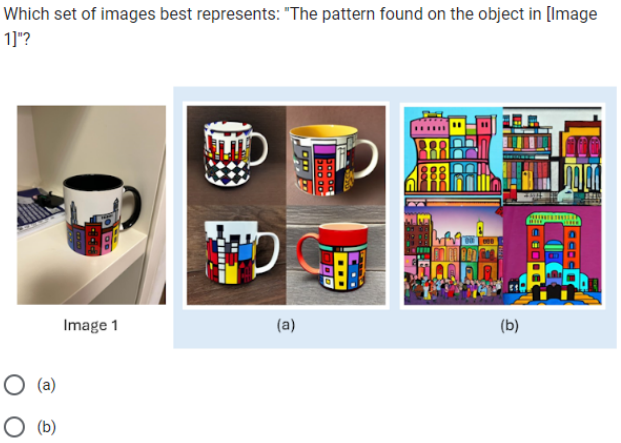
Again taking inspiration from Gal et al. (2023), to evaluate the fidelity to the surrounding textual prompt (and hence the editing capability), we measure *text similarity* between a set of images of a learned sub-concept and generated images using a naive method as a baseline. Concretely, for each dataset, we consider a set of varying prompts e.g. "A photo of a car", "A photo of a lunchbox". For each such prompt S, we generate a set of 64 images using the following:

(1) The prompt modified with S^*, e.g. "A photo of a S^* car". This represents our method.

(2) The prompt modified with S'_p, e.g. "A photo of a car with the pattern of S'_p". This serves as our baseline.

We then compute the semantic CLIP-space distance between the images generated by each of these prompts with a set of images generated by S.

In addition, we conduct a user study asking participants to rate the quality of generated images on several metrics compared to the naive baseline. An example question on the study is shown in Figure 1. The study consisted of 6 such questions.


Which set of images best represents: "The pattern found on the object in [Image 1]"?

○ (a)

○ (b)

Figure 1: Example survey question for user study.

5.3 Experimental details

For our experiments, we use Stable Diffusion (Rombach et al., 2021) as our LDM. To optimize the textual embedding, we use the Adam optimizer with an initial learning rate of 5×10^{-4}. To select the best sub-concepts according to Eq. 5, we optimize embeddings for 4 chosen seeds for 200 epochs, perform the consistency test on these 4 seeds, and further optimize the chosen seed for another 800 epochs. The entire process end-to-end for a single image set takes around 1.5 hours on a NVIDIA GeForce RTX 3090 GPU. The prompts used to decompose each dataset concept are listed in Table 1.

Dataset	Prompt
`mug-buildings`	"S_l with a S_r pattern"
`canada-bear`	"S_l with a S_r print"
`physics-mug`	"S_l with a S_r design"
`elephant`	"S_l made of S_r"
`red-teapot`	"red S_l with a S_r pattern"
`cat-sculpture`	"S_l in the style of S_r"

Table 1: Decomposition promopts. Prompt used for each dataset for concept decomposition.

5.4 Results

In Figure 2 we qualitatively show example comparisons of our method to the naive method described in Section 5.2. For each dataset, we show an example sub-concept learned with our method, and an attempt to naively capture that same concept using single-concept textual inversion and textual description. We perform further analysis of each dataset and limitations in Section 6, but we can see that our method is able to specify certain sub-aspects of a visual concept that the naive method is not able to.

In Table 2 we show the quantitative results of the *image similarity* and *text similarity* metrics described in Section 5.2. We see that on all datasets, our method achieves the higher image similarity than the baseline naive method. This indicates that as desired, our method is more effective at isolating specific sub-concepts of a set of images than the naive method trying to isolate a concept purely through textual description (e.g. "The pattern found on S_p"). Additionally, our method achieves higher text similarity on most of the datasets, with a higher average text similarity. This indicates that general, our learned sub-concepts remain as editable and as capable of expressing more complex ideas found in user prompts as the learned parent concepts from the naive textual inversion method.

In Table 3 we show the results of the user perceptual study described in Section 5.2. 33 participants responded to the survey. For each dataset, we show the percentage of participants who preferred the result from our method over the result from the baseline naive method. Participants preferred our method for the majority of datasets. However, several datasets exhibited notable failure cases. We

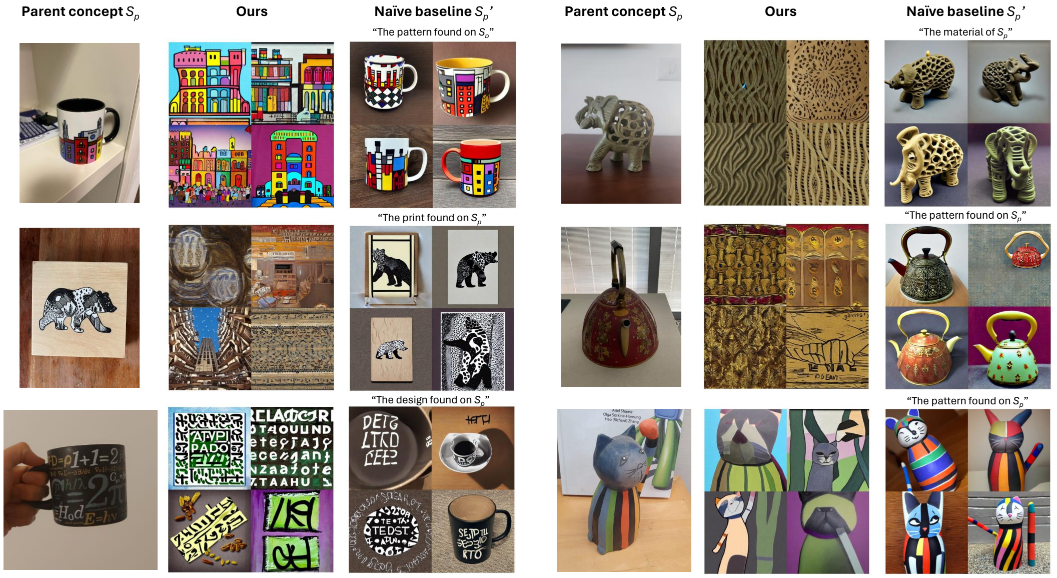

Parent concept S_p
Ours
Naïve baseline S_p’
Parent concept S_p
Ours
Naïve baseline S_p’
“The pattern found on S_p”
“The material of S_p”
“The print found on S_p”
“The pattern found on S_p”
“The design found on S_p”
“The pattern found on S_p”

Figure 2: Baseline comparison. For each dataset, we show an example sub-concept learned with our method, and an attempt to naively capture that same concept using single-concept textual inversion. Our method is able to specify certain sub-aspects of a parent concept that the naive method is not able to.

	Ours		**Baseline**	
Dataset	Image sim. (↑)	Text sim. (↑)	Image sim. (↑)	Text sim. (↑)
`mug-buildings`	**0.828**	**0.230**	0.822	0.199
`canada-bear`	**0.826**	0.214	0.822	**0.219**
`physics-mug`	**0.720**	0.220	0.705	**0.243**
`elephant`	**0.889**	**0.245**	0.888	0.245
`red-teapot`	**0.831**	**0.213**	0.744	0.207
`cat-sculpture`	**0.827**	**0.213**	0.741	0.207
Average	**0.820**	**0.222**	0.787	0.220

Table 2: Baseline comparison. The baseline we compare to is the naive textual inversion method described in Section 5.2. For each method, we show the image similarity and text similarity. Our method achieves higher image and text similarity, indicating higher expressivity and comparable editability to the baseline.

also notice that the datasets that participants do not prefer, they do so by an overwhelming amount. This indicates that the limitations are due to a choice of sub-concept decomposition, possibly at the consistency test stage. In Section 6, we further exhibit and discuss visual examples of sub-concepts learned from each dataset, and analyze the reasons for the failure cases seen here.

Dataset	**Preference for ours** (↑)
`mug-buildings`	**70%**
`canada-bear`	0 %
`physics-mug`	**66%**
`elephant`	**67%**
`red-teapot`	**81%**
`cat-sculpture`	9%

Table 3: User perceptual study. For each dataset, we show the percentage of participants who preferred the result from our method over the result from the baseline naive method. Participants preferred our method for the majority of datasets.

6 Analysis

In Figure 3 we show example concept decompositions from our method for each dataset. In column 1 we show a representative image from each dataset, and in columns 2 and 3 we show sub-concepts learned from each dataset using the prompts in Table 1. In column 4 we show generated images from example prompts to the resulting personalized generative model from our method, showing that the sub-concepts we learn can be used naturally in prompts to the generative image model. In general, the generated images appear faithful to the depicted sub-concept while still expressing the desired prompt. In general, qualitatively, our method is able to isolate plausible sub-concepts according to the decomposition prompts; notably in rows 1 and 3.

However, the method still exhibits some limitations. For example, in row 2, our method seems to decompose the print into a concept representing a pattern and a concept representing wood block prints in general. This could be due to our choice of consistency test (Eq. 5)—since we penalize sub-concepts comprising a large majority of the parent concept, and the bear print is the focal point of the image, our method is too heavily incentivized to serparate sub-concepts within the bear print. A limitation can also be seen in row 5. In the corresponding prompt (Table 1), we attempted to disentangle the color of the teapot from the sub-concepts. We can see in column 3 of this row that while the images still have some red, in general, sub-concept represents a much more gold-colored pattern, which is promising. However, in column 2 of this row, we see a sub-concept that seems unnatural for this image. This could be due to the increased complexity of the decomposition prompt compared to the other datasets. In row 6 we can also see some additional limitations. Similar to row 5, the sub-concept in column 2 is not very natural, as it seems the "style" of the sculpture in column 3 gets interpreted as encompassing the "cat"-like aspects of the sculpture, while column 2 is left to capture the idea of a 3D object. This could be due to a similar issue as in row 2 described above, or due to the lack of specificity in the decomposition prompt, since it may be unclear exactly to what the word "style" refers.

7 Conclusion

In this work, we presented a method for highly specific personalized image generation inspired by concept decomposition via textual inversion. Our method allows users to isolate specific desired sub-concepts from a larger visual concept of their choice, and generate images incorporating the desired sub-concept using natural-language prompts. We evaluate our method by comparing with the baseline naive method of involving only single-concept textual inversion and find that we are more successfully able to express desired sub-concepts of visual concepts for personalization.

Future work could address the existing limitations of our method. Namely, we have reason to believe the consistency test we propose may too strongly incentivize sub-concepts that comprise smaller proportions of the larger parent concept, even when this is not appropriate, resulting in unnatural sub-concept choices. Devising a better consistency metric, whether it is revising the relative weights on the current consistency test terms or proposing another metric entirely could be a direction for future work. Additionally, the method performs poorly on multi-object concepts e.g. decomposing "S_l next to S_r". This may be because it is difficult to obtain enough varying object poses (which Gal et al. (2023) shows helps with reconstruction quality) while having the model still understand the spatial arrangement of objects. Future work could also aim to improve in this direction.

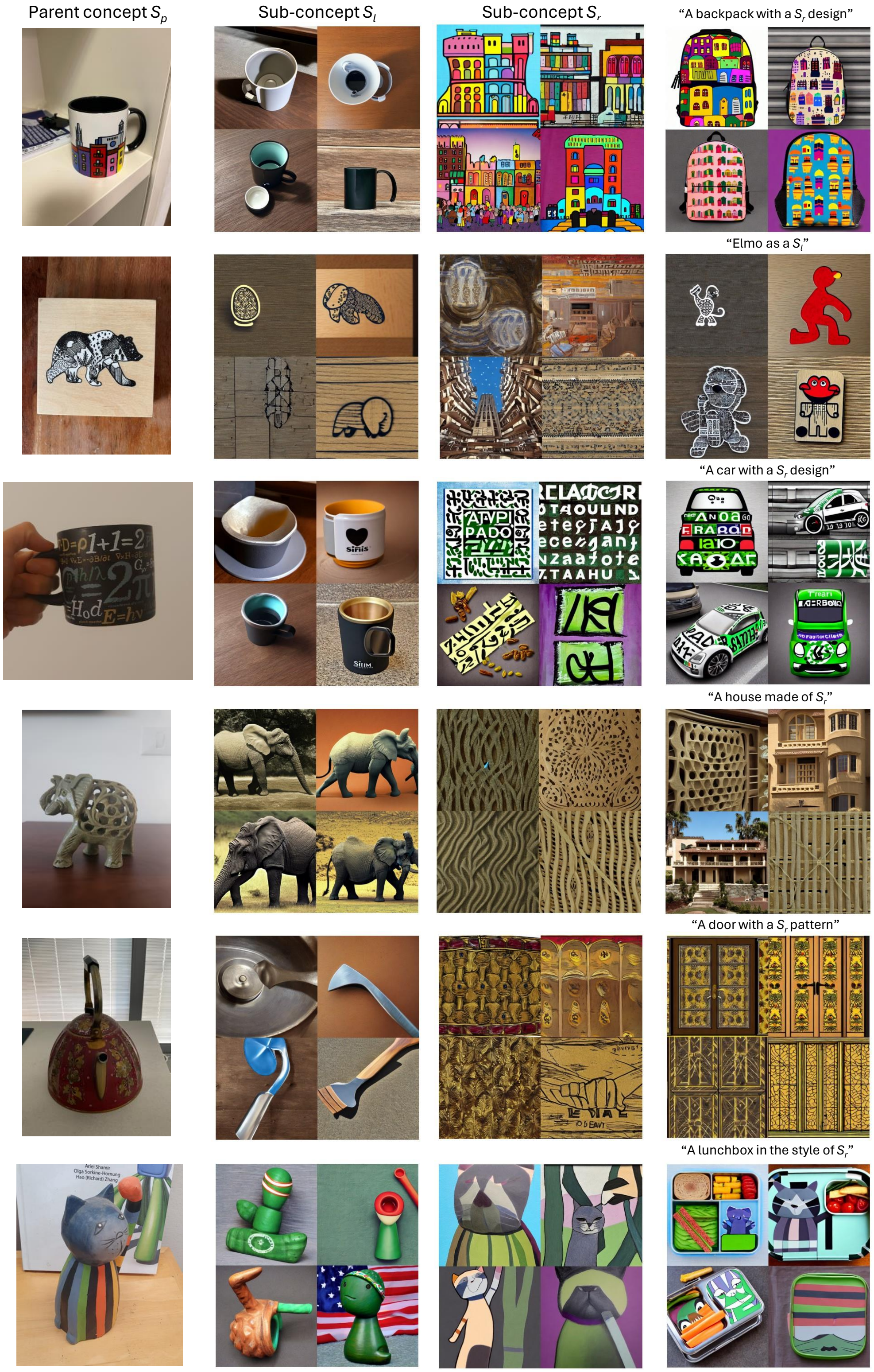

Parent concept S_p
Sub-concept S_l
Sub-concept S_r
“A backpack with a S_r design”
“Elmo as a S_l”
“A car with a S_r design”
“A house made of S_r”
“A door with a S_r pattern”
“A lunchbox in the style of S_r”

Figure 3: Textual inversion on sub-concepts. For each dataset we show sub-concepts learned according to the decomposition prompts in Table 1 and generated images from example prompts. Our method is able to isolate some plausible sub-concepts that can be used in further prompts.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*.

Jinjin Gu, Yujun Shen, and Bolei Zhou. 2020. Image processing using multi-code gan prior.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept decomposition for visual exploration and inspiration. *arXiv preprint arXiv:2305.18203*.