Concept Decomposition for Visual Exploration and Inspiration

YAEL VINKER*, Tel Aviv University ,Google Research, Israel
ANDREY VOYNOV, Google Research, Israel
DANIEL COHEN-OR, Tel Aviv University ,Google Research, Israel
ARIEL SHAMIR, Reichman University, Israel

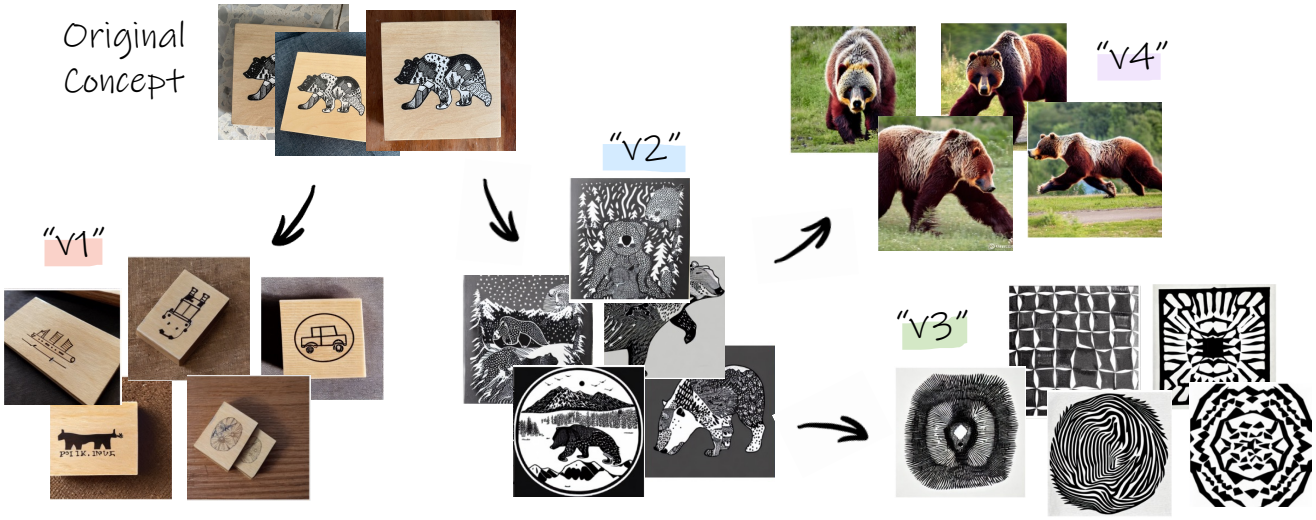


Fig. 1. Our method provides a tree-structured visual exploration space for a given unique concept. The nodes of the tree ("v_i") are newly learned textual vector embeddings, injected to the latent space of a pretrained text-to-image model. The nodes encode different *aspects* of the subject of interest. Through examining combinations within and across trees, the different aspects can inspire the creation of new designs and concepts.

A creative idea is often born from transforming, combining, and modifying ideas from existing visual examples capturing various concepts. However, one cannot simply copy the concept as a whole, and inspiration is achieved by examining certain aspects of the concept. Hence, it is often necessary to separate a concept into different aspects to provide new perspectives. In this paper, we propose a method to decompose a visual concept, represented as a set of images, into different visual *aspects* encoded in a hierarchical tree structure. We utilize large vision-language models and their rich latent space for concept decomposition and generation. Each node in the tree represents a sub-concept using a learned vector embedding injected into the latent space of a pretrained text-to-image model. We use a set of regularizations to guide the optimization of the embedding vectors encoded in the nodes to follow the hierarchical structure of the tree. Our method allows to explore and discover new concepts derived from the original one. The tree provides the possibility of endless visual sampling at each node, allowing the user to explore the hidden sub-concepts of the object of interest. The learned aspects in each node can be combined within and across trees to create new visual ideas, and can be used in natural language sentences to apply such aspects to new designs.

ACM Reference Format:
Yael Vinker*, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. *ACM Trans. Graph.* 1, 1 (September 2023), 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

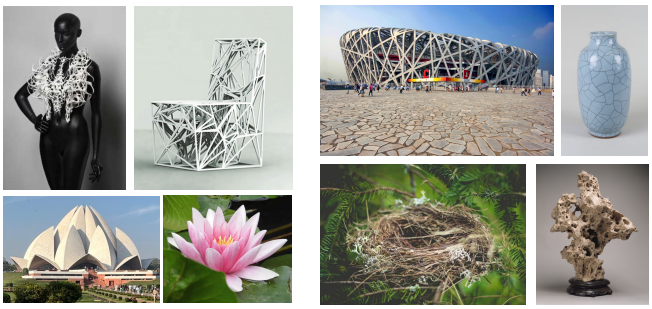
Modeling and design are highly creative processes that often require inspiration and exploration [Gonçalves et al. 2014]. Designers often draw inspiration from existing visual examples and concepts - either from the real world or using images [Eckert and Stacey 2000; Henderson 1999; Muller 1989]. However, rather than simply replicating previous designs, the ability to extract only certain aspects of a given concept is essential to generate original ideas. For example, in Figure 2a, we illustrate how designers may draw inspiration from patterns and concepts found in nature.

Additionally, by combining multiple aspects from various concepts, designers are often able to create something new. For instance, it is described [designingbuildings wiki 2021] that the famous Beijing National Stadium, also known as the "Bird's Nest", was designed by a group of architects that were inspired by various aspects of

Authors' addresses: Yael Vinker*, Tel Aviv University and ,Google Research, Tel Aviv, Israel, yaelvinker@mail.tau.ac.il; Andrey Voynov, Google Research, Tel Aviv, Israel, avoin@google.com; Daniel Cohen-Or, Tel Aviv University and ,Google Research, Tel Aviv, Israel, cohenor@gmail.com; Ariel Shamir, Reichman University, Tel Aviv, Israel, arik@runi.ac.il.

0730-0301/2023/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Work was done during an internship at Google



(a) (b)

Fig. 2. Examples of design inspired by visual concepts taken from other concepts. (a) top left - fashion design by Iris Van Herpen and Chair by Emmanuel Touraine inspired by nature patterns, bottom left - the Lotus Temple in India, inspired by the lotus flower (b) Beijing National Stadium is inspired by a combination of local Chinese art forms - the crackle glazed pottery that is local to Beijing, and the heavily veined Chinese scholar stones. ©Dress by Iris van Herpen, chair by Emmanuel Touraine from Wikimedia. Lotus flower, temple, cracked pottery, scholar stone, and bird nest are from rawpixel.com [Public Domain]. Beijing National Stadium photograph by Wojtek Gurak from Flickr.

different Chinese concepts (see Figure 2b). The designers combined aspects of these different concepts – the shape of a nest, porous Chinese scholar stones, and cracks in glazed pottery art that is local to Beijing, to create an innovative architectural design. Such a design process is highly exploratory and often unexpected and surprising.

The questions we tackle in this paper is whether a machine can assist humans in such a highly creative process? Can machines understand different aspects of a given visual concept, and provide inspiration for modeling and design? Our work explores the ability of large vision-language models to do just that - express various concepts visually, decompose them into different aspects, and provide almost endless examples that are inspiring and sometimes unexpected.

We rely on the rich semantic and visual knowledge hidden in large language-vision models. Recently, these models have been used to perform personalized text-to-image generation [Gal et al. 2022; Kumari et al. 2023; Ruiz et al. 2023], demonstrating unprecedented quality of visual concept editing and variation. We extend the idea presented in [Gal et al. 2022] to allow *aspect-aware* text-to-image generation, which can be used to visually explore new ideas derived from the original visual concept.

Our approach involves (1) decomposing a given visual concept into different aspects, creating a hierarchy of sub-concepts, (2) providing numerous image instances of each learned aspect, and (3) allowing to explore combinations of aspects within the concept and across different concepts.

We model the exploration space using a binary tree, where each node in the tree is a newly learned vector embedding in the textual latent space of a pretrained text-to-image model, representing different aspects of the original visual concept. A tree provides an intuitive structure to separate and navigate the different aspects of a given concept. Each level allows to find more aspects of the concepts in the previous level. In addition, each node by itself contains a plethora of samples and can be used for exploration. For example, in Figure 1, the original visual concept is first decomposed into its dominant semantic aspects: the wooden saucer in "v1" and the bear drawing in "v2", next, the bear drawing is further separated into the general concept of a bear in "v3" and its unique texture in "v4".

Given a small set of images depicting the concept of interest as input, we build the tree gradually. For each node, we optimize two child nodes at a time to match the concept depicted in their parent. We also utilize a CLIP-based [Radford et al. 2021a] consistency measurement, to ensure that the concepts depicted in the nodes are coherent and distinct. The different aspects are learned *implicitly*, without any external constraint regarding the type of separation (such as shape or texture). As a result, unexpected concepts can emerge in the process and be used as inspiration for new design ideas. For example the learned aspects can be integrated into existing concepts by combining them in natural language sentences passed to a pretrained text-to-image model (see Figure 3). They can also be used to create new concepts by combining different aspects of the same tree (intra-tree combination) or across different trees (inter-tree combination).

We provide many visual results applied to various challenging concepts. We demonstrate the ability of our approach to find different aspects of a given concept, explore and discover new concepts derived from the original one, thereby inspiring the generation of new design ideas.

Fig. 3. Combining the learned aspects in natural sentences to produce aspect-based variations. The original concept is shown on top, along with an illustration of the chosen aspects from the tree in Figure 1. Below are three random images generated by a pretrained text-to-image model, conditioned on the prompts above.

2 PREVIOUS WORK

Design and Modeling Inspiration. Creativity has been studied in a wide range of fields [Amabile 1996; Bonnardel and Cauzinille-Marmèche 2005; Elhoseiny and Elfeki 2019; Kantosalo et al. 2014; Runco and Jaeger 2012], and although defining it exactly is difficult, some researchers have suggested that it can be described as the act of evoking and recombinating information from previous knowledge to generate new properties [Bonnardel and Cauzinille-Marmèche 2005; Wilkenfeld and Ward 2001]. It is essential, however, to be able to associate ideas in order to generate original ideas rather than just mimicking prior work [Brown 2008]. Experienced designers and artists are more adept at connecting disparate ideas than novice designers, who need assistance in the evocation process [Bonnardel and Cauzinille-Marmèche 2005]. By reviewing many exemplars, designers are able to gain a deeper understanding of design spaces and solutions [Eckert and Stacey 2000]. In the field of human-computer interaction, a number of studies have been conducted to develop tools and software to assist designers in the process of ideation [Chilton et al. 2019; Ivanov et al. 2022; Kang et al. 2021; Koch et al. 2019, 2020]. They are focused on providing better tools for collecting, arranging, and searching visual and textual data, often collected from the web. In contrast, our work focuses on extracting different aspects of a given visual concept and *generating* new images for inspiration. Our work is close to a line of work on visualizing and exploring design alternatives for geometry [Denning and Pellacini 2013; Dobos and Steed 2012; Marks et al. 1997; Matejka et al. 2018], including utilizing evolutionary algorithms to inspire users' creativity [Averkiou et al. 2014; Cohen-Or and Zhang 2015; Xu et al. 2012]. However, they mostly work in the field of 3D content generation and do not decompose different aspects from existing concepts.

Hierarchical Structure of Images and Language. Humans are believed to comprehend and interpret intricate visual scenes by breaking them down into hierarchical parts and wholes [Hinton 1979]. Substantial research has focused on the hierarchical structure of images, involving capsule networks [Hinton 2021; Hinton et al. 2011; Sabour et al. 2017], and-or graphs [Tu et al. 2013, 2003], as well as scene and object parsing [Chen et al. 2014; Liang et al. 2016; Zhang et al. 2016; Zhou et al. 2017]. The relationship between the hierarchical nature of language and vision has also been explored in a variety of tasks, including image-text retrieval [Cao et al. 2022; Karpathy et al. 2014; Kiros et al. 2014], visual metaconcept learning [Han et al. 2020; Mei et al. 2022], and visual question answering [Aditya et al. 2018; Anderson et al. 2017].

Recently, the field of image generation and editing has undergone unprecedented evolution with the advancement of large language-vision models [Nichol et al. 2021; Radford et al. 2021a; Ramesh et al. 2022; Rombach et al. 2022]. These models have been trained on millions of images and text pairs and have shown to be effective in performing challenging vision related tasks [Amit et al. 2021; Avrahami et al. 2022; Gal et al. 2021; Patashnik et al. 2021; Sheynin et al. 2022]. Furthermore, the strong visual and semantic priors of these models have also been demonstrated to be effective for artistic and design tasks [Midjourney 2022; Oppenlaender 2022; Tian and Ha 2021; Vinker et al. 2022a,b]. In our work, we hypothesize that the latent space of such models also contain some hierarchical structure and demonstrate how large language-vision models can be used to decompose and transform existing concepts into new ones in order to inspire the development of new ideas.

Personalization. Personalized text-to-image generation has been introduced recently [Gal et al. 2022; Hu et al. 2021; Kumari et al. 2023; Ruiz et al. 2023], with the goal of creating novel scenes based on user provided unique concepts. In addition to demonstrating unprecedented quality results, these technologies enabled intuitive editing, made design more accessible, and attracted interest even beyond the research community. We utilize these ideas to facilitate the ideation process of designers and common users, by learning different visual aspects of user-provided concepts.

Current personalization methods either optimize a set of embeddings to describe the concept [Gal et al. 2022], or modify the denoising network to tie a rarely used word embedding to the new concept [Ruiz et al. 2023]. While the latter provides more accurate reconstruction and is more robust, it uses much more memory and requires a model for each object. In this regard, we choose to rely on the approach presented in [Gal et al. 2022]. It is important to note that our goal is to capture multiple *aspects* of the given concept, and not to improve the accuracy of reconstruction as in [Gal et al. 2023; Han et al. 2023; Shi et al. 2023; Tewel et al. 2023; Voynov et al. 2023; Wei et al. 2023].

3 PRELIMINARIES

Latent Diffusion Models. Diffusion models are generative models trained to learn data distribution by gradually denoising a variable sampled from a Gaussian distribution.

In our work, we use the publicly available text-to-image Stable Diffusion model [Rombach et al. 2022]. Stable Diffusion is a type of a latent diffusion model (LDM), where the diffusion process is applied on the latent space of a pretrained image autoencoder. The encoder $\mathcal{E}$ maps an input image x into a latent vector z, and the decoder $\mathcal{D}$ is trained to decode z such that $\mathcal{D}(z) \approx x$. As a second stage, a denoising diffusion probabilistic model (DDPM) [Ho et al. 2020] is trained to generate codes within the learned latent space. At each step during training, a scalar $t \in \{1, 2, ...T\}$ is uniformly sampled and used to define a noised latent code $z_t = \alpha_t z + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and α_t, σ_t are terms that control the noise schedule, and are functions of the diffusion process time t. The denoising network ϵ_θ which is based on a UNet architecture [Ronneberger et al. 2015], receives as input the noised code z_t, the timestep t, and an optional condition vector $c(y)$, and is tasked with predicting the added noise ϵ. The LDM loss is defined by:

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

For text-to-image generation the condition y is a text input and $c(y)$ represents the text embedding. At inference time, a random latent code $z_T \sim \mathcal{N}(0, I)$ is sampled, and iteratively denoised by the trained ϵ_θ until producing a clean z_0 latent code, which is passed through the decoder D to produce the image x.

We next discuss the text encoder and the inversion space.

Text embedding. Given a text prompt y, for example "A photo of a cat", the sentence is first converted into tokens, which are indexed into a pre-defined dictionary of vector embeddings. The

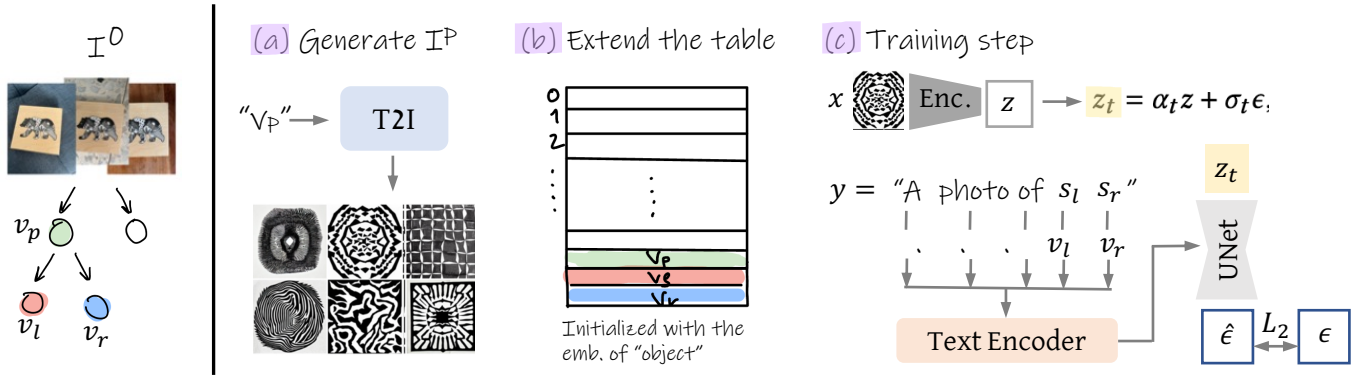


Fig. 4. High level pipeline of the "binary reconstruction" stage. We optimize two sibling nodes v_l, v_r at a time (marked in red and blue). (a) We first generate a small training set of images I^p depicting the concept in the parent node using a pretrained text-to-image model (T2I). At the root, we use the original set of images I^0. (b) We then extend the existing dictionary by adding the two new vectors, initialized with the embedding of the word "object". (c) Lastly, we optimize v_l, v_r w.r.t. the LDM loss (see details in the text).

dictionary is a lookup table that connects each token to a unique embedding vector. After retrieving the vectors for a given sentence from the table, they are passed to a text transformer, which processes the connections between the individual words in the sentence and outputs $c(y)$. The output encoding $c(y)$ is then used as a condition to the UNet in the denoising process. We denote words with S, and the vector embeddings from the lookup table with V.

Textual Inversion. We rely on the general framework proposed by [Gal et al. 2022], who choose the embedding space of V as the target for inversion. They formulate the task of inversion as fitting a new word s^* to represent a personal concept, depicted by a small set of input images provided by the user. They extend the predefined lookup table with a new embedding vector v_* that is linked to s^*. The vector v_* is often initialized with the embedding of an existing word from the dictionary that has some relation to the given concept, and then optimized to represent the desired personal concept. This process can be thought of as "injecting" the new concept into the vocabulary. The vector v_* is optimized w.r.t. the LDM loss in Equation (1) over images sampled from the input set. At each step of optimization, a random image x is sampled from the set, along with a neutral context text y, derived from the CLIP ImageNet templates [Radford et al. 2021b] (such as "A photo of s^*"). Then, the image x is encoded to $z = \mathcal{E}(x)$ and noised w.r.t. a randomly sampled timestep t and noise ϵ: $z_t = \alpha_t z + \sigma_t \epsilon$. The noisy latent image z_t, timestep t, and text embedding $c(y)$ are then fed into a pretrained UNet model which is trained to predict the noise ϵ applied w.r.t. the conditioned text and timestep. This way, v_* is optimized to describe the object depicted in the small training set of images. Note that while the supervision signal technically arrives from the reconstruction term in Equation (1), it encapsulates additional information from the knowledge of the pretrained model. The assumption behind why this process works for personalization is that the text-conditioned network could better denoise the image if it is provided with the correct descriptive object information.

4 METHOD

Given a small set of images $I^0 = \{I^0_1 ... I^0_m\}$ depicting the desired visual concept, our goal is to construct a rich visual exploration space expressing different aspects of the input concept.

We model the exploration space as a binary tree, whose nodes $V = \{v_1..v_n\}$ are learned vector embeddings corresponding to newly discovered words $S = \{s_1..s_n\}$ added to the predefined dictionary, representing different aspects of the original concept. These newly learned words are used as input to a pretrained text-to-image model [Rombach et al. 2022] to generate a rich variety of image examples in each node. We find a binary tree to be a suitable choice for our objective, because of the ease of visualization, navigation, and the quality of the sub-concepts depicted in the nodes (see supplemental file for further analysis).

4.1 Tree Construction

The exploration tree is built gradually as a binary tree from top to bottom, where we iteratively add two new nodes at a time. To create two child nodes, we optimize new embedding vectors according to the input image-set generated from the concept depicted in the parent node. During construction, we define two requirements to encourage the learned embeddings to follow the tree structure: (1) **Binary Reconstruction** each pair of children nodes together should encapsulate the concept depicted by their parent node, and (2) **Coherency** each individual node should depict a coherent concept which is distinct from its sibling. Next, we describe the loss functions and procedures designed to follow these requirements.

Binary Reconstruction. We use the reconstruction loss suggested in [Gal et al. 2022], with some modifications tailored to our goal. The procedure is illustrated in Figure 4 – in each optimization phase, our goal is to learn two vector embeddings v_l, v_r corresponding to the left and right sibling nodes, whose parent node is marked with v_p (illustrated in Figure 4, left). We begin with generating a new small training set of images $I^p = \{I^p_1 ... I^p_{10}\}$, reflecting the concept depicted by the vector v_p (Figure 4a). At the root, we use the

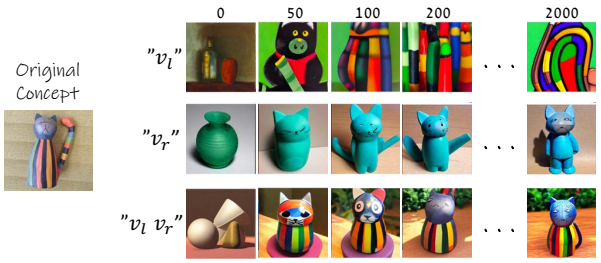


Fig. 5. Optimization iterations. The embedding of both children nodes v_l, v_r are initilized with the word "object". During iterations, they gradually depict two aspects of the original concept. Note that using both embedding together reconstructs the original parent concept.

original set of images I^0. Next, we extend the current dictionary by adding two new vector embeddings v_l, v_r, corresponding to the right and left children of their parent node v_p (Figure 4b). To represent general concepts, the newly added vectors are initialized with the embedding of the word "object". At each iteration of optimization (Figure 4c), an image x is sampled from the set I^p and encoded to form the latent image $z = \mathcal{E}(x)$. A timestep t and a noise ϵ are also sampled to define the noised latent $z_t = \alpha_t z + \sigma_t \epsilon$ (marked in yellow). Additionally, a neutral context text y is sampled, containing the new placeholder words in the following form "A photograph of s_l s_r", to reinforce that optimizing toward such concatenated text prompts ideally will capture naturally different concepts for the left and right nodes.

The noised latent z_t is fed to a pretrained Stable Diffusion UNet model ϵ_θ, conditioned on the CLIP embedding $c(y)$ of the sampled text, to predict the noise ϵ. The prediction loss is backpropagated w.r.t. the vector embeddings v_l, v_r:

$$\{v_l, v_r\} = \arg\min_{v} \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c(y))\|_2^2\right]. \quad (2)$$

This procedure encourages v_l, v_r together to express the visual concept of their parent depicted in the set I^p. Figure 5 illustrates how the two embeddings begin by representing the word "object", and gradually converge to depict two aspects of the input concept. We hypothesize that the nodes do not converge to a similar concept because of the model's training on natural sentences, where consecutive identical words are uncommon.

We use the timestep sampling approach proposed in ReVersion [Huang et al. 2023], which skews the sampling distribution so that a larger t is assigned a higher probability, according to the following importance sampling function:

$$f(t) = \frac{1}{T}(1 - \alpha \cos\frac{\pi t}{T}). \quad (3)$$

We set $\alpha = 0.5$. We find that this sampling approach improves stability and content separation. This choice is further discussed in the supplementary file.

Coherency. The resulting pair of embeddings described above together often capture the parent concept depicted in the original images well. However, the images produced by each embedding individually may not always reflect a logical sub-concept that is coherent to the observer.

Fig. 6. We demonstrate two sets of random images generated from two different vector embeddings. An example of a consistent set can be seen on the left, where the concept depicted in the node is clear. We show an inconsistent set on the right, where images appear to depict multiple concepts.

We find that such incoherent embeddings are frequently characterized by inconsistent appearance of the images generated from them, i.e., it can be difficult to identify a common concept behind them. For example, in Figure 6 the concept depicted in the set on the right is not clear, compared to the set of images on the left.

This issue may be related to the observation that textual inversion often results in vector embedding outside of the distribution of common words in the dictionary, affecting editability as well [Voynov et al. 2023]. It is thus possible that embeddings that are highly unusual may not behave as "real words", thereby producing incoherent visual concepts. In addition, textual-inversion based methods are sometimes unstable and depend on the seed and iteration selection.

To overcome this issue we define a consistency test, which allows us to filter out incoherent embeddings. We begin by running the procedure described above to find v_l, v_r using k different seeds in parallel for a sufficient number of steps (in our experiments we found that k=4 and 200 steps are sufficient since at that point the embeddings have already progressed far enough from their initialization word "object" as seen in Figure 5).

This gives us an initial set of k pairs of vector embeddings $V_s = \{v_l^i, v_r^i\}_{i=1}^k$. For each vector $v \in V_s$ we generate a random set I^v of 40 images using our pre-trained text-to-image model. We then use a pretrained CLIP Image encoder [Radford et al. 2021a], to produce the embedding $CLIP(I_i^v)$ of each image in the set.

We define the consistency of two sets of images I^a, I^b as follows:

$$C(I^a, I^b) = mean_{I_i^a\in I^a, I_j^b\in I^b, I_i^a\neq I_j^b}(sim(CLIP(I_i^a), CLIP(I_j^b))). \quad (4)$$

Note that $|C(I^a, I^b)| \leq 1$ because $sim(x, y) = \frac{x\cdot y}{||x||\cdot||y||}$ is the cosine similarity between a pair of CLIP embedding of two different images. This formulation is motivated by the observation that if a set of images depicts a certain semantic concept, their vector embedding in CLIP's latent space should be relatively close to each other. Ideally, we are looking for pairs in which each node is coherent by itself, and in addition, two sibling nodes are distinct from each other. We therefore choose the pair of tokens $\{v_l^*, v_r^*\} \in V_s$ as follows:

$$\{v_l^*, v_r^*\} = \arg\max_{\{v_l^i, v_r^i\}\in V_s}\left[C_l^i + C_r^i + (min(C_l^i, C_r^i) - C(I^{v_l^i}, I^{v_r^i}))\right], \quad (5)$$

where $C_l^i = C(I^{v_l^i}, I^{v_l^i}), C_r^i = C(I^{v_r^i}, I^{v_r^i})$. Note that we do not consider the absolute cross consistency score $C(I^{v_l^i}, I^{v_r^i})$, but we

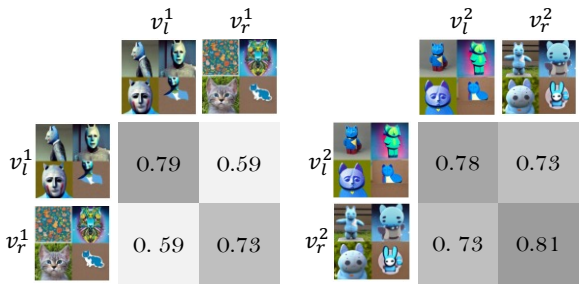


Fig. 7. Consistency scores matrix between image sample sets of nodes. The seed selection process favors pairs of siblings that have a high consistency score within themselves, and low consistency score between each other. In this example, the left pair is better than the right.

compute its relative difference from the node with the minimum consistency. We demonstrate this procedure in Figure 7. We optimized two pairs of sibling nodes $\{v_l^1, v_r^1\}, \{v_l^2, v_r^2\}$ using two seeds, w.r.t. the same parent node. Each matrix illustrates the consistency scores $C_l^i, C(I^{v_l^i}, I^{v_r^i}), C_r^i$ obtained for the sets of images of each seed. In both cases, the scores on the diagonal are high, which indicates that each set is consistent within itself. While the sets on the right obtained a higher consistency score within each node, they also obtained a relatively high score across the nodes (0.73), which means they are not distinct enough.

After selecting the optimal seed, we continue the optimization of the chosen vector pair w.r.t. the reconstruction loss in Equation (2) for 1500 iterations.

5 RESULTS

In Figures 1, 13 and 14, we show examples of possible trees. For each node in the tree, we use its corresponding placeholder word as an input to a pretrained text-to-image model [Rombach et al. 2022], to generate a set of random images. These images have been generated without any prompt engineering or additional words within the sentence, except for the word itself. For clarity, we use the notion "v" next to each set of images, illustrating that the presented set depicts the concept learned in that node. As can be seen, the learned embeddings in each node capture different elements of the original concept, such as the concept of a cat and a sculpture, as well as the unique texture in Figure 13. The sub-concepts captured in the nodes follow the tree's structure, where the concepts are decomposed gradually, with two sibling nodes decomposing their parent node. This decomposition is done *implicitly*, without external guidance regarding the split theme. For many more trees please see our supplementary file.

5.1 Applications

The constructed tree provides a rich visual exploration space for concepts related to the object of interest. In this section we demonstrate how this space can be used for novel combination and exploration.
Intra-tree combination – the generated tree is represented via the set of optimized vectors $V = \{v_1..v_n\}$. Once this set is learned we can use it to perform further exploration and conceptual editing *within* the object's "inner world". We can explore combinations of different aspects by composing sentences containing different subsets of V. For example, in the bottom left area of Figure 13, we have combined v_1 and v_5, which resulted in a variation of the original sculpture without the sub-concept relating to the cat (depicted in v_6). At the bottom right, we have excluded the sub-concept depicted in v_5 (related to a blue sculpture), which resulted in a new representation of a flat cat with the body and texture of the original object.

Such combinations can provide new perspectives on the original concept and inspiration that highlights only specific aspects.
Inter-tree combination – it is also possible to combine concepts learned across different trees, since we only inject new words into the existing dictionary, and do not fine-tune the model's weights as in other personalization approaches [Ruiz et al. 2023].

To achieve this, we first build the trees independently for each concept and then visualize the sub-concepts depicted in the nodes to select interesting combinations. In Figure 8 the generated original concepts are shown on top, along with an illustration of the concepts depicted in the relevant nodes. To combine the concepts across the trees, we simply place the two placeholder words together in a sentence and feed it into the pretrained text-to-image model. As can be seen, on the left the concept of a "saucer with a drawing" and the "creature" from the mug are combined to create many creative and surprising combinations of the two. On the right, the blue sculpture of a cat is combined with the stone depicted at the bottom of the Buddha, which together create new sculptures in which the Buddha is replaced with the cat.
Text-based generation – the placeholder words of the learned embeddings can be composed into natural language sentences to generate various scenes based on the learned aspects. We illustrate this at the top of Figure 9, where we integrate the learned aspects of the original concepts in new designs (in this case of a chair and a dress). At the bottom of Figure 9, we show the effect of using the learned vectors of the original concepts instead of specific aspects. We apply Textual Inversion (TI) [Gal et al. 2022] with the default hyperparameters to fit a new word depicting each concept, and choose a representative result. The results suggest that without aspect decomposition, generation can be quite limited. For instance, in the first column, both the dress and the chair are dominated by the texture of the sculpture, whereas the concept of a blue cat is almost ignored. Furthermore, TI may exclude the main object of the sentence (second and third columns), or the results may capture all aspects of the object (fourth column), thereby narrowing the exploration space.

5.2 Evaluations

Consistency Score Validation. We first show that our consistency test proposed in Equation (4) aligns well with human perception of consistency. We conducted a perceptual study with 35 participants in which we presented 15 pairs of random image sets depicting sub-concepts of 9 objects. We asked participants to determine which of the sets is more consistent within itself in terms of the concept it depicts (an example of such a pair can be seen in Figure 6). We also measured the consistency scores for these sets using our CLIP-based approach, and compared the results. The CLIP-based scores matched the human choices in 82.3% (stdv: 1%) of the cases.

Reconstruction and Separation. We quantitatively evaluate our method's ability to follow the tree requirements of reconstruction

Fig. 8. Examples of inter-tree combinations. We use our method to produce trees for the four concepts depicted in the first row. We then combine aspects from different trees to generate a set of inter-tree combinations (the chosen aspects are shown next to each concept). We also show combinations of three aspects from different trees at the bottom.

and sub-concept separation. We collected a set of 13 concepts (9 from existing personalization datasets [Gal et al. 2022; Kumari et al. 2023], and 4 new concepts from our dataset), and generated 13 corresponding trees. Note that we chose concepts that are complex enough and have the potential to be divided into different aspects (we discuss this in the limitations section). For each pair of sibling nodes v_l, v_r and their parent node v_p, we produced their corresponding sets of images – $I^{v_l}, I^{v_r}, I^{v_p}$ (where for nodes in the first level we used the original set of images I^0 as I^{v_p}). We additionally produced the set $I^{v_l v_r}$, depicting the joint concept learned by two sibling nodes.

We first compute $C(I^{v_p}, I^{v_l v_r})$ to measure the quality of reconstruction, i.e., that two sibling nodes together represent the concept depicted in their parent node. The average score obtained for this measurement is 0.8, which suggests that on average, the concept depicted by the children nodes together is consistent with that of their parent node. Second, we measure if two sibling nodes depict distinct concepts by using $C(I^{v_l}, I^{v_r})$. The average score obtained was 0.59, indicating there is larger separation between siblings, but they are still close.

Aspects Relevancy. We assess the ability of our method to encode different aspects connected to the input concept via a perceptual study. We chose 5 objects from the dataset above, and 3 random aspects for each object. We presented participants with a random set of images depicting one aspect of one object at a time. We asked the participants to choose the object they believe this aspect originated from, along with the option 'none'. In total we collected answers from 35 participants, and achieved recognition rates of 87.8% (stdv: 1%). These evaluations demonstrate that our method can indeed separate a concept into *relevant* aspects, where each new sub-concept is *coherent*, and the binary tree structure is valid - i.e., the combination of two children can *reconstruct* the parent concept.

5.3 Ablation

Binary Tree. Our choice to use a binary tree stems from two main reasons: (1) complexity, and (2) consistency.

Our method allows to build a tree with more than two children per node, however, this can add redundant complexity to the method. For example, using three children nodes, after only two levels we will get 12 aspects, which may be difficult to visualize and navigate. In addition, the use of more than two children will result in a longer running time in each level since we will have to split more nodes.

In terms of consistency, we observe that when optimizing more than two nodes at a time, the chance of receiving inconsistent nodes increases. Often, two nodes will be consistent, and the third node is inconsistent or may depict irrelevant concepts such as background. We visually demonstrate this in Figure 10, on the "red teapot" object. We present the aspects obtained from the optimal seed after 200 iterations, for the case of two nodes (left) and for the case of three nodes (right). As can be seen, the sub-concept in v_3 for the 3 nodes optimization does not appear to be consistent or comprehensible, and therefore is not useful in achieving our goal of extracting aspects from the parent concept.

The following quantitative experiment further confirms this observation. We obtained 52 trees for our set of 13 objects (using four seeds for each object as described in the main paper). Each tree

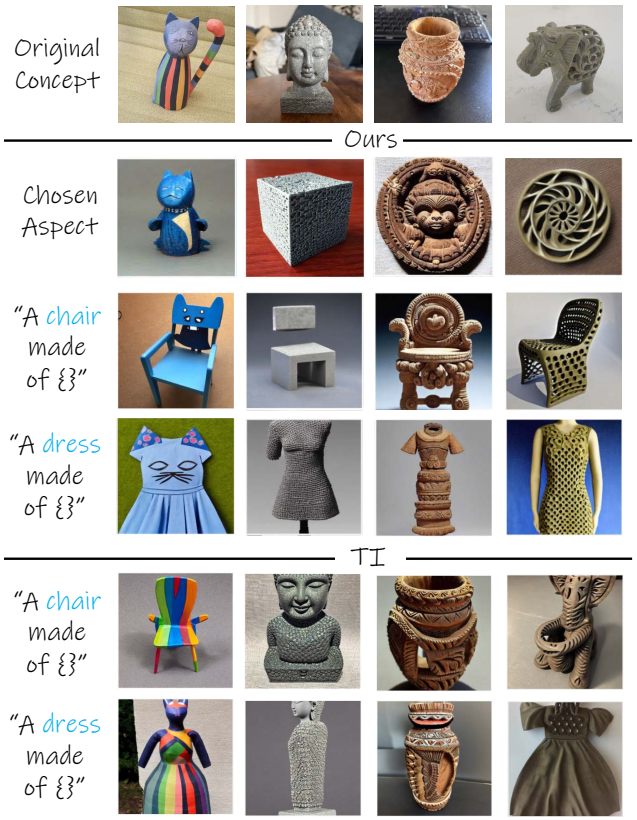


Fig. 9. Combining the learned aspects in natural sentences to produce aspect-based variations. The original concepts are shown at the top. In the third and fourth rows are our text-based generation results applied with the aspects depicted in the second row. Under "TI" we show image generation for the concepts in the first row (without our aspect decomposition approach), produced using [Gal et al. 2022].

is a 3-node tree with one level, resulting in a total number of 156 nodes. We measured our CLIP-based consistency test on each node to determine its average consistency score. Next, for each tree, we sorted the 3 nodes $\{v_1, v_2, v_3\}$ according to their consistency score, from the most consistent (v_1) to the least consistent (v_3). We then average the scores of $\{v_1, v_2, v_3\}$ across all trees, and received the final scores of: 0.804, 0.742, 0.633 for $\{v_1, v_2, v_3\}$ correspondingly. The noticeable consistency gap between the top 2 nodes and the third node indicates that, on average, two of the three nodes are consistent, while the last may contain incoherent information. This experiment correlates well with our visual observation (as demonstrated in Figure 10).

5.4 Vectors Initialization

As mentioned in Section 4.1, we use the embedding of the word "object" to initialize the new vectors v_l, v_r. In choosing the word "object" as a generic concept, we eliminate the requirement for user-defined input specific to the concept. Furthermore, the use of a most general concept for initialization allows for more unexpected decompositions to occur.

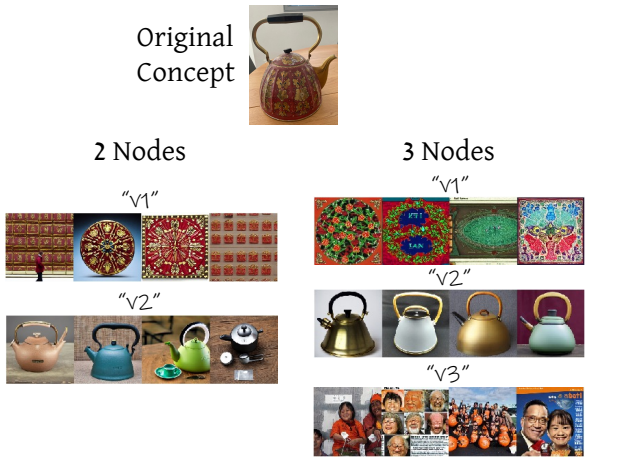


Fig. 10. Comparison of optimizing for two child nodes (left) v.s. three child nodes (right). Using three nodes increases the chance of arriving at inconsistent or irrelevant concepts.

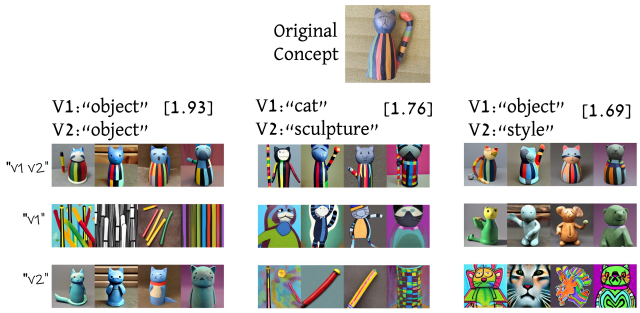


Fig. 11. Comparing different initialization approaches. The columns show the different initialization used, and the rows show the results after 200 iterations.

It is possible, however, for the user to select other words for initialization if they wish to encourage the generation of certain sub-aspects. In Figure 11, we demonstrate the impact of using different initialization approaches. The first column illustrates the baseline results, using "object" to initialize both nodes. The second column displays the results of initialization based on potential user defined inputs – a "cat" and a "sculpture" for the left and right nodes, respectively. In the third column, we present an alternative generic initialization option consisting of "object" and "style". Each experiment was conducted using four different seeds and 200 iterations, with similar results.

As can be seen in the first column the results we get are consistent, with a maximum consistency score of 1.93. In the second column, we can see that v1 represents the cat aspect and v2 represents a concept that is more akin to a sculpture, with a consistency score of 1.76. The third column indicates that using "style" instead of "object" negatively impacted the results, resulting in a consistency score of 1.69. More examples can be found in the supplemental file.

6 LIMITATIONS

Our method may fail to decompose an input concept. We divide the failure cases into four general categories illustrated in Figure 12:
(1) Background leakage - the training images should be taken from different perspectives and with varying backgrounds (this requirement also exists in [Gal et al. 2022]). When images do not meet these criteria, one of the sibling nodes often captures information from the background instead of the object itself.
(2) Incomprehensible aspects - some separations may not satisfy clear, interesting, aesthetic, or inspiring aspects, even when the coherency principle holds.
(3) Dominant sub-concept - we illustrate this in Figure 12c, where we show a split on the second level of the concept depicted under "v_1v_2". As shown, v1 has dominated the information, so even if the coherency term is held, decomposition to two sub-concepts has not really been achieved.
(4) Large overlap when two aspects share information – we illustrate this in Figure 12d, which is a split of the second level, where both concepts depicted in v1 and v2 appear to share too similar.

We hope that such limitations could be resolved in the future using additional regularization terms in the optimization process or through the development of more robust personalization methods.

Additionally, our method can have difficulties to create deeper trees. Our observations show two main factors influencing whether a node could be further split – the complexity of the concept depicted in the node and its' coherency. As we go deeper into the tree, the concepts become simpler and more challenging to decompose. When a concept reaches one of these conditions we stop the tree growth. This opens interesting avenues for future research to explore how can concept trees be further extended.

Currently the time for decomposing a node can reach up to approximately 40 minutes on a single A100 GPU. However, as textual inversion optimization techniques will progress, so will our method.

7 CONCLUSIONS

We presented a method to implicitly decompose a given visual concept into various aspects to construct an inspiring visual exploration space. Our method can be used to generate numerous representations and variations of a certain subject, to combine aspects across objects, as well as to use these aspects as part of natural language sentences that drive visual generation of novel concepts.

The aspects are learned implicitly, without external guidance regarding the type of separation. This implicit approach also provides another small step in revealing the rich latent space of large vision-language models, allowing surprising and creative representations to be produced. We demonstrated the effectiveness of our method on a variety of challenging concepts. We hope our work will open the door to further research aimed at developing and improving existing tools to assist and inspire designers and artists.

8 ETHICAL CONSIDERATIONS

One of the drawbacks of text-to-image models is their tendency to inherit biases from the large-scale internet data used during training. These biases naturally extend to our decomposition approach, potentially resulting in biased or stereotypical representations of specific objects or concepts. Hence, it is crucial to exercise caution when using our method and be mindful of these biases.

However, our approach can serve as a means of detecting biases by analyzing the tokens learned during the decomposition process. This underscores the need for further research into concept representations in text-to-image models, given the significant impact that bias can have on image generation.

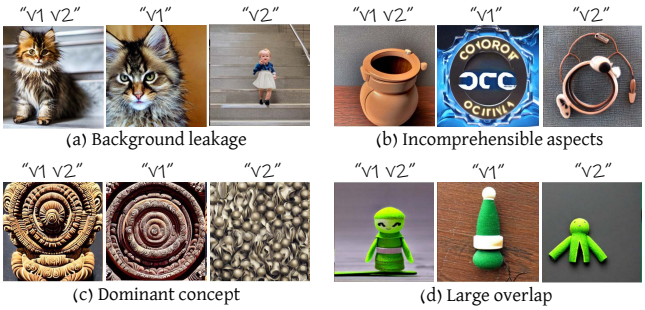


Fig. 12. We demonstrate four general cases of decomposition failure.

REFERENCES

Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. 2018. Image Understanding using vision and reasoning through Scene Description Graph. *Comput. Vis. Image Underst.* 173 (2018), 33–45. https://api.semanticscholar.org/CorpusID:13398739

Teresa M. Amabile. 1996. Creativity In Context: Update To The Social Psychology Of Creativity.

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. 2021. SegDiff: Image Segmentation with Diffusion Probabilistic Models. https://doi.org/10.48550/ARXIV.2112.00390

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 6077–6086. https://api.semanticscholar.org/CorpusID:3753452

Melinos Averkiou, Vladimir G. Kim, Youyi Zheng, and Niloy Jyoti Mitra. 2014. ShapeSynth: Parameterizing model collections for coupled shape exploration and synthesis. *Computer Graphics Forum* 33 (2014).

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18208–18218.

Nathalie Bonnardel and Evelyne Cauzinille-Marmèche. 2005. Towards supporting evocation processes in creative design: A cognitive approach. *Int. J. Hum. Comput. Stud.* 63 (2005), 422–435.

David C. Brown. 2008. GUIDING COMPUTATIONAL DESIGN CREATIVITY RESEARCH.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text Retrieval: A Survey on Recent Research and Development. *ArXiv* abs/2203.14713 (2022). https://api.semanticscholar.org/CorpusID:247763152

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Loddon Yuille. 2014. Detect What You Can: Detecting and Representing Objects Using Holistic Models and Body Parts. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 1979–1986. https://api.semanticscholar.org/CorpusID:2256682

Lydia B. Chilton, S. Petridis, and Maneesh Agrawala. 2019. VisiBlends: A Flexible Workflow for Visual Blends. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). https://api.semanticscholar.org/CorpusID:96456140

Daniel Cohen-Or and Hao Zhang. 2015. From inspired modeling to creative modeling. *The Visual Computer* 32 (2015), 7–14.

Jonathan D. Denning and Fabio Pellacini. 2013. MeshGit: diffing and merging meshes for polygonal modeling. *ACM Trans. Graph.* 32 (2013), 35:1–35:10. https://api.semanticscholar.org/CorpusID:16791519

designingbuildings wiki. 2021. Beijing National Stadium Design. https://www.designingbuildings.co.uk/wiki/Beijing_National_Stadium

Jozef Dobos and Anthony Steed. 2012. 3D diff: an interactive approach to mesh differencing and conflict resolution. *ACM SIGGRAPH 2012 Talks* (2012). https://api.semanticscholar.org/CorpusID:5661874

Claudia Eckert and Martin Stacey. 2000. Sources of inspiration: a language of design. *Design Studies* 21, 5 (2000), 523–538. https://doi.org/10.1016/S0142-694X(00)00022-3

Mohamed Elhoseiny and Mohamed Elfeki. 2019. Creativity Inspired Zero-Shot Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 5783–5792.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. https://doi.org/10.48550/ARXIV.2208.01618

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Designing an Encoder for Fast Personalization of Text-to-Image Models. *ArXiv* abs/2302.12228 (2023).

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. arXiv:2108.00946 [cs.CV]

Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. 2014. What inspires designers? Preferences on inspirational approaches during idea generation. *Design Studies* 35 (2014), 29–53.

Chi Han, Jiayuan Mao, Chuang Gan, Joshua B. Tenenbaum, and Jiajun Wu. 2020. Visual Concept-Metaconcept Learning. In *Neural Information Processing Systems*.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. *ArXiv* abs/2303.11305 (2023).

K Henderson. 1999. On line and on paper: visual representations, visual culture, and computer graphics in design engineering. Inside technology.

Geoffrey E. Hinton. 1979. Some Demonstrations of the Effects of Structural Descriptions in Mental Imagery. *Cogn. Sci.* 3 (1979), 231–250.

Geoffrey E. Hinton. 2021. How to Represent Part-Whole Hierarchies in a Neural Network. *Neural Computation* 35 (2021), 413–452.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. 2011. Transforming Auto-Encoders. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I* (Espoo, Finland) *(ICANN'11)*. Springer-Verlag, Berlin, Heidelberg, 44–51.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *CoRR* abs/2006.11239 (2020). arXiv:2006.11239 https://arxiv.org/abs/2006.11239

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. 2023. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).

Alexander Ivanov, David Ledo, Tovi Grossman, George Fitzmaurice, and Fraser Anderson. 2022. MoodCubes: Immersive Spaces for Collecting, Discovering and Envisioning Inspiration Materials. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 189–203. https://doi.org/10.1145/3532106.3533565

Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting Visual Metaphor Ideation through Multi-Dimensional Example-Based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 427, 15 pages. https://doi.org/10.1145/3411764.3445325

Anna Kantosalo, Jukka M. Toivanen, Ping Xiao, and Hannu (TT) Toivonen. 2014. From Isolation to Involvement: Adapting Machine Creativity Software to Support Human-Computer Co-Creation. In *International Conference on Innovative Computing and Cloud Computing*.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *ArXiv* abs/1406.5679 (2014). https://api.semanticscholar.org/CorpusID:2315434

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *ArXiv* abs/1411.2539 (2014). https://api.semanticscholar.org/CorpusID:7732372

Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI?: Design Ideation with Cooperative Contextual Bandits. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E. Mackay. 2020. ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 45 (may 2020), 27 pages. https://doi.org/10.1145/3392850

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. (2023).

Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. 2016. Semantic Object Parsing with Graph LSTM. *ArXiv* abs/1603.07063 (2016). https://api.semanticscholar.org/CorpusID:7886345

Joe Marks, Brad Andalman, Paul A. Beardsley, William T. Freeman, Sarah F. Frisken, Jessica K. Hodgins, T. Kang, Brian Mirtich, Hanspeter Pfister, Wheeler Ruml, Kathy Ryall, Joshua E. Seims, and Stuart M. Shieber. 1997. Design galleries: a general approach to setting parameters for computer graphics and animation. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997). https://api.semanticscholar.org/CorpusID:221894682

Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George W. Fitzmaurice. 2018. Dream Lens: Exploration and Visualization of Large-Scale Generative Design Datasets. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018). https://api.semanticscholar.org/CorpusID:5041597

Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, and Joshua B. Tenenbaum. 2022. FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations. *ArXiv* abs/2203.16639 (2022). https://api.semanticscholar.org/CorpusID:247839322

Midjourney. 2022. Midjourney.com. https://www.midjourney.com

W. Muller. 1989. Design discipline and the significance of visuo-spatial thinking. *Design Studies* 10, 1 (1989), 12–23. https://doi.org/10.1016/0142-694X(89)90021-5

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

Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. *Proceedings of the 25th International Academic Mindtrek Conference* (2022).

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

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*.

Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal* 24 (2012), 92 – 96.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic Routing Between Capsules. *ArXiv* abs/1710.09829 (2017).

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. KNN-Diffusion: Image Generation via Large-Scale Retrieval. arXiv:2204.02849 [cs.CV]

Jing Shi, Wei Xiong, Zhe L. Lin, and Hyun Joon Jung. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *ArXiv* abs/2304.03411 (2023).

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *ArXiv* abs/2305.01644 (2023).

Yingtao Tian and David Ha. 2021. Modern Evolution Strategies for Creativity: Fitting Concrete Images and Abstract Concepts. *CoRR* abs/2109.08857 (2021). arXiv:2109.08857 https://arxiv.org/abs/2109.08857

Kewei Tu, Maria Pavlovskaia, and Song-Chun Zhu. 2013. Unsupervised Structure Learning of Stochastic And-Or Grammars. In *NIPS*. https://api.semanticscholar.org/CorpusID:14785730

Zhuowen Tu, Xiangrong Chen, Alan Loddon Yuille, and Song-Chun Zhu. 2003. Image Parsing: Unifying Segmentation, Detection, and Recognition. *International Journal of Computer Vision* 63 (2003), 113–140. https://api.semanticscholar.org/CorpusID:1752880

Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, and Ariel Shamir. 2022a. CLIPascene: Scene Sketching with Different Types and Levels of Abstraction. https://doi.org/10.48550/ARXIV.2211.17256

Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022b. CLIPasso:

Semantically-Aware Object Sketching. *ACM Trans. Graph.* 41, 4, Article 86 (jul 2022), 11 pages. https://doi.org/10.1145/3528223.3530068

Andrey Voynov, Q. Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. *ArXiv* abs/2303.09522 (2023).

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *ArXiv* abs/2302.13848 (2023).

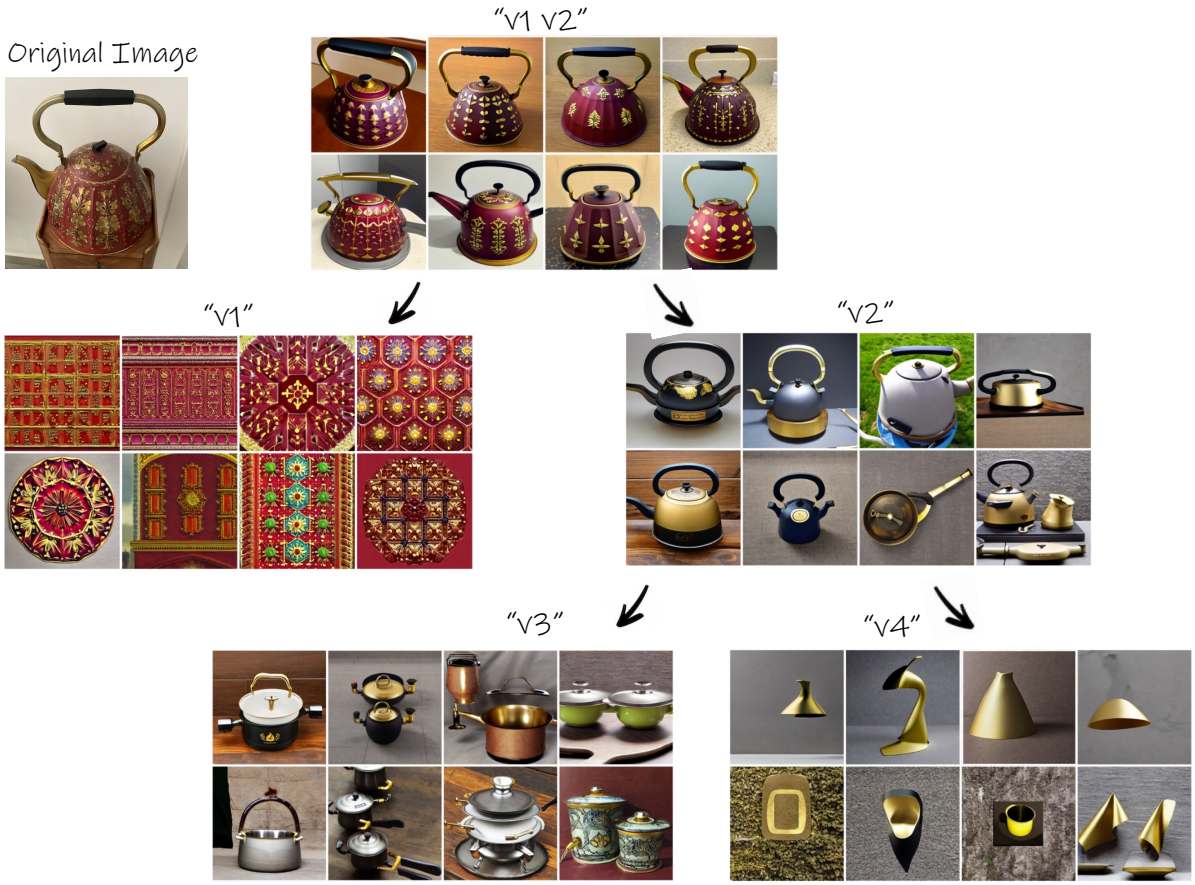
Merryl J. Wilkenfeld and Thomas B. Ward. 2001. Similarity and Emergence in Conceptual Combination. *Journal of Memory and Language* 45, 1 (2001), 21–38. https://doi.org/10.1006/jmla.2000.2772

Kai Xu, Hao Zhang, Daniel Cohen-Or, and Baoquan Chen. 2012. Fit and diverse. *ACM Transactions on Graphics (TOG)* 31 (2012), 1 – 10.

Quanshi Zhang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu. 2016. Growing Interpretable Part Graphs on ConvNets via Multi-Shot Learning. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:15371679

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5122–5130.

Fig. 13. Exploration tree for the cat sculpture. At the bottom we show examples of possible intra-tree combinations.



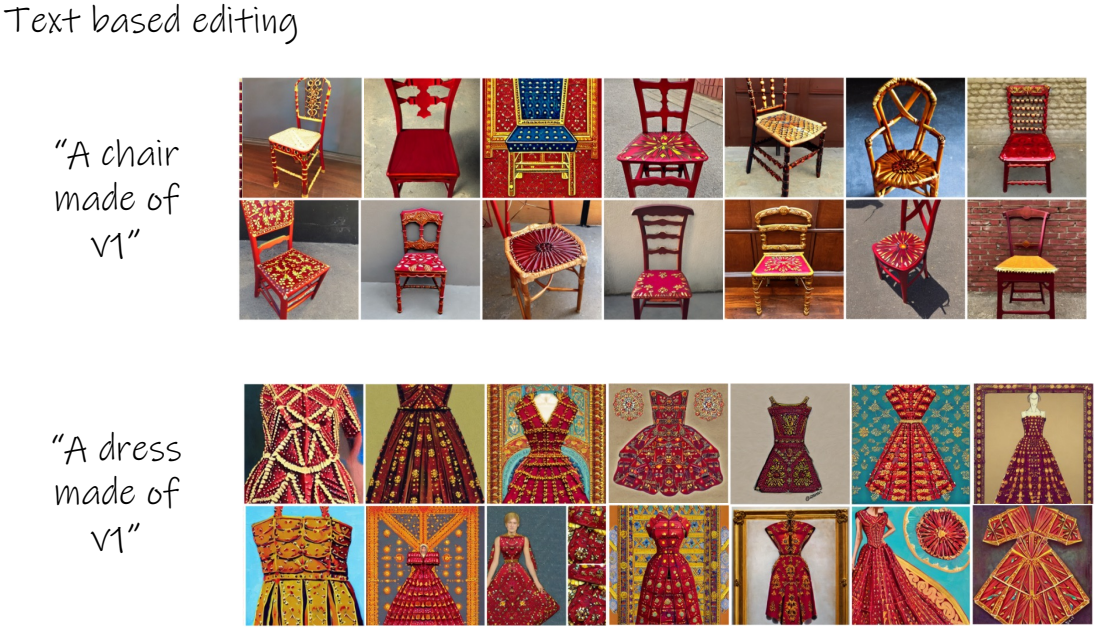


Fig. 14. Exploration tree for a decorated teapot. At the bottom we show examples of possible text-based generation.