Concept Decomposition for Visual Exploration and Inspiration – Supplementary Material

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

ACM Reference Format:
Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration – Supplementary Material. *ACM Trans. Graph.* 1, 1 (September 2023), 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Contents

This document provides additional details of our proposed method, including implementation details, analysis, and ablation. Additionally, we provide many additional qualitative results, including trees for more objects, inter- and intra-combinations, and text-based generation results.

1 IMPLEMENTATION DETAILS

We rely on the diffusers [von Platen et al. 2022] implementation of Textual Inversion [Gal et al. 2022], based on Stable Diffusion v1.5 text-to-image model [Radford et al. 2021]. We used the default training parameters provided in this implementation, except for changing the batch size to 2 (which scales the learning rate to 0.004 respectively). We used four different seeds $\{0, 1000, 1234, 111\}$ for each sibling nodes optimization. To generate the set of 10 images for each new node we first generated a random set of 40 images, and used our proposed CLIP consistency measurement to choose a subset of 10 images that are most consistent with each other. Our code will be made available to facilitate future research.

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

2 BASELINES

In the absence of existing works attempting to achieve our goal of decomposition into different aspects, we compare our performance in intra-tree combination with two existing relevant works.

We consider Textual Inversion [Gal et al. 2022], and its more advanced modification – Extended Textual Inversion [Voynov et al. 2023] – designed specifically for appearance mixing (which is most similar to our "intra-tree combination").

We provide a qualitative comparison to these methods in Figure 1. On the left we aim to combine the aspect of a wooden saucer and the creature on the cup from the objects presented on top. On the right we aim to combine a part of the stone statue with some specific style aspects of the cat sculpture.

Gal et al. [2022] propose a style transfer application, in which their method can be used to find pseudo words representing a specific style taken from a given concept, and can then be applied in combination with other concepts. To extract the style code from a given concept, they replace the training texts with prompts of the form: "A painting in the style of S^*".

For the TI baseline, we applied the original Textual Inversion for the first concept (from which we wish to take the structure), and for the appearance concept we used their proposed style extraction application described above. This results in a pair of textual tokens S_1^{TI}, S_2^{TI} that represent each concept. We explicitly combine these tokens in a sentence, providing the desired mixing description (e.g. "S_1^{TI} in the style of S_2^{TI}" and use it to generate an image.

Voynov et al. [2023] propose an extended textual conditioning space for a diffusion model that can be used to control style and geometry disentanglement. The main idea is to provide each diffusion UNet cross-attention layer with an independent textual prompt. The authors notice that low-resolution UNet layers are commonly responsible for geometrical attributes, while high-resolution input and output layers are responsible for style-related attributes.

For the task of style mixing, given a pair of objects, the method performs two independent Textual Inversions to this extended prompt space (called XTI in the paper). Then, the low-resolution layers are provided with the inversion of the object that donors the shape, and the high-resolution layers are provided with the inversions of the object that donors the appearance.

For the comparison to XTI [Voynov et al. 2023], we use their recommended hyperparameters. We apply two independent Textual Inversions to the extended prompt space, which brings the pair

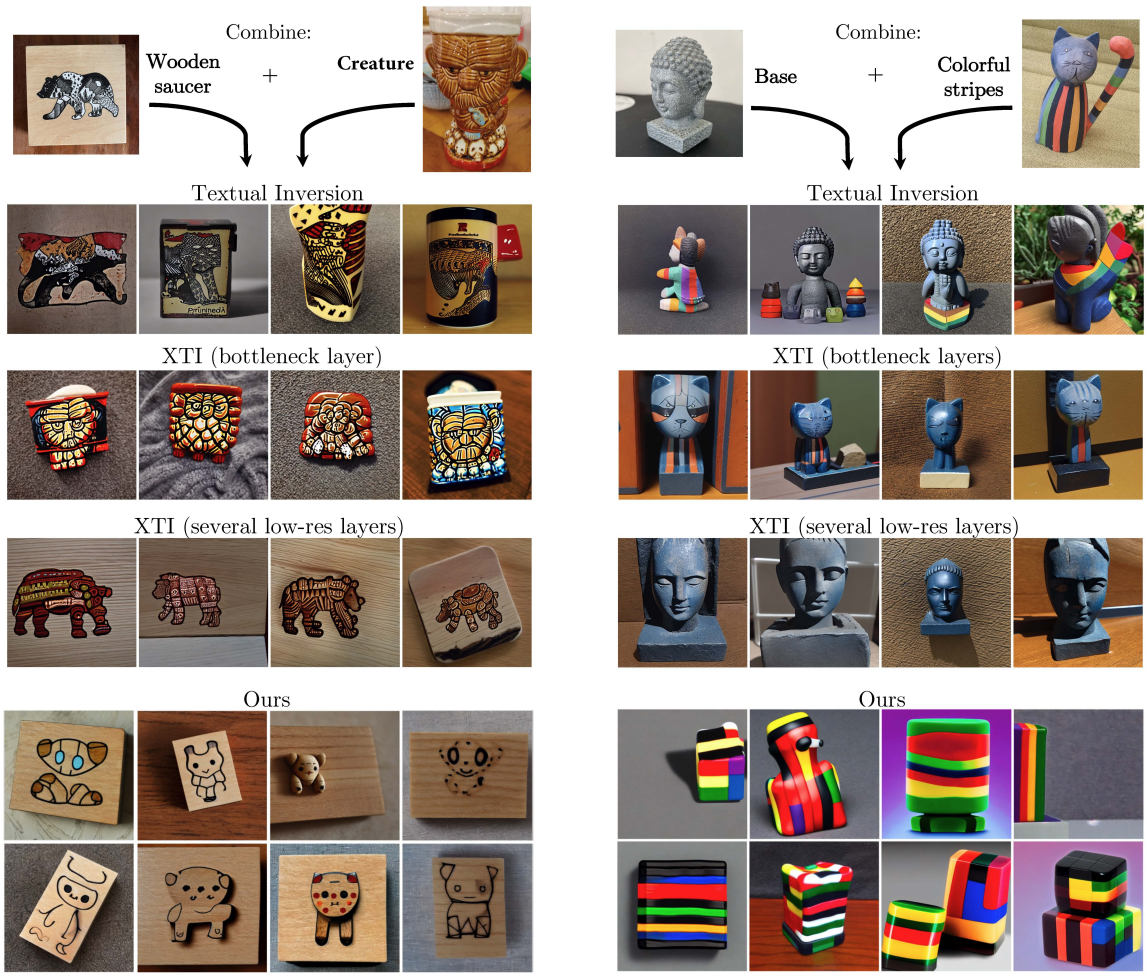


Fig. 1. Comparison of blending specific aspects of concepts. Top row are the source objects and a description of which aspect is taken. Second row are the results of blending the chosen concepts using Textual Inversion [Gal et al. 2022]. Third row are the results of blending with Extended Textual Inversion [Voynov et al. 2023] (XTI) when only bottleneck layers are provided with the left object. Fourth row are results of XTI where a wider range of the low-resolution layers are provided with the left object. Last two rows are images generated with our proposed approach.

of textual tokens S_1^{XTI}, S_2^{XTI}. To use the geometry from S_1^{XTI} and appearance of S_2^{XTI}, we provided the prompt "a photo of S_1^{XTI}" to the deeper (low-res) layers and "a photo of S_2^{XTI}" to the shallower (high-res) UNet layers. We tried to combine the concepts using different layers split to achieve the best possible performance.

From Figure 1, we can see that these baselines fail to combine the very specific aspects of the source objects. Textual Inversion commonly blends the attributes, while XTI is able to transfer either the whole creature's appearance, or texture only, failing to extract only the shape. In contrast, using our approach it is possible to pick the two distinct aspects and combine them naturally to depict a new concept.

3 ABLATION AND ANALYSIS

3.1 Timestep Sampling

As discussed in Section 4.1 of the main paper, we use the timestep sampling approach proposed in ReVersion [Huang et al. 2023], favoring larger t values. This sampling approach plays a significant role in the success of our method, as demonstrated in Figure 2.

The left side of Figure 2 shows the results obtained when using a uniform sampling approach (which is the more common approach in LDM-based optimization), the right side shows the results obtained when using the sampling method we selected from ReVersion.

In both cases, the results were obtained after 500 iterations with the same seed and settings. As can be seen, the uniform timestep sampling approach negatively affects both reconstruction quality (see "v1 v2") and decomposition quality, where for example for the cat sculpture the aspect depicted in "v1" is unrelated to the original concept.

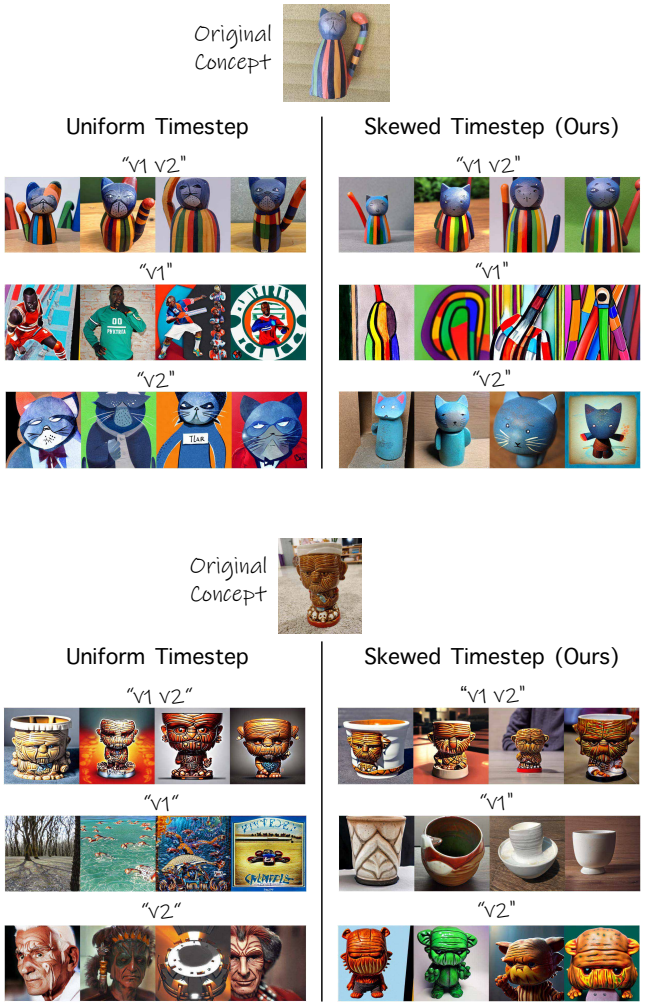


Fig. 2. Timestep sampling approach ablation. We show the effect of using a uniform sampling (left), compared to using the sampling approach from ReVersion [Huang et al. 2023], which favor larger values of t.

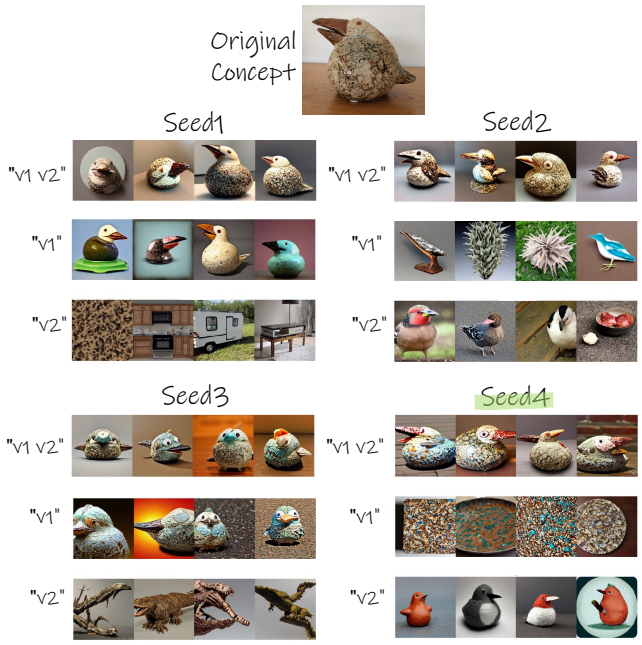


Fig. 3. Results of four different seeds after 200 steps. The best seed is marked in green.

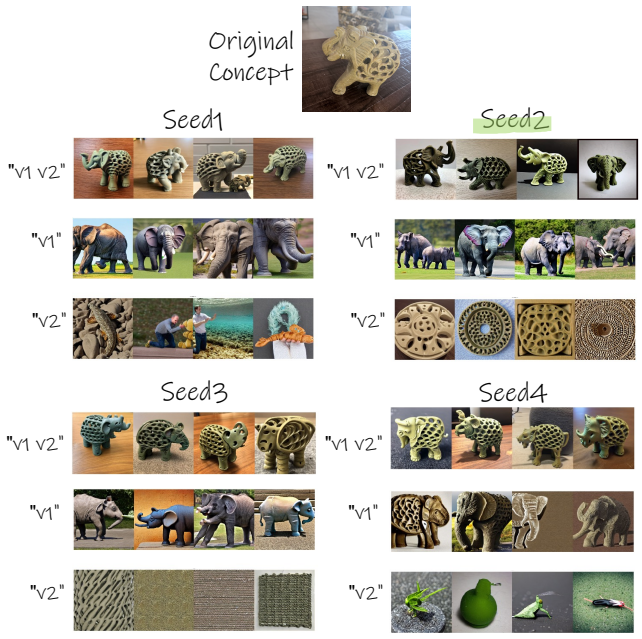


Fig. 4. Results of four different seeds after 200 steps. The best seed is marked in green.

3.2 Consistency Test

In this section we provide examples and details regarding our proposed CLIP-based consistency test presented in Section 4.1 in the main paper. First, we visually demonstrate the effect of using $k = 4$ seeds in each run. We observe that 4 seeds are generally enough for most of the concepts, and in most cases also 2 seeds may be good enough. However we do note that the variability in results among the different seeds can be quite meaningful in some cases. We demonstrate this in Figures 3 and 4, where we show the original concept on top, along with the random set of images generated for each nodes in each of the seeds.

The seed that was chosen using our CLIP-based consistency measurement is marked in green. While the results depicted in Figure 3 were reasonable for most of the seeds, in Figure 4 we can see that seed1 and seed2 are failure cases, where in seed 1 the concept depicted in v_2 is inconsistent and not interpretable, and in seed4 we have a case of one dominant node (v_1).

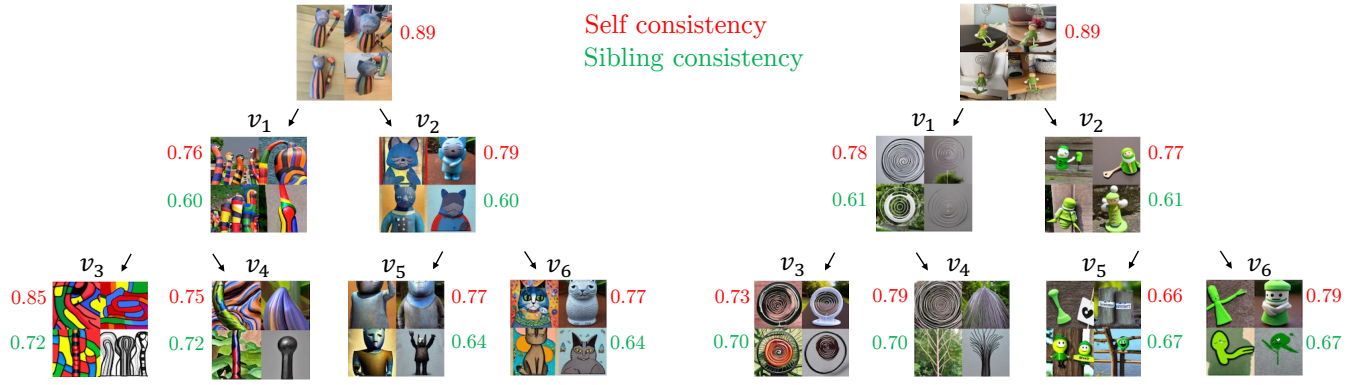


Fig. 5. An illustration of two trees with different characteristics. The original training set is depicted at the root of the trees. Next to each node we present its self-consistency score (in red) and the consistency score of that node with its brother node (in green). The scores were obtained using our CLIP-based consistency measurement described in the main paper.

Additionally, we provide an illustration to better clarify the significance of the consistency scores and their relationship to the patterns observed visually in the trees. In Figure 5 we show examples of two trees with different characteristics. Next to each node are the self-consistency score (marked in red) and the siblings consistency score (marked in green). The first score measures the degree to which the images depicted in a specific node are consistent withing themselves. The second score indicates the similarity between sibling nodes.

First, observe that the self consistency score for the root node (0.89) is the highest, since the images depicted in that node originated from the set provided by the user. This indicates the highest consistency score possible in our settings. In addition, we observe that the self consistency score across most nodes is relatively high and does not vary significantly as we go deeper in the tree. However, v_5 in the right tree obtained a self consistency score of 0.66, which is relatively low, and in our scale it means that the set is not considered consistent.

Considering that this node is not consistent with itself, it is obvious that it is not consistent with its sibling node, which is why, in such cases, we can ignore the score obtained in green for that node in this discussion.

We now examine the scores in green, which indicate consistency across siblings. First, note that in both trees, the consistency across siblings is low (0.6 and 0.61) in the first level, suggesting that a good separation has been achieved. However, at the second level we can see that this score generally increased, indicating that the quality of separation decreases as we go deeper in the tree. Additionally, the sibling similarity correlates well with the visual information, with v_3, v_4 in the left tree and v_3, v_4 and v_5, v_6 in the right tree appearing to be more consistent than the other pairs.

It is important to note that in these cases, when the consistency among siblings is high, or when one node is inconsistent within itself, the split will be stopped at this particular level.

In order to confirm this observation, we measured these scores for the set of 13 trees that were used for the other evaluations. For each node, we calculated the self consistency score as well as the sibling consistency score, and averaged these scores across the trees.

Table 1. Average self consistency (left) and sibling consistency (right) scores. The scores were obtained for 13 trees.

Node	Self Cons.	Avg. Level1	Node	Sibling Cons.	Avg. Level2
v1	0.790	0.792	v1	0.580	0.58
v2	0.794		v2	0.580	
v3	0.781	0.783	v3	0.711	0.69
v4	0.780		v4	0.711	
v5	0.768		v5	0.669	
v6	0.803		v6	0.669	

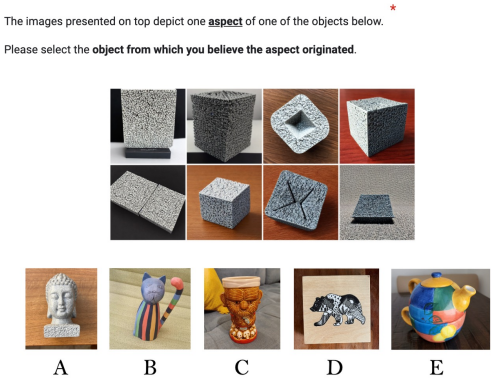


Fig. 6. Example of a question we presented in the aspect relevance survey.

The results are presented in Table 1. In both levels, the average self consistency score is high, while the average siblings consistency score increased with the transition from the first to the second level, indicating that the splits are less distinct on average. The reason for this is that as we go deeper into the tree, the components are becoming increasingly simple, making it more challenging to further split them.

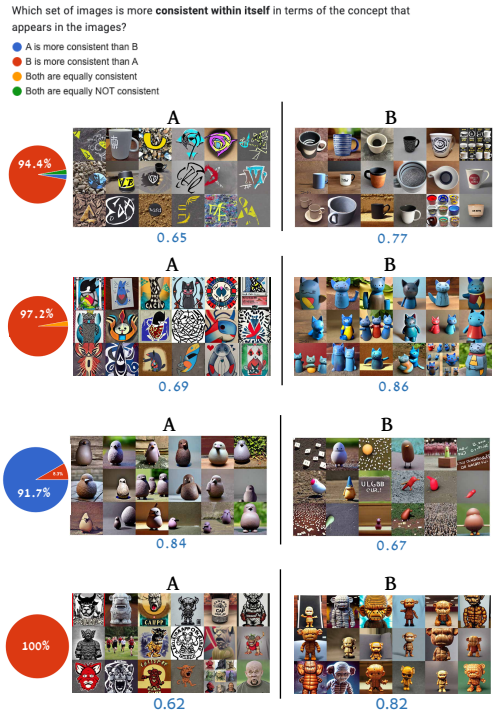


Fig. 7. Examples of questions asked in the consistency evaluation survey. On the left we show the results in percentages, indicating which answer was selected by the majority of people.

3.3 Perceptual Study

The following section provides additional details regarding our perceptual study described in section 5.2 of the main paper. For the consistency evaluation, we collected answers from 35 participants. Participants were presented with 15 pairs of random image sets, and they were asked to determine which set in each pair is more consistent. In order to handle cases where the sets are similar, we have also added two options to choose from - "Both sets are equally consistent", and "Both sets are equally not consistent". Figure 7 contains a few examples of the survey questions. On the left of each set, we also present the results in percentages, indicating which answer was selected by the majority of people. In the aspect relevance experiment, we collected answers from 35 participants and asked each participant 15 questions. Figure 6 provides an example of the questions. The question were obtained from 5 chosen objects, shown at the top of Figure 6.

4 ADDITIONAL QUALITATIVE RESULTS

In Figures 8 and 9 we show more examples of inter-tree combinations. At the top part of Figures 10 to 17 we show examples of trees on various objects.

At the bottom part of Figures 10 to 13 and in Figures 18 and 19 we show visual examples of intra tree combinations.

At the bottom part of Figures 10, 14 and 15 and in Figures 20 to 25 and 26 we show examples of text based generation.

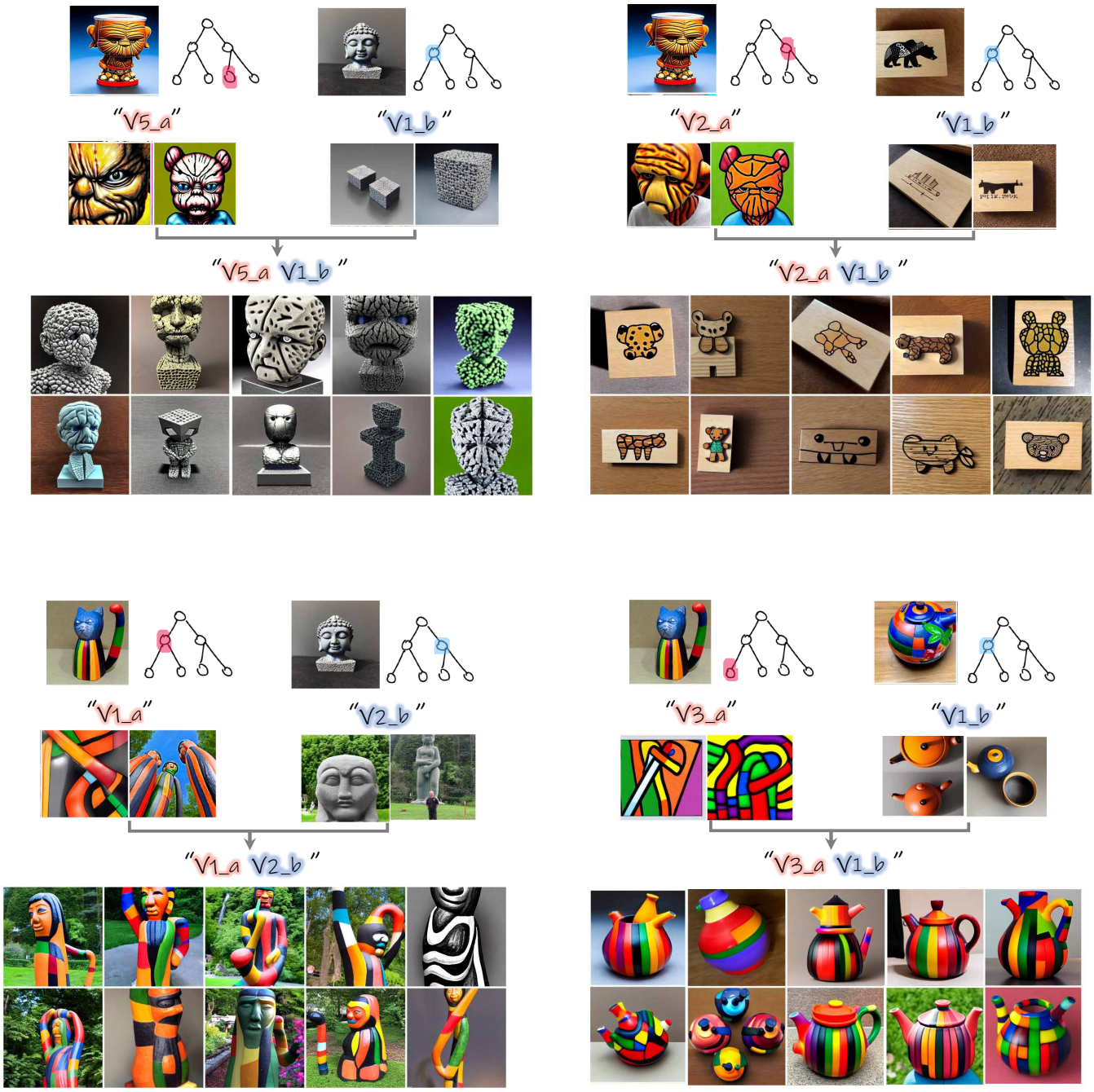


Fig. 8. More examples of inter-tree combinations.

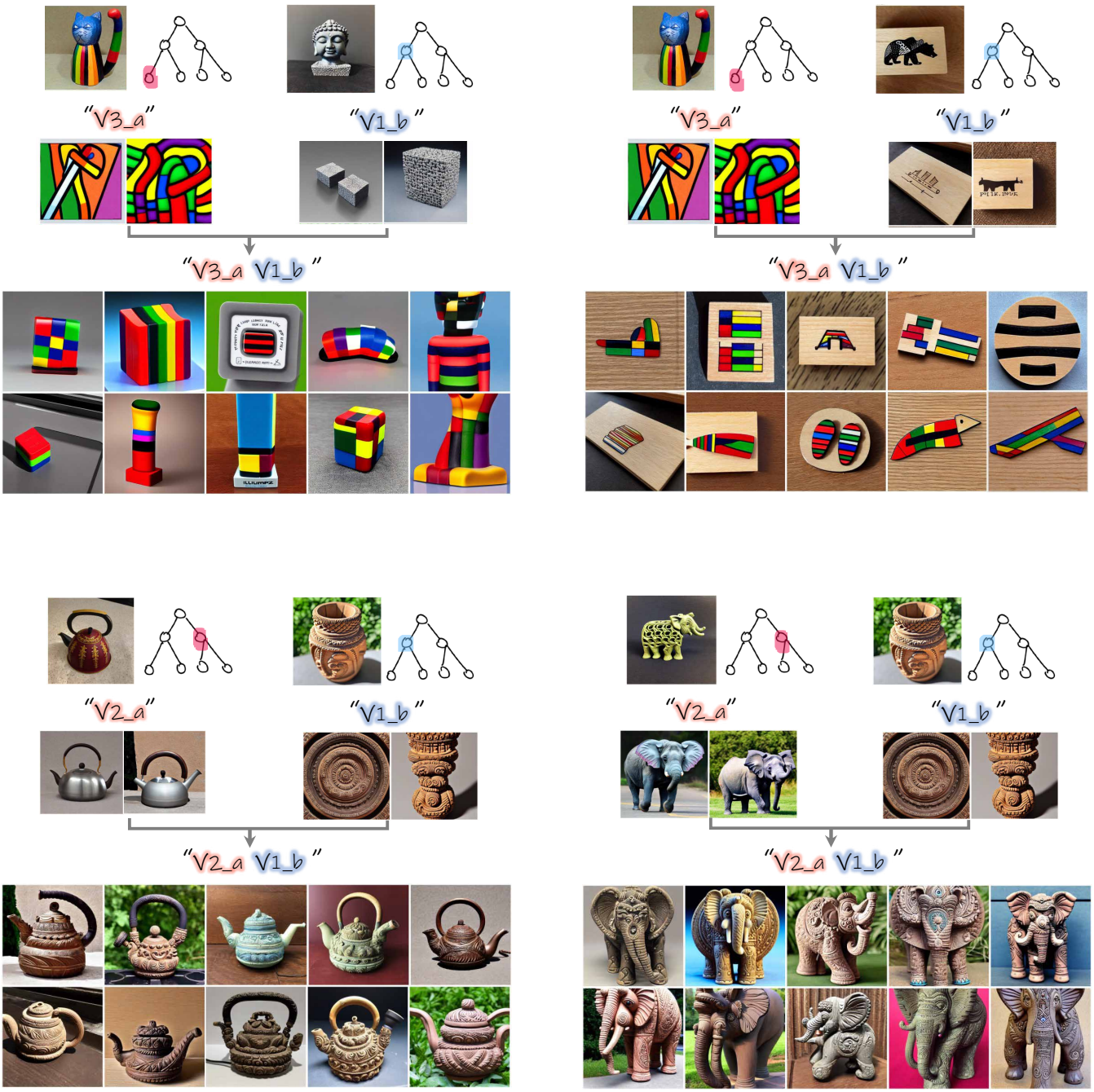
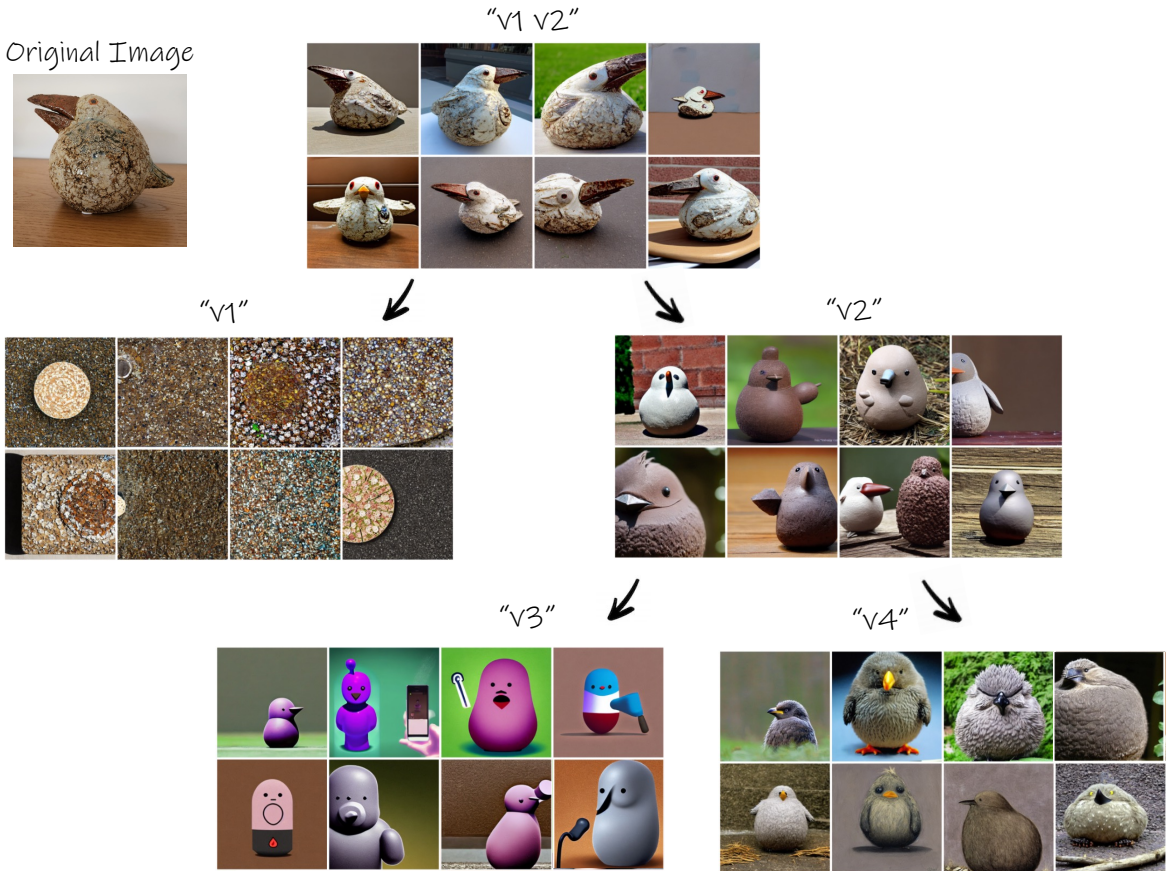


Fig. 9. More examples of inter-tree combinations.



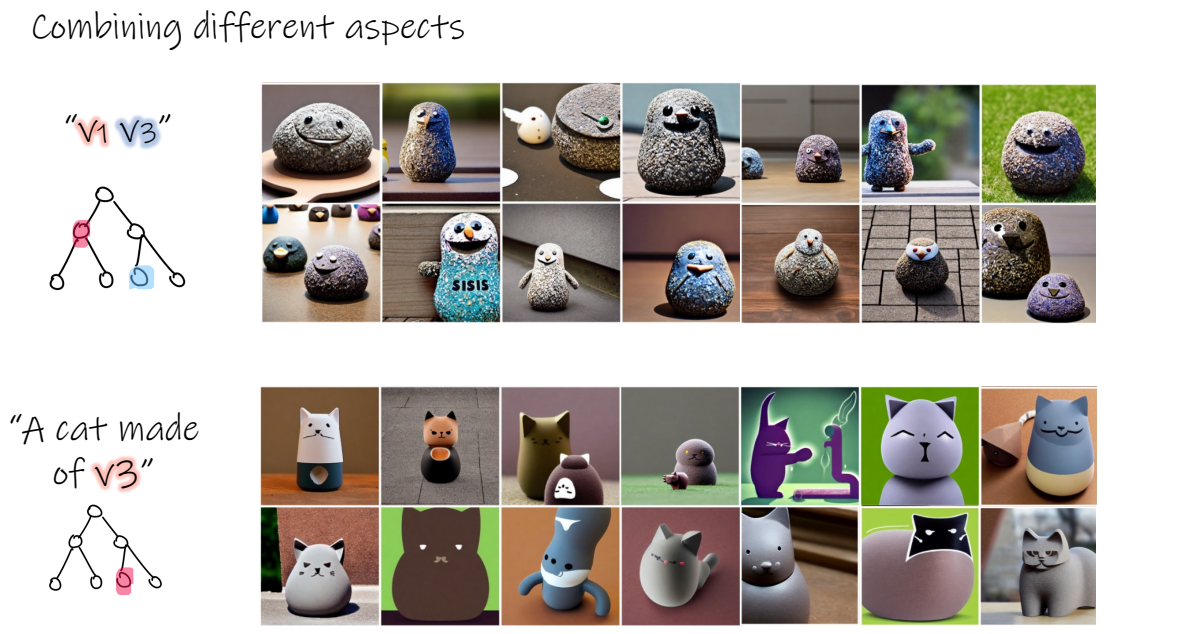
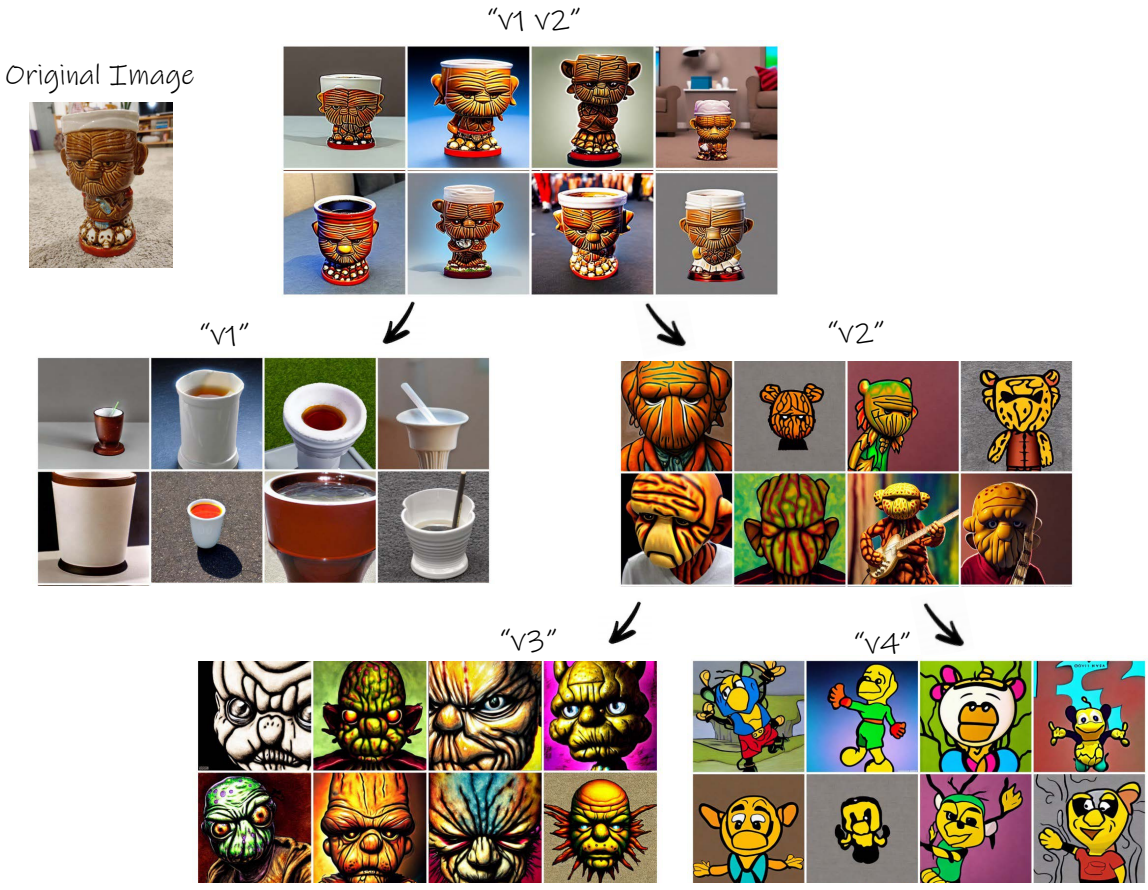


Fig. 10. Exploration tree for the "round bird" object. At the bottom we show examples of possible intra-tree combinations and text-based generation.



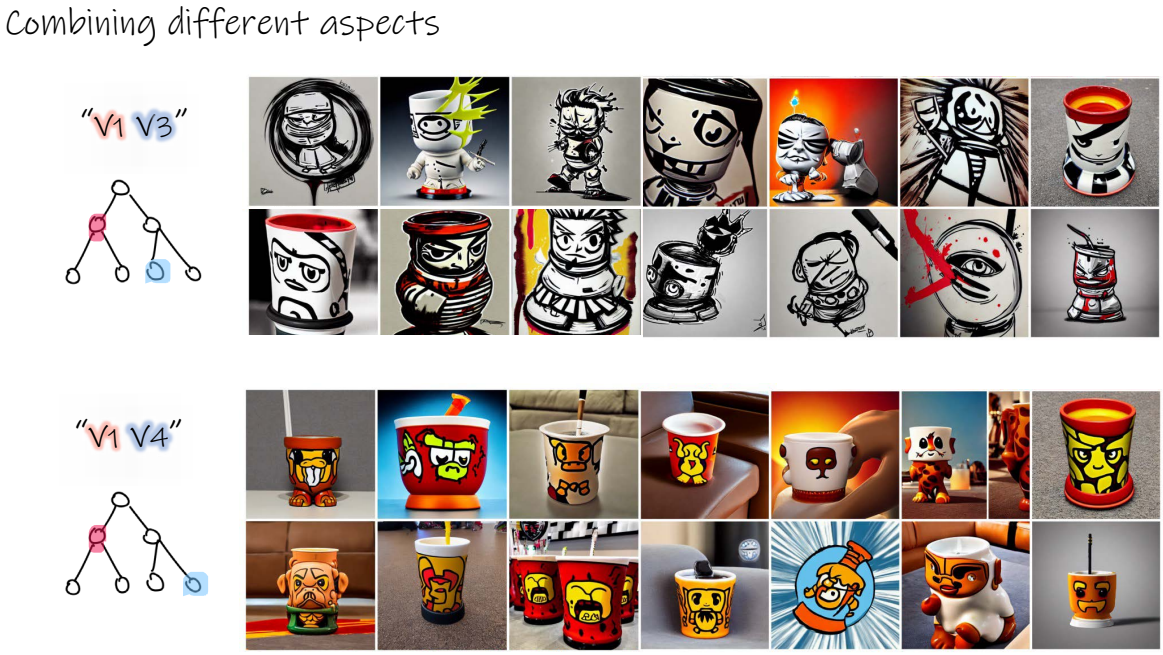
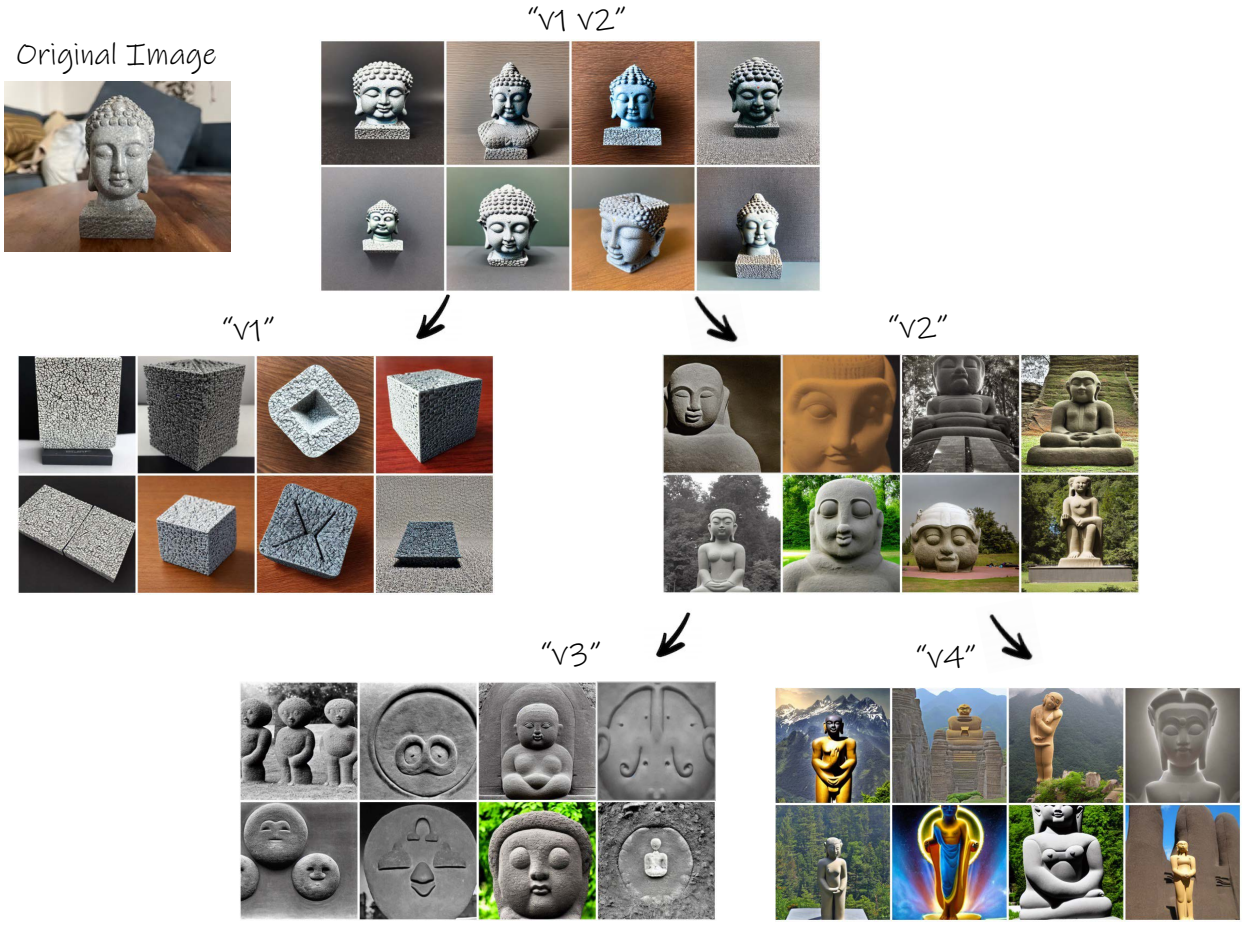


Fig. 11. Exploration tree for the "scary mug" object. At the bottom we show examples of possible intra-tree combinations.



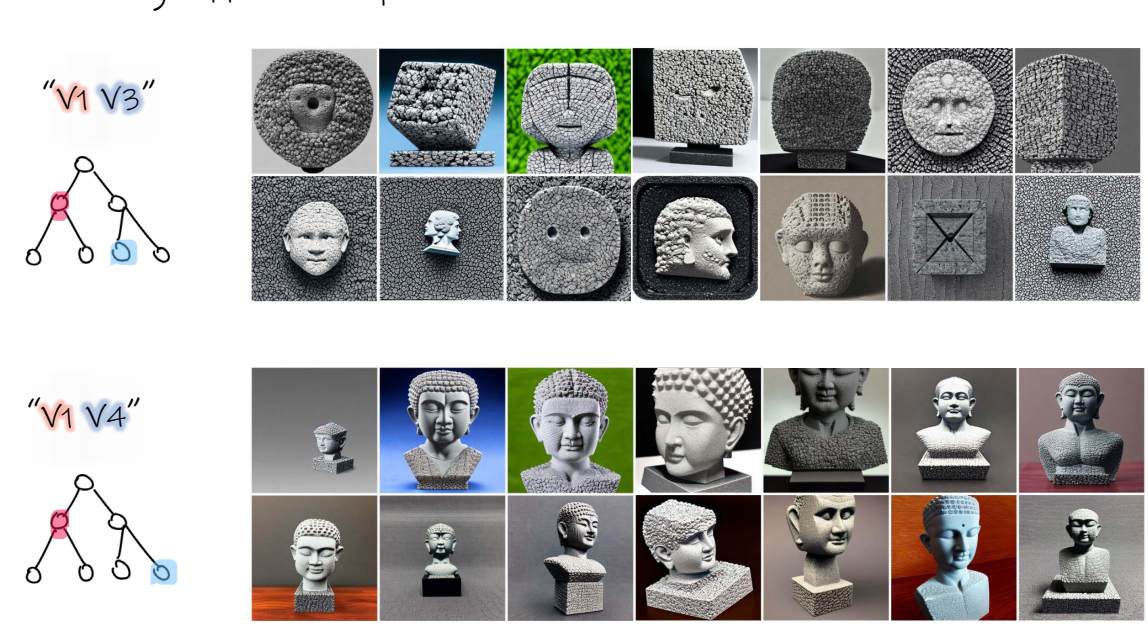


Fig. 12. Exploration tree for the "Buddha sculpture" object. At the bottom we show examples of possible intra-tree combinations.

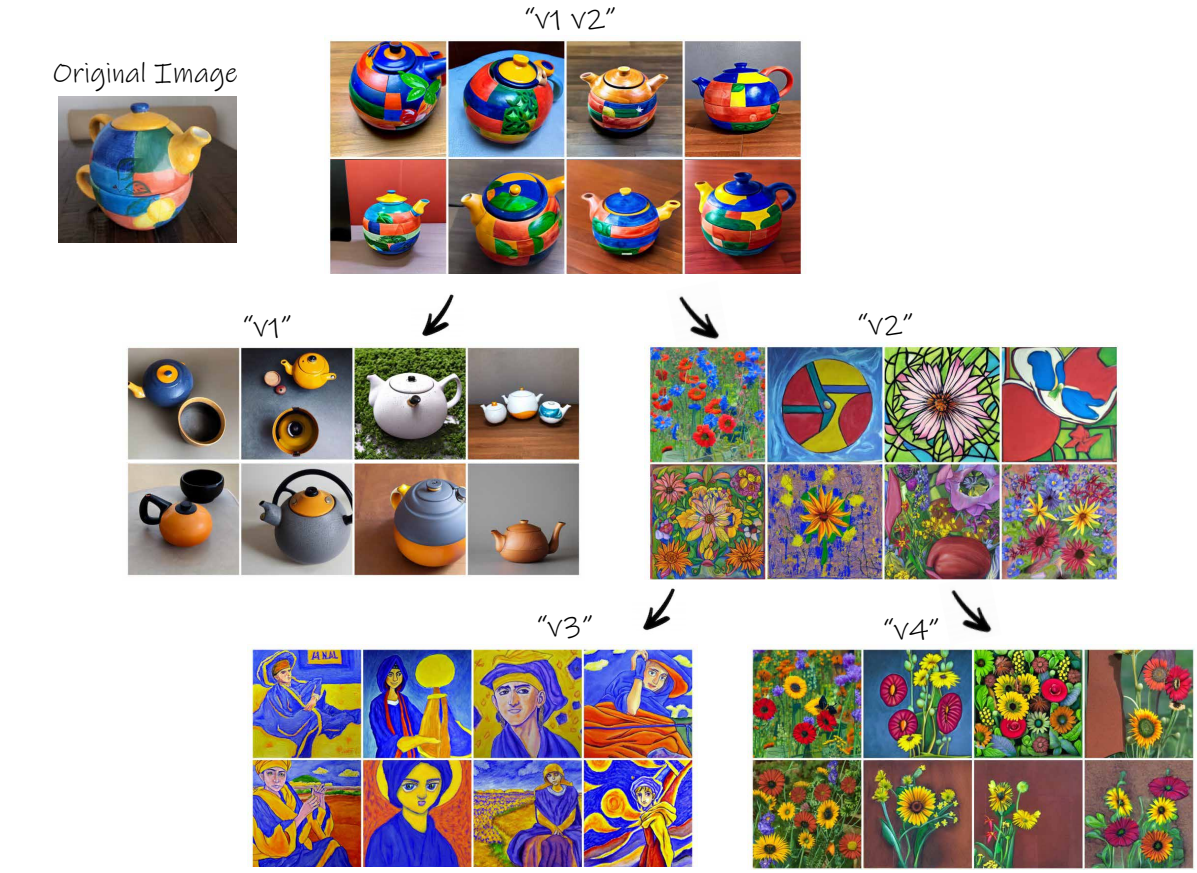


Fig. 13. Exploration tree for the "colorful teapot" object. At the bottom we show examples of possible intra-tree combinations.

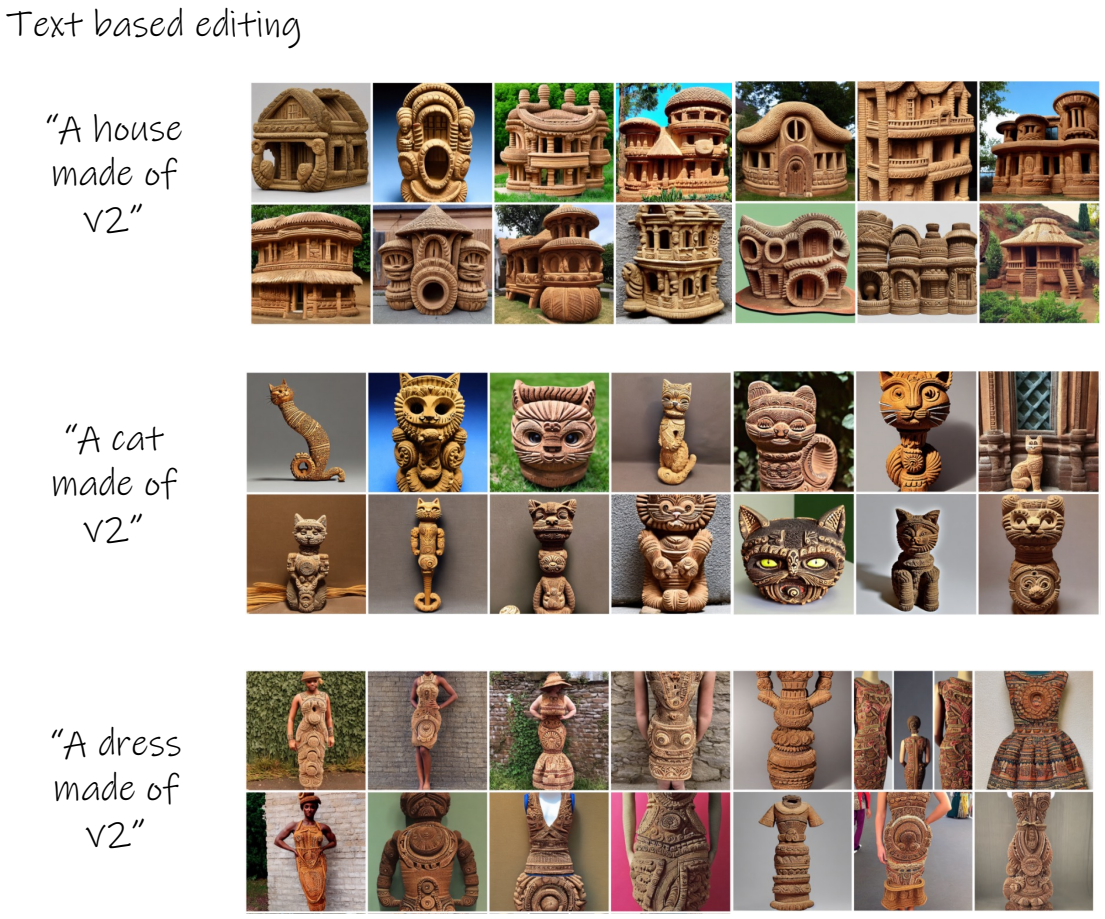
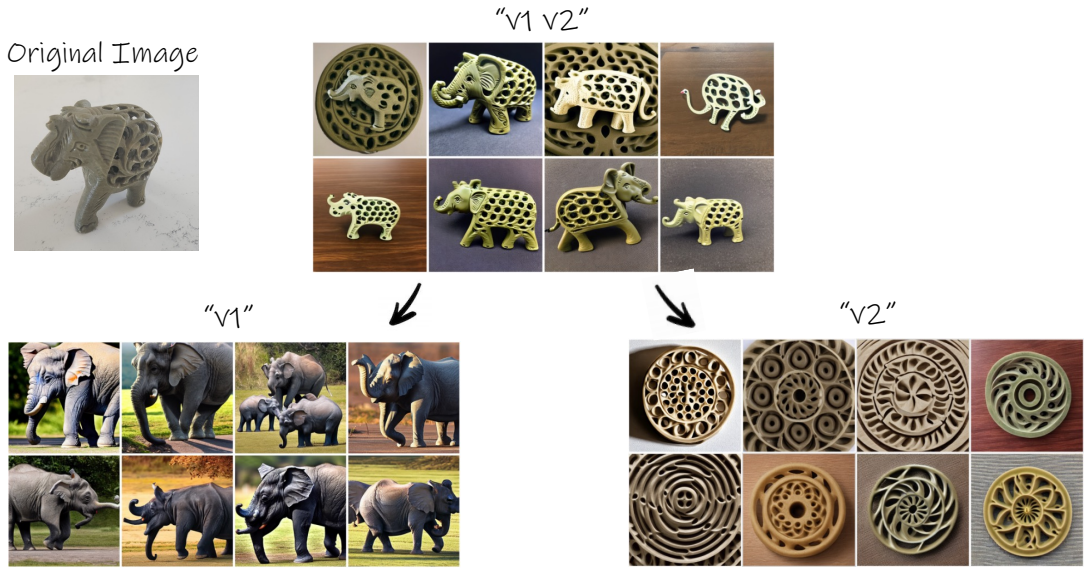


Fig. 14. Exploration tree for the "wooden pot" object. At the bottom we show examples of possible text-based generation.



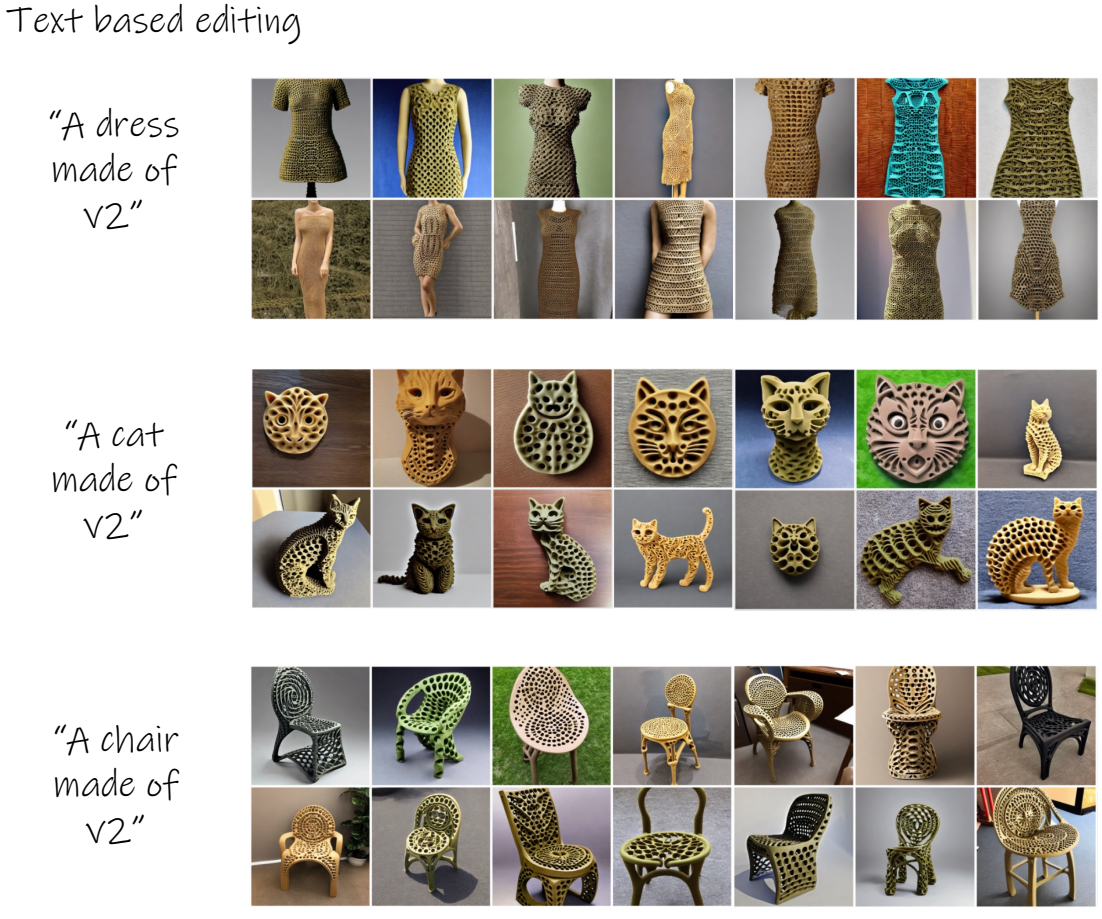
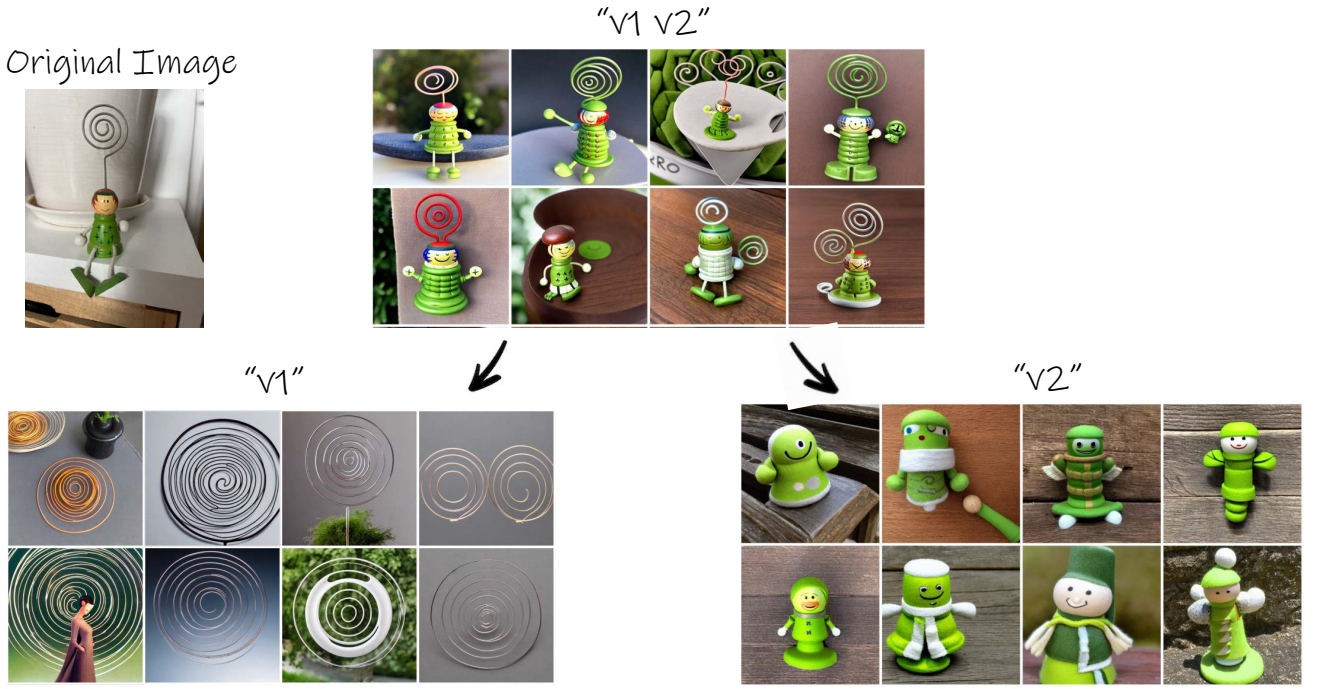


Fig. 15. Exploration tree for the "elephant" object. At the bottom we show examples of possible text-based generation.



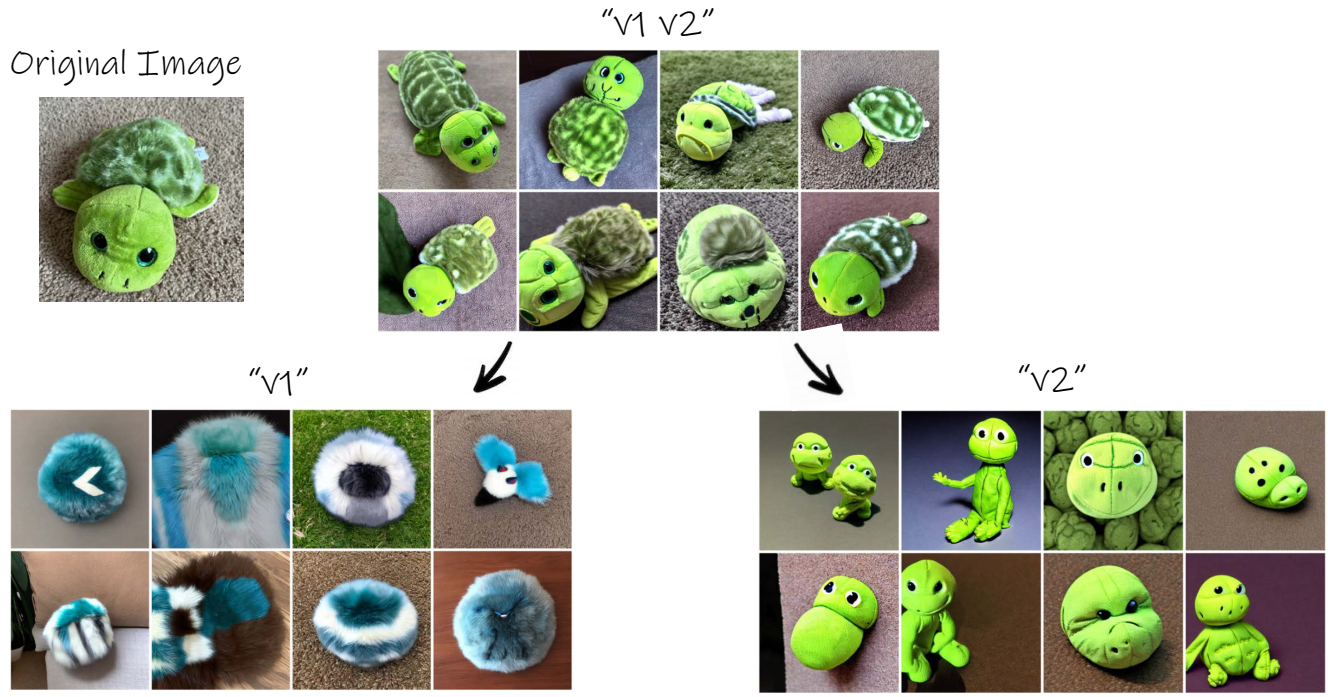


Fig. 16. Exploration trees for the "green doll" and the "turtle" objects.

Fig. 17. Exploration trees for the "Girona mug" and the "physics mug".

Fig. 18. More examples of intra-tree combinations for the "cat sculpture" object. The full original tree is shown in the main paper.

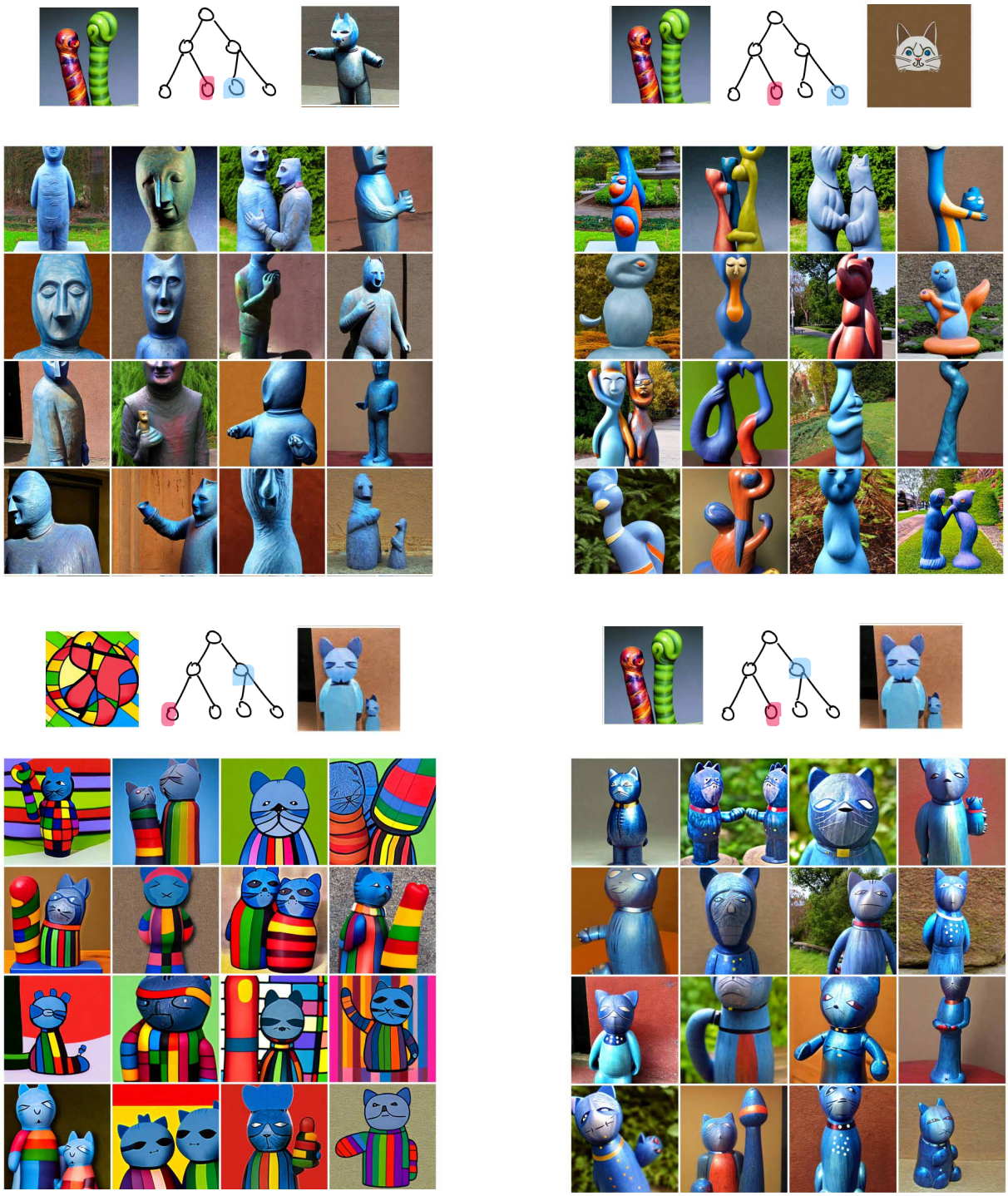


Fig. 19. More examples of intra-tree combinations for the “cat sculpture” object. The full original tree is shown in the main paper.

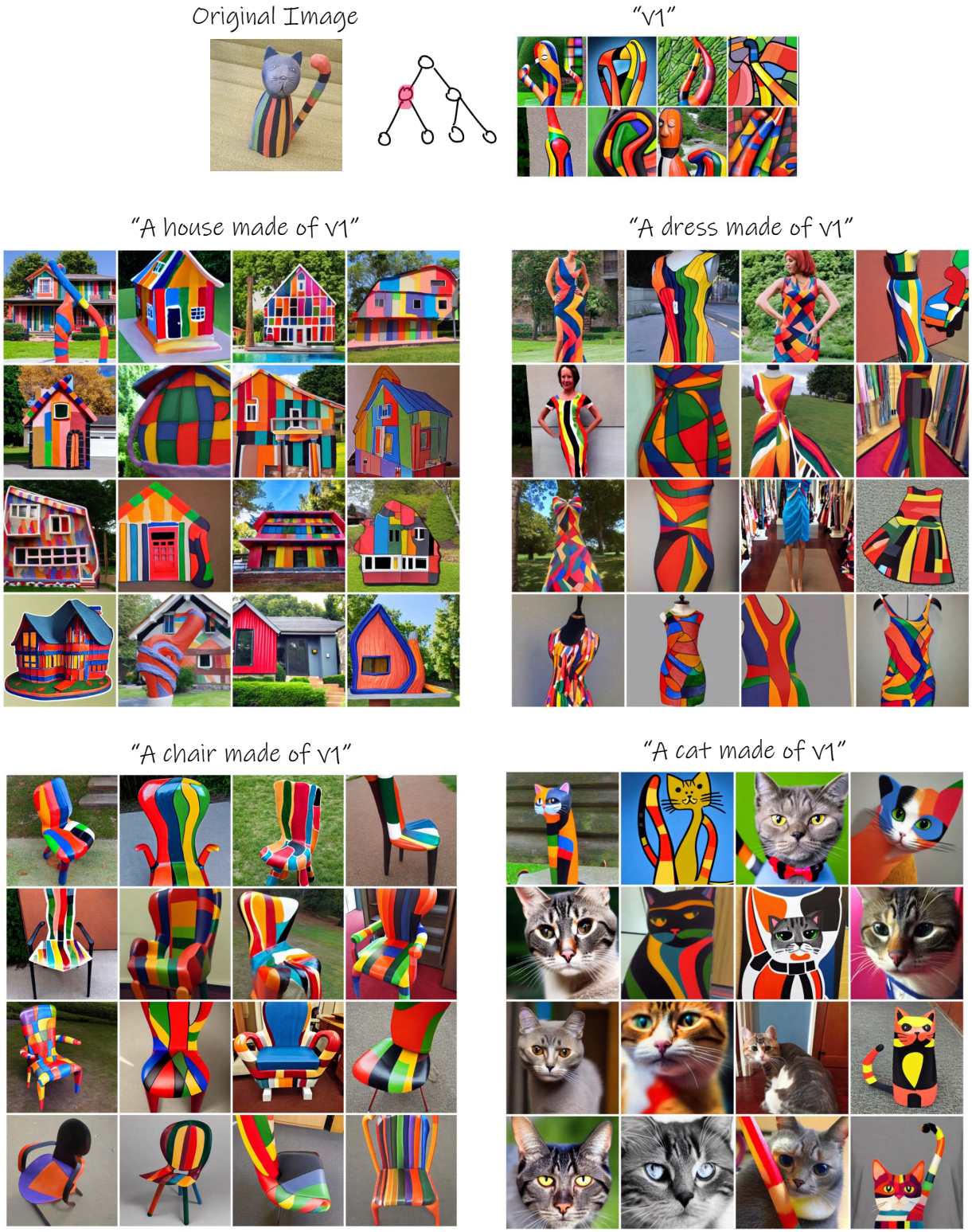


Fig. 20. More examples of text based generation for the "cat sculpture" object. The full original tree is shown in the main paper.

Fig. 21. More examples of text based generation for the "cat sculpture" object. The full original tree is shown in the main paper.

Fig. 22. More examples of text based generation for the "cat sculpture" object. The full original tree is shown in the main paper.

Fig. 23. More examples of text based generation for the "wooden saucer bear" object. The full original tree is shown in the main paper.

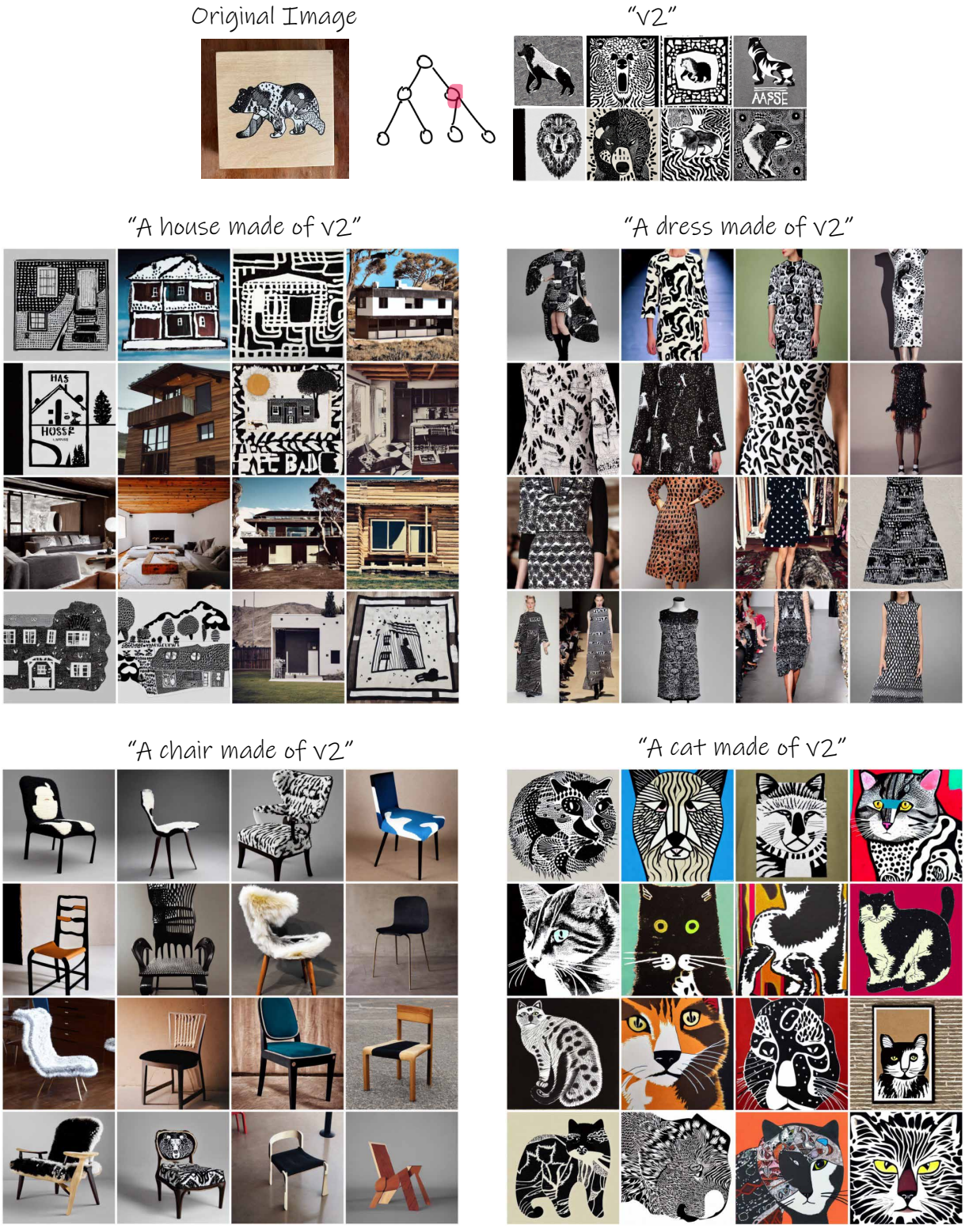


Fig. 24. More examples of text based generation for the "wooden saucer bear" object. The full original tree is shown in the main paper.

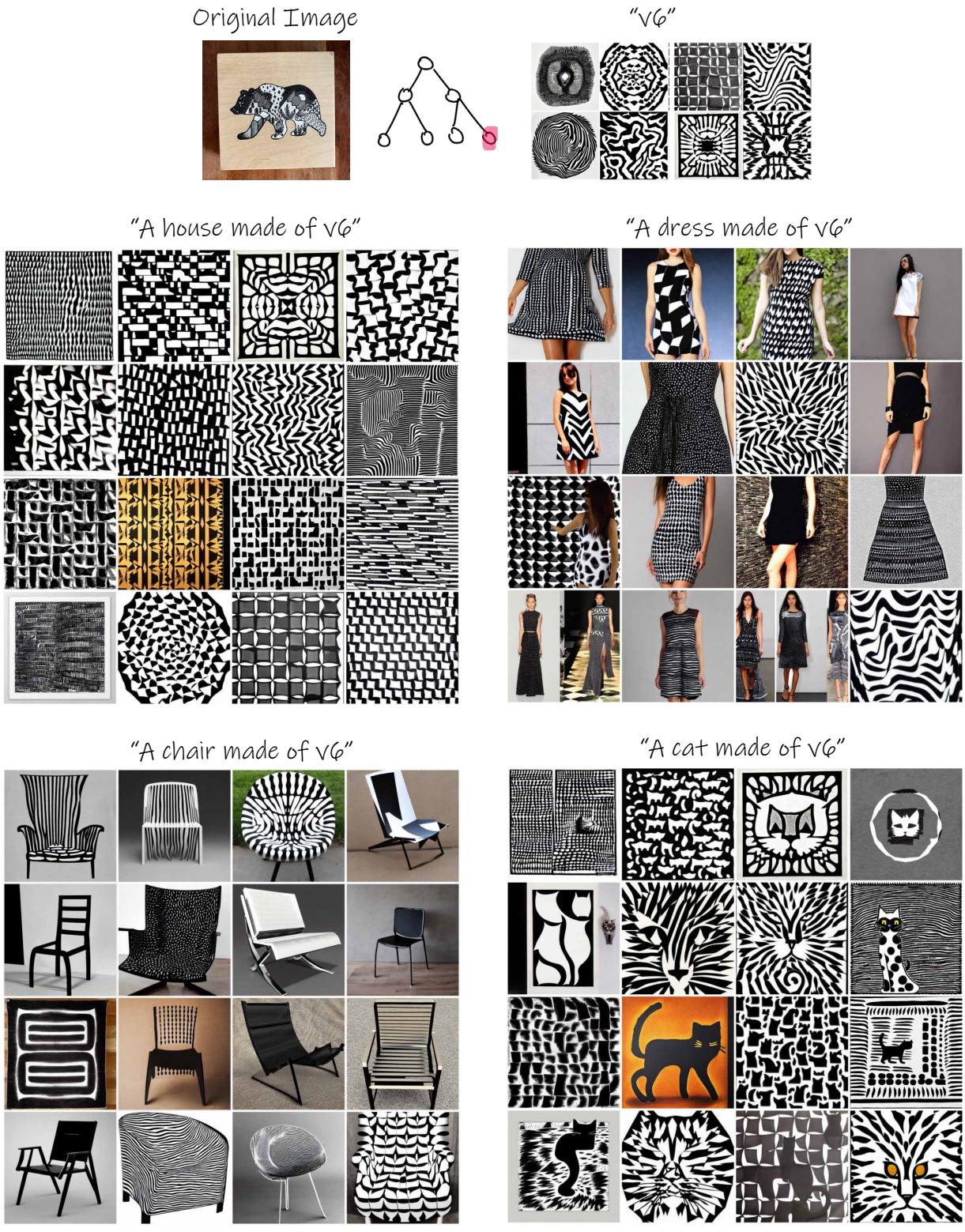


Fig. 25. More examples of text based generation for the "wooden saucer bear" object. The full original tree is shown in the main paper.

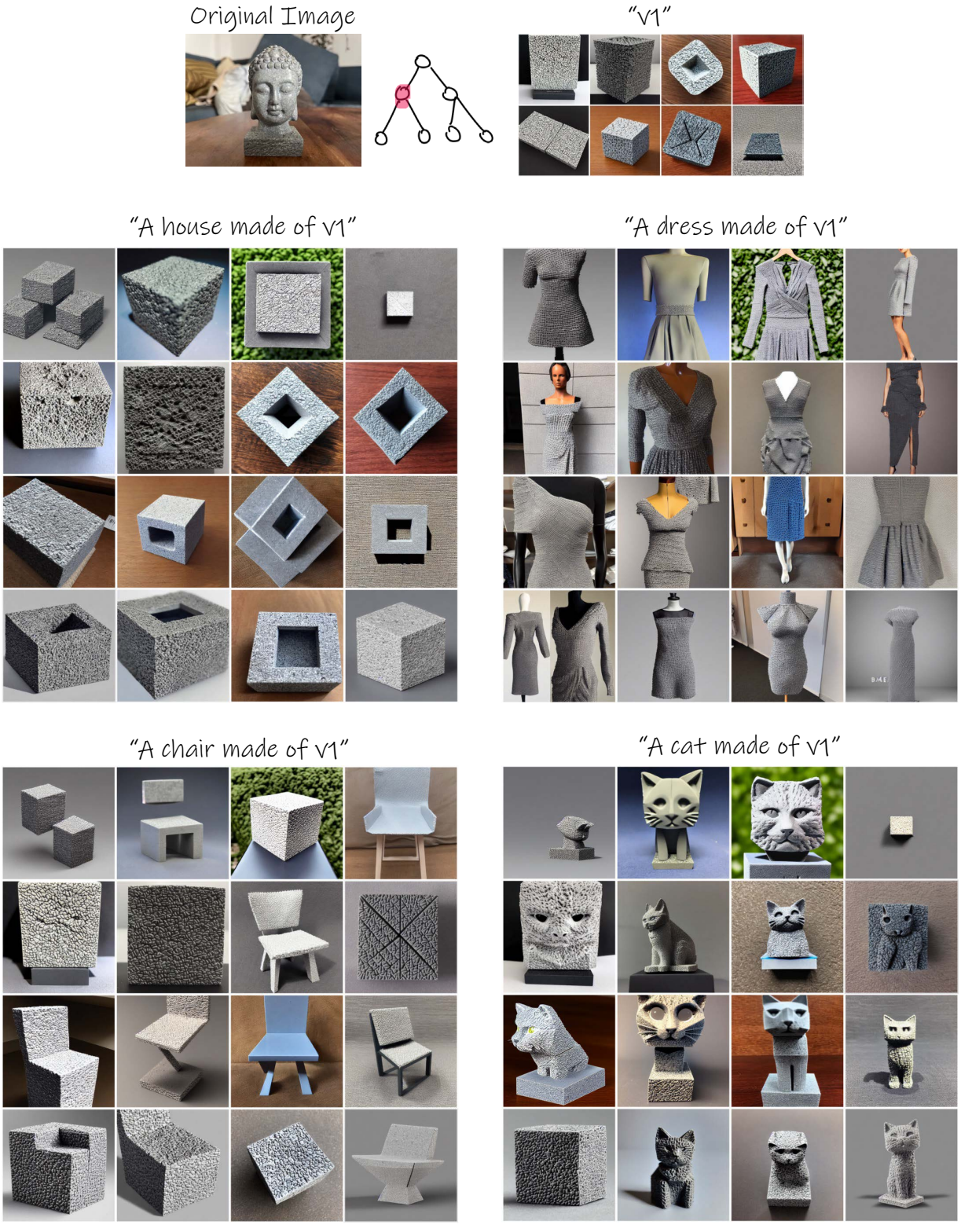


Fig. 26. More examples of text based generation for the "Buddha sculpture" object.

REFERENCES

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

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).

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

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).