Towards Synthetic Concept Activation Vectors via Generative Models

Supplementary Material

	Cast iron	**Fin**	**Glass**	**Grass**	**Leopard p.**	**Newspaper**	**Pen**
SD 1.5	0.52 ± 0.08	0.35 ± 0.10	0.62 ± 0.12	0.62 ± 0.11	0.77 ± 0.10	0.80 ± 0.03	0.32 ± 0.18
SD 1.5 + DB	0.54 ± 0.09	0.55 ± 0.10	0.78 ± 0.02	0.86 ± 0.08	0.90 ± 0.01	0.83 ± 0.02	0.93 ± 0.02
SD XL 1.0	0.53 ± 0.07	0.36 ± 0.19	0.52 ± 0.04	0.71 ± 0.12	0.77 ± 0.06	0.75 ± 0.07	0.66 ± 0.10
SD XL Turbo	0.43 ± 0.05	0.36 ± 0.16	0.71 ± 0.04	0.67 ± 0.15	0.86 ± 0.04	0.71 ± 0.04	0.67 ± 0.10

	Sphere	**Spotted**	**Taxi sign**	**Asparag.**	**Cloister**	**Feline**	**Grandfa.**
SD 1.5	0.40 ± 0.15	0.43 ± 0.08	0.53 ± 0.05	0.59 ± 0.06	0.89 ± 0.02	0.60 ± 0.08	0.75 ± 0.05
SD 1.5 + DB	0.48 ± 0.14	0.75 ± 0.73	0.58 ± 0.05	0.73 ± 0.04	0.87 ± 0.02	0.73 ± 0.05	0.84 ± 0.03
SD XL 1.0	0.47 ± 0.08	0.37 ± 0.07	0.46 ± 0.06	0.76 ± 0.02	0.85 ± 0.04	0.61 ± 0.09	0.61 ± 0.08
SD XL Turbo	0.67 ± 0.03	0.51 ± 0.04	0.67 ± 0.02	0.76 ± 0.04	0.83 ± 0.04	0.61 ± 0.09	0.59 ± 0.07

	Guitarist	**Kitchen**	**Lichen**	**Rodent**	**Steeple**	**Tattoo**	**Bubbly**
SD 1.5	0.82 ± 0.06	0.89 ± 0.02	0.86 ± 0.02	0.74 ± 0.07	0.85 ± 0.02	0.78 ± 0.03	0.72 ± 0.06
SD 1.5 + DB	0.82 ± 0.06	0.91 ± 0.01	0.84 ± 0.01	0.79 ± 0.05	0.75 ± 0.06	0.78 ± 0.03	0.73 ± 0.05
SD XL 1.0	0.74 ± 0.12	0.86 ± 0.03	0.86 ± 0.02	0.67 ± 0.11	0.76 ± 0.05	0.60 ± 0.05	0.70 ± 0.05
SD XL Turbo	0.82 ± 0.06	0.82 ± 0.03	0.62 ± 0.05	0.72 ± 0.08	0.85 ± 0.03	0.81 ± 0.03	0.76 ± 0.03

	Chequered	**Cracked**	**Crystalline**	**Dotted**	**Frilly**	**Honeycomb.**	**Meshed**
SD 1.5	0.78 ± 0.04	0.86 ± 0.02	0.45 ± 0.06	0.65 ± 0.07	0.67 ± 0.07	0.87 ± 0.02	0.78 ± 0.09
SD 1.5 + DB	0.82 ± 0.03	0.88 ± 0.03	0.74 ± 0.03	0.80 ± 0.03	0.72 ± 0.05	0.83 ± 0.02	0.81 ± 0.07
SD XL 1.0	0.67 ± 0.06	0.75 ± 0.08	0.66 ± 0.05	0.80 ± 0.04	0.70 ± 0.08	0.83 ± 0.04	0.75 ± 0.10
SD XL Turbo	0.84 ± 0.03	0.71 ± 0.09	0.53 ± 0.05	0.78 ± 0.05	0.72 ± 0.09	0.84 ± 0.04	0.84 ± 0.04

	Perforated	**Spiralled**	**Striped**	**Waffled**	**Woven**	**Wrinkled**
SD 1.5	0.83 ± 0.06	0.82 ± 0.03	0.48 ± 0.10	0.76 ± 0.06	0.84 ± 0.03	0.82 ± 0.03
SD 1.5 + DB	0.86 ± 0.04	0.81 ± 0.04	0.92 ± 0.02	0.79 ± 0.05	0.86 ± 0.03	0.85 ± 0.03
SD XL 1.0	0.82 ± 0.05	0.76 ± 0.05	0.62 ± 0.09	0.61 ± 0.09	0.91 ± 0.03	0.82 ± 0.04
SD XL Turbo	0.76 ± 0.04	0.76 ± 0.03	0.60 ± 0.10	0.74 ± 0.06	0.60 ± 0.09	0.73 ± 0.04

Table 4. Evaluation experiment results in terms of cosine similarity between generated and real CAVs. Concepts are presented separately.

(a) Real images from ImageNet

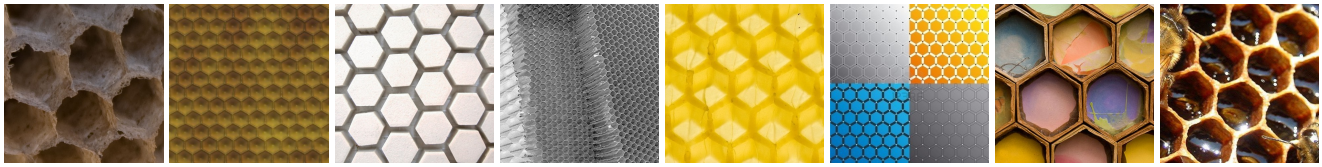
(b) SD 1.5 generated images

(c) SD 1.5 + DB generated images

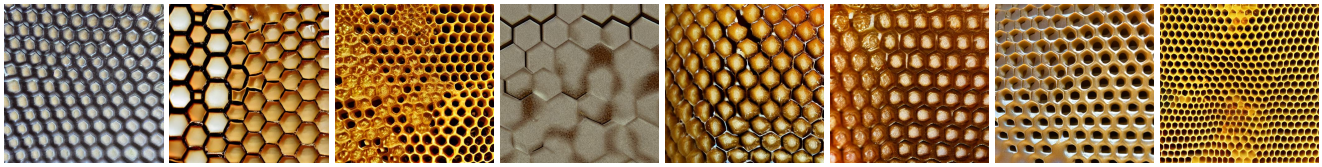
(d) SD XL 1.0 generated images

(e) SD XL Turbo generated images

Figure 6. Sample images of "cloister" concept. According to intra-similarity results, this concept is constituted by a homogeneous set of real images, and it is easy to learn a CAV from it (Table 3). Following the same trajectory, we can observe that generative models produce high-quality images. Evaluation experiment results confirm these findings with values ranging between 0.83 ± 0.04 and 0.89 ± 0.02 (Table 4).

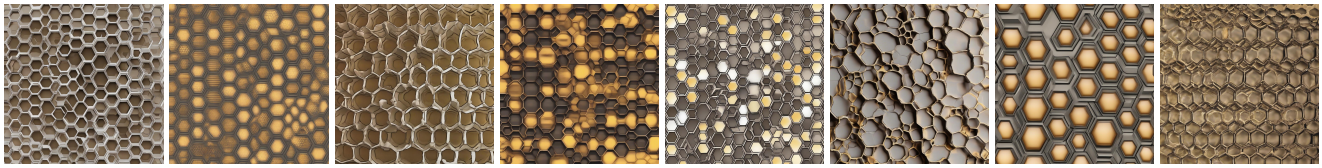


(a) Real images from ImageNet

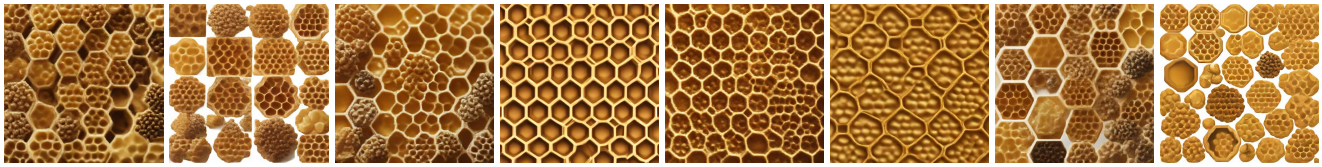


(b) SD 1.5 generated images

(c) SD 1.5 + DB generated images



(d) SD XL 1.0 generated images



(e) SD XL Turbo generated images

Figure 7. Sample images of "honeycombed" concept. This is one of the best performing textures concepts, as demonstrated by the evaluation results, ranging from 0.83 ± 0.04 to 0.87 ± 0.02 (Table 4). The intra-similarity value is also high, averaging at 0.94 (Table 3).

(a) Real images from ImageNet

(b) SD 1.5 generated images

(c) SD 1.5 + DB generated images

(d) SD XL 1.0 generated images

(e) SD XL Turbo generated images

Figure 8. Sample images of "newspaper" concept. The intra-similarity of this concept is quite low, at 0.89 (Table 3), indicating that it can be considered a complex concept. Despite this complexity, evaluation experiment results are promising with SD 1.5 (Figure 8b) and SD 1.5 + DB (Figure 8c) similarity values between 0.80 ± 0.03 and 0.83 ± 0.02 (Table 4).

(a) Real images from ImageNet

(b) SD 1.5 generated images

(c) SD 1.5 + DB generated images

(d) SD XL 1.0 generated images

(e) SD XL Turbo generated images

Figure 9. Sample images of "taxi sign" concept. While intra-similarity results are quite high (0.93, Table 3), evaluation results of generated CAVs are relatively low, ranging from 0.46 ± 0.06 to 0.67 ± 0.02 (Table 4). In this case, SD XL Turbo is the best generative model, as confirmed by the generated images (Figure 9e), which are visually more similar to the real ones (Figure 9a) than the others.