Interpretable Diffusion Models with B-cos Networks

Nicola Bernold*, Moritz Vandenhirtz*, Alice Bizeul*, Julia E. Vogt*
*Department of Computer Science, ETH Zürich

1 Motivation

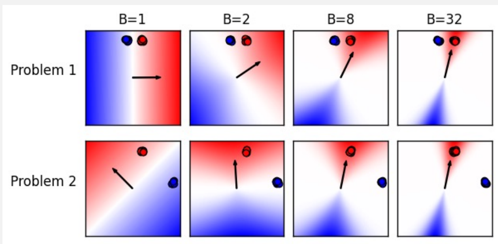
- Text-to-image diffusion models generate impressive visuals but fail to fully capture all semantic details in the prompt.
- These failures are difficult to detect automatically, hindering error detection, prompt refinement, and image-text alignment.
- Post-hoc explainability methods can be unfaithful or insufficient for interpreting complex generative models.
- In this work, we propose an inherently interpretable architecture that offers faithful explanations of its generations.

2 Background: B-cos

$$f_{\text{classic}}(x; w,b) = w^T x + b$$
$$f_{\text{B-cos}}(x; w) = w^T x\,|\cos(x,w)|^{B-1} \text{ with } ||w||=1$$
$$= w(x)^T x$$

- B-cos neuron as drop-in replacement for classical neurons.
- Only produces significant output if weights are aligned to input
- Summary of a deep B-cos network given by a dynamically linear transformation: $W(x)_{1\to L} = W_1(x) \ldots W_L(x)$

➔ $NN(x) = W(x)_{1\to L}x$



3 Method

Goal: Faithful explanation by dynamically linear model

- Remove all bias terms
- Deterministic DDIM sampling
- Interpret Cross Attention as dynamically linear

$$\text{Cross-Att}(X, Y; Q, K, V) = \text{softmax}\left(XQK^TY^T/\sqrt{d_k}\right)YV = \text{A}\,(X,Y)\,YV$$

- Encode color

$$\text{Enc}(r, g, b) = (r, g, b, 1-r, 1-g, 1-b)$$

At inference

- Visualize reconstructions via

$$R_{\text{normalized}}(x) = R_{rgb}(x)/(R_{rgb}(x) + R_{1-rgb}(x))$$

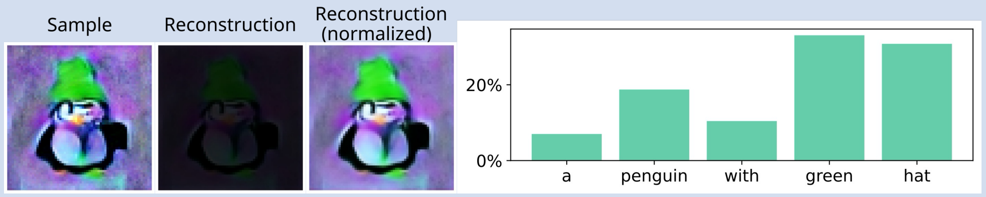
- Since $NN(x) = W(x)_{1\to L}x$, the i-th row of W(x) captures all contributions of token x_i to the output, and $W_{i,j}\,x_i$ corresponds to the contribution of x_i to the j-th output.
- As such, we define the normalized relevance score which faithfully quantifies the contribution of each token to the output

$$S_i(x) = \frac{|\sum_{h,w,c} W(x)_i x_i|}{\sum_j |\sum_{h,w,c} W(x)_j x_j|}$$

4 Generative Performance

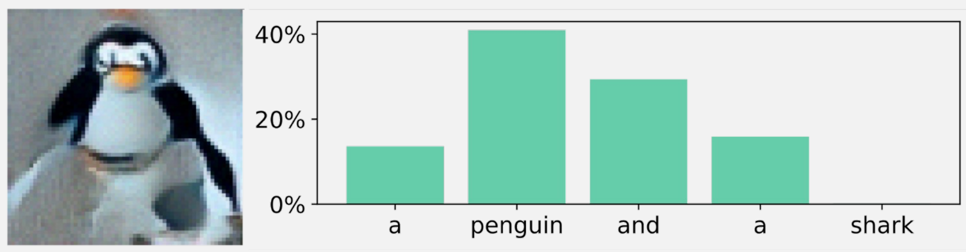
B-cos networks can produce similar results as vanilla networks. Predicting x_0 decreases the quality but omitting the CLIP encoder even slightly improves the FID score

5 Completeness of Explanation



Despite the bias terms, the reconstruction renormalized using the redundant channels is nearly perfect – the summary thus captures the complete diffusion process and can be used for interpretation.

6 Relevance Scores



The relevance score can be used to check prompt adherence.

Semantically meaningful tokens are typically more relevant.

Token	Mean-Relevance
penguin	17.1%
cat	15.6%
background	5.15%
or	1.87%
stock	1.26%

7 Conclusion

- B-cos networks can quantify the relevance and contribution of each token to the generation.
- Explanations faithfully capture alignment of image and prompt.
- This can provide actionable insights with respect to the prompt-adherence of generations
- Next steps: Improving generations and pixel-level attribution

References

- Böhle, Moritz, Mario Fritz, and Bernt Schiele. "B-cos networks: Alignment is all we need for interpretability." *CVPR* 2022
- Arya, Shreyash, Sukrut Rao, Moritz Böhle, and Bernt Schiele. "B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable." *NeurIPS* 2025