LATENT DIFFUSION U-NET REPRESENTATIONS CONTAIN POSITIONAL EMBEDDINGS AND ANOMALIES

Jonas Loos[1] **Lorenz Linhardt**[1,2]

ABSTRACT

Diffusion models have demonstrated remarkable capabilities in synthesizing realistic images, spurring interest in using their representations for various downstream tasks. To better understand the robustness of these representations, we analyze popular Stable Diffusion models using representational similarity and norms. Our findings reveal three phenomena: (1) the presence of a learned positional embedding in intermediate representations, (2) high-similarity corner artifacts, and (3) anomalous high-norm artifacts. These findings underscore the need to further investigate the properties of diffusion model representations before considering them for downstream tasks that require robust features. Project page: `https://jonasloos.github.io/sd-representation-anomalies`.

1 INTRODUCTION

Ever since diffusion models (Sohl-Dickstein et al., 2015b; Song & Ermon, 2019; Ho et al., 2020) superseded generative adversarial networks (Goodfellow et al., 2020) in image generation (Dhariwal & Nichol, 2021), diffusion methodology progressed steadily. Architectural and training improvements, such as latent diffusion (Rombach et al., 2022), transformer-based architectures (Peebles & Xie, 2023; Esser et al., 2024), and model distillation (Sauer et al., 2025; 2024) allow for more efficient training and generation of higher quality images.

Improvements in efficiency, together with remarkable image generation abilities have led to investigations into image diffusion models as embedding models (e.g. (Xiang et al., 2023)). Similar to DINO (Caron et al., 2021; Oquab et al., 2024) or CLIP (Radford et al., 2021) models, diffusion models may yield representations useful for downstream tasks (e.g. classification (Xiang et al., 2023) or semantic correspondence (Zhang et al., 2023)). Yet, attempts to use pretrained diffusion models as embedding models, as well as investigations of their general capabilities, have revealed limitations, such as texture bias in higher layers (Zhang et al., 2023), insufficient linguistic binding (Rassin et al., 2023), and left-right confusion (Zhang et al., 2024). We refer to Appx. A for additional related work.

In this work, we present three novel empirical phenomena in image diffusion model representations that do not encode spatially localized semantics and thus may deteriorate downstream task performance. We focus on the popular U-Net-based Stable Diffusion (SD) models, as they have been repeatedly investigated for their downstream utility (e.g. (Zhang et al., 2023; 2024; Tang et al., 2023; Baranchuk et al., 2022; Zhao et al., 2023; Ke et al., 2024)). The main contributions of this work are:

- (C1) We show that the representations of models of the SD family encode a *positional embedding*. This embedding is linearly extractable from the representations of lower blocks.
- (C2) We show that representations of lower blocks often contain corner *tokens of abnormally high similarity* to other corner tokens. This phenomenon is independent of the image content and can even be observed between tokens of different images.
- (C3) We show that representations of lower blocks sometimes contain *tokens of abnormally high norm* that do not appear to capture only the local image content.

[1]Machine Learning Group, Technische Universität Berlin, Berlin, 10623, Germany
[2]BIFOLD - Berlin Institute for the Foundations of Learning and Data, Berlin, 10623, Germany

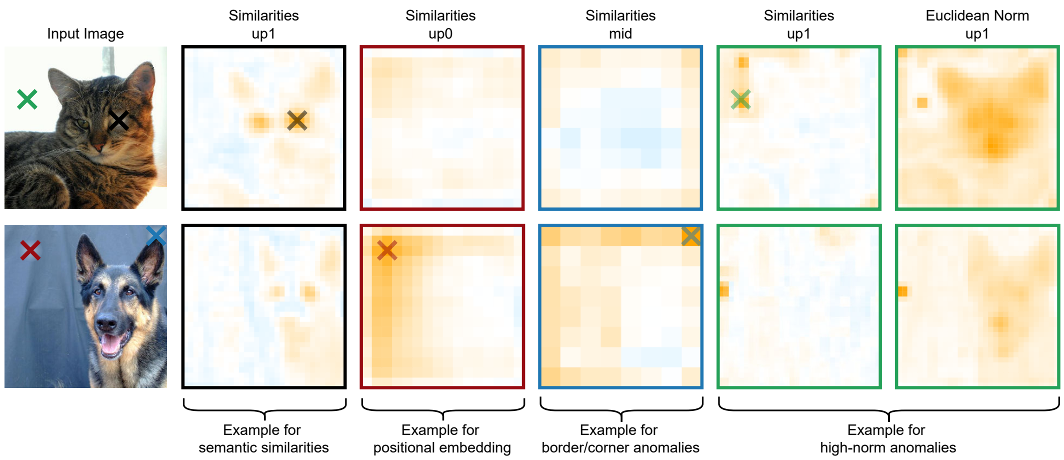

Figure 1: Cosine similarity and Euclidean norm across spatial positions of representations. Each column shows an example of one of the three observations, or of meaningful similarities. Similarities in each column are relative to the token highlighted by a marker of the matching color (×, ×, ×, ×) in one of the images. Representations are extracted from SD-1.5 at the blocks indicated at the top.

2 Methods

2.1 Representation Extraction

We extract intermediate representations from the U-Net layers of the evaluated models. The architecture consists of a series of four down-sampling (`dn0`-`dn3`) and four up-sampling (`up0`-`up3`) blocks, connected by skip connections, as well as a resolution-preserving `mid`-block at the lowest level. Each block consists of a combination of ResNet and attention layers, and a down- or up-sampling operation where applicable. We extract representations after each layer by noising a given image x in the latent space of the variational autoencoder $\mathcal{E}$ according to a given time step $t \in [1, 1000]$ and then recording the activations after each U-Net layer. More formally:

$$z_0 = \mathcal{E}(x), \quad z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad r_l = \text{U-Net}(z_t, t, \mathcal{C})_l, \tag{1}$$

where $\bar{\alpha}$ is defined by the noise scheduler and interpolates between the latent code z_0 of the image and the noise ϵ. We set conditioning $\mathcal{C}$ to an empty prompt and $t = 50$. The representation r_l at a layer l is of size $\mathbb{R}^{w_l \times h_l \times c_l}$, with w_l, h_l, and c_l being the width, height, and number of channels, respectively. The spatial dimensionality decreases in lower layers of the U-Net, while c_l increases. We refer to the representations at any spatial position as a token.

2.2 Position Estimation

To quantify the observation of positional embeddings in the representations of SD models, we train linear probe to estimate the token position. An estimator for layer l takes as input a token of dimensionality c_l and predicts the vertical and horizontal coordinate of the token, using labels formed by concatenating two one-hot vectors of dimensionality $w_l + h_l$. We minimize the cross-entropy loss using the Adam optimizer (Kingma & Ba, 2017) with learning rate 10^{-3} for 5 epochs.

3 Experiments and Analysis

In this section, we present our findings on positional embeddings, the influence of corner and border locations, and high norm anomalies. For each phenomenon, we provide a qualitative example, a quantification of the observation, and a brief discussion of potential implications.

For all experiments, we extract representations from the U-Net-based latent diffusion models SD-1.5, SD-2.1 (Rombach et al., 2022), and SD-Turbo (Sauer et al., 2025). For brevity, we show the

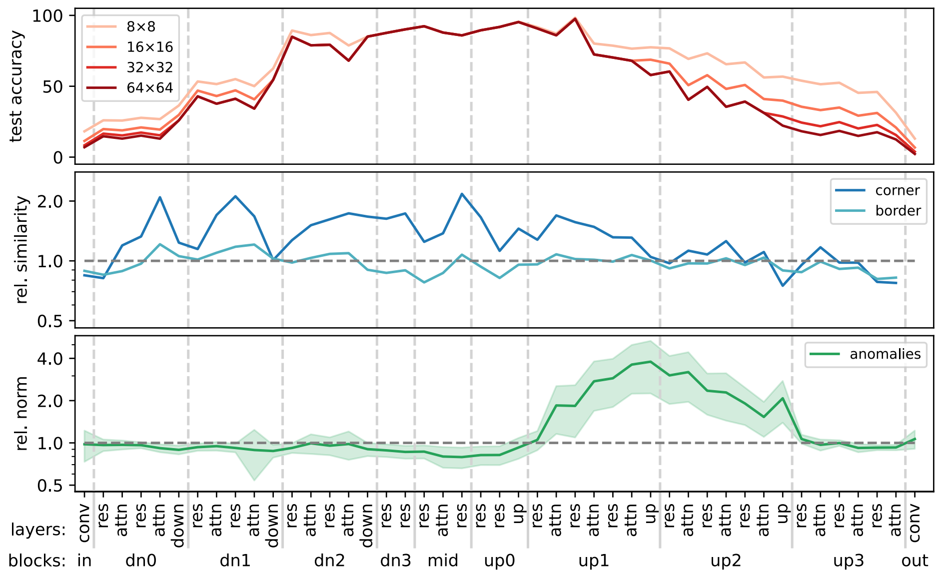

Figure 2: Quantitative results for position estimation, border/corner artifacts, and high-norm anomalies for SD-1.5. **Top row:** Linear probe accuracy for position estimation. Brighter shades indicate reduced resolution. **Middle row:** Relative similarity of tokens lying at a border/corner of the cropped images w.r.t. their similarity before cropping. (log-2 scale). **Bottom row:** Relative average norm of anomalous tokens w.r.t. to the mean norm of all tokens of the same representation (log-2 scale).

results for the latter two in Appx. B. We use a subset of ImageNet (Russakovsky et al., 2015), containing 100 randomly chosen images for each of 5 classes (*German shepherd* (235), *boxer* (242), *tiger cat* (282), *pickup* (717), *volcano* (980)), which were chosen to contain both concepts of high as well as low similarity. All images are center-cropped and resized to 512×512 pixels to match the default image size of SD-1.5.

3.1 SD U-Net Representations Contain Positional Embeddings

Qualitative observation. We observe that representations in SD models contain a positional embedding, which is visible in the third column of Fig. 1, where tokens of similar spatial locations show higher similarities, even across images. This implies that SD models saliently encode location in their representations. We find that this phenomenon is most apparent after the `up0` block and less visible in higher blocks. The following quantitative results support this observation.

Quantitative results. To quantify the positional embedding uncovered by inspecting token similarities, we train a linear probe to predict the spatial location of each token given a representation token as input. We use a random 80% split of the images for training and evaluate on the remaining 20% by calculating the fraction of correct width and height predictions. As shown in the first row of Fig. 2, the estimator achieves a test accuracy of over 90% for lower blocks (`down2` to `up1`), indicating that the positional information is more saliently encoded there. Part of the difference in performance across layers is due to the lower spatial resolution of the lower blocks. Yet, even when evaluating the performance of the higher blocks at a lower resolution by coarse-graining the prediction target, lower blocks still yield significantly higher accuracy.

Implications. SD models saliently encode spatial locations in their representations to generate images, which has immediate consequences for their use as representation learners. For example, in semantic or dense correspondence tasks, similarity between representation tokens is used to determine semantically matching image locations across two images. Saliently encoded position information may interfere with semantic matching, undermining task performance.

3.2 SD U-Net Representations Contain Corner Artifacts

Qualitative observation. We find that tokens located at the corners and borders often have unusually high cosine similarities to each other, even if there is no obvious correspondence of the image content. This anomalous behavior of the border and corner tokens is visualized in fourth column of Fig. 1, where all corners show a slightly increased similarity towards the reference token at the upper left corner of the second image, independent of their image content.

Quantitative results. To quantify our observation, we compare similarities between tokens, when they are at the border/corners and when they are not. To preclude confounding by image content, we create two versions of the representations for all images: (1) we center-crop the image before embedding, such that we arrive at a representation of shape $(w-2, h-2, c)$; and (2) we embed the original image and then discard the outermost tokens, arriving at the same shape. This allows us to compare the tokens representing the same image regions but where the outermost tokens (1) do and (2) do not lie on the border of the image during extraction. The second row of Fig. 2 shows the average cosine similarity between all border tokens and all corner tokens for both representations. In particular, the similarities for tokens at the image border/corners during extraction are shown relative to the baseline, when they are not at the border during extraction. Relative similarity among corner tokens is increased across multiple layers, while the results for border tokens are inconclusive.

Implications. The existence of corner artifacts may negatively affect dense prediction tasks that are based on similarity, such as dense correspondence. Similar to position embeddings, similarity caused by corner artifacts may obfuscate semantic (dis)similarity of image content at these locations.

3.3 SD U-Net Representations Contain High-Norm Artifacts

Qualitative observation. We identify anomalies, which consist of groups of neighboring tokens with high norm, and high mutual similarity. Several such anomalies can be seen in the last two columns of Fig. 1. They primarily consist of 2×2 token patches that have increased Euclidean norm and high mutual cosine similarity.

Quantitative results. To analyze high-norm anomalies, we manually label their occurrences in the L_2 norm maps of the `up1`-block of all images in the datasets. We find that for SD-1.5, about 25% of the images contain at least one such anomaly. In the bottom row of Fig. 2, it can be seen that the tokens at the location of the labeled anomalies have significantly higher norm in the layers of the `up1` and `up2` blocks than the average of the tokens in the respective representations. We find the locations of the anomalies to be consistent across different layers for the same image, but not across different images, time steps, or models.

Implications. Similar to the border artifacts in Sec. 3.2, high-norm anomalies may negatively affect dense prediction tasks. This includes tasks that are not similarity-based, such as depth estimation. Moreover, the observed anomalies affect the `up1` layer, commonly used for downstream tasks, and are not exclusively located at the image borders, thus potentially interfering with the representations of centered objects.

4 Conclusion

In this work, we presented idiosyncrasies of U-Net-based latent diffusion model representations that may provide challenges when using these representations for downstream tasks. We reported that these representations contain (1) a linearly extractable position embedding, (2) corner tokens of abnormally high similarity, and (3) high-norm anomalies in the up-sampling blocks. All findings are supported by both qualitative examples and quantitative analysis.

Future work may evaluate the concrete impact of these phenomena on large-scale and real-world applications. Furthermore, the causes of these phenomena, as well as their role in the generative process, should be established. For example, corner artifacts may be part of the position embedding mechanism, and high-norm tokens may function as register-like storage of global information (Darcet et al., 2024).

References

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SlxSY2UZQT.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024. URL https://arxiv.org/abs/2401.14404.

Yida Chen, Fernanda Vi'egas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023. URL https://arxiv.org/abs/2306.05720.

Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=N0xNf9Qqmc.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations ICLR 2024*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Banani_Probing_the_3D_Awareness_of_Visual_Foundation_Models_CVPR_2024_paper.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of diffusion features for semantic correspondence. *arXiv preprint arXiv:2412.03512*, 2024. URL https://arxiv.org/abs/2412.03512.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36:8266–8279, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1a074a28c3a6f2056562d00649ae6416-Paper-Conference.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23115–23127, 2024.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Yuxiang Ji, Boyong He, Chenyuan Qu, Zhuoyue Tan, Chuan Qin, and Liaoni Wu. Diffusion features to bridge domain gap for semantic segmentation. *arXiv preprint arXiv:2406.00777*, 2024. URL https://arxiv.org/abs/2406.00777.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024. doi: 10.1109/CVPR52733.2024.00907.

Jiwon Kim, Byeongho Heo, Sangdoo Yun, Seungryong Kim, and Dongyoon Han. Match me if you can: Semi-supervised semantic correspondence learning with unpaired images. In *Computer Vision – ACCV 2024*, pp. 462–479. Springer Nature Singapore, 2025. URL https://link.springer.com/chapter/10.1007/978-981-96-0960-4_28.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017. URL https://arxiv.org/abs/1412.6980.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pd1P2eUBVfq.

Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27558–27568, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SD4Match_Learning_to_Prompt_Stable_Diffusion_Model_for_Semantic_Matching_CVPR_2024_paper.html.

Lorenz Linhardt, Marco Morik, Sidney Bender, and Naima Elosegui Borras. An analysis of human alignment of latent diffusion models. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=PFnoxcKh33.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Vm1zeYqwdc.

Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19521–19530, 2024. doi: 10.1109/CVPR52733.2024.01846.

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification. *arXiv preprint arXiv:2307.08702*, 2023. URL https://arxiv.org/abs/2307.08702.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023. URL https://arxiv.org/abs/2302.12469.

Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28285–28295, 2024. doi: 10.1109/CVPR52733.2024.02672.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3536–3559. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0b08d733a5d45a547344c4e9d88bb8bc-Paper-Conference.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24. Association for Computing Machinery, 2024. ISBN 9798400711312. doi: 10.1145/3680528.3687625.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 87–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015a. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015b.

Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. *arXiv preprint arXiv:2412.03439*, 2024. URL https://arxiv.org/abs/2412.03439.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ypOiXjdfnU.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3554–3563. IEEE Computer Society, 2024. doi: 10.1109/CVPR52733.2024.00341.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15802–15812, 2023. doi: 10.1109/ICCV51070.2023.01448.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8e9bdc23f169a05ea9b72ccef4574551-Paper-Conference.pdf.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3076–3085, June 2024.

Manyuan Zhang, Guanglu Song, Xiaoyu Shi, Yu Liu, and Hongsheng Li. Three things we need to know about transferring stable diffusion to visual dense prediction tasks. In *Computer Vision – ECCV 2024*, pp. 128–145. Springer Nature Switzerland, 2025. ISBN 978-3-031-72946-1. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05837.pdf.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5706–5716. IEEE Computer Society, 2023. doi: 10.1109/ICCV51070.2023.00527.

A Related Work

Initially inspired by the physical process of diffusion, diffusion models iteratively transform a distribution of noise into a desired target distribution through a sequence of learned reverse steps Sohl-Dickstein et al. (2015a); Ho et al. (2020). Building on this, SD is a series of latent diffusion models for image generation Rombach et al. (2022), most of which employ a U-Net Ronneberger et al. (2015) in the latent space of a pretrained variational autoencoder. More recently, transformer-based models have been added to the series (Esser et al., 2024).

Diffusion Models for Representation Learning. Diffusion models, and SD in particular, have been analyzed and used for representation learning as a basis for a variety of downstream tasks. While some works modify the model architecture or training process specifically for representation learning (Hudson et al., 2024; Chen et al., 2024), many works use the intermediate representations of pretrained models. Common SD versions used in the literature are SD-1.5 and SD-2.1 (Luo et al., 2023; Zhao et al., 2023; Zhang et al., 2023; 2024; Stracke et al., 2024; Linhardt et al., 2024). Various works have investigated different aspects of the learned representations, finding that semantic information is captured in the bottleneck layers of the U-Net (Kwon et al., 2023; Park et al., 2023). Other works studied image diffusion models' alignment to human representations and human-like shape bias (Linhardt et al., 2024; Jaini et al., 2024).

SD Representations for Downstream Tasks. In recent years, there has been substantial interest in exploring the suitability of SD representations for downstream tasks, such as classification (Xiang et al., 2023; Mukhopadhyay et al., 2023; Stracke et al., 2024), semantic correspondence (Zhang et al., 2023; 2024; El Banani et al., 2024; Tang et al., 2023; Luo et al., 2023; Hedlin et al., 2023; Li et al., 2024; Stracke et al., 2024; Fundel et al., 2024; Mariotti et al., 2024; Kim et al., 2025), semantic segmentation (Baranchuk et al., 2022; Zhao et al., 2023; Ji et al., 2024; Couairon et al., 2024; Tian et al., 2024; Zhang et al., 2025), and depth estimation (Chen et al., 2023; Zhao et al., 2023; Patni et al., 2024; Stracke et al., 2024; Zhang et al., 2025).It has been observed that downstream task performance tends to increase with the number of pre-training iterations (Zhao et al., 2023; Zhang et al., 2025). Multiple works reported that `up`-blocks of the U-Net contain the most useful representations for downstream tasks (Zhang et al., 2023; El Banani et al., 2024; Stracke et al., 2024). Tang et al. (2023) suggest that `up`-blocks lower in the U-Net yield more semantically-aware representations, while `up`-blocks higher in the U-Net focus more on more low-level details.

B Results for SD-2.1 and SD-Turbo

Complementary to the results on SD-1.5 presented in the main text, we here provide results for SD-2.1 and SD-Turbo, which are based on the same model architecture (Rombach et al., 2022). Fig. 3 shows additional examples for the three phenomena described in the main text. Fig. 4 shows the results for the quantitative experiments on SD-2.1, and Fig. 5 on SD-Turbo. The results are overall consistent across all evaluated models, suggesting that our findings are not limited to a specific model.

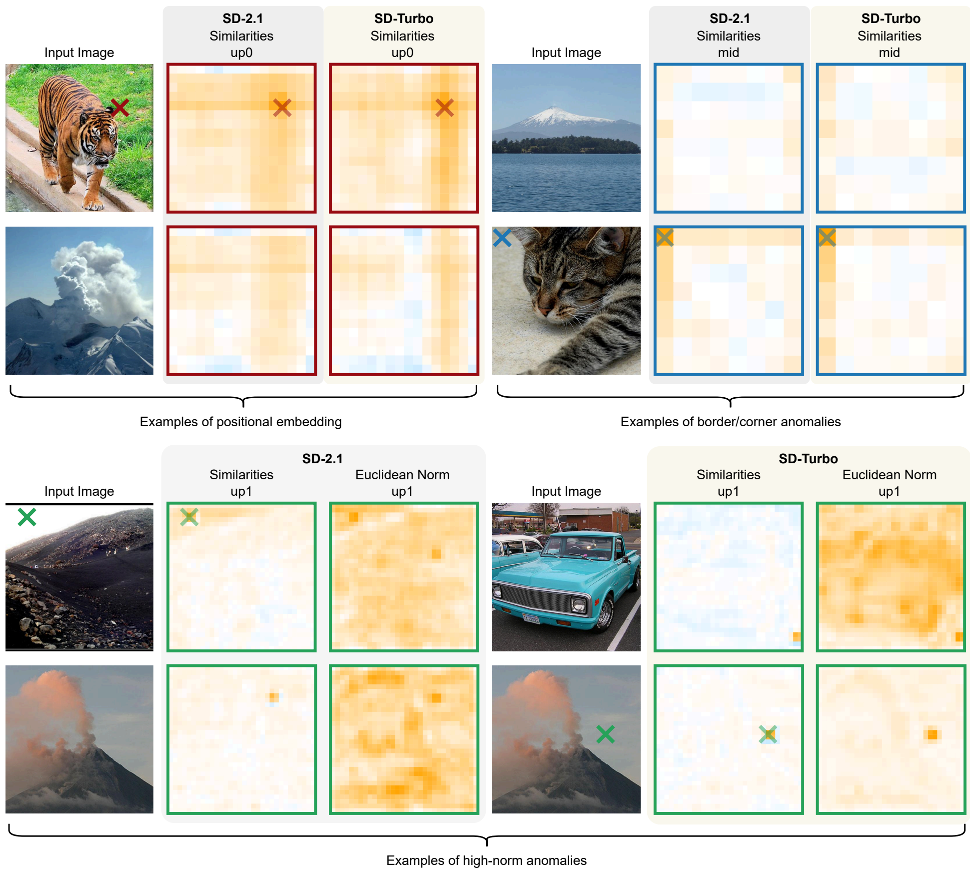

Figure 3: Cosine similarity and Euclidean norm for representations of SD-2.1 and SD-Turbo. The similarities are relative to the representation token at the image and location of the marker in the respective image pair. **Top left:** Positional embedding for SD-2.1 (left), and SD-Turbo (right). **Top right:** Corner/border anomalies for SD-2.1 (left), and SD-Turbo (right). **Bottom:** High-norm anomalies for SD-2.1 (left), and SD-Turbo (right).

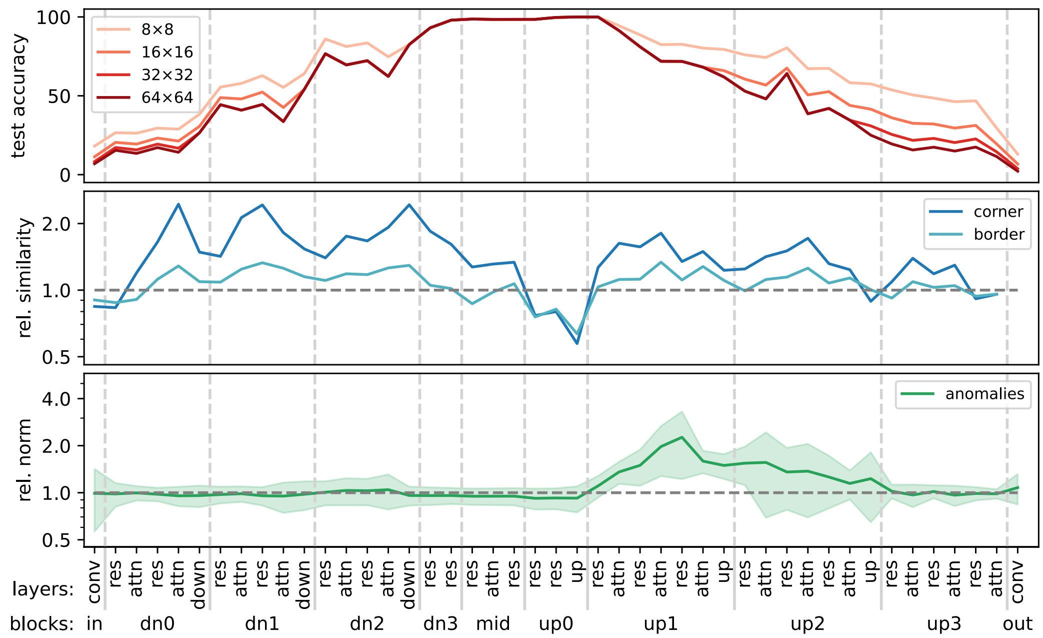

Figure 4: Quantitative results for SD-2.1. **Top row:** Linear probe accuracy for position estimation. Brighter shades indicate reduced resolution. **Middle row:** Relative similarity of tokens lying at a border/corner of the cropped images w.r.t. their similarity before cropping. (log-2 scale). **Bottom row:** Relative average norm of anomalous tokens w.r.t. to the mean norm of all tokens of the same representation (log-2 scale).

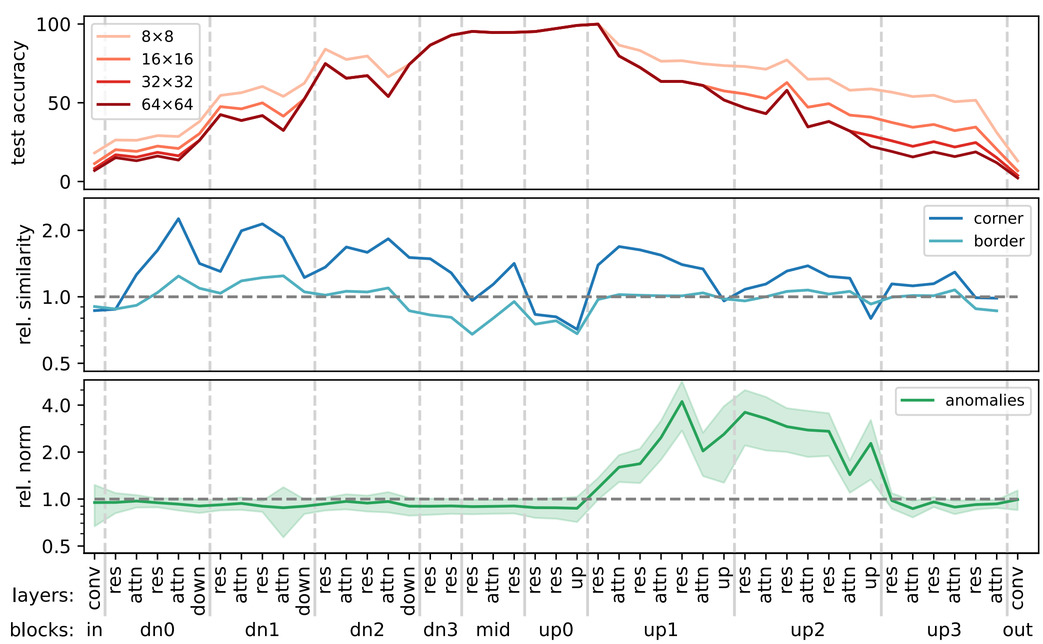

Figure 5: Quantitative results for SD-Turbo. **Top row:** Linear probe accuracy for position estimation. Brighter shades indicate reduced resolution. **Middle row:** Relative similarity of tokens lying at a border/corner of the cropped images w.r.t. their similarity before cropping. (log-2 scale). **Bottom row:** Relative average norm of anomalous tokens w.r.t. to the mean norm of all tokens of the same representation (log-2 scale).