scientific data

Check for updates

OPEN

DATA DESCRIPTOR

fruit-SALAD: A Style Aligned Artwork Dataset to reveal similarity perception in image embeddings

Tillmann Ohm[1,2] ✉, Andres Karjus[2,3,4], Mikhail V. Tamm[1,4] & Maximilian Schich[4,5]

The notion of visual similarity is essential for computer vision, and in applications and studies revolving around vector embeddings of images. However, the scarcity of benchmark datasets poses a significant hurdle in exploring how these models perceive similarity. Here we introduce Style Aligned Artwork Datasets (SALAD), and an example of fruit-SALAD with 10,000 images of fruit depictions. This combined semantic category and style benchmark comprises 100 instances each of 10 easy-to-recognize fruit categories, across 10 easy distinguishable styles. Leveraging a systematic pipeline of generative image synthesis, this visually diverse yet balanced benchmark demonstrates salient differences in semantic category and style similarity weights across various computational models, including machine learning models, feature extraction algorithms, and complexity measures, as well as conceptual models for reference. This meticulously designed dataset offers a controlled and balanced platform for the comparative analysis of similarity perception. The SALAD framework allows the comparison of how these models perform semantic category and style recognition task to go beyond the level of anecdotal knowledge, making it robustly quantifiable and qualitatively interpretable.

Background & Summary

Similarity perception is an abstract and complex concept that differs widely across mental and computational models, as explored in (computational) neuroscience[1,2], computer vision[3–6], or (computational) cognitive science[7,8]. For mental and conceptual models, similarity refers to resemblance or alikeness and describes groups with some shared properties, as prominently outlined in Wittgenstein's remarks on family resemblance[9,10]. Conversely, in computational models, similarity denotes proximity and is conventionally defined as inversely correlated with distance between data points in a metric space.

Computer Vision applications heavily rely on such visual similarity, often utilizing vector embeddings that set up a measurable multidimensional space to index images. In similarity learning the goal is to train models that can accurately capture the underlying similarities between data points, enabling tasks such as image retrieval or classification based on similarity metrics[11–17]. However, similarity in these contexts is often implied to be understood in a singular notion, overlooking the multifaceted nature of similarity perception crucial for informed decision-making in selecting models or methods. For instance, Ref. [18] utilizes CLIP[19] and DINO[20] to evaluate subject fidelity of generated images, acknowledging the varying importance of different similarity aspects. It is generally considered that CLIP captures semantic relationships, while DINO focuses more on visual features. Yet, validating such assumptions poses a significant challenge.

Research in quantitative and computational aesthetics[21–23], as well as the interplay of computation and human cultures[24,25], requires reliable benchmark datasets that are interpretable by machines and humans. Previous work has relied on embeddings of large amounts of well known artworks[26,27] or synthetic datasets of limited size[28–31].

Benchmark image datasets for perceptual similarity judgment exist, with some relying on annotated text captions of real-world images[32], while others utilize synthetic image triplets designed to better align with mental models[6]. However, these datasets primarily focus on specific tasks or aspects of similarity perception and alignment, such as zero-shot evaluation or similarity metric optimization.

Here we propose Style Aligned Artwork Datasets (SALAD), with the fruit-SALAD serving as an exemplar. This synthetic image dataset comprises 10,000 generated images featuring 10 easily recognizable fruit categories,

[1]Tallinn University, School of Digital Technologies, Tallinn, Estonia. [2]Tallinn University, School of Humanities, Tallinn, Estonia. [3]Estonian Business School, Tallinn, Estonia. [4]Tallinn University, ERA Chair of Cultural Data Analytics, Tallinn, Estonia. [5]Tallinn University, Baltic Film, Media and Arts School, Tallinn, Estonia. ✉e-mail: mail@tillmannohm.com

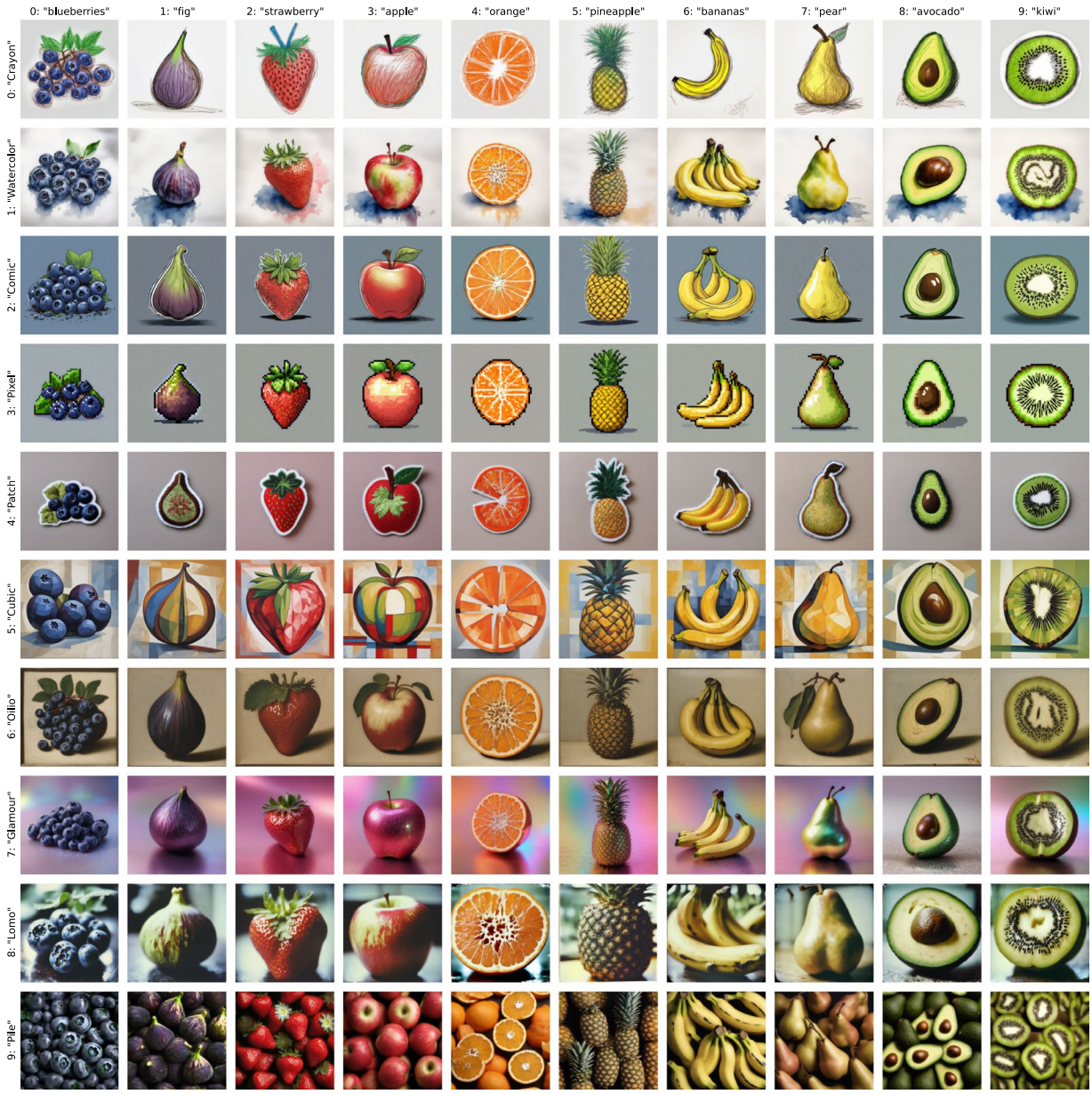

Fig. 1 Overview of the first instance of 10 fruit categories in 10 styles. Columns display fruit categories and rows display style categories with labels trying to describe the style prompts. The full dataset contains 100 instances of each category-style combination resulting in 10,000 unique fruit depictions. See Fig. 3 as an example for 100 instances of one combination.

each represented in 10 visually distinct styles, with 100 instances each (see example set of one instance in Fig. 1). Developed as a benchmark tool rather than for training purposes, the dataset is constructed on two highly controlled property dimensions – semantic (fruit category) and stylistic (artistic style) – that cannot be isolated at this level in existing real-world image datasets and therefore required image generation. The deliberate control over semantic and stylistic properties inherent to each image facilitates comparative analysis of different image embedding and complexity models, enabling an exploration of their similarity perception, only possible on scale through synthetic images.

We characterize the dataset through various machine learning models and measures of aesthetic complexity, showcasing how simple pairwise comparisons of image vectors can yield robust inter-comparable measures. Our examples reveal significant differences in similarity awareness across these methods and models, shedding light on anecdotal considerations stemming from differences in model or algorithm design, training data, parameter configuration, or similarity measures. In turn, this approach can be used to guide model training and alignment.

The fruit-SALAD offers opportunities for joint robust quantification and qualitative human interpretation, enhancing algorithmic and human perception regarding differences in measuring vector similarity and visual resemblance across computational, and statistical models. This approach allows for a more comprehensive

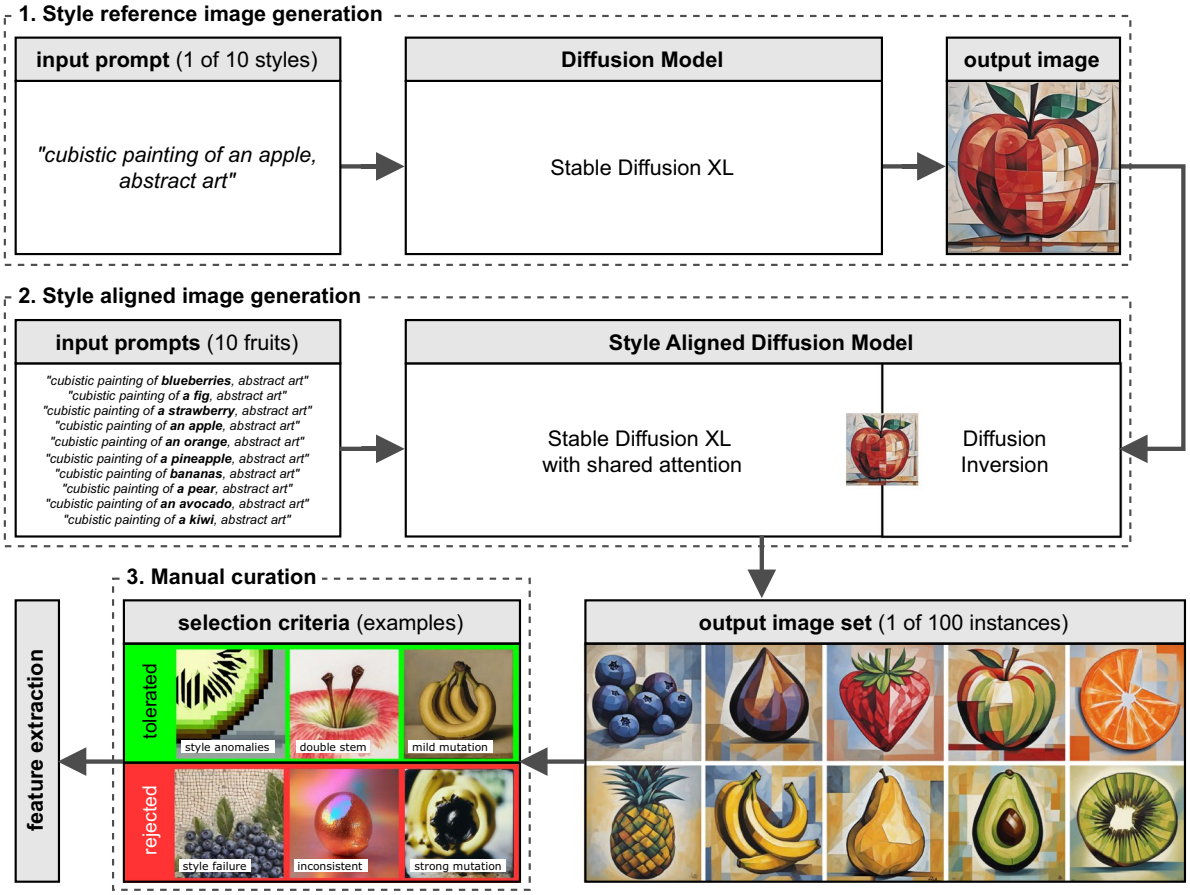

Fig. 2 Overview of the image generation process. **1.** Style reference image generation with Stable Diffusion XL[33] in manual trial-and-error fashion using text prompts of style description in combination with "an apple". **2.** Style aligned image generation[34] based on each style reference image using diffusion inversion and text prompts iterating over 10 fruit categories generating 100 instances each, resulting in 10,000 images. **3.** Manual curation with selection criteria examples: tolerated minor issues which do not impact recognition of category or style (green), and rejected major issues which are either unrecognizable or inconsistent across the style (red). The final step includes feature extraction to construct image embeddings for model comparison.

assessment of similarity perception, beyond the scope of existing benchmark datasets, ultimately contributing to a deeper understanding of computational and human similarity perception mechanisms.

Methods

Image generation. We used Stable Diffusion XL (SDXL)[33] and StyleAligned[34] to create the fruit-SALAD by carefully crafting image generation prompts and supervising the automation process. Diffusion probabilistic models[35] are typically trained with the objective of denoising blurred images. By leveraging their ability to iteratively refine images by processing random noise, these models can be used in conjunction with text prompts to generate images. Such Text-to-Image models, exemplified by DALL-E[36], Midjourney and StableDiffusion[37], have recently gained significant attention in various creative and commercial domains. These models and services have simplified the synthesis of high-quality individual images, enabling unprecedented ease of use through natural language. However, scaling the generation process or achieving stylistically consistent images remains challenging but can be improved by style alignment methods[34] to coordinate shared attention across multiple generations based on a reference style image.

We utilized a computational approach to scale the image generation process (see Fig. 2). Initially, we experimented in a trial-and-error fashion with different style prompts in conjunction with different fruit categories, using SDXL[33] for image generation. Successful results were selected as style references. We then used style alignment[34] to generate multiple instances of different fruits within the same style using diffusion inversion[38] of the reference image. Through several iterations and adjustments to the prompts, we refined the process and eventually automated the generation to produce 100 instances for each fruit-style combination (see all 100 instances of one fruit-style example in Fig. 3).

The fruit prompts and stylistic references we selected were carefully curated to improve the robustness of the style alignment generation method. Among the fruit prompts, we balanced between fruit prototypicality and

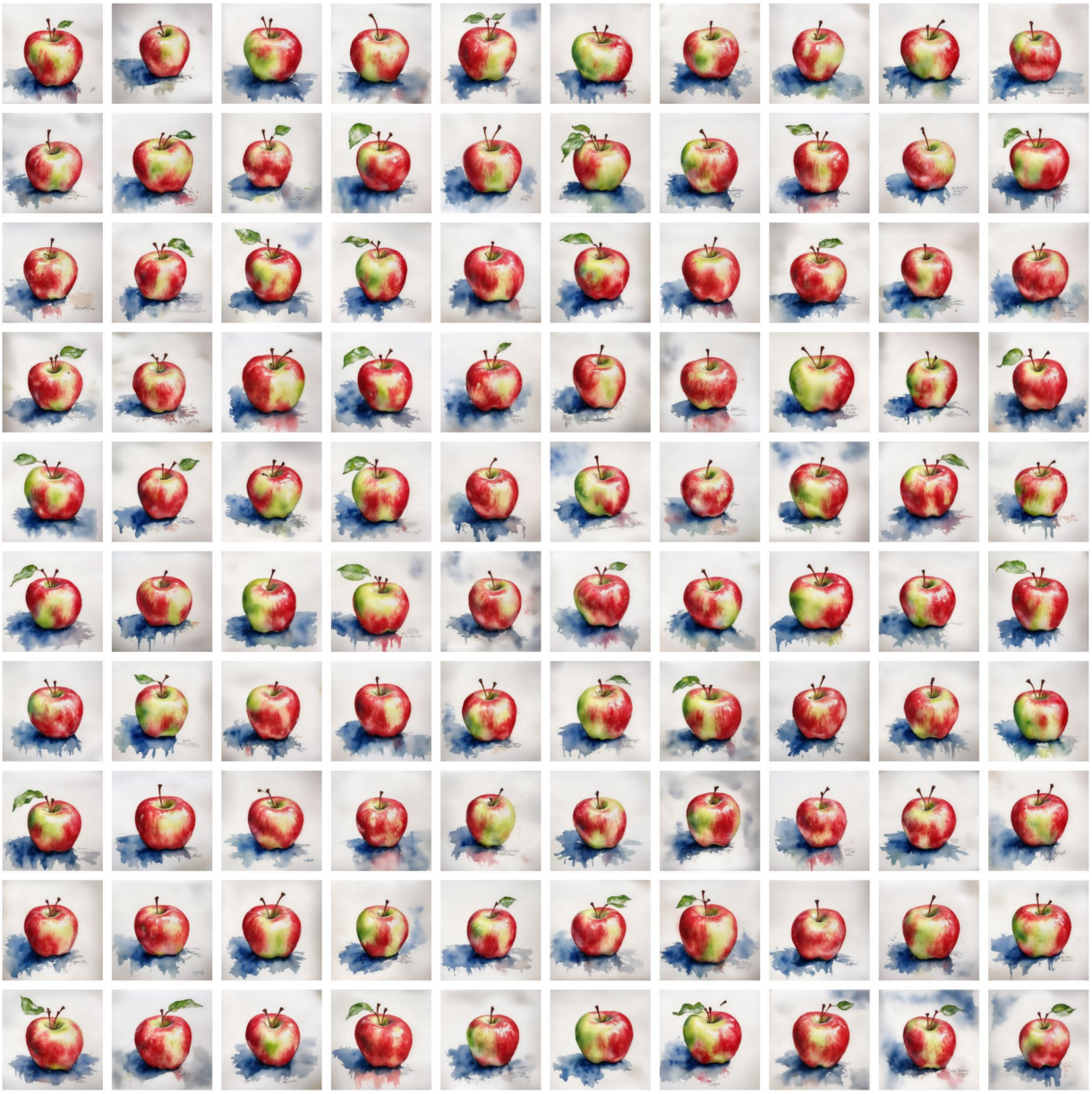

Fig. 3 All instances of fruit category *3* (apple) in style category *1* (Watercolor). Corresponds to 100 dataset files *3_1_0.png* to *3_1_99.png*. Text prompt: "watercolor sketch of a gala apple, aquarelle, wet paint".

variability across different stylistic prompts to ensure compatibility with generation on scale, while simultaneously covering a wide range of fruit shapes and colors. Similarly, our selection of stylistic references was based on their effectiveness in aligning with the generation space, focusing on those that demonstrated superior performance in achieving stylistic coherence.

We maintained dataset quality by visually assessing the entire dataset in 100 batches of 10 by 10 image grids and manually replaced images that were inconsistent across all instances (see examples of the manual selection criteria in Fig. 2). Therefore, the final dataset with category and style classes may be biased by our own aesthetic arbitration, which is akin to the inherent specificity of a chosen set of handwritten digits[39].

Image embeddings. Our exemplary vector embeddings are derived from machine learning models and compression algorithms through various commonly employed methods (Table 1). For[19,20,40–43] we extracted feature vectors using the flattened last hidden states. For[44–46] we used average pooling from the second to last layer.

As an example of a quantitative aesthetics measure, we used the Compression Ensembles method[21], which captures polymorphic family resemblance via a number of transformations (87 in our implementation). We used GIF image compression ratios, taking advantage of the Lempel–Ziv–Welch algorithm[47]. We also provide the PNG file sizes as comparison (Table 2).

To provide simple conceptual models for reference, we used binary, one-hot encoded vectors. In this encoding scheme, each vector represents a fruit category or style, with a value of 1 indicating the presence and 0

model short name	type	training set	dimensions
ViT-B-16_IN21k	Vision Transformer[40] (base, 16 × 16)	ImageNet-21k[53]	768
ViT-B-32_IN21k	Vision Transformer[40] (base, 32 × 32)	ImageNet-21k[53]	768
ViT-H-14_IN21k	Vision Transformer[40] (huge, 14 × 14)	ImageNet-21k[53]	1280
DINO_IN1k	DINO[20], Vision Transformer[40] (base, 16 × 16)	ImageNet-1k[53]	768
DINOv2-B_LVD	DINOv2[41] (base)	LVD-142M[41]	768
ResNet50_IN1k	ResNet[44]	ImageNet-1k[53]	2,048
VGG19_IN1k	VGG[45]	ImageNet-1k[53]	512
Xception_IN1k	Xception[46]	ImageNet-1k[53]	2,048
ConvNeXt_L400M	ConvNeXt[43] (base)	LAION-400M[54]	512
ConvNeXt-v2_L400M	ConvNeXt-V2[42]	LAION-400M[54]	320
CLIP-ViT-B-16_L2B	CLIP[19], Vision Transformer[40] (base, 16 × 16)	LAION-2B[55]	512
CLIP-ViT-B-32_L2B	CLIP[19], Vision Transformer[40] (base, 32 × 32)	LAION-2B[55]	512
CLIP-ViT-H-14_L2B	CLIP[19], Vision Transformer[40] (huge, 14 × 14)	LAION-2B[55]	1,024
CLIP-ViT-B-16_L400M	CLIP[19], Vision Transformer[40] (base, 16 × 16)	LAION-400M[53]	512
CLIP-ViT-B-16_OA	CLIP[19], Vision Transformer[40] (base, 16 × 16)	OpenAI (undisclosed)	512
CLIP-RN50_OA	CLIP[19], ResNet50[44]	OpenAI (undisclosed)	1,024
CLIP-RN101_OA	CLIP[19], ResNet101[44]	OpenAI (undisclosed)	512

Table 1. Pre-trained machine learning models used for feature extraction.

model short name	method	dimensions
CompressionEnsembles	Compression Ensembles[21]	87
GIF_compression	LZW[47] to PNG file size ratios	1
PNG_filesizes	original PNG file sizes	1
style_blind	one-hot encoding of fruit category only, ignoring styles	10
category_blind	one-hot encoding of styles only, ignoring fruit category	10
balanced	one-hot encoding of fruit category and styles	20

Table 2. Other methods used for feature extraction.

	blueberries	fig	strawberry	apple	orange	pineapple	bananas	pear	avocado	kiwi
Crayon	99.45					99.94		99.96	99.78	
Watercolor	96.01		99.98	94.96	94.23	96.87		99.91		
Comic	99.97					99.93		99.79		
Pixel										
Patch	99.99		99.98			99.74			99.62	
Cubic		99.99								
Ollio	97.03	99.90						99.95	99.78	99.80
Glamour	99.83			99.96	99.74	99.99				
Lomo	99.99					99.97		99.93	99.57	
Pile	99.99				99.99	99.91		99.99	99.63	

	blueberries	fig	strawberry	apple	orange	pineapple	bananas	pear	avocado	kiwi
Crayon	95.66	94.77	98.14	98.46	97.42	92.61	98.75	95.35	95.15	99.88
Watercolor	96.01	99.96	99.25	78.97	71.48	85.98	99.24	98.07	97.38	99.81
Comic	94.29	99.47	99.99	98.89	96.80	94.55	99.92	97.61	98.88	
Pixel	98.72	99.75	99.61	97.41	79.11	98.22	98.21		99.29	99.99
Patch	99.56	98.45	99.33	99.94	99.03	95.75	96.29	96.87	98.42	99.56
Cubic	98.31	99.76		97.86	99.60	96.89			99.99	99.42
Ollio	93.72	99.08	98.66	97.06	95.69	97.15	95.77	92.78	87.37	97.88
Glamour	99.74		99.78	83.95	98.02	94.50	99.82	99.89	95.96	99.77
Lomo	99.99	95.37	99.49	99.67	99.58	95.27	95.48	87.73	96.01	99.22
Pile	99.11	98.59	97.44	99.42	96.27	97.55	98.54	94.12	97.49	99.81

Fig. 4 Self-recognition tests. Each cell represents the mean number of same instances in the top 100 nearest neighbors of its fruit category (column) and style (row) combination images. White cells without values have a perfect score of 100 out of 100 correctly recognized instances. Left: Maximum values from all computational models, taking into account that high scores within 100 out of 10,000 images reflect higher than chance results. Right: *ResNet50_IN21k* as an example model.

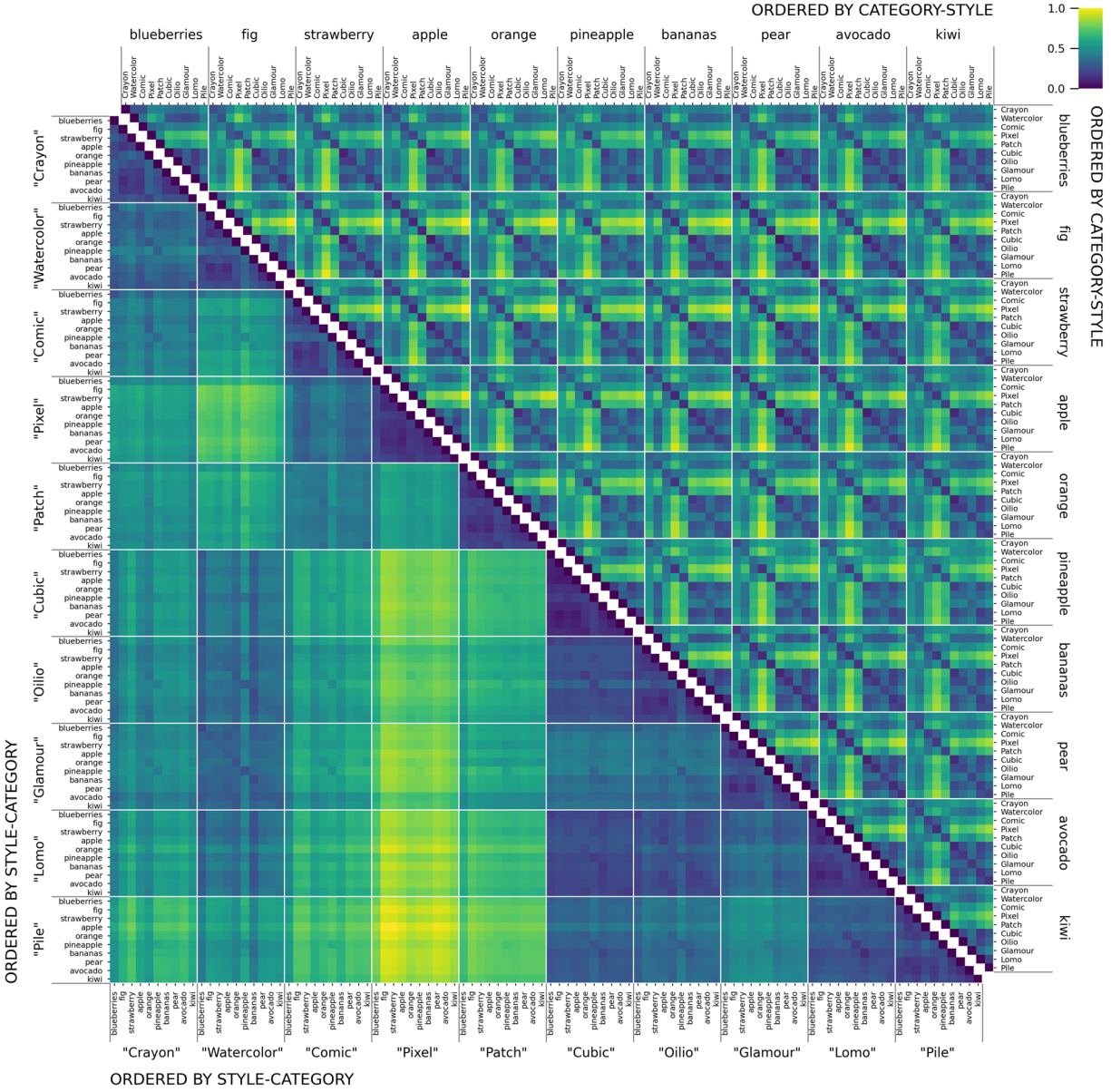

Fig. 5 *DINO-ViT-B-16_IN1k* heatmaps indicating the mutual Mahalanobis distances of fruit-SALAD images. The matrix cells correspond to the mean of all 10,000 distance pairs of 100 by 100 instances of fruit-SALAD_10k images. **Below the diagonal**: sorted by style first and fruit category second. **Above the diagonal**: sorted by fruit category first and style second. The color indicates the pairwise Mahalanobis distance of image embedding vectors obtained from the respective model or algorithm, from low to high (blue to yellow) while low values indicate higher similarity. The figure construction is comprehensive as the matrices are symmetric; diagonal cells can be left out. See all model heatmaps in Supplementary Fig. S1.

indicating the absence of the corresponding category or style (Table 2). We are consciously providing a simple conceptual reference, to avoid the complications of full blown conceptual reference models, such as the CIDOC-CRM[48].

Data Records

The *fruit-SALAD_10k* is available at Zenodo under record number 11158522[49] (https://zenodo.org/records/11158522). The repository includes 10,000 PNG files of fruit images (1024 × 1024 pixel), 10 PNG files of style reference images, 10 CSV files with text prompts, 100 PNG files of grid overview plots (10 × 10 images per instance), 23 CSV vector files, 23 PNG files of model heatmaps, 1 CSV file containing 23 model vectors and 1 CSV file with index labels. We provide a detailed overview of all dataset repository files in Supplementary Fig. S2.

The 10,000 fruit image filenames adhere to the following format: *fruit_style_instance.png*. For example, an image with the filename *8_1_42.png* signifies fruit category *8* (avocado) rendered in style category *1* (Watercolor), and it represents generation number *42*.

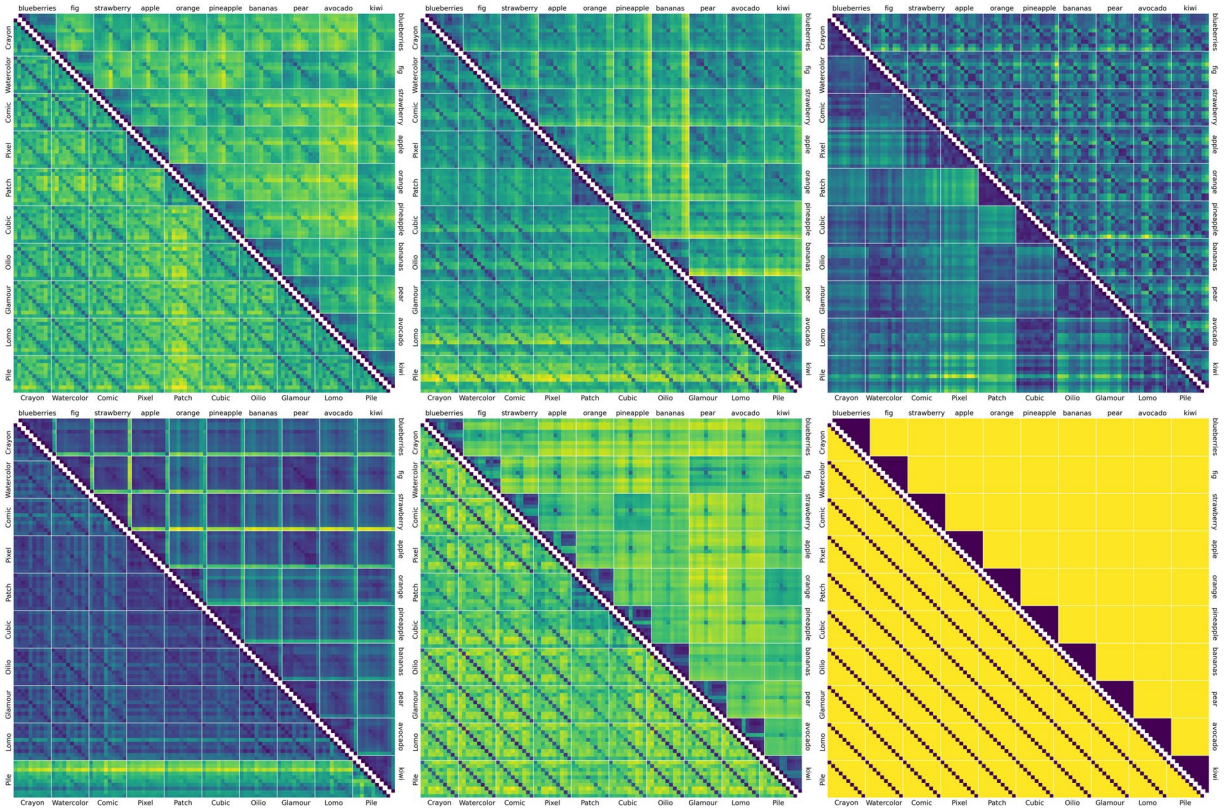

Fig. 6 Heatmaps indicating the mutual Mahalanobis distance of fruit-SALAD_10k images according to different models (see Fig. 5). Top row from left to right: *CLIP-ViT-B-16_L400M, DINOv2-B_LVD, CompressionEnsembles*. Bottom row from left to right: *VGG19_IN1k, ViT-B-32_IN21, style_blind*. The matrix ordering is identical.

For accessibility, we provide all vector files as comma-separated values (.csv) with image file names as indices.

Technical Validation

Self recognition test. One expects that, despite inevitable variation in similarity perception, the similarity of images from the same category-style combination should be systematically larger than between images of different categories and/or styles. To assess this, we conduct a self-recognition test on the fruit-SALAD_10k dataset. This test involves retrieving the top 100 nearest neighbors for each image and counting how many instances of the same category-style combination are found within this set. The average number of successful retrievals across all 100 instances per model is then calculated. To validate the self-recognition of image instances, we select the maximum values across all computational models (Fig. 4).

If a category-style combination cannot be sufficiently recognized in any of the computational models, we consider the self-recognition test failed. Notably, we found that "apples" and "oranges" in the "Watercolor" style pose the greatest challenge, achieving sufficient accuracy only after various iterations of image generation (see Fig. 3 for all 100 instances of the apple-Watercolor combination).

Model heatmaps. We characterize the dataset, and concurrently exemplify its possible future use by a set of category- and style-ordered distance matrices, which demonstrate salient differences in category and style similarity weights, across various computational models (Supplementary Fig. S1; see examples in Figs. 5 and 6). As a measure of similarity between two sets of images we calculate the average distances between all pairs of elements. To better generalize standardization, we use Mahalanobis distance[50,51], which normalizes and decorrelates the coordinates.

Model comparison. Each of the multiple embedding models can be characterized by a set of distances between images in this embedding. One can consider this set of distances as a multidimensional vector, characterizing a model. Thus, the different models are represented as vectors in a shared space, which enables their direct comparison. As coordinates we used standardized pairwise distances between all unique pairs of 100 fruit category-style combinations, i.e., all entries of the model heatmaps. The principle components of the resulting embedding are shown in Fig. 7.

Investigating the differences in similarity perception can also be accomplished by examining fruit categories and styles through the image embeddings of individual models (Fig. 8). We provide an interactive exploration

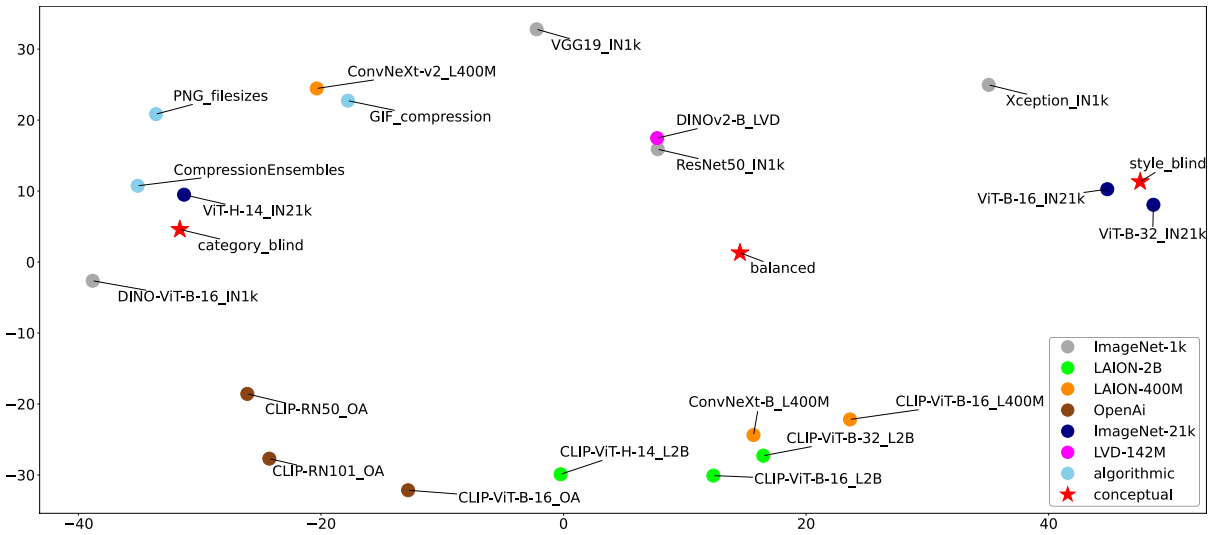

Fig. 7 Relative model comparison using principal component analysis (PCA) based on 23 standardized model vectors of 4,950 dimensions. These dimensions encompass the mutual Mahalanobis distances of all unique category-style combinations of the fruit-SALAD_10k images, excluding self-pairing. Each fruit category-style combination is the mean of all 10,000 mutual distances of 100 by 100 fruit-SALAD image instances.

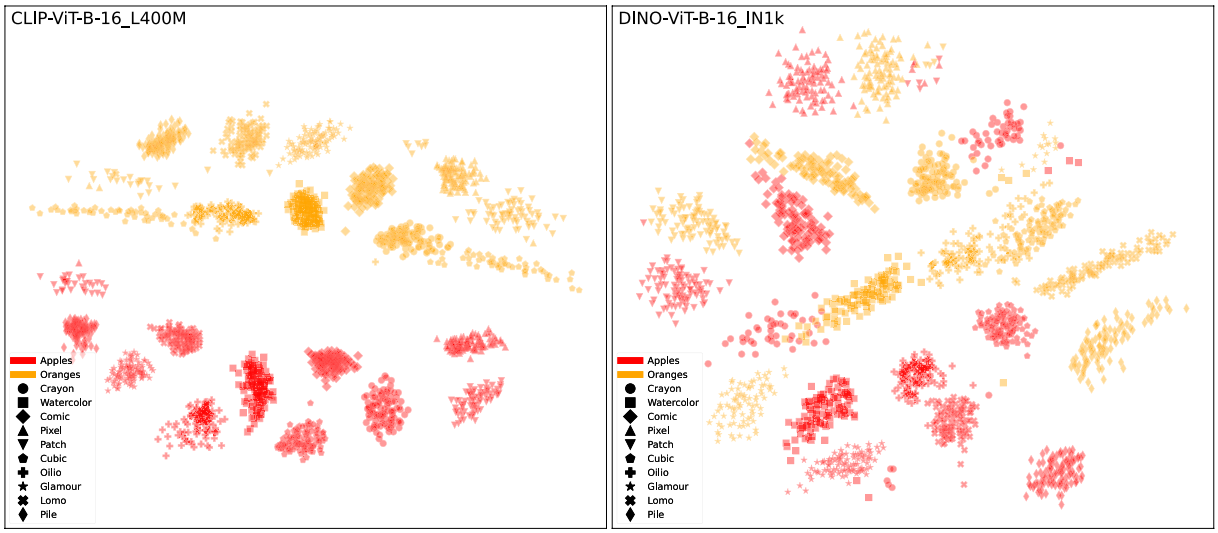

Fig. 8 Scatter plots of apples and oranges using multidimensional scaling (MDS) based on normalized image embedding vectors from two different models. **Left**: *CLIP-ViT-B-16_L400M*; **right**: *DINO-ViT-B-16_IN1k*. Colors indicate fruit categories and dot shapes indicate styles.

tool based on the Collection Space Navigator[52] to visually compare such projections of model embeddings (https://style-aligned-artwork-datasets.github.io/fruit-explorer).

Code availability

Code performed to generate the fruit images is available at https://github.com/Style-Aligned-Artwork-Datasets/fruit-SALAD. The GitHub repository entails all necessary files and implementations to reproduce the fruit-SALAD benchmark dataset.

Received: 9 May 2024; Accepted: 28 January 2025;
Published online: 12 February 2025

References

1. Kaiser, D., Jacobs, A. M. & Cichy, R. M. Modelling brain representations of abstract concepts. *PLoS Computational Biology* **18**, e1009837, https://doi.org/10.1371/journal.pcbi.1009837 (2022).

2. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences* **111**, 14565–14570, https://doi.org/10.1073/pnas.1402594111 (2014).
3. Lang, S. & Ommer, B. Attesting similarity: Supporting the organization and study of art image collections with computer vision. *Digital Scholarship in the Humanities* **33**, 845–856, https://doi.org/10.1093/llc/fqy006 (2018).
4. Wei, Z., Wang, S. & Thawonmas, R. Difference in perceived similarity between humans and machines. *Art Research* **22**, 2 (2022).
5. Muttenthaler, L. *et al*. Improving neural network representations using human similarity judgments. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 50978–51007, https://dl.acm.org/doi/10.5555/3666122.3668340 (2023).
6. Fu, S. *et al*. DreamSim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 50742–50768, https://dl.acm.org/doi/10.5555/3666122.3668330 (2023).
7. Hummel, J. E. & Doumas, L. A. A. Analogy and similarity. In *The Cambridge Handbook of Computational Cognitive Sciences*, 451–473, https://doi.org/10.1017/9781108755610.018 (2023).
8. Richie, R. & Bhatia, S. Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive science* **45**, e13030, https://doi.org/10.1111/cogs.13030 (2021).
9. Wittgenstein, L. Philosophical Investigations, *Basil Blackwell* (1968).
10. Rosch, E. & Mervis, C. B. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology* **7**, 573–605, https://doi.org/10.1016/0010-0285(75)90024-9 (1975).
11. Mishra, S. *et al*. Effectively leveraging attributes for visual similarity. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3899–3904, https://doi.org/10.1109/CVPRW53098.2021.00434 (2021).
12. Liu, W., Liu, Z., Rehg, J. M. & Song, L. Neural similarity learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 5025–5036, https://dl.acm.org/doi/10.5555/3454287.3454739 (2019).
13. Cheng, X., Zhang, L. & Zheng, Y. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **6**, 248–252, https://doi.org/10.1080/21681163.2015.1135299 (2018).
14. Veit, A., Belongie, S. & Karaletsos, T. Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 830–838, https://doi.org/10.1109/CVPR.2017.193 (2017).
15. Mathisen, B. M., Aamodt, A., Bach, K. & Langseth, H. Learning similarity measures from data. *Progress in Artificial Intelligence* **9**, 129–143, https://doi.org/10.1007/s13748-019-00201-2 (2020).
16. Song, H. O., Xiang, Y., Jegelka, S. & Savarese, S. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4004–4012, https://doi.org/10.1109/CVPR.2016.434 (2016).
17. Ma, W.-Y. & Manjunath, B. S. Texture features and learning similarity. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, 425–430, https://doi.org/10.1109/CVPR.1996.517107 (1996).
18. Ruiz, N. *et al*. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510, https://doi.org/10.1109/CVPR52729.2023.02155 (2023).
19. Radford, A. *et al*. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (2021).
20. Caron, M. *et al*. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640, https://doi.org/10.1109/ICCV48922.2021.00951 (2021).
21. Karjus, A., Solà, M. C., Ohm, T., Ahnert, S. E. & Schich, M. Compression ensembles quantify aesthetic complexity and the evolution of visual art. *EPJ Data Science* **12**, 21, https://doi.org/10.1140/epjds/s13688-023-00397-3 (2023).
22. Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J. & Sawey, M. Predicting beauty: Fractal dimension and visual complexity in art. *British journal of psychology* **102**, 49–70, https://doi.org/10.1348/000712610x498958 (2011).
23. Zhang, J., Miao, Y., Zhang, J. & Yu, J. Inkthetics: A comprehensive computational model for aesthetic evaluation of chinese ink paintings. *IEEE Access* **8**, 225857–225871, https://doi.org/10.1109/ACCESS.2020.3044573 (2020).
24. Brinkmann, L. *et al*. Machine culture. *Nature Human Behaviour* **7**, 1855–1868, https://doi.org/10.1038/s41562-023-01742-2 (2023).
25. McCormack, J. & Cruz Gambardella, C. Complexity and aesthetics in generative and evolutionary art. *Genetic Programming and Evolvable Machines* **23**, 535–556, https://doi.org/10.1007/s10710-022-09429-9 (2022).
26. Srinivasa Desikan, B., Shimao, H. & Miton, H. Wikiartvectors: Style and color representations of artworks for cultural analysis via information theoretic measures. *Entropy* **24**, 1175, https://doi.org/10.3390/e24091175 (2022).
27. Mao, H., Cheung, M. & She, J. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, 1183–1191, https://doi.org/10.1145/3123266.3123405 (2017).
28. Lakhal, S., Darmon, A., Bouchaud, J.-P. & Benzaquen, M. Beauty and structural complexity. *Physical Review Research* **2**, 022058, https://doi.org/10.1103/PhysRevResearch.2.022058 (2020).
29. Ostmeyer, J. *et al*. Synthetic images aid the recognition of human-made art forgeries. *Plos one* **19**, e0295967, https://doi.org/10.1371/journal.pone.0295967 (2024).
30. Duñabeitia, J. A. *et al*. Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly journal of experimental psychology* **71**, 808–816, https://doi.org/10.1080/17470218.2017.1310 (2018).
31. Ovalle-Fresa, R., Di Pietro, S. V., Reber, T. P., Balbi, E. & Rothen, N. Standardized database of 400 complex abstract fractals. *Behavior Research Methods* **54**, 2302–2317, https://doi.org/10.3758/s13428-021-01726-y (2022).
32. Liao, P., Li, X., Liu, X. & Keutzer, K. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404* https://doi.org/10.48550/arXiv.2206.11404 (2022).
33. Podell, D. *et al*. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* https://doi.org/10.48550/arXiv.2307.01952 (2023).
34. Hertz, A., Voynov, A., Fruchter, S. & Cohen-Or, D. Style aligned image generation via shared attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4775–4785, https://doi.org/10.1109/CVPR52733.2024.00457 (2024).
35. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 6840–6851, https://dl.acm.org/doi/10.5555/3495724.3496298 (2020).
36. Ramesh, A. *et al*. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning* **139**, 8821–8831 (2021).
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685, https://doi.org/10.1109/CVPR52688.2022.01042 (2022).
38. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* https://doi.org/10.48550/arXiv.2010.02502 (2020).
39. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324, https://doi.org/10.1109/5.726791 (1998).
40. Dosovitskiy, A. *et al*. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* https://doi.org/10.48550/arXiv.2010.11929 (2020).

41. Oquab, M. *et al*. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* https://doi.org/10.48550/arXiv.2304.07193 (2023).
42. Woo, S. *et al*. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16133–16142, https://doi.org/10.1109/CVPR52729.2023.01548 (2023).
43. Liu, Z. *et al*. A convnet for the 2020 s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11966–11976, https://doi.org/10.1109/CVPR52688.2022.01167 (2022).
44. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778, https://doi.org/10.48550/arXiv.1512.03385 (2016).
45. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* https://doi.org/10.48550/arXiv.1409.1556 (2014).
46. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258, https://doi.org/10.48550/arXiv.1610.02357 (2017).
47. Welch, T. A. A technique for high-performance data compression. *Computer* **17**, 8–19, https://doi.org/10.1109/MC.1984.1659158 (1984).
48. Doerr, M. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* **24**, 75–75, https://doi.org/10.1609/aimag.v24i3.1720 (2003).
49. Ohm, T. fruit-salad. *Zenodo*, https://doi.org/10.5281/zenodo.11158522 (2024).
50. Mahalanobis, P. On the generalised distance in statistics (reprint, 2018). *Sankhya A* **80**, 1–7, https://doi.org/10.1007/s13171-019-00164-5 (1936).
51. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* **50**, 1–18, https://doi.org/10.1016/S0169-7439(99)00047-7 (2000).
52. Ohm, T., Solà, M. C., Karjus, A. & Schich, M. Collection Space Navigator: An Interactive Visualization Interface for Multidimensional Datasets. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, 1–5, https://doi.org/10.1145/3615522.3615546 (2023).
53. Deng, J. *et al*. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
54. Schuhmann, C. *et al*. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* https://doi.org/10.48550/arXiv.2111.02114 (2021).
55. Schuhmann, C. *et al*. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 25278–25294, https://dl.acm.org/doi/10.5555/3600270.3602103 (2022).

Acknowledgements

All authors were supported by the CUDAN ERA Chair project for Cultural Data Analytics, funded through the European Union's Horizon 2020 research and innovation program (Grant No. 810961).

Author contributions

T.O. co-designed the research, generated the dataset, co-wrote the text, and created the figures. M.T., A.K. and M.S. also co-designed the research, co-wrote the text, and provided conceptual guidance. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04529-4.

Correspondence and requests for materials should be addressed to T.O.