ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

But Are These Images *Concept*ually Similar?

Isaac Roberts*, Riza Velioglu*, Inaam Ashraf, Luca Hermes and Barbara Hammer *†

CITEC – Center for Cognitive Interaction Technology
Bielefeld University – Faculty of Technology
Inspiration 1, 33619 Bielefeld – Germany

Abstract. Assessing the similarity between two images remains a core challenge in computer vision. Traditional full-reference image quality assessment (FR-IQA) metrics measure pixel-wise or low-level structural distortions and falter when human perception effortlessly recognizes equivalence. Perceptual metrics such as LPIPS and DISTS improve correlation with human judgments but remain opaque black boxes. We propose **Conceptual Similarity (CSIM)**, a transparent and steerable image similarity metric that operates directly on human-interpretable semantic concepts. The resulting metric is simultaneously (1) a FR-IQA metric that is robust to non-semantic distortions while remaining sensitive to meaningful semantic changes, and (2) a general image-to-image similarity measure. Most importantly, CSIM offers transparency—users can inspect which concepts drive the score—and gives rise to a novel capability, which we call **Human Similarity Steering** that permits user-determined per-concept weighting to influence the similarity score according to their preferences.

1 Introduction

The generation of high-fidelity images has seen a paradigm shift with the advent of diffusion models, enabling applications ranging from virtual try-on to complex image editing [1, 2]. However, the evaluation of these generative models remains a persistent bottleneck. Traditionally, Image Quality Assessment (IQA) has relied on Full-Reference (FR) metrics, such as the Structural Similarity Index Measure (SSIM) [3], to quantify the fidelity of a generated image against a ground truth. Yet, these pixel-level metrics often fail to align with human perception [4]. This limitation necessitated the development of perceptual metrics like LPIPS [4] and DISTS [5], which compute distances between deep image embeddings. Recently, diffusion models [6] and perceptual metric learning [7] have been used to measure visual similarity. All approaches output a score indicating *how* similar two images are but fail to explain *why*, offering no insight into specific features driving similarity.

A single scalar score cannot effectively communicate the multifaceted factors involved in image similarity [8]. Specifically, these scores lack transparency regarding *what* ultimately contributes to the final output. Recent works address this opacity using Multi-modal Large Language Models (MLLMs) to generate

*The first two authors contributed equally.

†This project received funding from the EU Horizon Europe programme under the Marie Skłodowska-Curie grant agreement No 101073307 and 101134447-ENFORCE.
Code available at: `https://github.com/rizavelioglu/ConceptSim`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

textual descriptions of differences [8]. However, we argue that relying on generative text is computationally heavier and indirect; instead, human-interpretable features can be extracted directly from the image embeddings themselves without requiring additional natural language.

To address these gaps, we introduce Conceptual Similarity (CSIM), a novel explainable similarity metric. Unlike traditional approaches that operate on opaque feature maps or pixel-wise differences, CSIM evaluates similarity based on semantically meaningful attributes characterized by Concept Activation Vectors (CAVs) [9]. By applying sparse dictionary learning to the embeddings of a foundation model, we decompose the latent space into distinct semantic components. Crucially, this maps the abstract dimensions of the embedding space to understandable concepts; by isolating *what* features are active in the vector, we can begin to derive *why* two images are deemed similar.

Our approach offers three distinct advantages over existing baselines. First, it ensures robustness to non-semantic transformations. Our experiments demonstrate CSIM maintains high similarity scores under geometric shifts like rotation and padding, while correctly penalizing semantic degradation such as noise or content removal. Second, it provides inherent explainability. Because each dimension corresponds to a distinct semantic concept, users can inspect *what* attributes are driving the distance score, thereby revealing *why* the model considers images similar or dissimilar. Finally, it enables Human Similarity Steering; by selectively weighting concepts, users align the metric with their intent, defining similarity based on attributes important to their domain.

2 Background

Recent work on fashion-oriented tasks revealed critical flaws in traditional metrics [10]. Particularly, SSIM assigns unreasonably high scores to comparisons between garments and plain white images, as the background dominates the image content. Perceptual metrics offer improvement but fail to reliably capture fine-grained fashion attributes such as fabric texture, pattern, neckline shape, or sleeve length that dominate human similarity judgments. These shortcomings motivate our focus on the fashion domain, where perceptual similarity is defined by semantic attributes rather than spatial correspondence. While we evaluate CSIM on fashion imagery, we expect the approach to extend to various data domains and should apply wherever interpretable concepts can be captured.

To address this, we map images into a more transparent vector space using Concept Activation Vectors (CAVs). While deep embeddings are generally opaque, unsupervised dictionary learning can extract interpretable attributes from them [11]. Recent automatic, post-hoc approaches typically use dictionary learning methods to extract concepts [11]: First, inputs are embedded into an activation space of a pre-trained model. Then, the embedding data matrix $\mathbf{A}$ is decomposed into a product of matrices $\mathbf{U}$ and $\mathbf{V}$, solved by reconstructing $\mathbf{A}$, i.e., $(\mathbf{U}, \mathbf{V}) = \arg\min_{\mathbf{U},\mathbf{V}} \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_F^2$. The decomposition yields: $\mathbf{V}$ the dictionary of concepts (or concept bank) and $\mathbf{U}$ a reduced representation of $\mathbf{A}$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

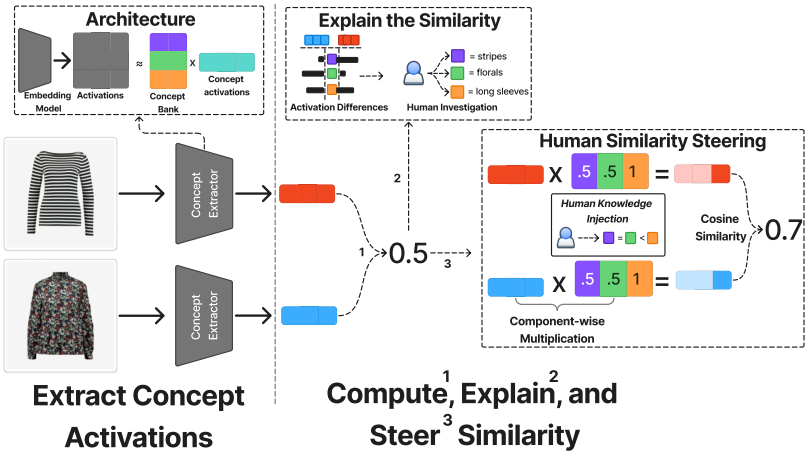

Fig. 1: Our pipeline for generating and explaining concept similarity scores.

according to the dictionary $\mathbf{V}$.

3 Methodology

In Fig. 1, we illustrate our Concept Similarity pipeline for performing image similarity with two images. Specifically, we use a trained deep embedding model $g : \mathcal{I} \to \mathbb{R}^d$ that maps a sample of the image space $\mathcal{I}$ onto a d-dimensional embedding vector, as well as a domain-trained concept extractor $f_{\mathbf{V}} : \mathbb{R}^d \to \mathbb{R}^{1\times c}$ (such as [11]). The latter effectively solves the optimization problem $f_{\mathbf{V}}(x) = u_x = \arg\min_{u_x} \|g(x) - u_x\mathbf{V}^\top\|_F^2$, representing the embedding of an image $x \in \mathcal{I}$ as a linear combination of the vectors in $\mathbf{V} \in \mathbb{R}^{d\times c}$. The concept dictionary $\mathbf{V}$ is precomputed over the whole dataset as described in Section 2. The vector u_x contains the *concept activations* of the image x.

Therefore, our proposed concept difference metric to compare two images $x_1, x_2 \in \mathcal{I}$ is formalized as $CSIM(x_1, x_2) = D(u_{x_1}, u_{x_2}) = \frac{u_{x_1} u_{x_2}^\top}{\|u_{x_1}\|\|u_{x_2}\|}$, where D can be any distance metric, for example the cosine similarity, as used in our experiments. To explain the resulting score, we can inspect the component-wise difference $u_{x_1} - u_{x_2}$, or compare the rank orderings σ_{x_1} and σ_{x_2} by computing $\sigma_x = \text{argsort}(u_x)$ to identify concepts whose relative importance differs across the two images. While defining an optimal explanation strategy is a compelling direction, it falls outside the current scope and is left as future work.

Human Similarity Steering. If the semantic meaning of the concepts is known, either through explicit supervision or human interpretation, we can adjust the similarity computation accordingly. In Fig. 1, we increase the similarity between the two images by reducing the floral and stripes concept contributions. Let

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

$w \in \mathbb{R}^c$ be a weight vector that nulls the contributions from these concepts. We apply the weights element-wise via the Hadamard product, $u_x \odot w$, recompute the distance between the weighted vectors $u_{x_1} \odot w$, $u_{x_2} \odot w$. This produces a similarity score influenced by the injected prior knowledge.

4 Experimental Evaluation

In this section, we empirically assess the robustness of our conceptual similarity metric under controlled image modifications. We evaluate whether the metric assigns high similarity to transformations that preserve semantic content and lower similarity to modifications that significantly alter the image content. We further benchmark our approach against established similarity measures from the literature. Finally, we demonstrate how our metric enables Human Similarity Steering, showcasing practical scenarios in which user-desired outcomes drive the concept contributions and influence the similarity score.

Experimental Setup. We use **DINOv3** [12] as the frozen backbone and train **TopK Archetypal SAEs** [13] on activations extracted from the VITON-HD dataset [1]. All training hyperparameters follow the default configuration of [13]. The number of archetypes is set to 30% of the number of data points, and the relaxation parameter(δ) is set to 1. For robustness experiments, we use a dictionary of 1000 concepts, selected via grid search and optimized for the robustness task. For steering demonstrations, we reduce this to 100 concepts and manually retain 59 relevant ones. We evaluate three variants: **CSIM-NI** (No Intervention), **CSIM-HI** (Human-Identified; 59 manually selected), and **CSIM-HS** (Human-Steered).

	Perceptual Quality↓		Structural Similarity↑			
Method	LPIPS	DISTS	SSIM	MS^{SSIM}	CW^{SSIM}	CSIM-NI↑
rotation	.129	.054	<u>.823</u>	.801	.799	**.946**
horizontal flip	.148	.070	<u>.827</u>	.801	.625	**.995**
shift	.448	.124	<u>.663</u>	.577	.353	**.953**
padding	.502	.170	<u>.590</u>	.547	.446	**.917**
JPEG (mild)	.010	.026	.988	<u>.998</u>	.989	**1**
grayscale	.066	.071	.999	.999	**.988**	<u>.997</u>
blur (heavy)	.374	.466	.830	<u>.786</u>	**.235**	.819
invert	.574	.371	**.118**	<u>.443</u>	1	.956
JPEG (heavy)	.230	.337	.877	.922	**.726**	<u>.850</u>
plain-white	.457	.695	.731	.661	**0**	<u>.069</u>
plain-black	.854	.756	<u>.004</u>	.269	**0**	.054
gaussian noise	1.32	.614	**.043**	.283	.161	<u>.051</u>

Table 1: Robustness to controlled transformations on VITON-HD test set. Best and second-best results per-row are highlighted in **bold** and <u>underline</u>. Perceptual metrics are reported for reference but operate on a different scale and range.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Robustness Analysis We apply mild transformations that preserve semantic content (small rotations ±3, horizontal flip, shift, padding, and light JPEG compression) and significant ones that destroy it (grayscale, heavy blur, inversion, heavy JPEG, plain white/black, and Gaussian noise). All tests are performed on the VITON-HD test split [1], using full-resolution (1024x768). Results are shown in Table 1. For mild transformations, our conceptual similarity metric (CSIM) outperforms all state-of-the-art metrics by considerable margins. On severe modifications, CSIM correctly produces near-zero similarity and matches or exceeds baselines in most cases, with few exceptions such as 'grayscale' and 'inversion'. We attribute these exceptions to the inherent properties of the current model's embeddings. Future work will explore alternative embeddings, including those that are domain-enriched, to extract more targeted concepts that may be capable of capturing properties related to such transformations.

Use Cases	CSIM-NI	CSIM-HI	CSIM-HS	
			(↑)	(↓)
Stripes vs Solids	0.55	0.53	**0.72**	**0.40**
V- vs Crew Neck	0.63	0.63	**0.64**	**0.43**
Florals vs Stripes	0.54	0.52	**0.71**	**0.40**
Long vs Short Sleeve	0.50	0.49	**0.59**	**0.09**
Slim vs Pleated	0.42	0.39	**0.57**	**0.12**

Stripes vs Solids

Florals vs Stripes

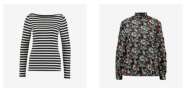

Table 2: Quantitative results (left) and qualitative example concept pairs (right).

Human Similarity Steering Example Here, we apply our trained concepts to images from the Dress Code dataset[14] and create five possible steering use cases, illustrating how humans can influence similarity scores. To increase similarity (indicated by ↑), we set the weights of the relevant concepts to 0; to reduce similarity (indicated by ↓), we retain only the relevant concepts by setting their weights to 1. For each concept, we select the 200 top-activating images (if possible) and generate 20 random pairs per image, recording the similarity. Table 2 reports the average similarity over the 20 random pairings.

CSIM-HS consistently achieves the intended behavior, demonstrating that human guidance can effectively tailor similarity scores to downstream requirements. Interestingly, we notice that for the 'V-vs Crew Neck', the similarity does not increase as dramatically as in other cases. This is most likely due to the influence of other distinctive attributes, such as patterns or designs, which dominate the similarity between garments. In such scenarios, more elaborate masking would be required to increase similarity. We leave the design of optimal steering strategies for such cases to future work.

Importantly, although the concept dictionary is learned exclusively on the VITON-HD training set, the resulting concepts transfer remarkably well to the Dress Code dataset [14] despite noticeable domain shifts in clothing style, poses,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

backgrounds, and resolution. This demonstrates that the extracted concepts are general and reusable rather than dataset-specific.

5 Conclusion

In this work, we introduced **Conceptual Similarity (CSIM)**, an explainable IQA metric designed to address the semantic limitations of traditional pixel-based and black-box perceptual methods. By decomposing deep embeddings into interpretable components, CSIM achieves superior robustness to non-semantic transformations. Uniquely, our framework enables **Human Similarity Steering**, allowing users to inject domain knowledge and tailor similarity assessments to specific downstream requirements. Future work will focus on extending this concept-based approach to other modalities, such as language for retrieval tasks, and investigating the alignment with human judgment through a user study.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021.

[2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *IJCV*, 2021.

[3] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 2004.

[4] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *PAMI*, 2020.

[6] Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity. In *ICCV*, 2025.

[7] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Neur*, 2023.

[8] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *ECCV*, 2024.

[9] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.

[10] Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. In *BMVC*, 2025.

[11] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *NeurIPS*, 2024.

[12] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. Dinov3, 2025. `https://doi.org/qpjq`.

[13] Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, et al. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. In *ICML*, 2025.

[14] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPRW*, 2022.