GEOMETRIC SIMILARITY IN VLM LOW-LEVEL VISION REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have emerged as powerful candidates for universal vision backbones, with representative architectures including autoregressive (AR) models and diffusion transformers (DiTs). Yet, adapting them efficiently for all-in-one low-level image restoration remains a challenge. Crucially, the field lacks an understanding of how VLMs organize hidden-layer representations and whether these structurally distinct paradigms share a common geometric organization for pixel-level perception. Such shared organization is a prerequisite for building highly transferable, unified restoration VLMs and adapters. In this paper, we systematically investigate representational similarity across 24 low-level tasks spanning 5 categories. We propose GeoSim, a unified four-level framework that analyzes task-conditioned representations from global similarity, local geometry, sparse feature decomposition, and topological verification perspectives. Our formulation applies to the analysis of hidden states in AR models and feature maps in DiTs across same- and cross-task/model settings. Our results reveal the organizing principles of low-level visual representations while exposing their limits in cross-task and cross-model agreement. Ultimately, GeoSim provides an interpretability lens for probing latent transferability in low-level vision and diagnosing model limitations in task- or model-specific scenarios.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated remarkable capabilities across a broad spectrum of visual understanding tasks (Zhang et al., 2024). In high-level vision, they have achieved success in image classification (Radford et al., 2021), object detection (Li et al., 2022c), semantic segmentation (Li et al., 2022a), instance segmentation (Huynh et al., 2022), and open-vocabulary recognition (Zhong et al., 2022). More recently, VLMs have also begun to extend into low-level vision, where a single model is expected to solve diverse perceptual tasks such as restoration, removal, enhancement, reconstruction, and photometric correction (Ai et al., 2024; Jiang et al., 2025a). Unlike high-level vision—which primarily maps inputs to invariant semantic categories—low-level tasks must carefully balance fine-grained pixel statistics, frequency components, and local image structures with high-level perceptual context. Consequently, modern all-in-one restoration systems increasingly rely on shared internal representations to transfer knowledge across diverse perceptual tasks. Understanding how these representations are organized is therefore becoming a fundamental problem, with important implications for model interpretability, efficient knowledge transfer, continual learning, and the principled design of scalable all-in-one architectures.

Despite rapid progress in low-level VLMs, existing studies have focused primarily on improving task performance through prompting (Potlapalli et al., 2023; Conde et al., 2024), agentic instruction (Jiang et al., 2024; Chen et al., 2024a), and multimodal representations (Luo et al., 2024). Comparatively little is known about the internal organization of the representations that enable these models to solve diverse perceptual tasks. In particular, it remains unclear whether different low-level tasks rely on shared computational mechanisms, whether they naturally organize into coherent perceptual families, or whether each task occupies an independent region of representation space. Without understanding this organization, it remains difficult to explain why knowledge transfers across some tasks but not others, how parameter sharing should be designed, or whether a single representation space can effectively support increasingly diverse low-level capabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The geometry of hidden representations provides a powerful lens for studying the internal organization of deep models. Classical representation analysis methods, including RSA (Kriegeskorte et al., 2008), SVCCA (Raghu et al., 2017), and CKA (Kornblith et al., 2019), characterize neural representations through pairwise similarity, shared linear subspaces, or relational structure. In VLMs, contrastive models such as CLIP and ALIGN learn a shared embedding space by encouraging matched image–text pairs to exhibit higher similarity than mismatched pairs (Radford et al., 2021; Jia et al., 2021). Subsequent work has shown that large multimodal models develop increasingly shared representational structures across architectures and, in some cases, across modalities (Huh et al., 2024; Jha et al., 2025). More recent layer-wise analyses further reveal that visual representations become progressively aligned with textual feature spaces (Venhoff et al., 2025), object-specific information is refined through subsets of visual tokens (Neo et al., 2025), and visual information is increasingly transferred into text-token representations in deeper layers (Lin et al., 2025).

Complementary advances in interpretability suggest that concepts, decisions, and intermediate computations are often organized as linear directions or low-dimensional manifolds in hidden space (Marks & Tegmark, 2024; Joshi et al., 2025; Gurnee et al., 2026; Ma et al., 2026). Sparse autoencoders (SAEs) further decompose hidden activations into interpretable sparse features (Huben et al., 2024; Gao et al., 2025; Pach et al., 2025; Shu et al., 2025), while recent evidence indicates that semantically equivalent visual and textual inputs can induce shared modality-invariant task representations (Luo et al., 2025a). Collectively, these studies suggest that neural representations often possess rich geometric structure. However, nearly all existing analyses focus on semantic concepts, language alignment, or simplified multimodal architectures. Whether similar geometric principles govern perceptual representations in low-level vision, and how they extend from object and modality comparisons to task-conditioned low-level transformations, remains largely unexplored. Understanding whether such tasks exhibit a coherent geometric organization is therefore important not only for interpretability but also as a foundation for predicting task transfer, designing modular architectures, identifying shared computational primitives, and scaling unified restoration models to increasingly diverse perceptual capabilities.

Motivated by these observations, this paper investigates the geometric organization of low-level visual task representations in modern VLMs. Specifically, we ask:

RQ1. Do low-level visual tasks exhibit meaningful geometric clustering in VLM hidden representations, and how does this organization evolve across network depth?

RQ2. Do low-level task representations exhibit shared and interpretable latent structures within and across task families?

RQ3. To what extent is the geometric organization of low-level tasks universal across different VLM architectures?

To answer these questions, we introduce **GeoSim**, a unified four-level framework for characterizing the geometry of low-level visual task representations. GeoSim analyzes task-conditioned representations from four complementary perspectives: global similarity structure, local geometric structure, sparse feature decomposition, and topological verification. This unified framework enables systematic comparisons across tasks, layers, architectures, and model families, providing a comprehensive view of how perceptual representations are organized within modern VLMs.

Our main contributions are summarized as follows:

- We present a systematic geometric study of **24 low-level vision tasks across multiple VLM families** and establish a reusable benchmark with unified prompt templates and evaluation protocols.
- We reveal how task geometry evolves across depth and quantify global, local, and cross-model agreement with Inter-task/Intra-task Cosine Similarity, Distance Correlation (dCor), Nearest Neighbor Graph Similarity (NNGS), Local Intrinsic Dimensionality (LID), Procrustes Alignment Residual (PAR), and topological analyses.
- We establish a comprehensive pipeline including Sparse Autoencoder (SAE)-based feature decomposition, Mutual Information (MI), and cross-model Maximum Mean Discrepancy (MMD) to distinguish task-specific and cross-task shared sparse representations.

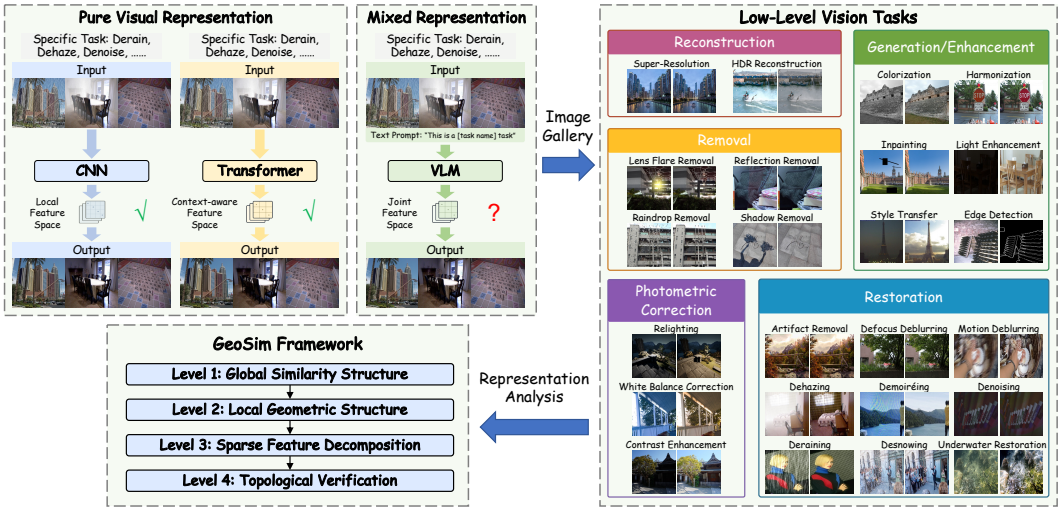

Figure 1: Overview of VLM representations for low-level tasks and the GeoSim framework.

2 Related Work

All-in-one image restoration. Image restoration has gradually evolved from task-specific networks toward unified models capable of handling diverse degradations within a single architecture (Jiang et al., 2025a). Recent work leverages VLM representations (Luo et al., 2024), visual instructions with diffusion priors (Luo et al., 2025b), and autoregressive visual token generation (Rajagopalan et al., 2026) for restoration. These advances demonstrate that increasingly diverse restoration tasks can be handled within a shared representation space, yet the internal organization of the learned representations remains largely unexplored. Beyond restoration, Li et al. (2025) used component-wise tuning to show that distortion recognition in an MLLM depends most strongly on adapting its visual encoder, with further gains from the alignment projector. This component-level evidence motivates our study of how low-level tasks are organized across layers and models.

Task transferability and task relationships. Understanding relationships among visual tasks has long been a central topic in computer vision because tasks with similar representations are expected to transfer knowledge more effectively. Taskonomy (Zamir et al., 2018) systematically quantified transfer relationships among a large collection of visual tasks, Task2Vec (Achille et al., 2019) proposed compact task embeddings for directly measuring task similarity, and later studies leveraged representation similarity to construct task taxonomies and guide multi-task learning (Dwivedi & Roig, 2019; Standley et al., 2020). Sachdeva et al. (2026) provided a comprehensive study of positive and negative transfer across visual perception tasks in VLMs. Beyond explicit transfer, recent work demonstrated that VLMs can spontaneously develop cross-modal task representations from in-context examples (Luo et al., 2025a) and perform cross-task visual in-context learning through implicit text-driven prompts (Xia et al., 2025). While these studies characterize *whether* tasks transfer knowledge, they provide limited insight into *why* such transfer occurs. In particular, how transferability is reflected in the geometric organization of hidden representations within modern VLMs remains largely unknown.

Geometric representation analysis. Geometric representation analysis seeks to understand neural networks by characterizing the structure of their hidden representations. These techniques have recently been applied to VLMs to study shared image–text embedding spaces (Radford et al., 2021), cross-model representational convergence (Huh et al., 2024), language-space interpretation of visual features (Jiang et al., 2025b), modality-invariant task representations (Luo et al., 2025a), and progressive visual–language alignment (Venhoff et al., 2025). Nevertheless, most existing studies focus on semantic concepts, language alignment, or individual geometric properties. A unified framework that systematically integrates global, local, distributional, sparse, and topological analyses for understanding *low-level perceptual task representations* remains lacking.

SAEs for model interpretability. SAEs have emerged as an effective tool for interpreting hidden representations in large neural networks (Huben et al., 2024; Gao et al., 2025; Templeton et al., 2024). Recently, SAE-based analysis has been applied to vision and vision-language models. Universal SAEs learn a single sparse concept space jointly across several vision encoders (Thasarathan et al., 2025), whereas Matryoshka SAEs optimize nested sparsity levels within CLIP representations (Zaigrajew et al., 2025). Pach et al. (2025) further demonstrated that SAEs can recover monosemantic and steerable features from VLM representations. These studies establish SAEs as an effective tool for interpreting individual concepts within VLMs. While in our work, we employ SAEs to characterize shared and task-specific sparse representations across multiple low-level vision tasks and multiple VLM architectures, providing a feature-level perspective on the geometric organization of perceptual task representations. Extended related work is provided in Appendix A.

3 Proposed Framework

3.1 Framework Overview

We introduce **GeoSim**, a unified framework for characterizing the geometric organization of low-level visual task representations in VLMs. Rather than relying on a single similarity measure, GeoSim analyzes representations from four complementary perspectives: Level 1 characterizes the global similarity structure among tasks and its evolution across depth, Level 2 examines local geometric structure, Level 3 decomposes representations into shared and task-specific sparse features, and Level 4 examines Cross-model Universality.

3.2 Level 1: Global Similarity Structure

Inter-Task and Intra-Task similarity. For an input image x, let $h_p^\ell(x) \in \mathbb{R}^d$ denote the hidden representation extracted for task p at layer ℓ. Given n sampled images, we stack their representations into $\mathbf{H}_p^\ell = [h_p^\ell(x_1); \ldots; h_p^\ell(x_n)] \in \mathbb{R}^{n \times d}$. We compute the average inter-task cosine similarity between every pair of tasks

$$S^\ell(p,q) = \frac{1}{b}\sum_{s=1}^{b}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos\left(h_p^\ell(x_i^s), h_q^\ell(x_j^s)\right), \tag{1}$$

where b denotes the number of sampling repeats. Given a single task t, we sample n pairs of distinct images. The average intra-task cosine similarity is defined as

$$S^\ell(t) = \frac{1}{b}\sum_{s=1}^{b}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos\left(h_t^\ell(x_i^s), h_t^\ell(x_j^s)\right). \tag{2}$$

Collecting all pairwise similarities produces the layer-wise task similarity matrix $\mathbf{S}^\ell = [S^\ell(p,q)]_{p,q\in\mathcal{T}}$, which summarizes the global organization of low-level task representations at layer ℓ.

Cross-model similarity. Cosine similarity characterizes relationships between tasks within a single model but does not directly compare the geometric structure of representations across architectures. We therefore additionally employ dCor (Székely et al., 2007), which measures statistical dependence between two representation matrices based on pairwise distance structure.

For the same task t, let $\mathbf{H}_t^{\ell,M_1}$ and $\mathbf{H}_t^{\ell,M_2}$ denote the representation matrices extracted from models M_1 and M_2, respectively. In cross-model expressions, ℓ refers to the selected layer ℓ_M^* of each model, which need not coincide across models.

The overall cross-model statistic is defined as the average dCor over the b sampling repeats:

$$\mathrm{dCor}^\ell(M_1, M_2) = \frac{1}{b}\sum_{s=1}^{b}\frac{\mathrm{dCov}_s}{\sqrt{\mathrm{dVar}_{1,s}\mathrm{dVar}_{2,s}}}. \tag{3}$$

Here, dCov_s and $\mathrm{dVar}_{1,s}, \mathrm{dVar}_{2,s}$ denote the distance covariance and distance variances in repeat s, computed from double-centered pairwise Euclidean distances as detailed in Appendix B.3.1.

Layer-wise evolution. Representation geometry evolves substantially across transformer layers. To identify where the global organization of low-level tasks stabilizes, we compute the Frobenius distance $D^{\ell}_{\text{full}} = \|\mathbf{S}^{\ell} - \mathbf{S}^{\ell+1}\|_F$ between consecutive task similarity matrices. A large value indicates substantial reorganization of task relationships between adjacent layers, whereas a consistently small value indicates a stable task similarity structure. We designate the final layer within this stable region as ℓ^* and use it for the subsequent analyses (Appendix C.3).

Visualization. Node positions in the task relation graph (Figure 5) are obtained by classical multi-dimensional scaling (MDS) (Torgerson, 1952) of the cross-model consensus dissimilarities derived from $S^{\ell}(p, q)$.

3.3 Level 2: Local Geometric Structure

Local intrinsic dimensionality (LID). We first characterize the local complexity of each task representation space using LID, estimated by maximum likelihood (Levina & Bickel, 2004) from the k nearest neighbors of each sample within the same task and averaged over samples and sampling repeats (Appendix B.3.2).

k-nearest-neighbor graph similarity (NNGS). To directly evaluate local neighborhood preservation across models, we construct k-nearest-neighbor graphs over corresponding samples and measure the fraction of shared neighbors. For representation matrices U and V ($\mathbf{H}_t^{\ell,M_1}, \mathbf{H}_t^{\ell,M_2}$) extracted from two models, the neighborhood similarity is

$$\text{NNGS}(U, V) = \frac{1}{b}\sum_{s=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\frac{|N_k(x_i^s; U) \cap N_k(x_i^s; V)|}{k}, \tag{4}$$

where $N_k(x_i^s; U)$ denotes the k nearest neighbors of sample i in representation space U.

Procrustes alignment residual (PAR). Neighborhood similarity evaluates local topology, but does not capture whether two representation spaces can be globally aligned after removing coordinate differences. Therefore, we further measure geometric alignment using orthogonal Procrustes analysis (Schönemann, 1966).

For task t and models M_1, M_2, we first center each representation matrix and apply principal component analysis (PCA) separately, retaining $d_{\text{PCA}} \le \min(d_{M_1}, d_{M_2}, n-1)$ dimensions. The resulting matrices are denoted as $\mathbf{H}_t^{\ell,M_1}, \mathbf{H}_t^{\ell,M_2} \in \mathbb{R}^{n \times d_{\text{PCA}}}$. The alignment residual is defined as

$$\text{PAR}(t, M_1, M_2) = \min_{Q \in \mathcal{O}(d_{\text{PCA}})} \|\mathbf{H}_t^{\ell,M_1} - \mathbf{H}_t^{\ell,M_2} Q\|_F. \tag{5}$$

A smaller PAR indicates stronger geometric agreement up to an orthogonal transformation.

3.4 Level 3: Sparse Feature Decomposition

SAE representation decomposition. At the stable layer ℓ^*, we write $z_p(x) = h_p^{\ell^*}(x) \in \mathbb{R}^d$. For each VLM, we train a single ReLU SAE with an ℓ_1 sparsity penalty on $z_p(x)$, shared across all low-level vision tasks, so that feature usage can be compared within a common learned dictionary; $a(x) \in \mathbb{R}^{D_{\text{SAE}}}$ denotes its sparse code and D_{SAE} the dictionary size (Appendix B.3.3; training details in Appendix C.4). The SAE is trained on task-balanced samples, and all feature statistics are computed on held-out samples.

Feature activation statistics. For SAE feature f, we define its activation rate for task t as

$$\rho_f(t) = \frac{1}{N_t}\sum_{x \in \mathcal{D}_t} \mathbf{1}[a_f(x) > \tau], \tag{6}$$

where $\mathcal{D}_t$ is the held-out sample set for task t, $N_t = |\mathcal{D}_t|$, $a_f(x)$ denotes the activation value of feature f, and τ is an activation threshold. The activation rate $\rho_f(t)$ measures how frequently a feature is used by a given task.

Task selectivity index (TSI). We quantify whether a feature is concentrated on a small number of tasks using TSI:

$$\text{TSI}(f) = \frac{\max_{t \in \mathcal{T}} P(f(t))}{\sum_{t \in \mathcal{T}} P(f(t))}. \tag{7}$$

A high $\mathrm{TSI}(f)$ indicates that feature f is preferentially activated by a particular task, whereas a lower value indicates broader usage across tasks. Here, $P(f(t)) = \rho_f(t)$ is the activation probability defined in Eq. (6). Features are categorized as task-specific or shared according to predefined thresholds θ_{high} and θ_{low}.

Cross-model sparse representation comparison. To evaluate whether different VLMs develop similar sparse representations, we compare feature activation-rate distributions across models. For a given task t, let $A_t^{M_1}$ and $A_t^{M_2}$ denote the empirical distributions of the activation rates $\rho_f(t)$ obtained from the SAEs of models M_1 and M_2, respectively. Each feature active on at least one held-out task contributes one scalar rate, while features inactive across all held-out tasks are excluded. We measure their distributional similarity using the MMD (Gretton et al., 2012) with a radial basis function (RBF) kernel κ:

$$\mathrm{MMD}^2(A_t^{M_1}, A_t^{M_2}) = \left\| \mathbb{E}_{A_t^{M_1}} \phi - \mathbb{E}_{A_t^{M_2}} \phi \right\|_{\mathcal{H}_\kappa}^2, \tag{8}$$

where ϕ denotes the feature mapping induced by kernel κ, $\mathbb{E}_{A_t^{M_1}} \phi$, $\mathbb{E}_{A_t^{M_2}} \phi$ are shorthand for $\mathbb{E}_{a \sim A_t^{M_1}}[\phi(a)]$, $\mathbb{E}_{a \sim A_t^{M_2}}[\phi(a)]$, with $a \in [0, 1]$ denoting a feature activation rate. We report the squared MMD between these empirical distributions. Lower values indicate more similar task-conditioned activation-rate distributions, without establishing correspondence between individual SAE features across models.

3.5 Level 4: Topological Verification

For each task t, we project the stable-layer representations $\mathbf{H}_t^{\ell^*}$ to a low-dimensional point cloud $\widetilde{\mathbf{H}}_t \subset \mathbb{R}^{d_{\mathrm{PH}}}$ with a fixed UMAP (McInnes et al., 2018) configuration ($d_{\mathrm{PH}} \ll d$), which mitigates distance concentration in the ambient space. Using persistent homology (PH), we summarize the birth and death of connected components and one-dimensional loops (homology degrees H_0 and H_1) in persistence diagrams, and compare models on the same task with the 2-Wasserstein and bottleneck distances between their diagrams. Level 4 serves as a supplementary cross-check; construction details and results are given in Appendices B.3.4 and D.2.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate 24 low-level vision tasks spanning five perceptual families, including Restoration, Removal, Generation/Enhancement, Reconstruction, and Photometric Correction. Each task is paired with a representative public dataset covering a diverse range of image degradations and transformations. Dataset sources and construction details are provided in Appendix C.

Implementation Details. We analyze unified VLMs spanning visual understanding and image generation across multiple architectural families and parameter scales (see Table 1). All experiments were conducted using eight NVIDIA RTX 6000 GPUs, each with 48 GB of memory, and eight NVIDIA B200 GPUs, each with 180 GB of memory.

4.2 Results

Our main results are organized according to the four task/model conditions, with the metrics in Appendix Table 4.

4.2.1 Condition I: Same Task, Same Model

Condition I fixes the task and model while varying the input image. As shown in Figure 2, all seven models exhibit aligned instruction-induced displacements, with mean cosine similarities of 0.80–0.96 at the selected analysis layer. Despite architectural differences, these transformations are stable across images at that layer, while their consistency can vary across depth.

Table 1: VLMs Used in Our Analysis.

Scale	Models
Base (∼7–8B)	Emu3-Chat (8B) (Wang et al., 2024; 2026), Anole-7B (Chern et al., 2024), Janus-Pro-7B (Chen et al., 2025), InternVL3.5-8B (Wang et al., 2025), Qwen3-VL-8B (Bai et al., 2025)
Large (>13B)	Qwen-Image-Edit (∼20B) (Wu et al., 2025), Emu3.5-Image (∼34B) (Cui et al., 2025)

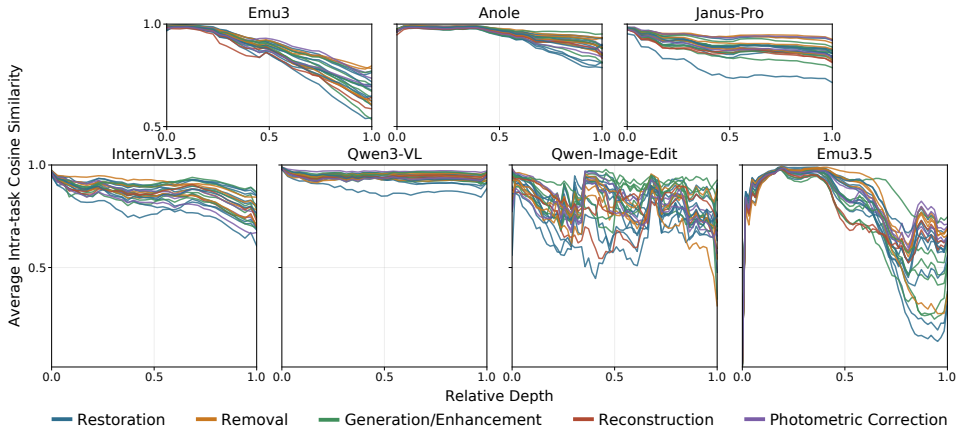

Figure 2: **Condition I:** Consistency of the instruction-induced displacement across network depth. Each panel is one model; each of the 24 curves is one task, colored by perceptual family.

Appendix Figure 9 shows that local geometric complexity is task-dependent. The 24 tasks follow a consistent ordering across models ($\rho = 0.80$), with Denoising having the lowest LID in every model. Each task exhibits a stable global direction and a characteristic local manifold complexity.

Key Takeaway I

Low-level visual tasks exhibit stable within-task geometry, with aligned instruction-induced transformations across images and consistently organized local task manifolds across models.

4.2.2 Condition II: Same Task, Different Models

Condition II compares the same task across different models to assess cross-model geometric consistency. As shown in Figure 4, the 21 model pairs exhibit strong global agreement (mean $\mathrm{dCor} = 0.77$), but weaker local consistency (mean chance-adjusted $\mathrm{NNGS} = 0.27$). The held-out, normalized PAR exceeds the mean-prediction baseline of 1.0 for 16 of 21 pairs. Overall, models share global relational geometry but not local neighborhoods or a common coordinate frame.

Level-3 SAE evidence further narrows the scope of this cross-model agreement by demonstrating that same-task cross-model activation-rate MMD is lower than the matched within-model, different-task reference for only 6 of 21 model pairs when aggregated over 24 tasks and three SAE seeds, with only Anole–Qwen3-VL, Emu3–Janus-Pro, and Emu3.5–Qwen3-VL satisfying this relation across all three seeds. Moreover, model pairs with high global agreement do not generally show lower same-task MMD than the different-task reference. Across models, task-specific features account for only 0.9–15.1% of live features, while the shared-feature fraction varies from 13.4% to 72.4%, indicating that sparse features are broadly reused across tasks within each model.

Key Takeaway II

Cross-model universality is primarily global in that models preserve within-task relational geometry despite limited local-neighborhood agreement and the absence of a shared coordinate frame. Level-3 SAE evidence shows the same boundary at the activation-rate distribution level, where 6 of 21 model pairs exhibit lower same-task MMD than the different-task reference.

4.2.3 Condition III: Different Tasks, Same Model

Condition III examines how low-level visual tasks are organized within a shared model representation. The consensus graph in Figure 5 identifies 36 reliable task relations, including 20 cross-family links. Removal and Photometric Correction tasks form the most densely connected region, with Contrast Enhancement the most connected task, while Reconstruction lies nearby and the nine Restoration tasks spread more thinly across the graph. In contrast, most Generation/Enhancement tasks (e.g., Colorization, Harmonization, and Style Transfer) are more isolated from the main cluster.

The heatmaps in Figure 12 (Appendix) show reproducible task structure across models, but the within- versus across-family gap is only 0.135 standard deviations (SD), below the typical cross-

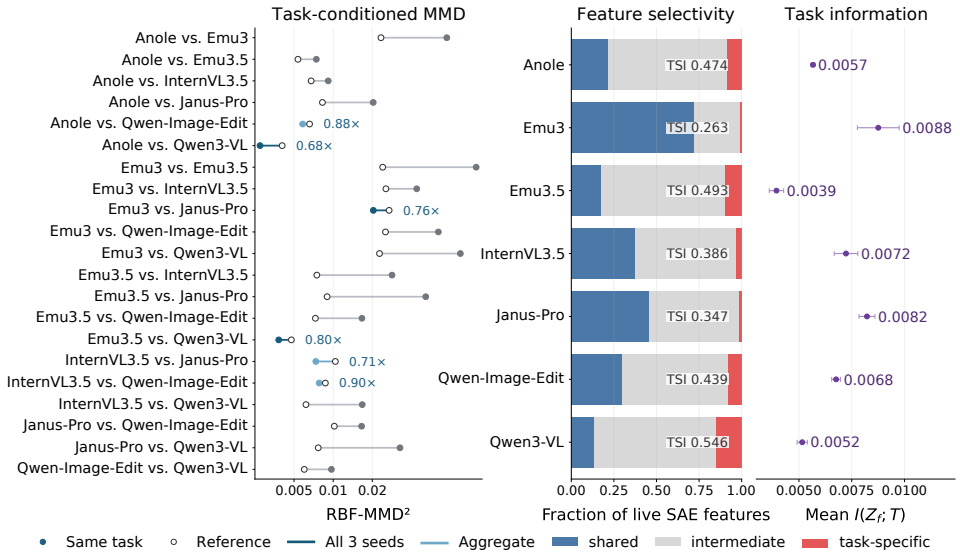

Figure 3: **Condition II:** Task-conditioned cross-model and within-model reference MMD (left), per-model TSI composition (center), and feature–task MI (right).

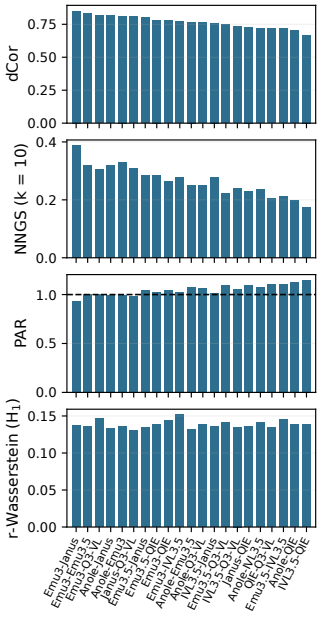

Figure 4: **Condition II:** All 21 model pairs, averaged over 24 tasks and ordered by dCor.

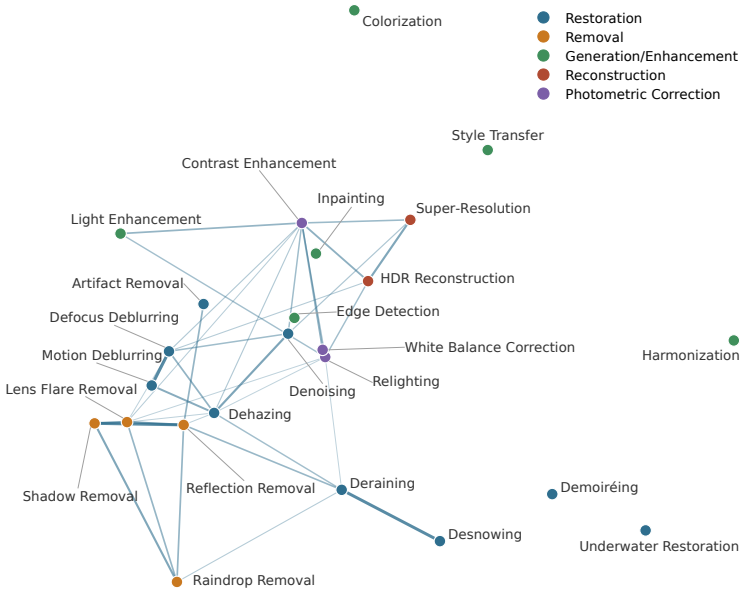

Figure 5: **Condition III:** Consensus relation graph over the 24 tasks. Node positions are visualized with MDS; edge width increases with consensus similarity.

model disagreement of 0.71-SD. Thus, task geometry is better explained by shared visual mechanisms than by broad perceptual families.

Key Takeaway III

Cross-task geometry is meaningful but mechanism-centric, with tasks sharing degradation or restoration operations forming the strongest neighborhoods, whereas the predefined perceptual families only weakly capture the representation structure.

4.2.4 Condition IV: Different Tasks, Different Models

Condition IV explores whether feature activation-rate distributions are similar across tasks and models through an exhaustive Level-3 MMD sweep. In Figure 6, the left panel summarizes reciprocal

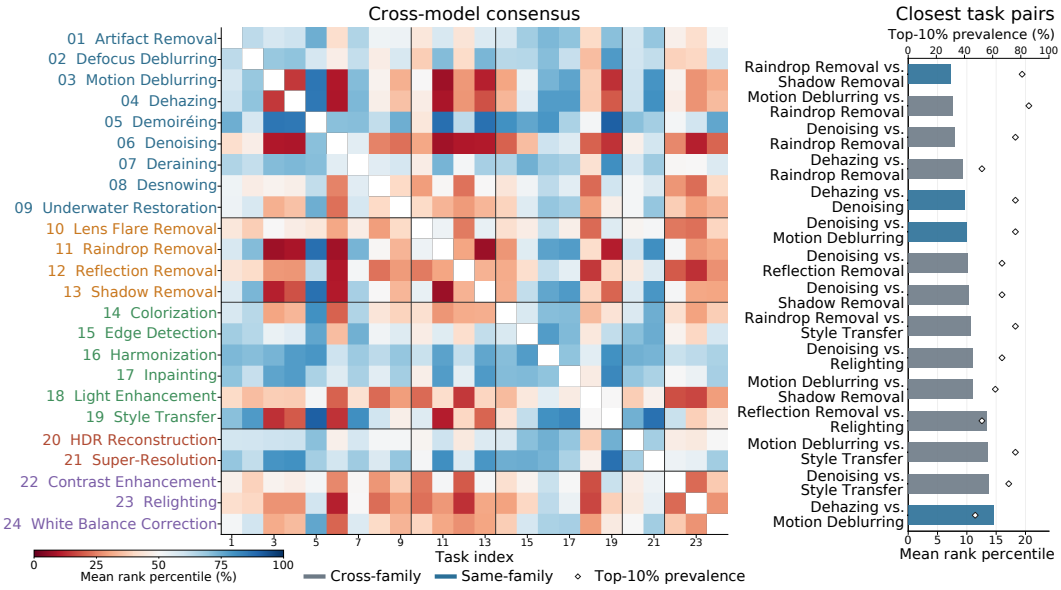

Figure 6: **Condition IV:** Cross-task, cross-model activation-rate distribution similarity from the Level-3 sweep. Results are aggregated using MMD across seven models and 24 tasks.

percentile-rank agreement across 21 model pairs, while the right panel presents the 15 most consistently close task pairs using mean within-model-pair percentile and top-decile prevalence. Since raw MMD values are not comparable across SAE model pairs, all results are rank-normalized within each model pair before aggregation. Same-family task pairs also rank 3.1 percentile points worse than cross-family pairs on average, with the same trend in 16 of 21 model pairs.

Key Takeaway IV

Despite the difficulty of cross-task, cross-model comparison, a small set of specific task/model combinations still exhibits consistently low activation-rate MMD under the joint screening criteria. These similarities are pair-specific, and same-family task pairs are not closer on average.

4.3 Discussion

Our results reveal meaningful geometric organization of low-level visual tasks (RQ1). Within each model, task instructions induce consistent global displacement directions across images, while LID follows a similar task-dependent ordering across architectures. This organization is also interpretable and partly shared at the feature level (RQ2): sparse features are broadly reused across tasks within each model, and the inter-task geometry is better explained by shared degradation and transformation mechanisms than by the predefined perceptual families. Overall, global similarity and local geometry provide complementary views of a mechanism-centric task structure. Cross-model universality is substantial but limited (RQ3). Different VLMs preserve similar global distance relationships for the same task, but exhibit weaker agreement in local neighborhoods and coordinate alignment. This suggests that universality emerges mainly through shared relational organization rather than pointwise feature correspondence. Sparse feature analysis may reveal recurring visual primitives, but feature-level sharing should be established through explicit cross-model validation.

5 Conclusion

In this study, we present a comprehensive geometric analysis of how VLMs represent low-level visual tasks and the relations among cross-task representations. We propose GeoSim, a four-level analysis framework comprising global similarity, local geometry, sparse feature decomposition, and topological verification. This work establishes a benchmark for future VLM research on representation interpretability, transfer learning, and representation alignment across low-level tasks. More importantly, our analysis provides actionable insights into how low-level task representations are organized in VLMs, offering guidance for the development of VLMs in all-in-one image restoration.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic data, develop the conceptual framework, formulate mathematical claims, or implement methods, and the remaining required-disclosure tasks are not applicable to this work. We used generative AI tools to polish and edit parts of the manuscript, check and verify citations, review code, and create or modify figures. We have reviewed all AI-assisted work; in particular, every AI-assisted description of prior work was checked by us against the original paper, and every AI-suggested methodological or textual change was evaluated by the authors before adoption. We take full responsibility for the final text, claims, and artifacts in this work.

REPRODUCIBILITY STATEMENT

Section 3 defines the analysis metrics. Appendix C describes dataset sources, feature extraction, stable-layer selection, and SAE training. Table 4 maps the evaluation conditions to the reported analysis levels and metrics. We will publicly release all code and data used in this study upon publication, including experimental configurations, sampling manifests, and analysis scripts.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2018.

Abdullah Abuolaim and Michael S. Brown. Defocus deblurring using dual-pixel data. In *Proceedings of the European Conference on Computer Vision*, pp. 111–126, 2020.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6430–6439, 2019.

Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S. Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1535–1544, 2019.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1122–1131, 2017.

Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25432–25444, 2024.

Cosmin Ancuti, Codruta O. Ancuti, and Christophe De Vleeschouwer. D-HAZY: A dataset to evaluate quantitatively dehazing algorithms. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 2226–2230, 2016.

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *Advances in Neural Information Processing Systems*, volume 37, pp. 84298–84328, 2024.

Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. RestoreAgent: Autonomous image restoration agent via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110643–110666, 2024a.

Tianwei Chen, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Hajime Nagahara. Learning more may not be better: Knowledge transferability in vision-and-language tasks. *Journal of Imaging*, 10(12):300, 2024b.

Xiangyu Chen, Zhengwen Zhang, Jimmy S. Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from SDRTV to HDRTV. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4500–4509, 2021.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. ANOLE: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have L_p-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision*, pp. 1–21, 2024.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8391–8400, 2020.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yihang Luo, and Chen Change Loy. Flare7K++: Mixing synthetic and real datasets for nighttime flare removal and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7041–7055, 2024.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12387–12396, 2019.

Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020a.

Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. AIM 2020: Scene relighting and illumination estimation challenge. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 499–518, 2020b.

Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1715–1723, 2017.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *arXiv preprint arXiv:2601.04480*, 2026.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024.

Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7031, 2022.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45963–45987, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11892–11911, 2025a.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025b.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. AutoDIR: Automatic all-in-one image restoration with latent diffusion. In *Proceedings of the European Conference on Computer Vision*, pp. 340–359, 2024.

Yeying Jin, Xin Li, Jiadong Wang, Yan Zhang, and Malu Zhang. Raindrop clarity: A dual-focused dataset for day and night raindrop removal. In *Proceedings of the European Conference on Computer Vision*, pp. 1–17, 2024.

Abhinav Joshi, Divyanshu Bhatt, and Ashutosh Modi. Geometry of decision making in language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 144139–144196, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Alexey Kravets, Da Li, Chuan Li, Da Chen, and Vinay P. Namboodiri. Interpretability transfer from language to vision via sparse autoencoders. In *Proceedings of the International Conference on Machine Learning*, 2026.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A. Sakla, and Kowshik Thopalli. Interpretable and steerable concept bottleneck sparse autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2918–2927, 2026.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 33(4): 149:1–149:11, 2014.

Sergio Lanza, Jae Hee Lee, and Stefan Wermter. Extraction and analysis of multimodal concepts in vision language models through sparse autoencoders. In *Proceedings of the International Conference on Artificial Neural Networks*, pp. 312–323, 2026.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022a.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, 2022b.

Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022c.

Yuan Li, Zitang Sun, Yen-Ju Chen, and Shin'ya Nishida. Investigate the low-level visual perception in vision-language based image quality assessment. *arXiv preprint arXiv:2512.09573*, 2025.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 24444–24469, 2025.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025.

Jiaying Liu, Dong Liu, Wenhan Yang, Sifeng Xia, Xiaoshuai Zhang, and Yuanying Dai. A comprehensive benchmark for single image compression artifact reduction. *IEEE Transactions on Image Processing*, 29:7845–7860, 2020.

Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. DesnowNet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41157–41176. PMLR, 2025a.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12764–12777, 2025b.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Yanbiao Ma, Fei Luo, Linfeng Zhang, Chuangxin Zhao, Mingxuan Wang, Yinan Wu, Zhe Qian, Yang Lu, Long Chen, Zhao Cao, Xiaoshuai Hao, Ji-Rong Wen, and Jungong Han. Reasoning emerges from constrained inference manifolds in large language models. *arXiv preprint arXiv:2605.08142*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95706–95742, 2025.

Arghya Pal and Vineeth N. Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2189–2198, 2019.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71275–71293, 2023.

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 International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sudarshan Rajagopalan, Kartik Narayan, and Vishal M. Patel. RestoreVAR: Visual autoregressive generation for all-in-one image restoration. In *International Conference on Learning Representations*, 2026.

Yusung Ro, Jaehyun Choi, and Junmo Kim. Beyond semantics: Disentangling information scope in sparse autoencoders for CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3252–3261, 2026.

Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

Bhuvan Sachdeva, Karan Uppal, Abhinav Java, and Vineeth N. Balasubramanian. Understanding task transfer in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28754–28763, 2026.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1690–1712, 2025.

Huizhen Shu, Xuying Li, Hongxu Lin, Wenjie Sun, and Hui Li. Steering vision-language models with joint sparse autoencoders. *arXiv preprint arXiv:2606.25657*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision*, pp. 746–760, 2012.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132. PMLR, 2020.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59304–59325. PMLR, 2025.

Anthony Tiong, Junqi Zhao, Boyang Li, Junnan Li, Steven Hoi, and Caiming Xiong. What are we measuring when we evaluate large vision-language models? an analysis of latent factors and biases. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3427–3454, 2024.

Warren S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2363, 2022.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal LLMs. In *Explainable AI for Computer Vision Workshop at the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3942–3950, 2017.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1788–1797, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei

Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650(8101):327–333, 2026.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep Retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference*, 2018.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Shao-Jun Xia, Huixin Zhang, and Zhengzhong Tu. T2T-VICL: Cross-task visual in-context learning via implicit text-driven VLMs. *arXiv preprint arXiv:2511.16107*, 2025.

Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *Proceedings of the European Conference on Computer Vision*, pp. 646–662, 2022.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73918–73956. PMLR, 2025.

Amir R Zamir, Alexander Sax, William B Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5825–5835, 2023.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5122–5130, 2017.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.