The Geometry of Material Representation Across Vision, Language, and Action Models

Gene Tangtartharakul & Katherine R. Storrs

Waipapa Taumata Rau University of Auckland

COMPUTATIONAL PERCEPTION LAB

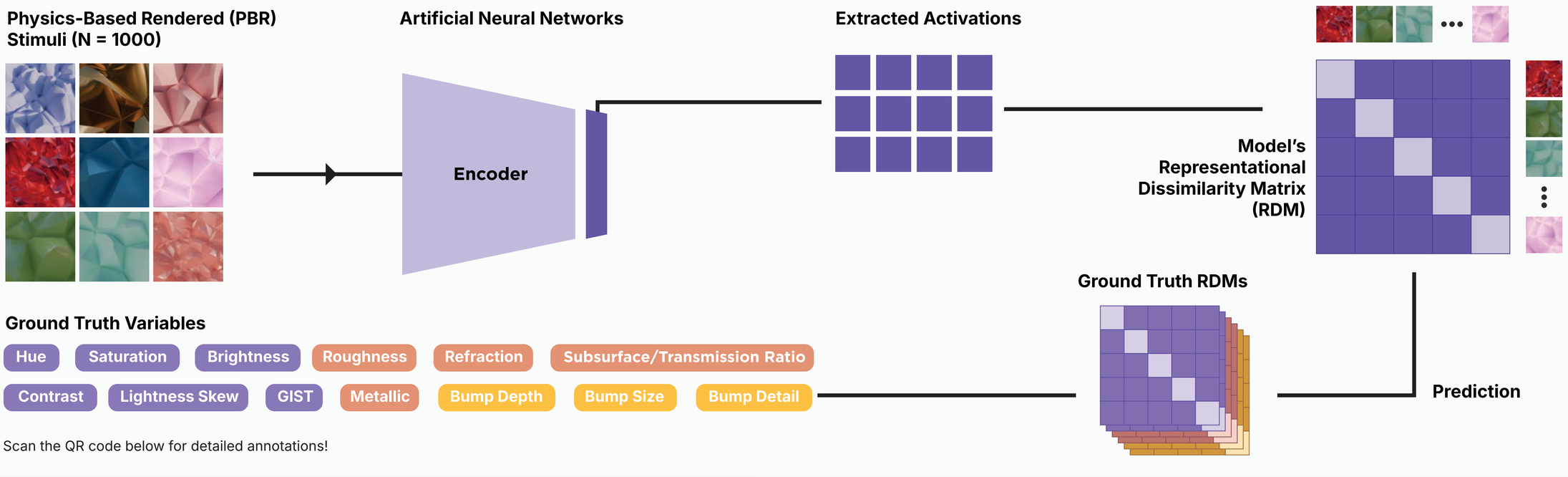
Background

Humans intuitively cluster complex physical material space into meaningful categories, drawing on multimodal knowledge like visual features, semantic knowledge, and affordances [1-2]. How different factors shape the way humans represent material remains poorly understood.

Work on human–ANN alignment of material representations is limited [3-5]. Studies largely relied on photographic or synthetic material images, which raise problems: 1) a lack of ground-truth parameters and 2) the conflation of material appearance with object identity.

Addressing this gap and leveraging the recent emergence of large-scale multimodal ANNs, we examine the representational geometry of diverse ANNs and the visual properties that shape it.

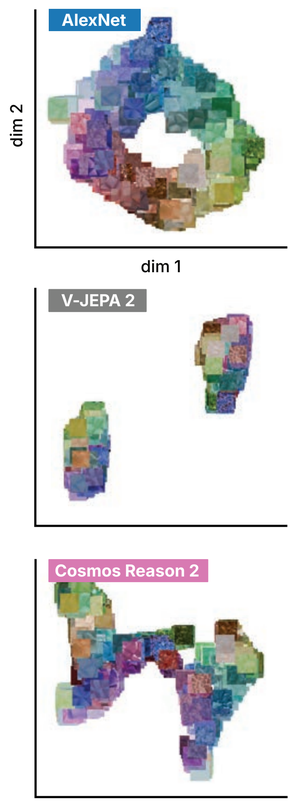
Methods



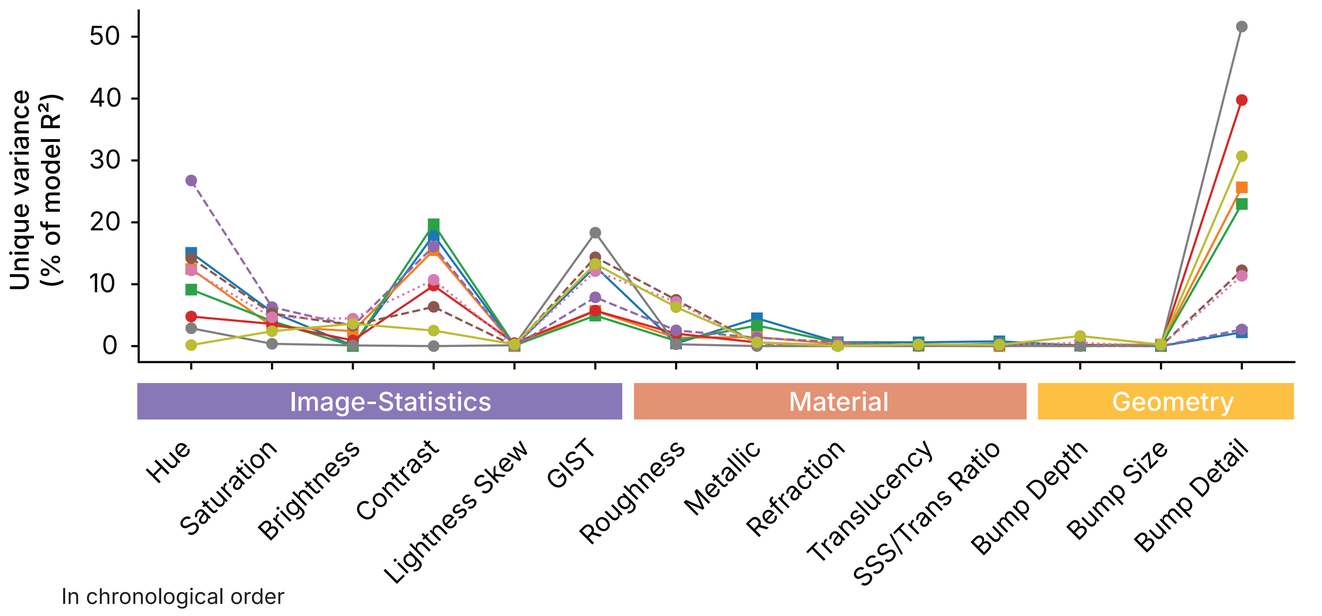
Scan the QR code below for detailed annotations!

Results

1 Geometry varies across models



2 The representation captures image statistics and surface geometry



In chronological order

- [alexnet] R^2 = 0.41
- [resnet50] R^2 = 0.28
- [cornet-s] R^2 = 0.40
- [vit_l_16] R^2 = 0.33
- [clip_vit_b_16] R^2 = 0.29
- [vjepa2_vitl] R^2 = 0.56
- [Qwen3_VL_2B_Instruct] R^2 = 0.42
- [dinov3_vitl16] R^2 = 0.33
- [nvidia_cosmos_reason2] R^2 = 0.36

- Transformer
- Convolutional
- Vision
- Vision-language
- Vision-language-action*

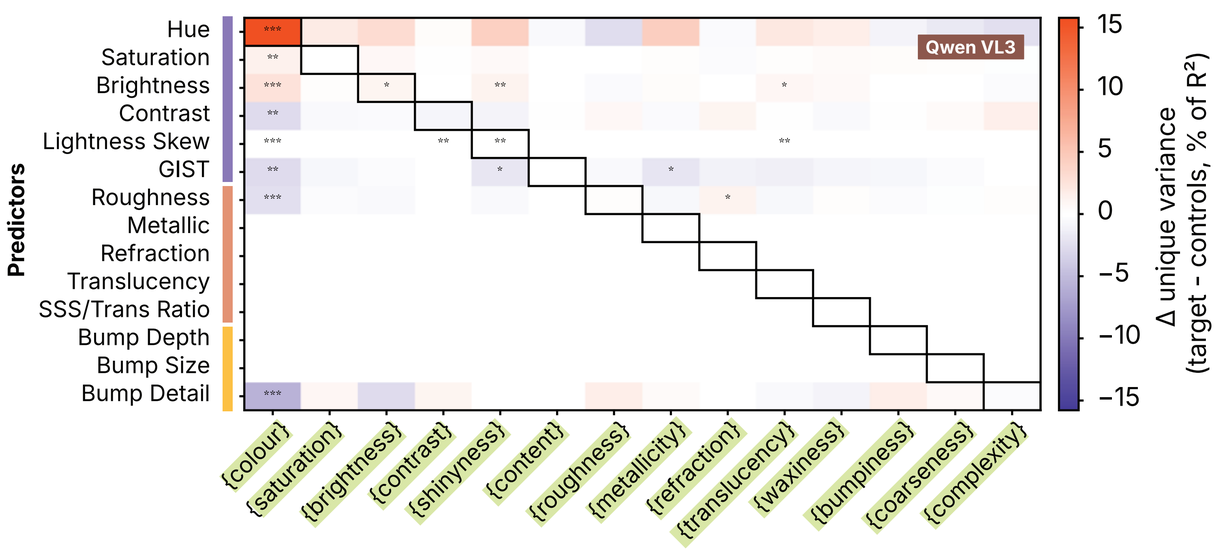
*Nvidia-Cosmos-Reason2 is a VLM fine-tuned on action reasoning

3 Text prompting can steer a model's representation

Target Prompt: Describe the {target_word} of this surface e.g., *["colour"]*.

Control Prompt: Describe the {control_word} of this surface.

control_word ∈ *["bread", "fortune", "obligation", "injury", "freedom", ..., "description"]*



1 Two-dimensional UMAP derived from selected model representational similarity matrices (RDMs).

2 Unique variance explained by ground-truth image-level, material, and surface-geometry predictors across model RDMs. Computed as the difference in adjusted R^2 between full and leave-one-out ordinary least squares models.

3 Heatmap displays changes in unique variance explained per predictor (y), induced by each targeting prompt (x). Significance is relative to the control distribution (n = 20; two-sided; t-dist; FDR-corrected).

Discussion

ANNs represent material images similarly, with their geometry largely captured by image statistics and surface geometry. In multimodal models, some representational axes can be influenced by task-relevant prompts, revealing inference-time dynamics of models' representation.

Our next steps

1) collect human perceptual ratings on the stimuli, 2) examine models trained natively on action tasks, and 3) investigate how multimodal models' representations unfold across layers and at test time (e.g. chain-of-thought and output generation).

References & more