Probing the Link Between Vision and Language in Material Perception

Chenxi Liao[1*], Masataka Sawayama[3], and Bei Xiao[2]

[1]American University, Department of Neuroscience, Washington DC, 20016, USA

[2]American University, Department of Computer Science, Washington, DC, 20016, USA

[3]The University of Tokyo, Graduate School of Information Science and Technology, Tokyo, 113-0033, Japan

[*]cl6070a@american.edu

Abstract

Materials are the building blocks of our surroundings. Material perception enables us to create a vivid mental representation of our environment, fostering the appreciation of the qualities and aesthetics of things around us and shaping our decisions on how to interact with them. We can visually discriminate and recognize materials and infer their properties, and previous studies have identified diagnostic image features related to perceived material qualities. Meanwhile, language reveals our subjective understanding of visual input and allows us to communicate relevant information about the material. To what extent do words encapsulate the visual material perception remains elusive. Here, we used deep generative networks to create an expandable image space to systematically create and sample stimuli of familiar and unfamiliar materials. We compared the representations of materials from two cognitive tasks: visual material similarity judgments and verbal descriptions. We observed a moderate correlation between vision and language within individuals, but language alone cannot fully capture the nuances of material appearance. We further examined the latent code of the generative model and found that image-based representation only exhibited a weak correlation with human visual judgments. Joining image- and semantic-level representations substantially improved the prediction of human perception. Our results imply that material perception involves the semantic understanding of scenes to resolve the ambiguity of the visual information and beyond merely relying on image features. This work illustrates the need to consider the vision-language relationship in building a comprehensive model for material perception.

Introduction

We often describe what we see with words. Language reveals how we interpret and communicate our sensory experiences and provides critical information about our mental representation of the environment[1]. The interaction between language and perception has long been debated, mainly in visual cognition, such as color categorization[2–4] and scene interpretation[5,6]. Jointly modeling visual and natural language features expands the capability of artificial intelligence systems (e.g., image-classification[7,8], image-retrieval[9,10], and text-to-image generation[11,12]) and provides valuable tools for investigating the neural correlates of object and scene recognition[13,14]. Little is known about how and what aspects we communicate about materials,

which are the building blocks of objects and the environment. Material perception facilitates us to form a vivid and rich representation of the external world, which in turn guides our interaction with it. Although we can visually recognize and discriminate a broad range of materials, we might find it challenging to precisely and effectively describe their appearances and properties with words. To what extent do words encapsulate the richness of visual material perception? What are the communicable dimensions of materials?

Based on visual input, we can often distinguish materials, and infer their diverse optical properties (e.g., surface glossiness[15–18], translucency[19–27] or transparency[28]), surface properties (e.g., roughness[29]), mechanical properties (e.g., softness[30], stiffness[31]) and states (e.g., freshness[32], wetness[33]). Previous works actively examined how visual estimates of material attributes are related to the statistical image features[34], as well as seeking to probe the neural representation of material perception in cortical areas of the ventral visual pathway[35–38]. Along with visual discrimination, verbalizing what we see reflects, to a certain degree, how we process and organize visual information into semantic-level representation. Verbal description could serve as an interpretable representation that encodes the salient features of material qualities. While a plethora of works scrutinized the visual estimation of specific material properties related to physics[16,34,39,40], few studies shined the light on more subjective material perception from both visual judgment and language expression. With a large-scale image dataset of materials, Schmidt et al. (2023)[41] used visual triplet similarity judgments from crowd-sourcing to distill a representational space, which was later annotated by humans to find conceptual and perceptual dimensions of materials. Cavdan et al. (2023)[42] studied the structure of the representational space of perceptual softness triggered by material name with a cross-group analysis and suggested that verbally activated softness representation correlates with that derived from vision[30]. However, participants in these studies were often limited to judging materials based on predetermined categories and attributes without being given the opportunity to express their personal semantic interpretations. Further, previous works typically focused on the group-level analysis and downplayed the potential individual variances.

To definitively assess the link between vision and language in material perception, it is crucial to measure visual judgment and verbal description within the individual participants, as well as allow them to freely articulate their unique visual experiences. Such verbal reports also serve as an accessible information channel for individuals' understanding of materials. For example, when looking at the photograph of a chunk of tofu (Asian food) under a particular lighting, different participants might describe it differently (Figure 1): individuals identifying the object as tofu may describe it as soft, whereas those recognizing it as plastic may describe it as hard. How do we systematically design stimuli that couple with object-level realism and also elicit semantic-level ambiguity?

Here, we developed an effective approach to create an extensive range of plausible visual appearances of familiar and novel materials (see Figure 2). We use an unsupervised image synthesis model, StyleGAN2-ADA[43], to generate images of diverse materials based on the learning of real-world photos. As a result, the model parameterizes the statistical structures of material appearances [27] and facilitates linear interpolation between image data points, allowing us to morph between different material categories (e.g., morphing between a soap to a rock results in an ambiguous translucent object shown in Figure 2C).

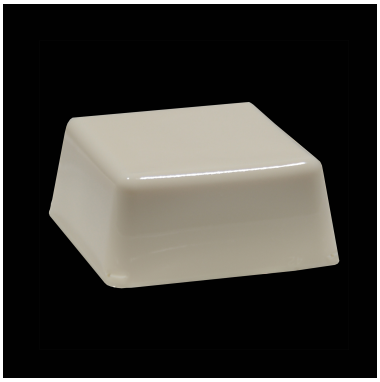

Fig. 1. Given a photograph of an object, different participants might perceive its material properties differently. Those who recognize it as tofu might expect it to be soft, while those who see it as plastic might perceive it as hard. Image features alone might not predict individual perception.

This approach enables us to continuously vary the multidimensional structural features of materials (e.g., the combination of shape and color variation) and build an expanded Space of Morphable Material Appearance. The morphed materials resemble the visual characteristics of both original materials (e.g., soap and rock), potentially resulting in the ambiguity of perceived material identity. This offers an opportunity to investigate the influence of semantic-level interpretation on material perception. Furthermore, our models' learned multi-scale latent space allows us to construct image-feature-based representations and compare them to human visual judgments.

We measured material perception with two behavior tasks involving vision and language within individuals: Multiple Arrangement and Verbal Description (Figure 3). Stimuli were systematically sampled from the Space of Morphable Material Appearance (Figure 2D and E). In the Multiple Arrangement task, participants arranged materials based on visual similarities[44]. For the verbal description task, the participants described the same images with texts. With the recent advancements in Large Language Models (LLMs), it is now possible to create a representation based on verbal reports provided by the participants. We discovered a moderate vision-language correlation within individual participants by quantitatively comparing the behavioral representations derived from two tasks. We observed that material naming, colorfulness, and softness could be critical in explaining the participants' visual similarity judgments. We also analyzed how the representations based on image features expressed through the model's latent code relate to human visual judgments. Our findings imply that material perception extends beyond the analysis of image features in a feed-forward manner, encompassing also the semantic interpretation of visual scenes, likely shaped by the individual's prior experience and knowledge.

Results

Space of Morphable Material Appearance

Employing the unsupervised learning model, StyleGAN2-ADA, we generated images of diverse materials with perpetually convincing quality by training on real-world photos (Figure 2). With its multi-scale generative network (G) and scale-dependent latent space (W), the model learns the statistical regularity of the images at multiple spatial scales, spontaneously disentangling

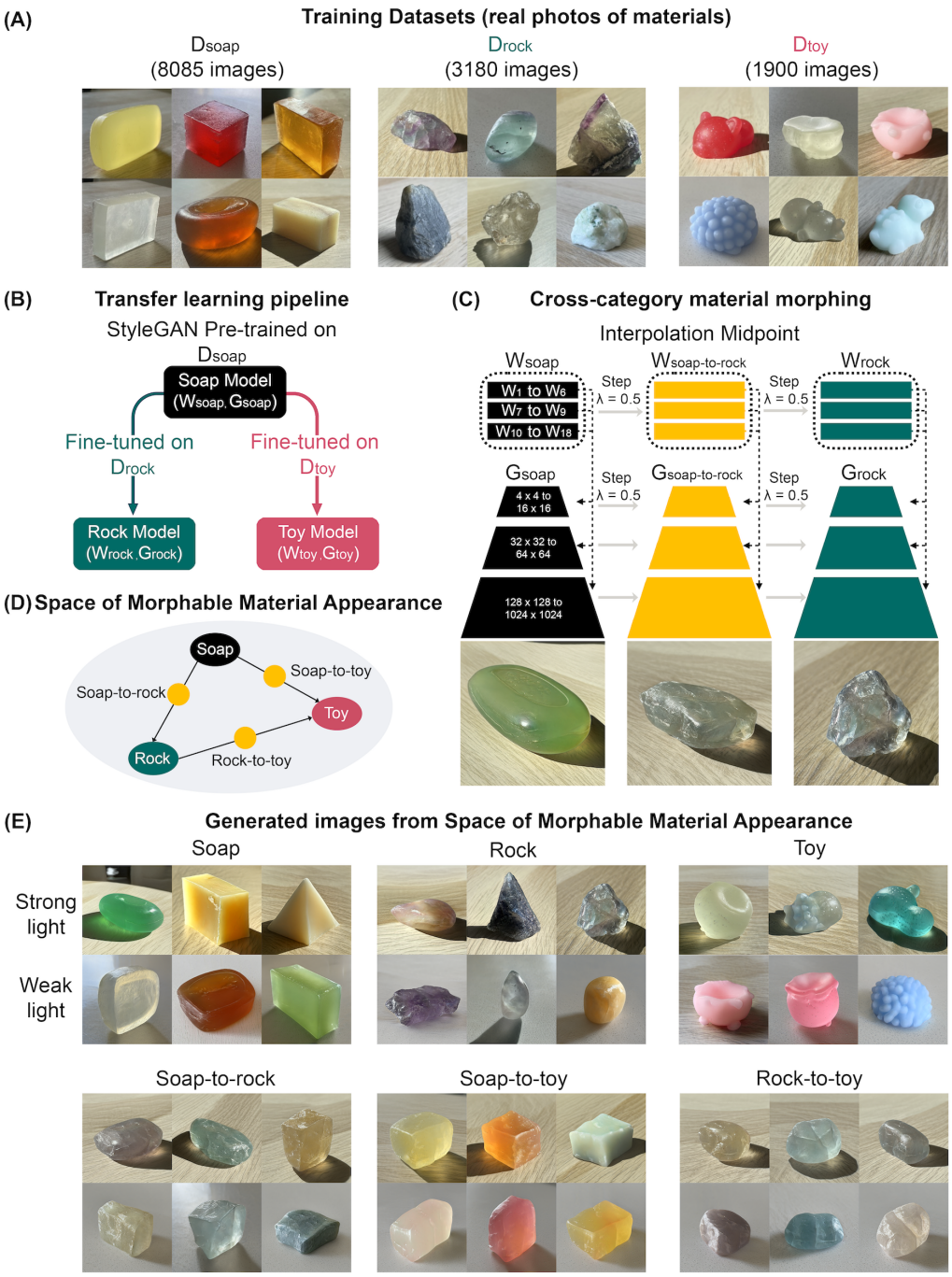

Fig. 2. Overview of the synthesis pipeline for morphable material appearances. (A) Training datasets. (B) Transfer learning pipeline. Upon training, we obtained models to generate images from three material classes. We can generate images of a desired material (e.g., soaps) by injecting the latent codes (e.g., $w_{soap} \in W_{soap}$) into the corresponding material generator (e.g., G_{soap}). (C) Illustration of cross-category material morphing. By linearly interpolating between a soap and a rock, we obtain a morphed material, "soap-to-rock," produced from its latent code $w_{soap-to-rock}$ and generator $G_{soap-to-rock}$. (D) Illustration of the Space of Morphable Material Appearance. (E) Examples of generated images from the Space of Morphable Material Appearance. These images are a subset of stimuli used in our psychophysical experiments, covering two major lighting conditions (i.e., strong and weak lighting).

semantically meaningful visual attributes, such as the object's shape, texture, and body color[27]. Here, we built our own image datasets that include three materials: soaps (D_{soap}), rocks (D_{rock}), and squishy toys (D_{toy}) (Figure 2A). We fine-tuned the StyleGAN pre-trained on the large soap dataset D_{soap} on the smaller datasets D_{rock} and D_{toy} (Figure 2B). With a short training time, the Soap Model (W_{soap}, G_{soap}) turned into Rock (W_{rock}, G_{rock}) and Toy Models (W_{toy}, G_{toy}) and can synthesize images of realistic and diverse rocks/crystals and squishy toys, under the broad variation of three-dimensional (3D) shapes, colors, textures, and lighting environments (Figure 2E Top Row). The effectiveness of transfer learning also suggests that the different categories of materials have common visual characteristics, such as color variation, specular highlight, and surface geometry; thus, learning features from one material benefits learning new materials.

We can produce novel material appearances without additional training, by morphing between existing learned materials. Given the images of a pair of source and target materials, we can linearly interpolate between their layer-wise latent codes (e.g., w_{soap} and w_{rock}) while interpolating all convolution layers' weight parameters of the corresponding material generators (e.g., G_{soap} and G_{rock}) (see Method). At a given step size, we can synthesize the image of a morphed material with the interpolated latent code (e.g., $w_{soap-to-rock}$) and generator (e.g., $G_{soap-to-rock}$). Figure 2C illustrates the method of creating a morphed material between soap and rock.

Combining transfer learning and model morphing, we constructed an expandable Space of Morphable Material Appearance, from which we can systematically sample and create existing and novel material appearances with object-level realism (Figure 2D). In this study, we focused on the material appearances at the morphing midpoints (i.e., step $\lambda = 0.5$). With this technique, we generated morphed materials, soap-to-rock (midpoint from soap to rock), soap-to-toy (midpoint from soap to squishy toy), and rock-to-toy (midpoint from rock to squishy toy) (Figure 2E Bottom row). We sampled 72 images from the Space of Morphable Material Appearance as stimuli for both of our behavioral experiments (see Method).

Visual Material Judgment and Verbal Description are Moderately Correlated within Individuals

Using the above-mentioned stimuli, we measured material perception with Multiple Arrangement and Verbal Description tasks. In the Multiple Arrangement task, participants were instructed to place the images within the circled region based on the "similarity of material properties" (Figure 3A). The task prompted the consideration of various aspects of the materials, allowing for the capturing of a multidimensional representation of how visual material discrimination is processed. During the Verbal Description task, the same group of participants provided unrestricted descriptions of the material with texts covering five aspects: material name, color, optical properties, mechanical properties, and surface texture. These aspects have been found useful in characterizing the mental representations of materials[41].

We constructed the Representational Dissimilarity Matrices (RDM) from each participant's behavioral results for both tasks. A Vision RDM is created based on the on-screen Euclidean distances of pairwise comparisons of material similarity[44] from the Multiple Arrangement. Meanwhile, we also built a Text RDM by encoding the images' text descriptions provided by the participant into an embedding space with a large pre-trained LLM (see Methods). We tested four publicly accessible LLMs, CLIP[7], BERT[45], Sentence-BERT[46], and GPT-2[47], whose embedding spaces were shown to capture the semantic similarity of

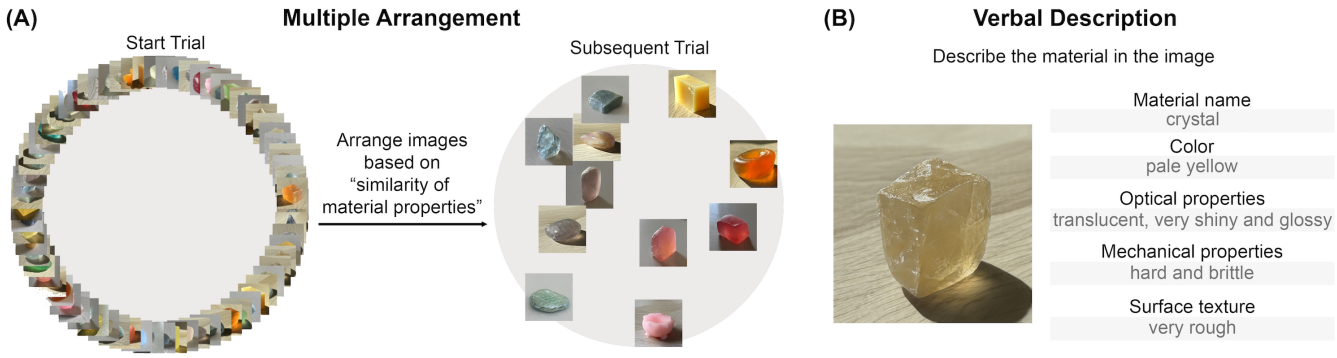

Fig. 3. Illustrations of psychophysical experiment interface. (A) The Multiple Arrangement task. Participants (N =16) arranged images within a circle based on their judgment of the visual similarity of material properties. In the first trial, participants were presented with all 72 images of materials. In each subsequent trial, a subset of images was iteratively presented based on an adaptive sampling algorithm [44]. (B) The Verbal Description task. With free-form text input, participants were asked to describe the material shown in the image from five aspects: material name, color, optical properties, mechanical properties, and surface texture. The gray font texts are example responses.

textual information. The primary analysis in this paper is conducted using the CLIP embedding unless otherwise noted.

Across individuals, we found a moderate correlation between the RDMs of the two tasks within each participant. Figure 4A displays the RDMs of three participants. The comprehensive statistical results encompassing all participants are provided in Figure 6. While the participants used different numbers of unique words (Figure 6A, mean = 128 unique words, max = 288 words, min = 37 words), we found that all of the participants' verbal responses exhibited a significant correlation (all $p < 0.001$, FDR-corrected) with their own multiple arrangement behavior, signifying the presence of inherent cross-task consistency within an individual (Figure 6B blue bars). These moderate correlations reflect that participants' own Vision and Text RDMs share similarities in their overall structures, while also underpinning differences in their local patterns. We observed a stronger correlation when comparing the group average Vision and Text RDMs (Spearman's correlation $r_s = 0.75, p < 0.001$) (rightmost column in Figure 4A). By applying classical multidimensional scaling (MDS) on the group average RDMs, we found that Vision and Text embeddings exhibit similar organizations, forming three major clusters: squishy-like (squishy toys, top left cluster in MDS), soap-like (soap and soap-to-toy, bottom left cluster in MDS), and rock-like (rock, rock-to-toy, and soap-to-rock, bottom right cluster in MDS).

Despite overall similarities in general patterns among participants' behavioral results, the Vision RDMs across participants reveal substantial individual differences in finer material discrimination (Figure 4A). We assessed the inter-participant consistency of the behavioral tasks with the leave-one-out test by iteratively correlating one participant's data with the group average of the rest of the participants. We found that Vision RDMs (mean Spearman's correlation $r_s = 0.41$) show higher variance than the Text RDMs (mean Spearman's correlation $r_s = 0.57$). Compared with the ways of articulating materials with words, participants tended to be more diverse in how they visually judge material similarities. The varied degree of correlation between vision- and language-based representations across participants underscores the complex nature of individual differences in material reasoning.

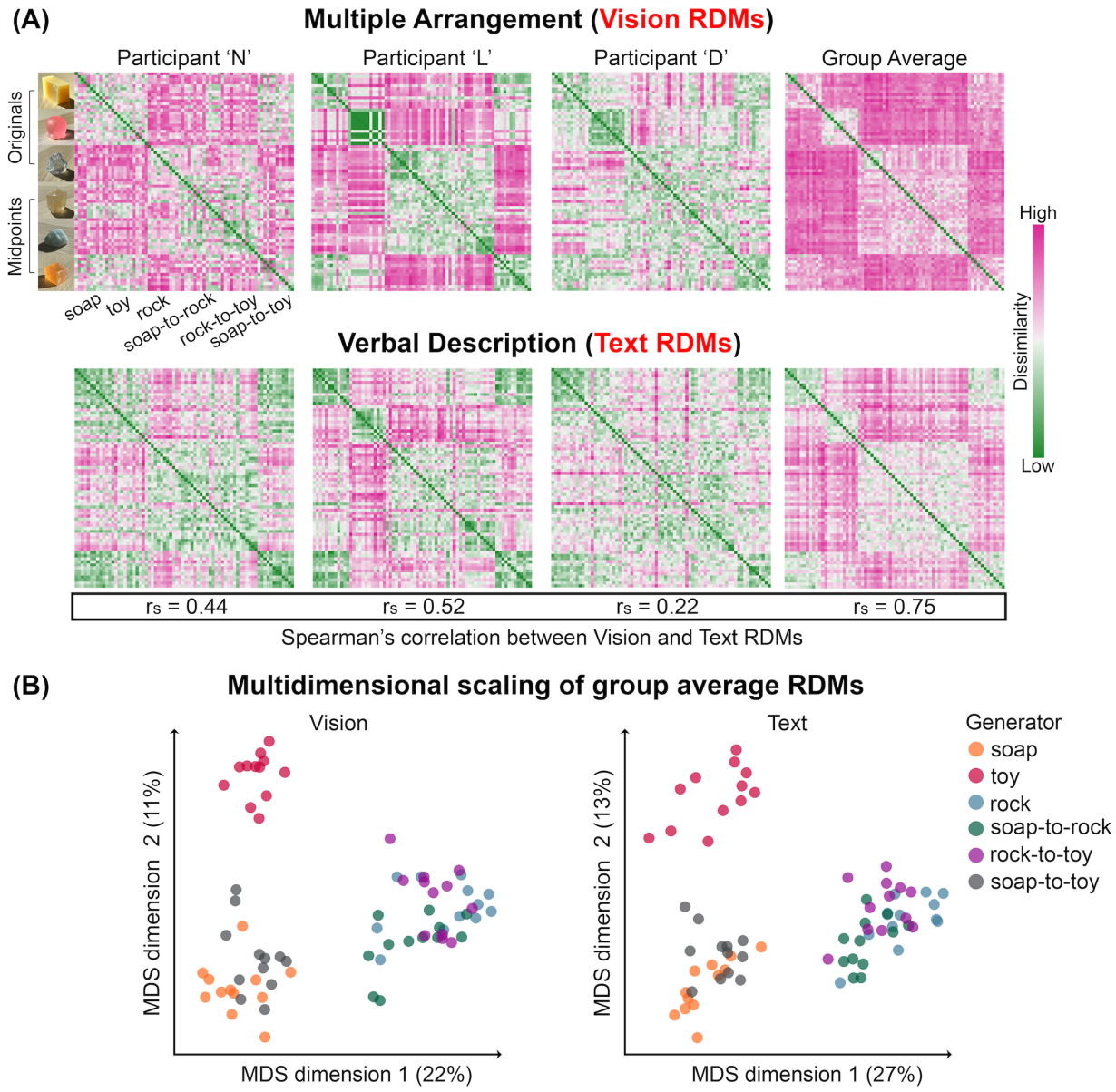

Fig. 4. Vision-based similarity judgment and verbal description of materials are moderately correlated. (A) Representational Dissimilarity Matrices (RDMs) of vision-based material similarity judgment via Multiple Arrangement (Vision RDMs) and Verbal Description (Text RDMs). Top: Vision RDMs. Bottom: Text RDMs. From left to right: RDMs for three participants and the group average RDM across all participants. In each RDM, on both x- and y-axis, the images are organized by the type of material generator, spanning from the learned original materials (i.e., soap, toy, rock) to the morphed midpoint materials (i.e., soap-to-rock, rock-to-toy, and soap-to-toy). The green colors indicate low dissimilarity between pairwise combinations of materials, whereas the pink colors indicate high dissimilarity. The Spearman's correlation (r_s) between the corresponding Vision and Text RDMs are annotated in the box below. (B) Two-dimensional embedding from the MDS of the group average Vision and Text RDMs, color-coded based on the six types of material generator depicted in (A).

Vision- and Language-based Representations Reveal Salient Perceptual Features

Next, we sought to interpret the representative dimensions of materials expressed through the behavioral tasks. We annotated the MDS results of the Vision RDMs with the image stimuli and the participants' verbal descriptions. Colorfulness, material

name, and softness are the key features across participants. On the group average level, as shown in Figure 5A, we labeled the most frequent word that participants used in the description of "material name" and "mechanical properties" for each stimulus. Materials with vivid body colors and high saturation (left side in MDS, e.g., red, pink) are separated from less saturated colors (e.g., light blue and light gray) (leftmost column in Figure 5A). Materials' chemical and physical properties determine the specific range of their colors and surface textures[32]. This innate connection may facilitate visual material categorization. We also observed that participants tend to group "hard" materials (e.g., rock, glass, or crystal) away from "soft" ones (e.g., soap, wax, or rubber) (middle column in Figure 5). Here, perceived softness might be a notable attribute associated with the material category. At the individual level, although participants used different sets of material names to depict the stimuli, they grouped the materials with similar names close to each other. This suggests that visual judgments are influenced by the interpretation of material identity (rightmost column in Figure 5B).

We found that removing the "material names" from the text embeddings ("No material name" condition in Figure 6B) significantly decreased the correlation between Vision and Text RDMs for almost all participants (Wilcoxon one-sided signed-rank test, all $p<0.0005$ across three language models). Different LLMs produced similar results, except GPT-2 embedding led to lower vision-language correlations. Material naming may serve as a high-level feature that envelops the particular structural combination of various material characteristics, providing critical information to the perceptual inference of material attributes. Together, our results suggest that participants actively use semantic-level features to distinguish and group materials in visual assessment, which might explain the correlation between the two behavioral tasks.

Image-level Representations from the Generative Model Weakly Correlate with Human Perception

We investigated how image-level features extracted from our image-generative model contribute to participants' visual judgment of materials. To generate an image with StyleGAN, a latent code $w \in W$ is transformed to the channel-wise latent features, StyleSpace features[48], with different learned affine transformations at each convolution layer (i.e., b4.conv1 to b1024.conv1) of the generator G. Figure 7A illustrates the structure of the StyleSpace features. For a given image, we can retrieve 17 StyleSpace-feature vectors representing image features across nine resolutions (from 4 pixels $\times$ 4 pixels to 1024 pixels $\times$ 1024 pixels) in its image generation process. StyleSpace features represent visual attributes in a scale-specific manner (Figure 7B). Manipulating coarse-scale features (4×4 to 16×16 resolutions) mainly changes the rough contour of the source material to that of the target material. Middle-scale features (32×32 to 64×64 resolutions) correspond to the material surface properties (e.g., translucency) and local variation of geometric complexity. Fine-scale features (128×128 to 1024×1024 resolutions) represent the more refined details, such as the microstructure of surface texture and the object's body color. Based on these latent features, we built 17 StyleSpace-feature RDMs to represent the image-level dissimilarity of materials.

We computed the Spearman's correlation between a participant's Vision RDM and each of the StyleSpace-feature RDMs. The individual participants' data with statistical significance ($p < 0.005$ with FDR correction) are shown in Figure 7D. Our analysis indicated weak but significant correlations (mean correlations range from 0.09 to 0.24) between participants' visual perception of material similarity and image features across all StyleSpace layers. The coarse-to-middle-scale StyleSpace

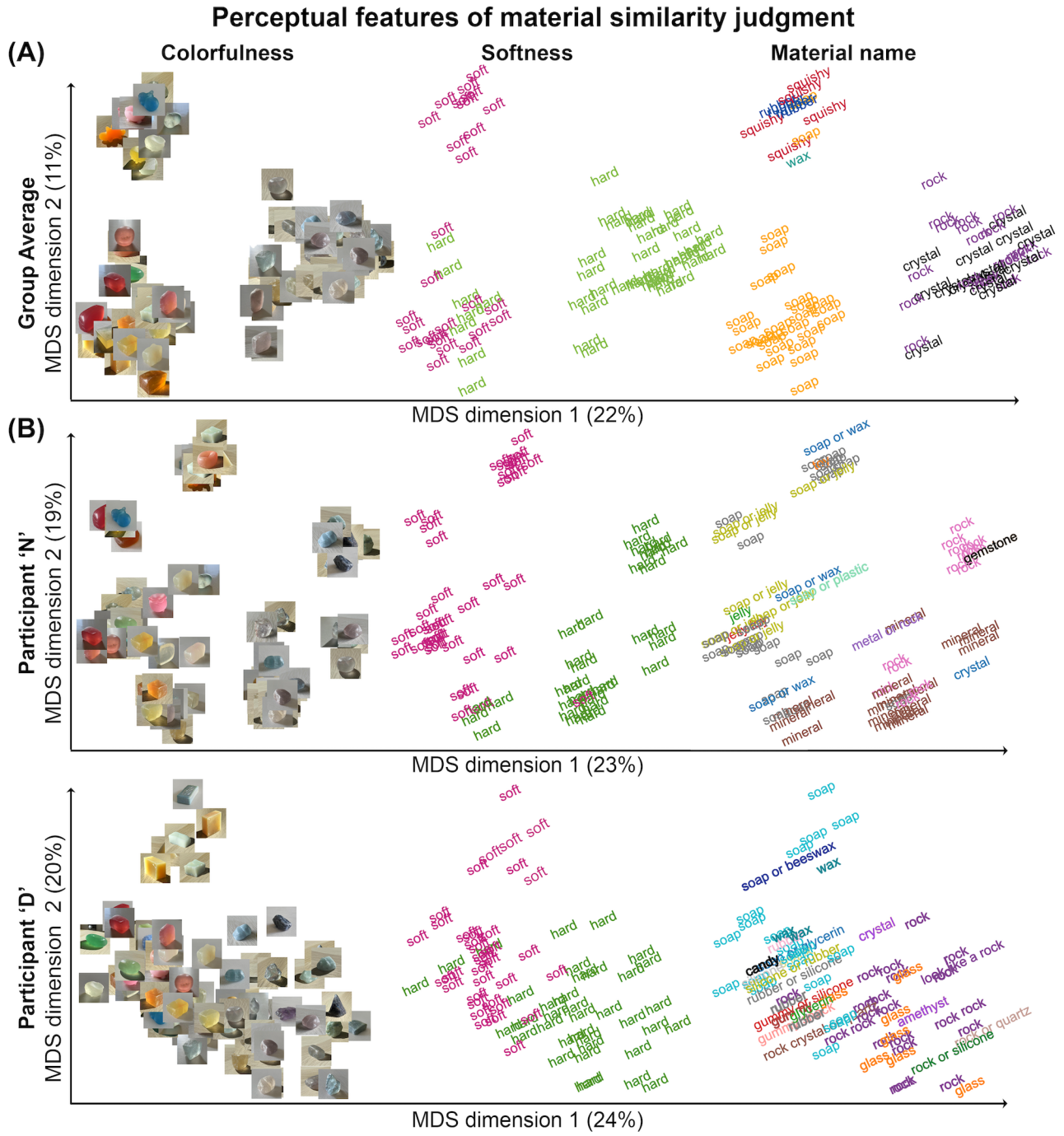

Fig. 5. Colorfulness, softness, and material name are critical perceptual features in material similarity judgment. (A) Annotated MDS of the group average Vision RDM. For "colorfulness", we marked the image stimuli. For "softness" and "material name", we marked the most frequently used word aggregated across all participants. (B) MDS of two individual participants' Vision RDMs, annotated with their own use of words.

features generally show low correlations with the participants' Vision RDMs, implying that the object's rough contour and middle-scale surface texture alone are insufficient for material discrimination. In contrast, the fine-scale StyleSpace features show relatively higher correlations with the multiple arrangement data of some participants. Nevertheless, these representations (e.g., b256.conv1 RDM in Figure 7C) still exhibit substantially different patterns from human perceptual results. Our findings

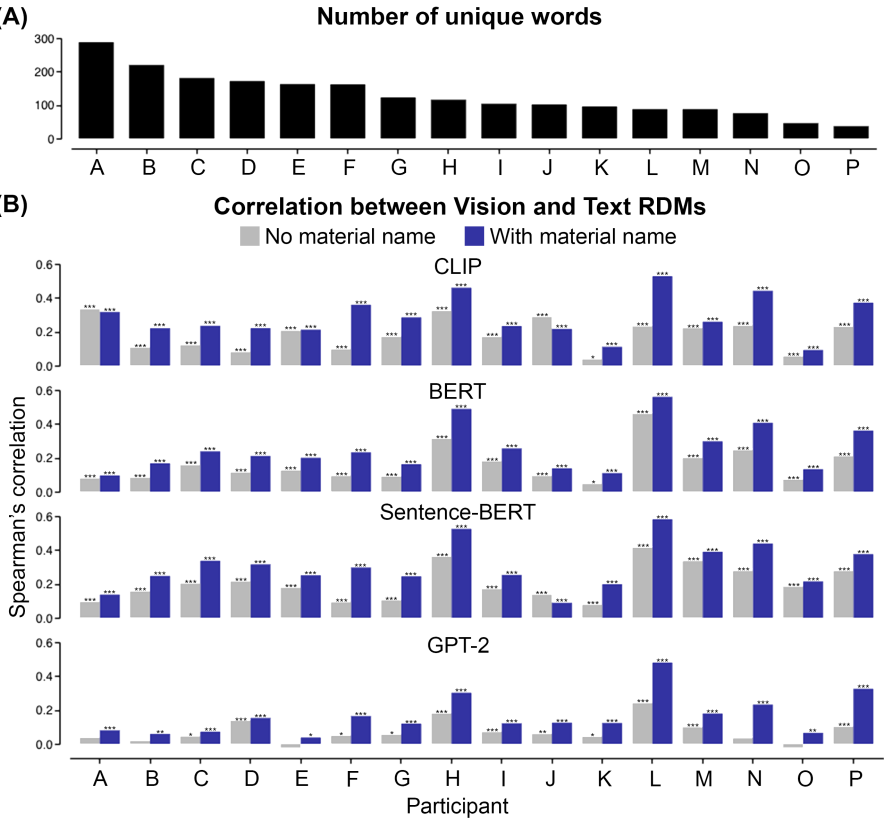

Fig. 6. (A) Distribution of the number of unique words participants used in the Verbal Description task. (B) Comparison of vision-language correlations across different language models. For each individual, we computed within-person Spearman's correlation between the Vision and Text RDMs. The Text RDM is built by embedding verbal descriptions with four different pre-trained LLMs: CLIP, BERT, Sentence-BERT, and GPT-2. The blue bars indicate the correlation values when all text features are included to construct the Text RDM. The gray bars indicate the correlation values when the "material name" is excluded from constructing the Text RDM. Asterisks indicate FDR-corrected p-values: *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

demonstrate that participants do not solely rely on image features when evaluating material similarity. Instead, they incorporate high-level cognition to ensure that their interpretations of the images are consistent.

Joining Text and Image-level Representations Improves Prediction of Human Visual Judgments

Lastly, we tested whether combining the image representations and human verbal descriptions improves the prediction of human visual judgments of materials. For each individual, we used the 17 different StyleSpace-feature RDMs and the participant's Text RDM together (i.e., full model) or only the Text RDM (i.e, reduced model) as predictors in a multiple regression model to predict the participant's Vision RDM. Compared with the reduced model, the model incorporating StyleSpace features and verbal descriptions performs significantly better ($p < 0.0001$ with ANOVA test) and increases the explained variance (adjusted R^2) across all participants (Figure 7E). This provides compelling evidence that visual judgment of materials involves integrating and leveraging multiple spatial scales' image features and that participants' verbal reports provide complementary information to the prediction of their visual perception.

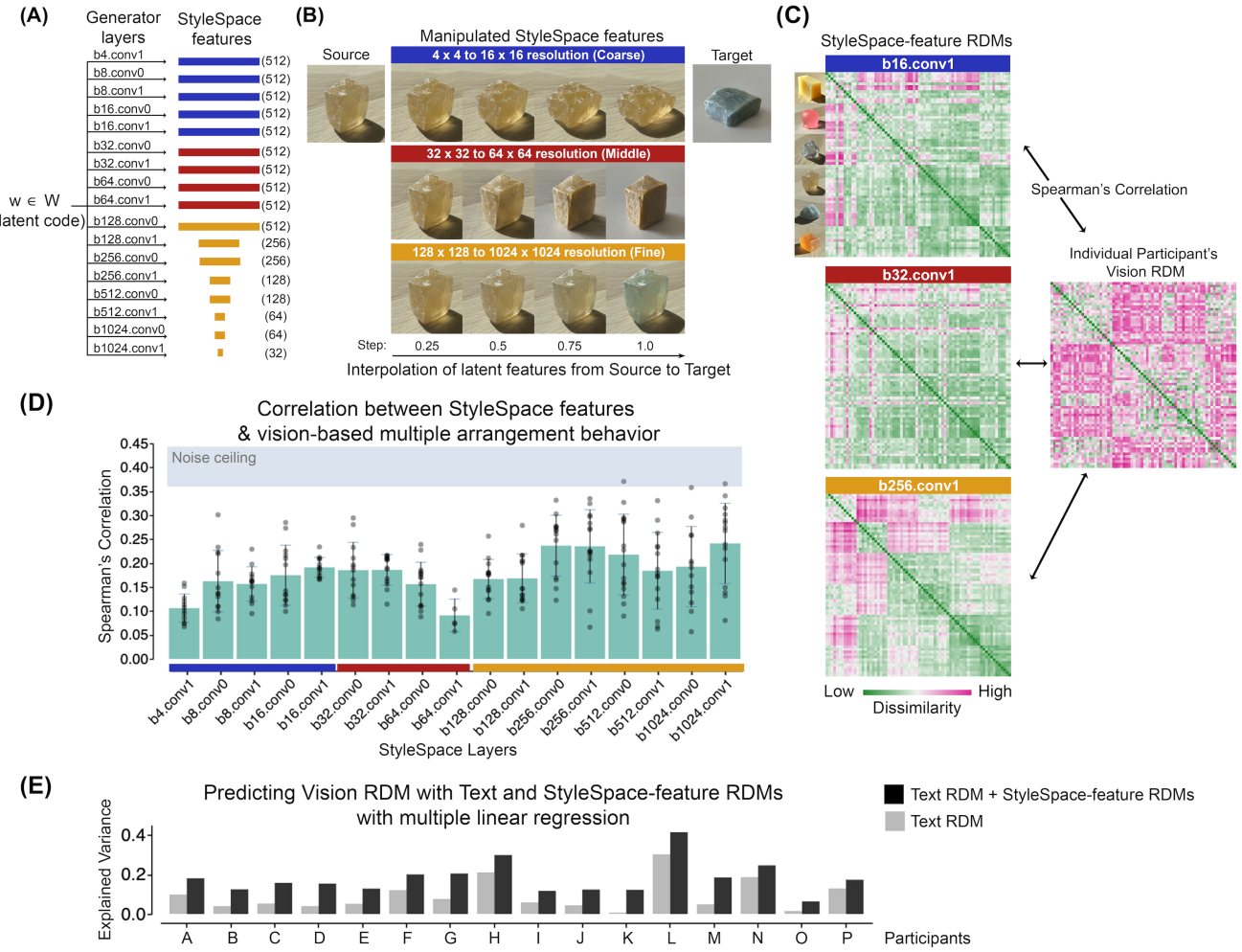

Fig. 7. Comparing representations from the image features and human visual similarity judgment of materials. (A) Structure of the StyleGAN's StyleSpace features. The blue, red, and yellow bars illustrate the StyleSpace-feature vectors at coarse-, middle-, and fine-spatial scales, respectively. The number in the parentheses denotes the dimensionality of the StyleSpace-feature vector at a specific layer. (B) Manipulating StyleSpace features at different spatial scales leads to changes in various visual attributes. From left to right of each row, a subset of StyleSpace features of the source image is interpolated with those of the target image, while keeping the rest of the source image's StyleSpace features unchanged. In this example, both the source and target images are produced from the soap-to-rock generator. (C) Examples of StyleSpace-feature RDMs constructed from the StyleSpace features at different spatial scales. We compute the correlation between each participant's Vision RDM and StyleSpace-feature RDM corresponding to each layer. (D) The Spearman's correlation between an individual's Vision RDM and StyleSpace-feature RDM from each of the 17 layers (black dots). Among the 272 pair comparisons (16 participants $\times$ 17 layers), 83% of them demonstrate statistical significance and are plotted. The green bars represent the average correlations across participants. The blue-shaded region indicates the upper and lower bounds of the noise ceiling. (E) Jointing verbal description with StyleSpace features improves the prediction of multiple arrangement behavior for all participants with multiple linear regression. The x-axis represents individual participants. The y-axis represents the explained variance.

Discussion

We probed how vision and language are connected in material perception. Using an unsupervised image synthesis model, we developed a novel approach to create a diverse array of plausible visual appearances of familiar and novel materials. With these images, we measured and analyzed behavioral tasks alongside image features derived from our material appearance synthesis

network. We found a moderate but significant correlation between visual judgments and verbal descriptions of materials within individual participants, signifying both the efficacy and limitation of language in describing materials. Along with the image-level analysis of material similarity, our findings suggest that material perception goes beyond extracting information solely from low-to-mid-level image statistics; instead, it may actively integrate semantic-level representation to resolve the ambiguous visual information. Demonstrating the potential link between vision- and language-based representations, our study invites further investigation of material perception by considering it an avenue to explore the language-perception relationship across a broad range of visual cognition tasks.

The lack of precise alignment between the representations from two behavioral tasks pinpoints the gap between visual judgment and verbal description of materials (Figure 7E). On the one hand, combining vision- and text-based representations reveals informative features for material discrimination, such as the object's color, softness, and material name. This would be challenging to manifest when limited to a single modality. At the same time, the verbal descriptions do not fully capture the nuances of material appearances the participants visually perceive. One potential explanation could be that participants faced difficulty in describing subtle visual characteristics, such as spatial color variation and surface geometric complexity, with accuracy and consistency. Nevertheless, these visual attributes could be crucial in finely distinguishing samples within the general clustering of materials. Our findings reveal that combining latent image features and text representations enhances the explained variance in participants' visual assessment of materials. This implies that semantic representation and visual perception collaboratively facilitate material judgment. Our finding is consistent with the recent research demonstrating that assembling distilled image features from Deep Neural Networks (DNN) and textual features from LLMs reduces the gap to approximate human similarity judgment[49,50]. The misalignment between language and vision in material perception draws attention to the potential limited expressiveness of language in communicating about materials. This notion is crucial to consider in developing computer vision applications, spanning from material-related scene annotation to text-guided material synthesis.

Our results show that when the material name is removed from the text embedding, the correlation between Vision and Text RDMs systematically decreases across participants (Figure 6). This may stem from the functional roles that nouns (material names) play in everyday language usage[51]. With material names (e.g., crystal or soap), we can label materials that possess an array of unique and/or related attributes, such as softness, translucency, glossiness, and the object's shape. During this labeling process, material names can encapsulate the perceptual similarity of materials across multiple dimensions and partition samples of materials into a system of semantic categories[52], potentially helping us communicate about materials in an efficient way. In future works, we might investigate material perception from the perspective of effective communication[4,52,53], such as examining the structure and complexity of material naming and comparing material naming across various language systems.

The computational mechanisms of material perception remain an area of ongoing exploration without a unified consensus. Over the past decade, a prevailing view has portrayed material perception as a mid-level visual process anchored in the feed-forward-based visual processing hierarchy. Instead of explicitly estimating a material's physical parameters, the visual system may derive a representation by learning the statistical image structures that materials exhibit[54–57]. Material perception

starts by extracting low-level image features and gradually integrating them into a mid-level representation, which indicates material properties. Subsequently, the mid-level features are combined to form a high-dimensional representational space that provides the basis for high-level processing, such as material categorization[34].

Our discovery underscores the significance of acknowledging the top-down influence in material perception, challenging the conventional notion of strictly perceiving materials as a feed-forward process. Notably, we found that participants' visual judgments of material similarity align more closely with their own semantic representations of materials, as evidenced by verbal reports, rather than with image-level representations extracted from the image generation model. We observed substantial individual differences in material description. For the same image, participants could describe it differently and even oppositely. For example, one participant described the yellow soap-to-toy midpoint material (Figure 2E, Soap-to-toy panel, Leftmost in Top Row) as "soap", "easy to deform", and "translucent and glossy". In contrast, another participant described the same image as "stone", "hard", and "opaque and cloudy". When presented with identical image features, variations in high-level cognitive factors, such as material recognition, may determine how participants interpret and integrate information at the image level. This highlights the intricate interplay between low-level visual features and higher-level cognitive functions in perception[37,58].

Evaluating the link between visual judgment and verbal description with behavioral data is the first step for probing their neural representations in material perception. Neuroscience research actively examined the neural representation of language and non-linguistic processing (e.g., music, working memory) and investigated the specificity and interrelationship of brain regions responsible for these cognitive skills[59]. Efforts were also made to explore how the brain encodes certain conceptual representations (e.g., objects, actions) elicited by visual and linguistic stimuli[14,60]. Recent work suggested that incorporating language feedback is crucial for explaining neural responses in high-level visual brain regions[61]. Following these practices, a plausible future direction could be to examine whether and how brain regions' engagement for visual judgment differs from those activated by semantic descriptions of materials. Addressing such cortical representations across modalities may help to unravel the open questions: What is the causal relationship between material recognition and attribute estimation? At a more fundamental level, how does the functional mechanism of perceiving materials differ from and connect to that of perceiving textures and objects?

Unlike previous works that usually examined material perception with real or rendered zoom-in surfaces, we intentionally synthesized images of materials coupling with object-level realism. Our approach of constructing a Space of Morphable Material Appearance with transfer learning and model interpolation methods could be extended to a broader range of materials (e.g., metal, glass). Beyond sampling at the interpolation midpoint, the expressiveness of our model and its latent representation offers a unique capability to manipulate material-related attributes (e.g., translucency and surface geometry) of the object while facilitating controlled and continuous adjustments of visual characteristics linked to material categories. This enables us to potentially design visual stimuli for psychophysical experiments to examine individual differences in material perception and related scene understanding. Given the lack of labeled image datasets of materials, our image synthesis framework can also serve as a versatile tool for data augmentation, providing an ample supply of additional samples for training in material-related

computer vision tasks, including material classification, text-to-image generation, and semantic material attribute editing.

Methods

Image Datasets

We created our training datasets of high-resolution images (1024 pixels $\times$ 1024 pixels) by taking photographs of real-world materials with an iPhone 12 Mini smartphone. Overall, our training data consists of three subcategories: soap (D_{soap}), crystal (D_{rock}), and squishy toy (D_{toy}) datasets, including 8085 (60 objects), 3180 (24 objects), and 1900 (15 objects) images, respectively.

StyleGAN and Transfer Learning

We used the style-based generative adversarial network, StyleGAN2-ADA, as the backbone model. Our previous work, Liao et al. (2023)[27], provides a detailed description of the model and the training process. StyleGAN2-ADA inherently applies a variety of data augmentation during training, and the length of training is defined by the total number of real images seen by the network. We obtained a Soap Model by training the StyleGAN2-ADA from scratch on D_{soap} for a total length of 3,836,000 images, with a learning rate of 0.002 and R_1 regularization of 10.

We fine-tuned the Soap Model separately on the D_{rock} and D_{toy}, which allows all model parameters to adjust to the new datasets. Full-model fine-tuning processes on D_{rock} and D_{toy} used the same hyperparameters as the training on D_{soap}. The lengths of fine-tuning were 1,060,000 and 960,000 images for D_{rock} and D_{toy}, respectively. We used the models with the lowest Fréchet Inception Distance (FID) scores for the rest of our study. The FID scores for Rock and Toy Models are 22.22 and 23.38, respectively. All training was performed on a Tesla V100 GPU on Google Colab.

Cross-category Material Morphing

The morphing of images of materials requires applying linear interpolation of the layer-wise latent codes $w \in W$, as well as the StyleGAN's generator weights[62]. To morph from a source to a target material, we first sample two latent codes (i.e., w_{source} and w_{target}) from the corresponding learned W latent spaces (e.g., $w_{soap} \in W_{soap}$ as source, $w_{rock} \in W_{rock}$ as target). As illustrated in Figure 2C, w is a tensor with the dimension of 18×512. With equation (1), we can compute the interpolated latent code w_λ, at any desired step size λ. The dimension of w_λ is also 18×512. Similarly, we implement linear interpolation between the convolutional weights of each convolution layer in the source material generator and the corresponding weights in the target material generator. The weights are multidimensional tensors. With the same λ, we calculate the interpolated generator weights G_λ (equation (2)).

$$w_\lambda = w_{source} + \lambda(w_{target} - w_{source}) \tag{1}$$

$$G_\lambda = G_{source} + \lambda(G_{target} - G_{source}) \tag{2}$$

We insert w_λ into G_λ to generate the image of morphed material. Specifically, each of 18 slices of w_λ is injected into the convolution layer at the corresponding spatial resolution (from 4 pixels $\times$ 4 pixels to 1024 pixels $\times$ 1024 pixels) (Figure 2C).

Psychophysical Experiments

Participants Sixteen participants (13 female, median age = 22) from the American University (AU) were given informed consent and were reimbursed for their participation. All were Native English speakers and had a normal or corrected-to-normal vision. The experiments were approved by the ethics board at AU, and were conducted in adherence to the Declaration of Helsinki.

Stimulus Selection We first generated 30 images for each of the "original" materials: soaps, rocks, and squishy toys, by sampling from their corresponding latent spaces, W_{soap}, W_{rock}, and W_{toy} and synthesizing with their paired material generators, G_{soap}, G_{rock}, and G_{toy}. We balanced the images in two lighting conditions for each material category: strong and weak (Figure 2E).

We randomly paired up two different "original" materials under the same lighting conditions and then synthesized the image corresponding to the linear interpolation midpoint (step $\lambda = 0.5$). We initially generated 1000 images of morphed materials through the corresponding midpoint material generators ($G_{soap-to-rock}$, $G_{soap-to-toy}$, and $G_{rock-to-toy}$). We picked 12 images synthesized from six material categories: soap, rock, squishy toy, soap-to-rock midpoint, rock-to-toy midpoint, and soap-to-toy midpoint. For each material category, half of the selected images are from strong lighting conditions (i.e., sunny indoor scene), and the remaining half are from weak lighting conditions (i.e., overcast indoor scene). We selected 72 images and tried to make the range of visual appearances as diverse and natural as possible. These images were then used as stimuli for Multiple Arrangement and Verbal Description tasks.

Multiple Arrangement Task We conducted the multiple arrangement experiment using Meadows.com (https://meadows-research.com/). Participants were instructed to arrange the images (180 pixels $\times$ 180 pixels) of materials based on the "similarity of material properties" by dragging and dropping them in the circled region (Figure 3A). In the first trial, the participants roughly arranged all 72 images into groups. In the subsequent trials, more refined subsets were chosen and displayed by an adaptive lift-the-weakest algorithm to reduce the remaining uncertainty of the similarity judgment of materials[44]. The average duration of the experiment was about 60 minutes. The pairwise on-screen Euclidean distances between the arranged images were computed upon the completion of the experiment, producing a Vision RDM with inverse MDS.

Verbal Description Task With the same 72 images used in the Multiple Arrangement task, participants described the material in the image by freely inputting texts based on five aspects (see Figure 3B). They had unlimited time on each trial and were not restricted regarding the order in which they could enter their responses.

Experiment Procedures All participants first completed the Multiple Arrangement task, and then the Verbal Description task in a separate session. All experiments were conducted in a dimly lit laboratory room. The stimuli were presented on an Apple

iMac computer with a 21.5-inch Retina Display, with a resolution of 1920 pixels × 1080 pixels.

Creating Text RDMs from Verbal Description Data We used a fixed template to concatenate the five aspects that participants described an image: "It is a material of [material name] with the color of [color], it is [optical properties], it is [mechanical properties], and it is [surface texture]." Next, we encoded the concatenated text into a feature vector through a pre-trained LLM. The four commonly used pre-trained transformer-based language models (i.e., CLIP[7], BERT[45], Sentence-BERT[46], and GPT-2[47]) can embed a sentence or paragraph of text into a high-dimensional feature space. For CLIP, Sentence-BERT, and GPT-2, we extracted the feature vector at the last hidden layer. For BERT, we concatenated the feature vectors from its last four hidden layers. The size of the embedded text feature vector varies across different language models: 512 for CLIP, 3072 for BERT, 384 for Sentence-BERT, and 768 for GPT-2. For each participant, we built a 72 × 72 Text RDM by computing the pairwise cosine dissimilarity between the resulting feature vectors of the verbal descriptions (Figure 4A, Bottom Row).

To investigate the effect of removing the "material name" on the embedding of the verbal descriptions, we used the following template to form the image caption: "It is a material with the color of [color], it is [optical properties], it is [mechanical properties], and it is [surface texture]." Hence, we used the same procedure described above to encode the descriptions without material names as feature vectors.

Creating StyleSpace-Feature RDMs We used StyleGAN's StyleSpace[48] to describe the innate image features of materials (Figure 7A). For a StyleGAN generator, a 4 × 4 input is progressively expanded to a 1024 × 1024 output image. For each major resolution (every resolution from 4 pixels × 4 pixels to 1024 pixels × 1024 pixels), there are two convolution layers (conv0 and conv1) for feature map synthesis (except that the 4 × 4 resolution only has conv1) and a single convolution layer (i.e., tRGB layer) that converts the output to an RGB image. The latent code $w \in W$ is injected into different learned convolution layers to obtain the channel-wise style vectors, namely StyleSpace. Here, we focused on the style vectors for feature map representation.

Overall, we can extract 17 style vectors representing latent image features across different generative steps for a particular generated image. The dimensionality of these style vectors varies depending on their corresponding spatial resolution of the convolution layer, gradually decreasing from 512-dimension at 4 × 4 resolution to 32-dimension at 1024 × 1024 resolution.

For the 72 images used in the psychophysical experiments, we extracted their StyleSpace features and computed the pairwise Euclidean distances between images, thus generating 17 StyleSpace-feature RDMs (i.e., one for each StyleSpace layer) (see examples in Figure 7C).

Predicting Vision RDM with Text and StyleSpace-feature RDMs To determine the contribution of verbal description and image-level features to the multiple arrangement behavior, we use multiple linear regression based on the Text and StyleSpace-feature RDMs[63].

We first converted the 72 × 72 RDMs into 2556-dimensional feature vectors, by extracting the off-diagonal elements of an RDM. We set the participant's own Vision RDM feature vector as the predicted variable. Two first-order multiple regression models were fitted for each participant: one "full" model that included the Text RDM and all 17 StyleSpace-feature RDMs; and

one "reduced" model that only included the Text RDM. For each model, we computed the adjusted R^2 to indicate the explained variance of the Vision RDM. We also used ANOVA to test whether the "full" model improves the fit of the data compared to the "reduced" model, with statistical significance (95% confidence level).

References

1. Pinker, S. & Bloom, P. Natural language and natural selection. *Behav. brain sciences* **13**, 707–727 (1990).
2. Brown, R. W. & Lenneberg, E. H. A study in language and cognition. *The J. Abnorm. Soc. Psychol.* **49**, 454 (1954).
3. Regier, T., Kay, P. & Cook, R. S. Focal colors are universal after all. *Proc. Natl. Acad. Sci.* **102**, 8386–8391 (2005).
4. Zaslavsky, N., Kemp, C., Regier, T. & Tishby, N. Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci.* **115**, 7937–7942 (2018).
5. Dils, A. T. & Boroditsky, L. Processing unrelated language can change what you see. *Psychon. bulletin & review* **17**, 882–888 (2010).
6. Lupyan, G., Rahman, R. A., Boroditsky, L. & Clark, A. Effects of language on visual perception. *Trends cognitive sciences* **24**, 930–944 (2020).
7. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
8. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
9. Vo, N. *et al.* Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6439–6448 (2019).
10. Hu, R. *et al.* Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4555–4564 (2016).
11. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**, 3 (2022).
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695 (2022).
13. Hebart, M. N. *et al.* Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* **12**, e82580 (2023).
14. Bonner, M. F. & Epstein, R. A. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nat. communications* **12**, 4081 (2021).
15. Xiao, B. & Brainard, D. H. Surface gloss and color perception of 3d objects. *Vis. neuroscience* **25**, 371–385 (2008).

16. Schmid, A. C., Barla, P. & Doerschner, K. Material category of visual objects computed from specular image structure. *Nat. Hum. Behav.* **7**, 1152–1169 (2023).

17. Olkkonen, M. & Brainard, D. H. Joint effects of illumination geometry and object shape in the perception of surface reflectance. *i-Perception* **2**, 1014–1034, DOI: 10.1068/i0480 (2011).

18. Toscani, M., Valsecchi, M. & Gegenfurtner, K. R. Lightness perception for matte and glossy complex shapes. *Vis. Res.* **131**, 82–95, DOI: 10.1016/j.visres.2016.12.004 (2017).

19. Fleming, R. W. & Bülthoff, H. H. Low-level image cues in the perception of translucent materials. **2**, 346–382, DOI: 10.1145/1077399.1077409 (2005).

20. Motoyoshi, I. Highlight–shading relationship as a cue for the perception of translucent and transparent materials. *J. Vis.* **10(9):**, 1–11, DOI: 10.1167/10.9.6 (2010).

21. Nagai, T. *et al.* Image regions contributing to perceptual translucency: A psychophysical reverse-correlation study. *i-Perception* **4**, 407–428, DOI: 10.1068/i0576 (2013).

22. Xiao, B. *et al.* Looking against the light: How perception of translucency depends on lighting direction. *J. Vis.* **14(3):**, 1–22, DOI: 10.1167/14.3.17 (2014).

23. Xiao, B., Zhao, S., Gkioulekas, I., Bi, W. & Bala, K. Effect of geometric sharpness on translucent material perception. *J. Vis.* **20(7):**, 1–17, DOI: 10.1167/jov.20.7.10 (2020).

24. Gkioulekas, I. *et al.* Understanding the role of phase function in translucent appearance. *ACM Transactions on Graph. (TOG)* **32**, 1–19 (2013).

25. Marlow, P. J. & Anderson, B. L. The cospecification of the shape and material properties of light permeable materials. *Proc. Natl. Acad. Sci.* **118**, e2024798118, DOI: 10.1073/pnas.2024798118 (2021).

26. Liao, C., Sawayama, M. & Xiao, B. Crystal or jelly? effect of color on the perception of translucent materials with photographs of real-world objects. *J. Vis.* **22**, 6–6, DOI: 10.1167/jov.22.2.6 (2022).

27. Liao, C., Sawayama, M. & Xiao, B. Unsupervised learning reveals interpretable latent representations for translucency perception. *PLOS Comput. Biol.* **19**, e1010878, DOI: 10.1371/journal.pcbi.1010878 (2023).

28. Fleming, R. W., Jäkel, F. & Maloney, L. T. Visual perception of thick transparent materials. *Psychol. Sci.* **22**, 812–820 (2011).

29. Di Stefano, N. & Spence, C. Roughness perception: A multisensory/crossmodal perspective. *Attention, Perception, & Psychophys.* **84**, 2087–2114 (2022).

30. Cavdan, M., Drewing, K. & Doerschner, K. The look and feel of soft are similar across different softness dimensions. *J. vision* **21**, 20–20 (2021).

411 **31.** Bi, W., Jin, P., Nienborg, H. & Xiao, B. Manipulating patterns of dynamic deformation elicits the impression of cloth with
412 varying stiffness. *J. Vis.* **19**, 18–18, DOI: 10.1167/19.5.18 (2019).

413 **32.** Zaidi, Q. Visual inferences of material changes: color as clue and distraction. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 686–700
414 (2011).

415 **33.** Sawayama, M., Adelson, E. H. & Nishida, S. Visual wetness perception based on image color statistics. *J. Vis.* **17**, 7–7,
416 DOI: 10.1167/17.5.7 (2017).

417 **34.** Fleming, R. W. Material perception. *Annu. Rev. Vis. Sci.* **3**, 365–388, DOI: 10.1146/annurev-vision-102016-061429
418 (2017).

419 **35.** Hiramatsu, C., Goda, N. & Komatsu, H. Transformation from image-based to perceptual representation of materials along
420 the human ventral visual pathway. *Neuroimage* **57**, 482–494 (2011).

421 **36.** Nishio, A., Goda, N. & Komatsu, H. Neural selectivity and representation of gloss in the monkey inferior temporal cortex.
422 *J. Neurosci.* **32**, 10780–10793, DOI: 10.1523/JNEUROSCI.1095-12.2012 (2012).

423 **37.** Baumgartner, E. & Gegenfurtner, K. R. Image statistics and the representation of material properties in the visual cortex.
424 *Front. Psychol.* **7**, 1185 (2016).

425 **38.** Komatsu, H. & Goda, N. Neural mechanisms of material perception: Quest on shitsukan. *Neuroscience* **392**, 329–347,
426 DOI: 10.1016/j.neuroscience.2018.09.001 (2018).

427 **39.** van Assen, J. J. R., Barla, P. & Fleming, R. W. Visual features in the perception of liquids. *Curr. Biol.* **28**, 452–458, DOI:
428 10.1016/j.cub.2017.12.037 (2018).

429 **40.** Serrano, A., Gutierrez, D., Myszkowski, K., Seidel, H.-P. & Masia, B. An intuitive control space for material appearance.
430 *arXiv preprint arXiv:1806.04950* (2018).

431 **41.** Schmidt, F., Hebart, M. N., Fleming, R. W. *et al.* Core dimensions of human material perception. DOI: 10.31234/osf.io/
432 jz8ks (2022).

433 **42.** Cavdan, M., Goktepe, N., Drewing, K. & Doerschner, K. Assessing the representational structure of softness activated by
434 words. *Sci. Reports* **13**, 8974 (2023).

435 **43.** Karras, T. *et al.* Training generative adversarial networks with limited data. *Adv. Neural Inf. Process. Syst.* **33**, 12104–12114
436 (2020).

437 **44.** Kriegeskorte, N. & Mur, M. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Front.*
438 *psychology* **3**, 245 (2012).

439 **45.** Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language
440 understanding. *arXiv preprint arXiv:1810.04805* (2018).

46. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

47. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

48. Wu, Z., Lischinski, D. & Shechtman, E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12863–12872, DOI: 10.1109/CVPR46437.2021.01267 (2021).

49. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).

50. Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N. & Griffiths, T. L. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308* (2023).

51. Gentner, D. & Rattermann, M. J. *Language and the career of similarity* (University of Illinois at Urbana-Champaign, Center for the Study of Reading, 1991).

52. Regier, T., Kemp, C. & Kay, P. Word meanings across languages support efficient communication. *The handbook language emergence* 237–263 (2015).

53. Witzel, C. & Gegenfurtner, K. R. Color perception: Objects, constancy, and categories. *Annu. Rev. Vis. Sci.* **4**, 475–499 (2018).

54. Fleming, R. W. Visual perception of materials and their properties. *Vis. Res.* **94**, 62–75, DOI: 10.1016/j.visres.2013.11.004 (2014).

55. Nishida, S. Image statistics for material perception. *Curr. Opin. Behav. Sci.* **30**, 94–99 (2019).

56. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417, DOI: 10.1038/s41562-021-01097-6 (2021).

57. Storrs, K. R. & Fleming, R. W. Learning about the world by learning about images. *Curr. Dir. Psychol. Sci.* **30**, 120–128, DOI: 10.1177/0963721421990334 (2021).

58. Fleming, R. W., Wiebel, C. & Gegenfurtner, K. Perceptual qualities and material classes. *J. vision* **13**, 9–9 (2013).

59. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci.* **108**, 16428–16433 (2011).

60. Wurm, M. F. & Caramazza, A. Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nat. communications* **10**, 289 (2019).

61. Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat. Mach. Intell.* 1–12 (2023).

62. Wu, Z., Nitzan, Y., Shechtman, E. & Lischinski, D. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323* (2021).

472 **63.** Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of
473 scenes in human brain and behavior. *Elife* **7**, e32962 (2018).

474 Acknowledgments

475 This work was supported by the National Institutes of Health grant 1R15EY033512.