# Evidence for Hypodescent in Visual Semantic AI

Robert Wolfe
University of Washington
Seattle, WA, USA
rwolfe3@uw.edu

Mahzarin R. Banaji
Harvard University
Cambridge, MA, USA
mahzarin_banaji@harvard.edu

Aylin Caliskan
University of Washington
Seattle, WA, USA
aylin@uw.edu

## ABSTRACT

We examine the state-of-the-art multimodal "visual semantic" model CLIP ("Contrastive Language Image Pretraining") for the rule of hypodescent, or one-drop rule, whereby multiracial people are more likely to be assigned a racial or ethnic label corresponding to a minority or disadvantaged racial or ethnic group than to the equivalent majority or advantaged group. A face morphing experiment grounded in psychological research demonstrating hypodescent indicates that, at the midway point of 1,000 series of morphed images, CLIP associates 69.7% of Black-White female images with a Black text label over a White text label, and similarly prefers Latina (75.8%) and Asian (89.1%) text labels at the midway point for Latina-White female and Asian-White female morphs, reflecting hypodescent. Additionally, assessment of the underlying cosine similarities in the model reveals that association with White is correlated with association with "person," with Pearson's $\rho$ as high as 0.82, $p < 10^{-90}$ over a 21,000-image morph series, indicating that a White person corresponds to the default representation of a person in CLIP. Finally, we show that the stereotype-congruent pleasantness association of an image correlates with association with the Black text label in CLIP, with Pearson's $\rho = 0.48$, $p < 10^{-90}$ for 21,000 Black-White multiracial male images, and $\rho = 0.41$, $p < 10^{-90}$ for Black-White multiracial female images. CLIP is trained on English-language text gathered using data collected from an American website (Wikipedia), and our findings demonstrate that CLIP embeds the values of American racial hierarchy, reflecting the implicit and explicit beliefs that are present in human minds. We contextualize these findings within the history of and psychology of hypodescent. Overall, the data suggests that AI supervised using natural language will, unless checked, learn biases that reflect racial hierarchies.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

## KEYWORDS

multimodal, bias in AI, visual semantics, language-image models, racial bias, hypodescent

**ACM Reference Format:**
Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. 2022. Evidence for Hypodescent in Visual Semantic AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533185



*FAccT '22, June 21–24, 2022, Seoul, Republic of Korea*

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533185

## 1 INTRODUCTION

Recent progress in multimodal "visual semantic" artificial intelligence (AI) has produced CLIP ("Contrastive Language Image Pretraining"), the first language-and-image model that allows the definition of image classes in natural language and generalizes to datasets on which it was not explicitly trained [56]. For the field of computer vision, CLIP was able to realize many of the benefits of large-scale self-supervised pretraining first seen in natural language processing (NLP). While its designers primarily evaluate CLIP in the context of image classification [56], features derived from the model have been used to train zero-shot object detection models and zero-shot image generation AI [23, 60]. Investigations of biases in CLIP have uncovered under-representation of women in image retrieval tasks [70], the presence of biased and offensive content in an open-source training corpus similar to the one on which CLIP was trained [6], and semantic biases (*e.g.,* association of Muslims with terrorism), in multimodal neuron activations in the CLIP image encoder [20]. In this research, we investigate whether the manner in which CLIP forms categorical boundaries between images of humans reflects a particular racial bias: the rule of hypodescent, or one-drop rule.

In the paper introducing CLIP, Radford et al. [56] report that in the zero-shot setting, CLIP associates with a White text label only 58.3% of people with the White racial label in the FairFace dataset [35], while associating 91.3% of individuals belonging to the six other racial and ethnic groups represented in FairFace with their FairFace label. While the FairFace dataset consists of visually noisy in-the-wild photographs, and relies not on self-identification of race but on the potentially biased labels of human perceivers (Amazon Mechanical Turkers), the results of Radford et al. [56] indicate that a substantial minority of people who are perceived by other humans as White are associated with a race other than White by CLIP. The inverse does not appear to be true, as individuals who are perceived to belong to other racial or ethnic labels in the FairFace dataset are associated with those labels by CLIP. This result suggests the possibility that CLIP has learned the rule of "hypodescent," as described by social scientists: individuals with multiracial ancestry are more likely to be perceived and categorized as belonging to the minority or less advantaged parent group than to the equally legitimate majority or advantaged parent group. In other words, the child of a Black and a White parent is perceived to be more Black than White; and the child of an Asian and a White parent is perceived to be more Asian than White. Psychological research finds that a rule of hypodescent reflects a belief in racial hierarchy, and maintains and enhances the racial status quo [31].

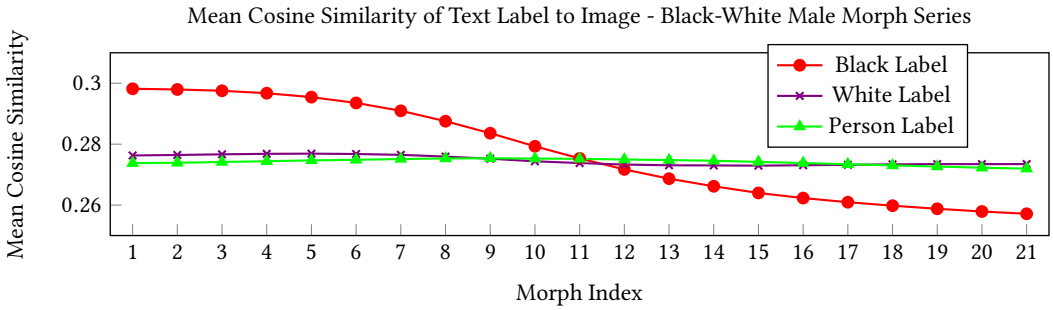


**Figure 1: Across 5% increments in mixing ratio between Black source images (index 1) and White target images (index 21), the mean probability of a White text label and the mean probability of a person text label are nearly equivalent and nearly invariant. On the other hand, the mean probability of the Black text label varies with the mixing ratio of the images, indicating that White is the default race in CLIP, with other racial and ethnic groups defined based on difference from White.**

We systematically test for evidence of hypodescent in CLIP. This is, to our knowledge, the first analysis of hypodescent in AI. We make research code public at https://github.com/wolferobert3/evidence_for_hypodescent. Three main results are highlighted:

(1) **CLIP shows hypodescent, or the one-drop rule, by biased association of images of multiracial people with minority ancestry.** Faces of individuals who self-identified as Black, Asian, and Latino/a were morphed into faces of people who self-identified as White. At the halfway point of the morph series, the critical juncture, CLIP associates the majority of the morphed images with a Black, Asian, or Latino/a text label rather than a White text label. Hypodescent is more pronounced for images of women: at the halfway point, 69.7% of Black-White female multiracial images are associated with Black; 75.8% of Latina-White female multiracial images are associated with Latina; and 89.1% of Asian-White female multiracial images are associated with Asian. 53.8% of Black-White male multiracial images at the halfway point are associated with Black. We do not observe hypodescent for the Asian-White male or for the Latino-White male morph series.

(2) **White is the default race in CLIP, with other racial and ethnic groups defined by their deviation from White.** For each morph series, we obtain Pearson's $\rho$ for the cosine similarity of a "person" label (with no text indicating race or ethnicity) and the cosine similarity of White, Black, Asian, and Latino/a text labels. In all cases, the correlation between person and White is higher ($\rho \in [0.60, 0.82]$) than the correlation between person and Asian, Black, or Latino ($\rho \in [0.11, 0.40]$), indicating that the most similar representation to a person is a White person in CLIP. As shown in Figure 1, the mean cosine similarity for the White label at each step of the morph series is nearly invariant, and nearly indistinguishable from that of the person label. On the other hand, the mean cosine similarity of the Black label varies as the mixing ratio of the source and target images changes, such that images are associated with Black based on deviation from the White default.

(3) **Valence bias (association with bad or unpleasant vs. good or pleasant concepts) correlates with association with the minority group.** For each morph series, the SC-WEAT [9] is used to measure the association of each embedded image with unpleasant vs. pleasant concepts in the CLIP embedding space, represented using 25 valenced words each (*e.g.*, grief, agony, disaster vs. love, peace, cheer). The cosine similarity of each image with the minority race or ethnicity text label is obtained. For the 21,000-image Black-White male morph series, the association with unpleasant concepts for an embedded image correlates with the image's association with the Black label, with Pearson's $\rho = 0.48$, $p < 10^{-90}$. The correlation is similar for the Black-White female morph series, with $\rho = 0.41$, $p < 10^{-90}$.

Although this work will discuss racial categories such as Asian, Black, Latino, and White, we use them in the manner in which the social and psychological sciences use them, *i.e.*, to refer to social constructions, not biological realities [3].

## 2 RELATED WORK

This research draws on multiple strands of previous research on hypodescent, CLIP and multimodal AI, and bias in AI.

**Hypodescent.** Hypodescent refers to the association of people who have multiracial ancestry with the race of their minority parent group. The way multiracial individuals are perceived by others is of sociopolitical importance, as it can serve as an indication of whether intermarriage and miscegenation will result in the disturbance of racial hierarchy[1] [33]. For example, if a society has a high number of multiracial births, those at the top of a racial hierarchy may make the hierarchy durable to demographic change by categorizing multiracial individuals as belonging to a minority racial or ethnic group [33]. The minority group increases in size, but can be denied privileges providing access to life's opportunities and outcomes, ranging from housing and loans, to education and jobs, to treatment by law and law enforcement [29].

[1]The term "racial hierarchy" is common in the social and psychological literature, and is the term we will employ here. Others such as Wilkerson [71] have characterized the racial ordering of political and economic advantages in America as a caste system. In referring to racial hierarchy, we do not intend to lend legitimacy to this system, but to draw attention to the ways it operates to prevent its encoding in AI.

The primary psychological reference for the present research is the work of Ho et al. [33], who uncovered evidence for a rule of hypodescent in human perceivers based on measures of both relatively more implicit and explicit cognition [33]. To study the implicit bias of hypodescent, Ho et al. [33] used a face morphing experiment, wherein the face of a Black person or an Asian person (subjectively strongly judged to be so) is "morphed" across a series of images into a face of a White person. They find that human perceivers are more likely to classify intermediate images according to minority ancestry, reflecting a rule of hypodescent [33]. Moreover, they show that perceivers presented with the family tree of a multiracial person with one Black or Asian parent and one White parent were more likely to classify that person as Black or Asian, respectively, than as White [33]. Ho et al. [33] find that biases of hypodescent are more significant for multiracial people with Black ancestry, but are also observed for multiracial people with Asian ancestry, a result reflecting the existence of American racial hierarchy [29]. As described in the Section 4, our research design draws insight from the work of Ho et al. [33] to test for the bias of hypodescent in AI. If the result of Ho et al. [33] is limited to a bias in human minds, it should not be observed in the analyses performed using visual semantic CLIP associations. If, on the other hand, hypodescent is reflected in the CLIP embedding space, the result will add to a growing body of research demonstrating the manner in which human biases are transduced into machine learning architectures.

In a subsequent study of the psychological underpinnings of hypodescent and racial hierarchy, Ho et al. [31] find that a rule of hypodescent is employed when a minority group makes significant gains in social realms such as business and politics (a threat condition to the racial status quo), and that individuals high in social dominance orientation (an individual difference measure of preference for group-based hierarchies [32]) are more likely to categorize multiracial individuals according to a minority parent group. Ho et al. [31] find that hypodescent can be seen as a "hierarchy-enhancing" social categorization, in that it amplifies perception of racial differences such that racial boundaries and advantages are preserved despite changing demographics and social structures. More recently, Krosch et al. [38] find that White perceivers who read about impending demographic shifts which will render the U.S. a majority-minority country are more likely to characterize multiracial faces as belonging to a minority group, reflecting a rule of hypodescent.

Other psychological studies have suggested that hypodescent is related to attention and to frequency of observation. Halberstadt et al. [26] suggest that hypodescent is related to selective attention, wherein humans pay more attention to physical characteristics infrequently and first observed later in life, such that people who possess a mixture of majority-group and minority-group traits are perceived as belonging to the minority group. This has particular salience for computer vision, given prior research that finds that people who have darker skin tones, and in particular women who have darker skin tones, are significantly underrepresented in computer vision training datasets [8]. Lewis [43] finds that people who have more experience seeing Black faces are more likely to classify multiracial faces as belonging to a White person. Ho et al. [30] find that Black perceivers may use hypodescent inclusively, extending Blackness to multiracial people with Black ancestry "because they perceive that Black–White biracials face discrimination and consequently feel a sense of linked fate with them." Peery and Bodenhausen [55] finds that automatic (reflexive/implicit) racial categorizations by White perceivers reflect hypodescent, but that "more complex racial identities may be acknowledged upon more thoughtful reflection." Zarate and Smith [79] find that white perceivers classify members of their own race more quickly, and that speed of categorization is linked to attribution of racial stereotypes.

While hypodescent has been observed for both males and females, Ho et al. [33] observed stronger effects for men than women. This is consistent with research by Sidanius and Pratto [64] and Navarrete et al. [51], who find that histories of intergroup conflicts involving primarily male aggressors have resulted in stronger exclusionary biases toward outgroup males. The disparity in hypodescent patterns observed in experimental psychology may be a result of broader intersectional gender biases, wherein Black faces are less readily classified as female [19], and male faces are more likely to be classified as Black [10]. We are not aware of studies examining hypodescent beyond a gender binary.

***American Context of Hypodescent.*** A notable result that has emerged from the present work is that hypodescent in AI resembles hypodescent in human cognition, as observed by experimental psychologists. Our work focuses on hypodescent in an American context, for two reasons: because most modern psychological research of hypodescent studies American subjects and locates findings within the American context [78]; and because CLIP is trained on English-language internet data, the majority of which is based on a query list collected from an American website (English-language Wikipedia) [56], suggesting that what CLIP has learned about hypodescent is likely to be best understood within the American context. The historical context of hypodescent in the United States - as a mechanism for expanding slavery [4], for enforcing segregation [29], and for preserving the stability of racial hierarchy [47] - has been noted by psychologists as a likely factor in modern Americans' implicit biases with regard to the race of multiracial individuals [33].

**CLIP and Visual Semantic AI.** We now describe the technical architecture of CLIP, the AI model we will test for evidence of hypodescent. CLIP was designed to be the first "zero-shot" image classifier, meaning that it has learned to associate images in computer vision datasets with their text labels without seeing the training data provided for those datasets [56]. This constituted a significant step forward in computer vision and multimodal AI, as CLIP advanced the zero-shot state-of-the-art on ImageNet from 11.5% [44] to 76.2% [56]. CLIP is known as a "visual semantic" or "multimodal" AI model, in that it encodes text and images and projects them into the same embedding space, such that the text most similar to an image can be matched to it by measuring the cosine similarity of the encoded text and the encoded image [56]. While CLIP has been discussed primarily in terms of its performance as a zero-shot image classifier, the model learns "transferable" visual features, meaning that the representations formed by CLIP can be easily adapted for many computer vision tasks, such as image retrieval [70], and can be used to train specialized zero-shot computer vision models for tasks such as object detection [23] and image generation [53, 60]. Moreover, the word and sentence representations formed by CLIP

have been shown to be highly semantic, to the point that they set or match state-of-the-art on intrinsic evaluations [75].

CLIP jointly trains a computer vision model (a ResNet [27] or a Vision Transformer (ViT) [17]) and a contextualizing language model (a smaller version of GPT-2 [57]), and projects representations from each model into the same embedding space. Both GPT-2 and ViT are based on the transformer architecture, a deep learning network which uses self-attention to draw information from anywhere in its input window [69]. CLIP's success in generalizing across datasets is the result of applying a novel training objective to an internet-scale language-and-image corpus. CLIP trains on the WebImageText (WIT) corpus, a collection of 400 million images and their associated captions scraped from the internet [56]. While WIT is not publicly available, Radford et al. [56] note that it contains roughly as many words as the text corpus used to train GPT-2. The query list for WIT consists of all words occurring at least 100 times in English language Wikipedia, bigrams from Wikipedia which have high pointwise mutual information, the names of Wikipedia articles, and all WordNet synsets not included in the list [56]. CLIP employs a "contrastive learning" pretraining objective, which maximizes the cosine similarity of an encoded caption with its encoded image while minimizing its cosine similarity with all of the other encoded images in the batch [56, 67]. When this research refers to an image or text embedding, this means the multimodal representation formed by CLIP after projection to this joint embedding space. Our supplementary materials include a discussion of advances in deep learning visual semantic AI beginning with its origins in 2013.

This research tests for hypodescent in CLIP by comparing the cosine similarities of embedded images and embedded text labels corresponding to four racial or ethnic groups, an approach which reflects the typical zero-shot use of the model during inference [56]. Biases observed in the embedding space will be reflected in the many zero-shot settings in which CLIP is used, including image retrieval [70], image classification [56], and other settings which may benefit from transferable visual features. We examine the CLIP-ViT-Base-Patch32 model, which was downloaded from the Transformers library [73] more than one million times during the month prior to this writing alone, accounting for more than 98% of the downloads of any CLIP model from the library.

**Bias in AI.** Social biases related to gender [7], age [37], religion [1], and sexuality [63] have been observed in AI systems. Our review of the extensive related work in this area focuses on racial bias in AI and on biases observed in CLIP.

***Implicit Bias and AI.*** CLIP offers the first opportunity to study implicit biases in machine-learned visual semantic representations. Where a supervised algorithm is provided with explicit class labels, CLIP is provided, like word embeddings and language models, with text (and accompanying images) collected from the internet. While unsupervised and self-supervised algorithms free machine learning from using a relatively small number of human-curated labels, and allow for the expansion of training datasets without the need for supervised labeling, unsupervised models learn the social biases of the data on which they are trained [5, 9]. The most direct adaptation of an implicit bias measurement to machine learned representations is the Word Embedding Association Test (WEAT) of Caliskan et al. [9], which replicated in word embeddings an array of widely shared human associations and social biases previously observed in human subjects in the Implicit Association Test (IAT) [22]. The WEAT was expanded to uncover racial, gender, and intersectional biases in contextualized word embeddings formed by language models like BERT and GPT-2 [25, 76], and in the generative computer vision model Image GPT [65]. Our research is most similar to these studies, in that we test an AI model for a bias well-known in psychology.

***GANs for Bias Assessment.*** Our study builds on prior work which uses GAN-generated synthetic images to study bias in AI. Denton et al. [15] use a generative adversarial network (GAN) to generate series of counterfactual images for analyzing unintended biases in facial analysis AI. Li and Xu [45] use series of images produced by a GAN to discover unknown biases in image classifiers. Ramaswamy et al. [58] use a GAN to generate synthetic training data which is more balanced based on protected attributes such as race and gender. Most recently, Lang et al. [40] train a GAN to explain the decisions of an image classifier by discovering the visual attributes which lead to its predictions.

***Racial Bias in Computer Vision.*** Buolamwini and Gebru [8] find that state-of-the-art AI facial recognition systems fail to detect the faces of women with darker skin, and that this problem is traceable to the underrepresentation of women and people with darker skin in widely used facial recognition datasets. Wilson et al. [72] finds that this problem is not limited to facial recognition, and that state-of-the-art object detection systems also perform poorly for people with darker skin. Moreover, emotion detection AI systems have been shown to disproportionately detect unpleasant emotions such as anger in emotionally ambiguous or neutral images of members of minority racial or ethnic groups [61].[2] Steed and Caliskan [65] show that Image GPT, which generates image completions given a visual prompt [12], implicitly associates images of Black individuals with images of weapons.

***Racial Bias in NLP.*** May et al. [49] find evidence of sentence-level racial bias in language models, while Sheng et al. [63] show that the text output of language models like GPT-2 contains biases demonstrating low regard for members of marginalized groups, including racial minorities. Wolfe and Caliskan [74] find that underrepresentation of women and marginalized racial and ethnic groups in the training corpora of language models produces representations more likely to be biased and overfit to stereotypical pretraining contexts. Dodge et al. [16] show that filtering algorithms used to construct the C4 training corpus remove information by and about people belonging to marginalized populations.

***Bias in CLIP.*** Radford et al. [56] and Agarwal et al. [2] catalogue social biases in CLIP, including that images of people who are labeled as Black in the FairFace dataset [35] are the most likely to be mislabeled as animals. In highlighting multimodal neurons in CLIP which respond to photographic, symbolic, or textual depictions of a concept, Goh et al. [20] find that the activation of these neurons reflect stereotypes, such as association of terrorism with Muslims. Birhane et al. [6] find that LAION-400m [62], a multimodal image-and-language dataset similar to CLIP's training corpus, contains

[2]There are ethical questions related to whether technologies like facial recognition and emotion detection are beneficial, as they are used for purposes of surveillance and social control [42], and recent research focuses on ways to circumvent mass surveillance [24]. Where such systems are used, the research of scholars such as Buolamwini and Gebru [8] makes clear the over-representation of men and people who have lighter skin in AI training data, and the failure of commercial AI systems to perform equally for women and for people who have darker skin.

racial and ethnic slurs and stereotypes, along with images of sexual violence and other obscene adult content.

## 3 DATA

We use the Chicago Face Database (CFD), a dataset of images produced for studies of race in psychology [48]. The CFD includes 597 high-resolution (2,444 x 1,718 pixel) images of male and female study volunteers, with labels (Asian, Black, Latino/a, and White) based on self-identified race or ethnic group [48]. The subjects of the photos faced the camera, and images are standardized such that they are all set against a White background, with the face occupying the same area of the image. CFD subjects were recruited in the United States. The CFD includes photos of all subjects with a "neutral" facial expression, and of a subset with "happy (open mouth)," "happy (closed mouth)," "angry," and "fearful" facial expressions. The experiments of Ho et al. [33] used images of subjects with a neutral facial expression, and for consistency we use only the images with neutral facial expressions for this research.

**GAN-Generated Images.** Consistent with the research design of Ho et al. [33], who produce face morphs between racial or ethnic groups in 5% increments, we morph faces of people who self-identify as Black, Asian, and Latino/a into faces of people who self-identify as White in the CFD. Specifically, for each morph series, an image of a person who self-identifies as Black, Asian, or Latino/a is selected, and nineteen intermediate images are generated which blend facial features between that image and the image of a person who self-identifies as White. At each morph step, the morphed image becomes less similar to the image of the person who self-identifies as Black, Asian, or Latino/a, and more similar to the person who self-identifies as White. Using a 21-step series allows the definition of 75%, 50%, and 25% mixture ratios between the source and target images at the $6^{\text{th}}$, $11^{\text{th}}$, and $16^{\text{th}}$ morph indices respectively.

For each pair of racial or ethnic groups, we produce 1,000 unique morph series to ensure the statistical significance of our results. Let $n$ denote the number of images for a "source" racial or ethnic group in the Chicago Face Database, *i.e.,* for the group which serves as the initial image in the morph series, rather than the final image in the morph series. For each image in the source group, we randomly select $\lceil \frac{1,000}{n} \rceil$ (*i.e.,* the largest integer $\geq \frac{1,000}{n}$) images from the target group to serve as final images in the morph series. This method enables maximum variety in the morph series. Because this produces slightly more than 1,000 morph series per paired source and target group, we randomly downsample to 1,000 morph series (a total of 21,000 images) per pair to ensure uniformity between morph series comparisons. To produce highly realistic images, we use StyleGAN2-ADA[3] pretrained on the FFHQ dataset [36]. Images are normalized by cropping around the face, such that facial features appear in positions similar to the positions of the facial features in the pretraining dataset. Failing to do this produces distorted and distended faces bearing little similarity to either of the images between which the morph occurs. To produce morph series using the GAN, we train on the source image and the target image. We use the default hyperparameters of StyleGAN2-ADA, and train for 125 steps per image, beyond which we observe no benefit given the high-resolution, standardized images we provide to the GAN. This produces a source embedding, and a target embedding, for which we use the generator to produce the first and last images in the series. For the rest of the morph series, we take the difference between these two embeddings and divide it by the number of intermediate images to be produced.[4] At each step, we add the divided difference, and create a new image using the generator. This creates a series of 21 high-resolution (1,024x1,024 pixel) synthetic images, such as those seen in Figure 2. Our supplementary materials include an example of a morph series for each source and target group pair.

**Figure 2: Images of the five middle steps (40-60% mixing ratio) of a Black-White male GAN-generated morph series.**

A projected image embedding is obtained from CLIP for each of the 21 images in each morph series. Then, projected text embeddings reflecting the self-identified race or ethnicity of the source group image, the self-identified race or ethnicity of the target group image, three multiracial group labels ("multiracial," "biracial," and "mixed race"), and a "person" label (with race omitted) are obtained from CLIP. Following the prompt of Radford et al. [56], who recommend using "a photo of [[image class]]" for best performance with zero-shot CLIP, we use "a photo of a [[race or ethnicity]] person" as the text input to the model for each of the text labels.

Note that we do not morph faces across genders. That is, we morph images of men into men, and images of women into women. The reasons for this are three-fold: first, psychological research has tested morphs of men into men and women into women, and we are able to compare our results to prior work in this light. Second, prior research indicates that the rule of hypodescent is applied more to men than to women, and we intend to test this hypothesis with CLIP. Third, our experiments test for a bias based on race, rather than gender. It is unclear whether gender-related stimuli changing across a series of morphs would introduce noise which distorts the model's association based on race. Future research designs might study how CLIP's biases change when gender stimuli vary along with racial stimuli.

## 4 APPROACH AND EXPERIMENTS

We conduct three experiments. The first tests whether CLIP associates images of multiracial people with minority ancestry, according to a rule of hypodescent. The second tests whether CLIP embeds White as a default racial group. The third tests a relationship between valence association (good vs. bad) and association with a minority group label.

**Test of Hypodescent.** At each step in the morph series, we measure the percentage of morphed images for which the image's cosine

[3] A GAN may itself exhibit bias based on its training dataset; for example, if images of White individuals are overrepresented in the training data, the GAN may lighten the skin tone of generated images, or produce noisy or lower quality images of individuals with physical characteristics underrepresented in the training dataset [34]. We note that inspection of several thousand generated images by multiple domain experts suggests inter-annotator reliability of these series for quantifying bias.

[4] Steps for morphing between images are derived from the work of Heaton [28].

similarity with the minority group label is higher than with the White label. If hypodescent is reflected in CLIP, we would expect that the majority of images which blend source features and target features most equally (*i.e.,* those in the middle of the morph series) would be more similar to the embedded text corresponding to the disadvantaged group. We also characterize the skewness of each distribution of associations, defined as $\frac{m_3}{m_2^{3/2}}$, with the biased $i$th central moment given by $m_i = \frac{1}{N}\sum_1^N (x[n]-\bar{x})^i$, and $\bar{x}$ referring to the sample mean. Skewness measures of the symmetry of a distribution: negative skewness indicates that the distribution of associations leans toward the minority racial or ethnic group, while positive skewness indicates a distribution that leans toward White.

**Test of White as a Default Race.** For the 21,000 images of each morph series (1,000 images at each morph step), we obtain the cosine similarity with the source (Asian, Black, or Latino/a) group text label, the White text label, and a person label with no race or ethnicity included. We then take Pearson's $\rho$ between the cosine similarities for the person label with those for the minority group label, and with those for the White label. If CLIP encodes White as a default race, we would expect to observe stronger correlations between White and person than between other racial or ethnic groups and person. Next, we take the mean cosine similarity at every morph index (corresponding to a 5% morph increment) for each racial or ethnic group label with the 1,000 GAN-generated images at that morph index. Mean cosine similarity can be thought of as the average association with a text class at a morph step. To validate that results using the mean are representative of the full series, we report the standard deviation of 21,000 cosine similarities for each text label. If CLIP encodes White as a default race, we would expect to observe correspondence between the cosine similarity for White and person, with other racial and ethnic groups defined in relation to White.

**Test of Relationship Between Valence and Race or Ethnicity.** The third analysis focuses on the association of race or ethnicity with valence (good or pleasant vs. bad or unpleasant), which is fundamental to any psychological analysis of social cognition and social interaction [22, 54]. Previous research has shown that static word embeddings encode the concept of Black (represented using 25 African-American names) such that it is more associated with unpleasantness than the concept of White (represented using 25 European-American names), a result of training on large language corpora scraped from the internet [9]. This result confirmed demonstrations of human implicit social cognition showing evidence that even those individuals who report no racial bias nevertheless demonstrate automatic association of Black with bad and White with good [22]. The WEAT, and the IAT before it, measure racial bias using two sets of attribute words: one good or pleasant group, and one bad or unpleasant group. These groups correspond to the psycholinguistic property of valence, or the pleasantness or unpleasantness of a stimulus [54]. Below are the pleasant and unpleasant attribute words used in the IAT and the WEAT, which we also employ:

**Pleasant:** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

**Unpleasant:** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

In addition to measuring the standardized difference in valence between two groups of target words (*e.g.,* European-American names and African-American names), the WEAT is also able to measure the valence of a single target stimulus using the single-category WEAT (SC-WEAT). We adapt the SC-WEAT to measure the valence of a visual semantic representation of an encoded image, rather than a word. The formula for the SC-WEAT is:

$$\frac{\text{mean}_{a\in A}\cos(\vec{i},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{i},\vec{b})}{\text{std\_dev}_{x\in A\cup B}\cos(\vec{i},\vec{x})} \tag{1}$$

where $\vec{i}$ refers to the multimodal image representation, and $A$ and $B$ refer to multimodal text representations of sets of attribute words. The SC-WEAT returns an effect size, Cohen's $d$ [13], and a $p$-value denoting statistical significance. We obtain Pearson's $\rho$ between the SC-WEAT effect size (association with unpleasant vs. pleasant) for the 21,000 images, with the cosine similarities between the images and a text label corresponding to the minority racial or ethnic group (*i.e.,* the degree of association with the minority group). For this experiment, a positive effect size denotes association with unpleasantness, such that a positive correlation denotes association of a minority racial or ethnic group with unpleasantness, and a negative correlation denotes association with pleasantness.

By relying on visual, face-based stimuli, our research overcomes the limitation imposed by words, as there is little question about whether a face is a face or not, whereas words may contain multiple meanings. Moreover, both research on word embeddings and human subjects research using pictorial stimuli can use only limited inputs, typically 8 faces or 25 words, to represent social groups [14]. Using the current methodology expands the stimulus input to be two orders of magnitude greater than the inputs used in previous research with humans or word embeddings. If the present analyses support previous small-scale input studies, the inferences that can be drawn about the strength and generality of group valence associations will be significantly strengthened.

**Validating the SC-WEAT for Visual Semantics.** The WEAT and SC-WEAT are well-established methods for measuring biases in linguistic representations [9, 25, 74, 76], and have been adapted to measure bias in image embeddings by Steed and Caliskan [65]. As this is the first application of the SC-WEAT to a visual semantic model, we test that using the SC-WEAT produces human-interpretable results. For each image included in the OASIS norms, a dataset of 900 images which includes labels reflecting the human-rated valence of each image [39], we obtain an SC-WEAT effect size measuring the pleasantness of the image's embedding in CLIP. Comparing the SC-WEAT effect sizes with the human-rated valence norms yields Pearson's $\rho = 0.77$, $p < 10^{-30}$, on par with similar semantic evaluations of the best static and contextualized word embeddings on valence-labeled linguistic lexica of comparable size [68, 76]. This means, concretely, that CLIP associates images depicting war, sickness, and so on with unpleasantness; and images related to nature, family, and so on with pleasantness, and that the SC-WEAT can be used to detect these associations.

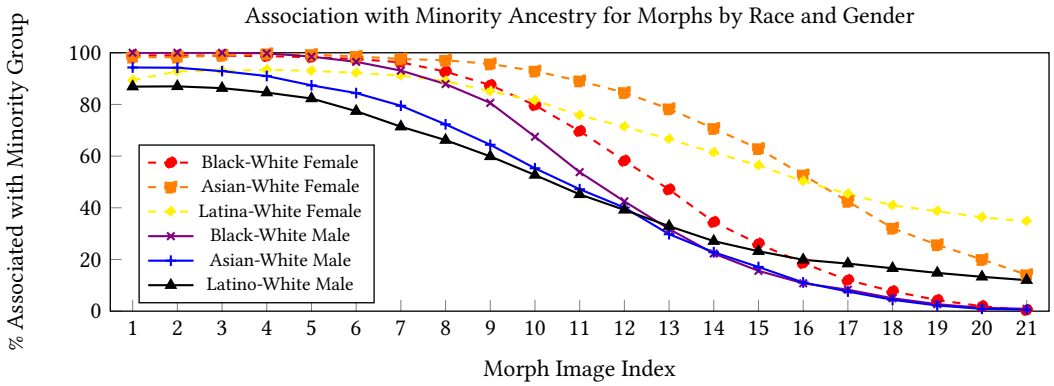


**Figure 3: Across 5% increments in mixing ratio between source (minority group) images at index 1 and target (White) images at index 21, CLIP associates a majority of images with the minority group until the image is only 20% similar to the source group for Asian-White and Latina-White Female morph series, and 40% similar to the source group for the Black-White Female morph series, indicating evidence of hypodescent for Female images. 53.5% of Black-White Male images are associated with Black at the 50% point.**

## 5 RESULTS

Our results provide evidence for hypodescent in the CLIP embedding space, a bias applied more strongly to images of women. Results further indicate that CLIP associates images with racial or ethnic labels based on deviation from White, with White as the default. Finally, an image's valence association correlates with its association with a minority racial label. For readability, some tables below refer to series of Asian-White morphed images as "A-W," of Black-White morphed images as "B-W," and of Latino/a-White morphed images as "L-W."

| Minority Group Associations by Mixing Ratio | | | | | |
|---|---|---|---|---|---|
| Morph Series | 75% | 55% | 50% | 45% | 25% |
| Asian-White Female | 98.6 | 92.9 | 89.1 | 84.6 | 52.7 |
| Black-White Female | 97.6 | 79.7 | 69.7 | 58.2 | 18.6 |
| Latina-White Female | 92.2 | 81.5 | 75.8 | 71.4 | 50.3 |
| Asian-White Male | 84.4 | 55.3 | 47.2 | 40.1 | 11.1 |
| Black-White Male | 96.5 | 67.5 | 53.8 | 42.5 | 10.8 |
| Latino-White Male | 77.4 | 52.7 | 45.2 | 39.2 | 19.9 |

**Table 1: At 50% mixing ratio, the point denoting equal similarity between the source (Asian, Black, or Latino/a) and target (White) images, CLIP associates a higher percentage of 1,000 morphed Female images with Asian (89.1%), Latina (75.8%), and Black (69.7%) text labels than with a corresponding White label, indicating hypodescent. Hypodescent is also observed for Black-White Male images (53.8% at 50% mixing ratio), but not for Asian-White Male or Latino-White Male images.**

**Evidence for hypodescent in CLIP.** For the 1,000 morph series of 21 images each between Black males and White males, CLIP assigns a higher cosine similarity to the Black text label 67.5% of the time at a 55% mixing ratio (MR) (*i.e.*, step 9 out of 20, wherein the image is 55% similar to the source (Black) image and 45% similar to the target (White) image), 53.8% of the time at 50% MR, and 42.5% of the time at 45% MR. Hypodescent is not observed at 50% MR for the Asian-White male or Latino-White male morph series. Female morph series reflect stronger biases of hypodescent, with 89.1% of Asian-White female morphed images associated with Asian at 50% MR, 69.7% of Black-White female morphed images associated with Black at 50% MR, and 75.8% of Latina-White female morphed images associated with Latina at 50% MR. The Black-White female morph series yields a majority of Black associations until 40% MR, while the Asian-White and Latina-White morph series yield a majority of Asian and Latina associations, respectively, until 20% MR. Full results for 25%, 45%, 50%, 55%, and 75% MR steps are provided in Table 1, and association curves across the six morph series are visualized in Figure 3.

| Skew by Morph Series | | | |
|---|---|---|---|
| Female Series | Skew | Male Series | Skew |
| A-W Female | -0.84 | A-W Male | 0.00 |
| B-W Female | -0.30 | B-W Male | -0.06 |
| L-W Female | -0.42 | L-W Male | 0.12 |

**Table 2: Skew is negative for the distribution of minority group vs. White label associations across all three Female morph series and the Black-White Male morph series. This reflects an asymmetry which leans toward the minority group, indicating hypodescent.**

Four of the morph series exhibit negative skewness, indicating asymmetry in the distribution of associations which leans toward a racial minority group label. Table 2 shows that results also reflect a stronger bias of hypodescent for women than for men, with a largest male series skew of -0.06 (Black-White male series), and a largest female series skew of -0.84 (Asian-White female series). The smallest female series skew is -0.30 (Black-White female series).

**Evidence that White is the Default Racial Group in CLIP.** As shown in Table 3, the correlation between the cosine similarity

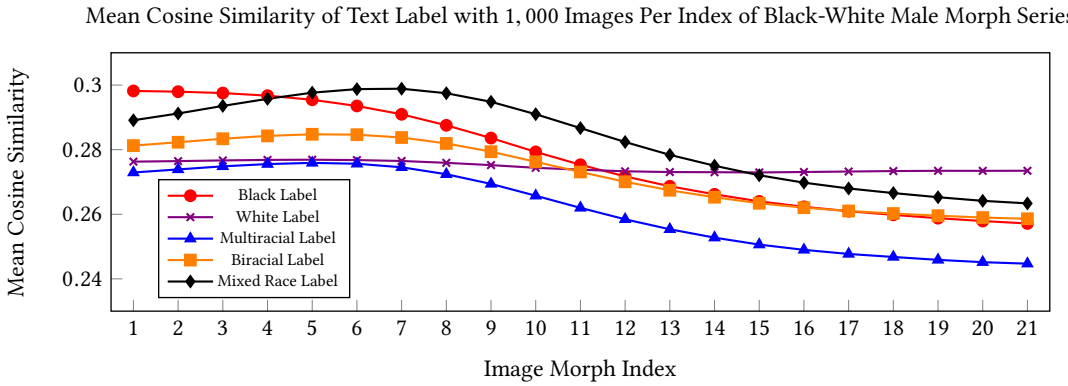


**Figure 4: Across 5% increments in mixing ratio between Black source images (index 1) and White target images (index 21), the mean cosine similarity of Multiracial, Biracial, and Mixed Race labels varies with the mixing ratio of the images in a similar manner as the Black label, instead of increasing to a maximum value at index 11 (50% mixing ratio), as would be expected if CLIP were detecting a mix of racial features. The Mixed Race label is preferred at the intermediate steps of the morph series.**

| Correlation (Pearson's $\rho$) of race or ethnicity label with "person" label (over 21,000 images) | | | | | | |
|---|---|---|---|---|---|---|
| Racial or Ethnic Group | A-W Female | B-W Female | L-W Female | A-W Male | B-W Male | L-W Male |
| Source Racial or Ethnic Group | 0.13 | 0.33 | 0.21 | 0.11 | 0.40 | 0.33 |
| White | 0.81 | 0.75 | 0.82 | 0.60 | 0.74 | 0.64 |

**Table 3: Over 21,000 images, the cosine similarity of the White label and the images strongly correlates with the cosine similarity of the person label and the images, with Pearson's $\rho$ up to 0.82. Correlation does not exceed 0.40 for any other label.**

| Standard Deviation of Cosine Similarities with 21,000 Images by Morph Series | | | | | | |
|---|---|---|---|---|---|---|
| Racial or Ethnic Group Label | A-W Female | B-W Female | L-W Female | A-W Male | B-W Male | L-W Male |
| Source Group | 0.025 | 0.019 | 0.013 | 0.028 | 0.019 | 0.014 |
| White | 0.008 | 0.010 | 0.008 | 0.007 | 0.008 | 0.007 |
| Person (Race Omitted) | 0.006 | 0.008 | 0.006 | 0.006 | 0.006 | 0.006 |

**Table 4: The standard deviation of associations between the White label and a series of 21,000 images is smaller than for any other race or ethnicity. This indicates that the probability of the White label, like the person label, is nearly constant across the steps of the morph series, suggesting that White is a default race against which other races and ethnicities are defined.**

of an image with person and cosine similarity of an image with White is higher in all morph series than the correlation between the cosine similarity of an image with person and with a minority racial group. Correlations are for 21,000 images each, and $p$-values are $< 10^{-50}$ in all cases. While the mean cosine similarities with 1,000 encoded images at each morph step vary according to mixing ratio for minority racial and ethnic groups, the mean cosine similarity of the White text label with encoded images is nearly invariant across the morph series, and nearly indistinguishable from the person text label. Table 4 captures standard deviation by text label for 21,000 images in all six morph series, and shows that standard deviations for the White label fall within $[0.007, 0.010]$, while standard deviations for Black, Asian, and Latino/a labels fall within $[0.013, 0.028]$, with all Black and Asian labels $\geq 0.019$. Standard deviations are small because cosine similarity ranges from 0 to 1, and has a narrow span for one kind of image (faces).

Adding Multiracial, Biracial, and Mixed Race labels reveals that, on average, the model associates intermediate morph steps with the Mixed Race text label rather than with Black or White, as shown in Figure 4. This result corresponds well to the research of Peery and Bodenhausen [55], who found evidence of hypodescent as an automatic, reflexive bias, but also found that humans acknowledge more complex racial identities when presented with option to choose them. However, cosine similarity with the Multiracial, Biracial, and Mixed Race text labels does not increase uniformly from both sides to a maximum at morph index 11 (50% mixing ratio), as might be expected if the label denoted that the model has perceived a blend of features. Rather, these labels have higher mean cosine similarity when the images are more similar to Black source images, and lower mean cosine similarity as images are more similar to White target images. That CLIP prefers the Mixed Race label of the three, and assigns lower probability to Multiracial than even the default White label, may be a consequence of the model being trained on internet image captions, such that more colloquial descriptions are preferred.

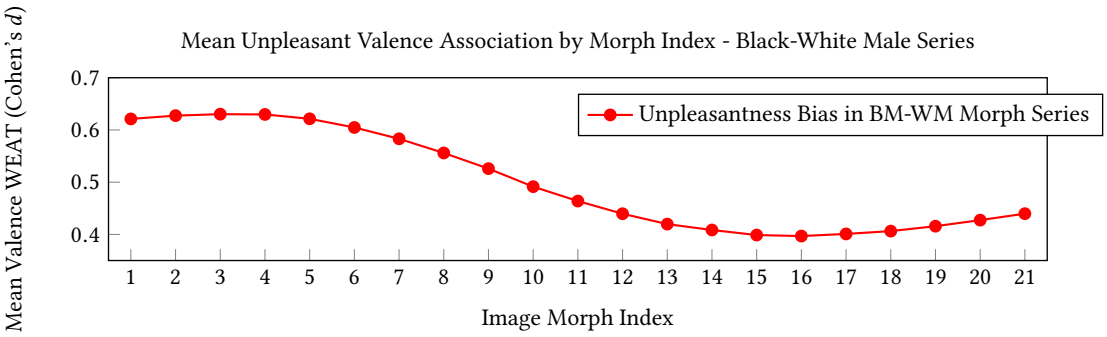


**Figure 5: Images more similar to the Black Male source images (index 1) are more associated with unpleasantness than images similar to the White Male target images (index 21). Correlation between the mean association with the "Black" label and mean valence association across the morph series is Pearson's $\rho = .97, p < 10^{-12}$. This indicates that multiracial people are subject to valence bias in CLIP, and suggests a link between the model's perception of an image as "Black" and its automatic association with unpleasantness.**

**Evidence that Valence Correlates with Minority Group Association.** As depicted in Figure 5, mean valence association (association with bad or unpleasant vs. with good or pleasant) varies with the mixing ratio over the Black-White male morph series, such that CLIP encodes associations with unpleasantness for the faces most similar to CFD volunteers who self-identify as Black. As shown in Table 5, for the 21,000 images of the Black-White male morph series, Pearson's $\rho$ between valence association and the cosine similarity of an image with the Black text label is 0.48, $p < 10^{-90}$; $\rho$ between mean valence association and mean cosine similarity with the Black label at each morph step is $\rho = 0.97, p < 10^{-12}$. A similar effect is observed for the Black-White female morph series, with $\rho = 0.41, p < 10^{-90}$ over 21,000 images. The negative correlations for Asian-White morph series indicate that Asian faces are more associated with pleasantness relative to White faces.

| Correlation of Minority Group Label with Unpleasantness | | | |
|---|---|---|---|
| Female Series | Pearson's $\rho$ | Male Series | Pearson's $\rho$ |
| A-W Female | -0.22 ($10^{-90}$) | A-W Male | -0.43 ($10^{-90}$) |
| B-W Female | 0.41 ($10^{-90}$) | B-W Male | 0.48 ($10^{-90}$) |
| L-W Female | 0.05 ($10^{-13}$) | L-W Male | 0.01 (.26) |

**Table 5: The valence association of an image correlates with the probability that it will be associated with a minority group text label. Most significantly, the unpleasantness association of an image increases with the probability that the model will associate the image with Black, with Pearson's $\rho = 0.48, p < 10^{-90}$ across 21,000 images for the Black-White Male morph series.**

## 6 DISCUSSION

**The evidence indicates a rule of hypodescent, or one-drop rule, in CLIP.** The effect occurs in four of six morph series, and is more pronounced for images of women than images of men, such that the largest negative skew for male images is -0.06, for the Black-White male morph series, while the skew of associations for female images ranges between -0.30 (Black-White female series) and -0.84 (Asian-White female series). The three female morph series do not see a majority of associations with White until 40% mixing ratio (Black-White series) or 20% mixing ratio (Asian-White and Latina-White series). We describe a few possible causes for a primary effect based on gender. First, Radford et al. [56] observe that CLIP directs more attention to the physical features of women (such as hair), likely the result of training captions which describe women in terms of physical appearance. Such attention to physical features may render CLIP more sensitive to what the model perceives as race in women. On the other hand, Halberstadt et al. [26] have suggested that hypodescent in human subjects is related to attention, such that a perceiver notices deviation from the physical features most frequently observed. Prior research in computer vision and in NLP finds that women with darker skin [8] or who are Black, Asian, or Latina [74] are underrepresented in machine learning training data. Hypodescent as it pertains to women may thus be related to underrepresentation in CLIP's WIT training corpus.

**The evidence indicates that CLIP encodes White as a default race.** This is supported by the stronger correlations between White cosine similarities and person cosine similarities than for any other racial or ethnic group, with Pearson's $\rho \in [0.60, 0.82]$. Standard deviations for both person and White labels are also much lower than for Asian, Black, or Latino/a labels. This accords with the near invariance of the means of the White label across all steps of the morph series, such that images are associated with a minority racial or ethnic group based on deviation from White. That White is a default race in CLIP is further supported by image associations with Multiracial, Biracial, and Mixed Race text labels, which do not increase from both sides of the series to a peak at 50% mixing ratio, but are higher when the image is similar to an image of a Black person, and lower when more similar to a White person.

**The evidence indicates that the valence of an image correlates with racial association**, with Pearson's $\rho = 0.48, p < 10^{-90}$ for images in the Black-White male series. More concretely, our results indicate that the more certain the model is that an image reflects a Black individual, the more associated with the unpleasant

embedding space the image is. We note three consequences of this result. First, it indicates that multiracial people whom the model associates more strongly with Black are subject to valence biases in visual semantic AI. Second, it suggests that an implicit bias - of Black individuals with unpleasantness - is linked to the perception of race in visual semantic AI. This accords with research of Zarate and Smith [79], who found that the speed of racial categorization by human perceivers is linked to the attribution of racial stereotypes to a subject. Third, this result affirms the results of smaller studies of implicit bias in word embeddings and in human subjects.

Observing a correlation between pleasantness and probability of the Asian text label may correspond to the "model minority" stereotype, wherein people of Asian ancestry are lauded for their upward mobility and assimilation into American culture, and even associated with "good behavior" [77]. This is likely a stereotype held in conjunction with Asian individuals being perceived and marked as outsiders [41], and the strong negative response during the pandemic [66]. While this is consistent with CLIP's association of faces of Asian people relative to the White default group, further work is needed to confirm the underlying positive association to Asian and its dissociation from acts of discrimination.

**Impact on AI.** As indicated by the title of the paper introducing CLIP, the model is designed to learn "transferable" visual features, *i.e.,* features which can be used in a wide variety of settings and applications [56]. One of these uses is to serve as ground truth for creating other specialized multimodal models: among the first uses of CLIP was to train the zero-shot image generation model DALL-E [56, 60]. A larger, non-public version of the CLIP architecture was used in the training of DALL-E 2 [59]. Commensurate with the findings of the present research, the Risks and Limitations described in the DALL-E 2 model card note that it "produces images that tend to overrepresent people who are White-passing" [59]. Such uses demonstrate the potential for the biases learned by CLIP to spread beyond the model's embedding space, as its features are used to guide the formation of semantics in other state-of-the-art AI models. Moreover, due in part to the advances realized by CLIP and similar models for associating images and text in the zero-shot setting, multimodal architectures have been described as the foundation for the future of widely used internet applications, including search engines [52]. Our results indicate that additional attention to what such models learn from natural language supervision is warranted.

**Impact on the Sciences.** Our results suggest that visual semantic AI models like CLIP may prove useful tools for studying societal-level biases. For example, word embeddings have been used to study the consistency of gender stereotypes across child and adult language corpora [11], as well as changes in gender and ethnic stereotypes over one hundred years [18]. Visual semantic AI may facilitate similar scientific research enabling the study of human biases which can be observed in the statistical relationship between language and images.

**Impact on Society.** As human interaction with AI systems becomes more common, hypodescent and valence bias in visual semantic AI may affect the human perception of multiracial individuals. Greenwald et al. [21] find that implicit racial biases in humans can "explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons." Given their ubiquity and ease of use, modern AI applications may, on an unprecedented scale, affect the way human beings perceive other human beings.

**Limitations and Future Work.** This research examines only one visual semantic model, which prevents assessment of whether our results generalize across architectures. CLIP is the first in a new generation of multimodal AI, and until late December 2021 was the only English-language zero-shot visual semantic model available open source for scientific study. Further work will be necessary to test hypodescent in new zero-shot visual semantic models such as SLIP [50], and in other CLIP architectures. Moreover, CLIP creates highly contextual text representations, and using different labels may produce different results. We employ a principled method by adhering to the prompt outlined by Radford et al. [56], and operationalizing only the racial and ethnic categories specified in the CFD. 2019 Google N-Gram frequency statistics indicate that these are also by far the most common ways to describe each racial or ethnic group in all English N-Gram sources [46]. Nonetheless, using different text prompts is likely to affect associations in the model, and might be explored in future work. Future work might also use a GAN-based approach such as that of Lang et al. [40] to test what characteristics of an image directly influence hypodescent and valence bias. Such a study might test our hypothesis that increased attention to physical features results in stronger bias of hypodescent for women, a finding which varies from studies of hypodescent in human perceivers, for which hypodescent is observed more strongly for men [33, 51, 64].

## 7 CONCLUSION

The primary result of the present research is the demonstration of hypodescent in a state-of-the-art visual semantic AI system. Additionally, the results reflect that women are more likely to be classified according to a rule of hypodescent, that White is a default race in the model, and that stereotype-congruent pleasantness bias correlates with association with Black. This work adds to a body of literature showing that AI encodes the implicit and explicit biases of the society and language on which it trains.

## ACKNOWLEDGMENTS

This material is based on research partially supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB-20D212T. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST.

## REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783* (2021).

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *arXiv preprint arXiv:2108.02818* (2021).

[3] American Medical Association et al. 2020. New AMA policies recognize race as a social, not biological, construct. *Published November* 16 (2020).

[4] Carlos A Ball. 2007. The Blurring of the Lines: Children and Bans on Interracial Unions and Same-Sex Marriages. *Fordham L. Rev.* 76 (2007), 2733.

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[10] Colleen M Carpinella, Jacqueline M Chen, David L Hamilton, and Kerri L Johnson. 2015. Gendered facial cues influence race categorizations. *Personality and Social Psychology Bulletin* 41, 3 (2015), 405–419.

[11] Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science* 32, 2 (2021), 218–240.

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International Conference on Machine Learning*. PMLR, 1691–1703.

[13] Jacob Cohen. 1992. Statistical power analysis. *Current directions in psychological science* 1, 3 (1992), 98–101.

[14] Nilanjana Dasgupta, Debbie E McGhee, Anthony G Greenwald, and Mahzarin R Banaji. 2000. Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology* 36, 3 (2000), 316–328.

[15] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. 2019. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439* (2019).

[16] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the english colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[18] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[19] Phillip Atiba Goff, Margaret A Thomas, and Matthew Christian Jackson. 2008. "Ain't I a woman?": Towards an intersectional approach to person perception and group-based harms. *Sex Roles* 59, 5 (2008), 392–403.

[20] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill* 6, 3 (2021), e30.

[21] Anthony G Greenwald, Mahzarin R Banaji, and Brian A Nosek. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. (2015).

[22] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. *arXiv preprint arXiv:2104.13921* 2 (2021).

[24] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, and Yuval Elovici. 2021. Dodging Attack Using Carefully Crafted Natural Makeup. *arXiv preprint arXiv:2109.06467* (2021).

[25] Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 122–133.

[26] Jamin Halberstadt, Steven J Sherman, and Jeffrey W Sherman. 2011. Why Barack Obama is Black: A cognitive account of hypodescent. *Psychological Science* 22, 1 (2011), 29–33.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[28] Jeff Heaton. 2020. Applications of deep neural networks. *arXiv preprint arXiv:2009.05673* (2020).

[29] Christine B Hickman. 1996. The devil and the one drop rule: Racial categories, African Americans, and the US census. *Mich. L. Rev.* 95 (1996), 1161.

[30] Arnold K Ho, Nour S Kteily, and Jacqueline M Chen. 2017. "You're one of us": Black Americans' use of hypodescent and its association with egalitarianism. *Journal of personality and social psychology* 113, 5 (2017), 753.

[31] Arnold K Ho, Jim Sidanius, Amy JC Cuddy, and Mahzarin R Banaji. 2013. Status boundary enforcement and the categorization of black–white biracials. *Journal of Experimental Social Psychology* 49, 5 (2013), 940–943.

[32] Arnold K Ho, Jim Sidanius, Nour Kteily, Jennifer Sheehy-Skeffington, Felicia Pratto, Kristin E Henkel, Rob Foels, and Andrew L Stewart. 2015. The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new $SDO_7$ scale. *Journal of Personality and Social Psychology* 109, 6 (2015), 1003.

[33] Arnold K Ho, Jim Sidanius, Daniel T Levin, and Mahzarin R Banaji. 2011. Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. *Journal of personality and social psychology* 100, 3 (2011), 492.

[34] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. 2020. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528* (2020).

[35] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.

[36] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. In *Proc. NeurIPS*.

[37] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. 2021. Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 638–644.

[38] Amy R Krosch, Suzy J Park, Jesse Walker, and Ari R Lisner. 2022. The threat of a majority-minority US alters white Americans' perception of race. *Journal of Experimental Social Psychology* 99 (2022), 104266.

[39] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. 2017. Introducing the open affective standardized image set (OASIS). *Behavior research methods* 49, 2 (2017), 457–470.

[40] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. 2021. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. *arXiv e-prints* (2021), arXiv–2104.

[41] Stacey J Lee, Nga-Wing Anjela Wong, and Alvin N Alvarez. 2009. The model minority and the perpetual foreigner: Stereotypes of Asian Americans. (2009).

[42] James Leibold. 2020. Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement. *Journal of Contemporary China* 29, 121 (2020), 46–60.

[43] Michael B Lewis. 2016. Arguing that Black is White: Racial categorization of mixed-race faces. *Perception* 45, 5 (2016), 505–514.

[44] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2017. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*. 4183–4192.

[45] Zhiheng Li and Chenliang Xu. 2021. Discover the Unknown Biased Attribute of an Image Classifier. *arXiv preprint arXiv:2104.14556* (2021).

[46] Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*. 169–174.

[47] Jordan Liz. 2018. "The Fixity of Whiteness": Genetic Admixture and the Legacy of the One-Drop Rule. *Critical Philosophy of Race* 6, 2 (2018), 239–261.

[48] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

[49] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 622–628.

[50] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. SLIP: Self-supervision meets Language-Image Pre-training. *arXiv preprint arXiv:2112.12750* (2021).

[51] Carlos David Navarrete, Melissa M McDonald, Ludwin E Molina, and Jim Sidanius. 2010. Prejudice at the nexus of race and gender: an outgroup male target hypothesis. *Journal of personality and social psychology* 98, 6 (2010), 933.

[52] Pandu Nayak. 2021. MUM: A new AI milestone for understanding information. https://blog.google/products/search/introducing-mum/

[53] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv preprint arXiv:2112.10741* (2021).

[54] Charles E Osgood. 1964. Semantic differential technique in the comparative study of cultures 1. *American Anthropologist* 66, 3 (1964), 171–200.

[55] Destiny Peery and Galen V Bodenhausen. 2008. Black+ White= Black: Hypodescent in reflexive categorization of racially ambiguous faces. *Psychological Science* 19, 10 (2008), 973–977.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[58] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9301–9310.

[59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092* (2021).

[61] Lauren Rhue. 2018. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765* (2018).

[62] Christoph Schuhmann. 2021. LAION-400-Million Open Dataset. https://laion.ai/laion-400-open-dataset/

[63] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3407–3412.

[64] Jim Sidanius and Felicia Pratto. 2001. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

[65] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 701–713.

[66] Hannah Tessler, Meera Choi, and Grace Kao. 2020. The anxiety of being Asian American: Hate crimes and negative biases during the COVID-19 pandemic. *American Journal of Criminal Justice* 45, 4 (2020), 636–646.

[67] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

[68] Autumn Toney-Wails and Aylin Caliskan. 2021. ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. *Empirical Methods in Natural Language Processing (EMNLP)* (2021).

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[70] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. *arXiv preprint arXiv:2109.05433* (2021).

[71] Isabel Wilkerson. 2020. *Caste (Oprah's Book Club): The origins of our discontents*. Random House.

[72] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[73] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[74] Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 518–532.

[75] Robert Wolfe and Aylin Caliskan. 2022. Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations. *Association for Computational Linguistics* (2022).

[76] Robert Wolfe and Aylin Caliskan. 2022. VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

[77] Ellen D Wu. 2015. *The color of success: Asian Americans and the origins of the model minority*. Vol. 100. Princeton University Press.

[78] Danielle M Young, Diana T Sanchez, Kristin Pauker, and Sarah E Gaither. 2021. A meta-analytic review of hypodescent patterns in categorizing multiracial and racially ambiguous targets. *Personality and Social Psychology Bulletin* 47, 5 (2021), 705–727.

[79] Michael A Zarate and Eliot R Smith. 1990. Person categorization and stereotyping. *Social cognition* 8, 2 (1990), 161–185.