Embodiment in Multimodal Semantics: Comparing Sensory, Emotional, and Visual Features in Chinese Color Metaphors

Wu Yufeng
City University of Hong Kong
18, Tat Hong Avenue, Kowloon, Hong Kong
Yufenwu2-c@my.cityu.edu.hk

Liu Meichun
City University of Hong Kong
18, Tat Hong Avenue, Kowloon, Hong Kong
meichliu@cityu.edu.hk

Abstract

This study examines how sensory-motor experience, emotional valence and arousal, and visual image statistics contribute to multimodal alignment in Chinese color metaphors. Using 184 metaphorical lexemes from six basic color terms, we combined textual data from the Chinese Corpus Internet (CCI 3.0) with image sets from Baidu, embedding both with Chinese-CLIP and measuring alignment using robust pooled cosine and set-to-set Chamfer metrics. Sensory-motor ratings, especially effector exclusivity and tactile strength, correlated negatively with alignment, emotional valence showed strong positive correlations, and visual color statistics (variability, entropy) correlated positively but yielded modest generalization primarily under Chamfer. Under strict 5-fold Ridge cross-validation, emotion was the only feature group with consistently non-negative out-of-fold performance, whereas sensory ratings did not generalize. The findings indicate that affective salience and perceptual richness captured by image statistics are the principal drivers of multimodal grounding for metaphorical color words, with visual contributions emerging when alignment is evaluated many-to-many.

Keywords: Embodied cognition, Multimodal semantics, Chinese color metaphors, Text-image alignment

1 Introduction

The theory of embodied cognition proposes that word meaning is grounded in perceptual and motor experience (Barsalou, 2010; Glenberg & Kaschak, 2002; Pulvermüller, 2005). Decades of behavioral and neuroimaging research have shown that accessing word meaning can reactivate sensory-motor systems, and large-scale sensory-motor rating norms have quantified embodiment across thousands of concepts (Barsalou, 2010; Connell & Lynott, 2012; Glenberg & Kaschak, 2002; Lynott et al., 2020; Zhong et al., 2022). However, embodiment is multidimensional. Alongside sensory-motor grounding, emotion provides another pathway, with evidence that affective dimensions such as valence and arousal strongly shape lexical processing and memory (Vigliocco et al., 2009; Xu et al., 2022). Vision adds yet another layer: beyond whether a concept has visual attributes, measurable image statistics such as color entropy and variability can influence semantic representation and multimodal alignment (Jonauskaite et al., 2020; Palmer & Schloss, 2010; Radford et al., 2021; Vigliocco et al., 2009).

Despite these advances, it remains unclear how sensory, emotional, and visual factors compare in their relative contribution to cross-modal semantics, particularly in metaphorical language. Color metaphors in Chinese provide an ideal testing ground: they are rich in cultural meanings, widely represented in textual and visual data, and closely tied to both perceptual and affective associations. For instance, *lán tú* 蓝图 ‘blueprint’

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 249–257

conveys planning and foresight, while *lǜ mào* 绿帽 'green hat' carries strong emotional connotations of betrayal.

This study therefore asks: Which embodied dimensions provide reliable signal, correlationally and out-of-fold, for alignment between linguistic and visual representations of Chinese color metaphors? To address this, we constructed a multimodal dataset of 184 color-derived lexemes, combining textual contexts from the Chinese Corpus Internet (CCI 3.0) (Liangdong Wang et al., 2024) with images retrieved from Baidu Image. Using Chinese-CLIP embeddings, we measured text–image alignment with robust pooled cosine and set-to-set Chamfer metrics, and integrated three types of features: sensory-motor ratings, emotional ratings, and image-based visual statistics. Through correlation and strict 5-fold Ridge cross-validation, we assessed both association and out-of-fold predictive power of these dimensions, including ΔR^2 contrasts to test incremental contributions, aiming to clarify the relative roles of sensory, affective, and visual factors under robust pooling and set-to-set alignment metrics, and to quantify their generalization with cross-validated models.

2 Literature review

2.1 Sensory Experience and Embodied Semantic Representations

Embodied cognition theory posits that semantic representations are partly grounded in past sensory-motor experiences, and activating word meaning will (re)engage perceptual and motor systems. A wealth of evidence supports this view:

Actions or percepts congruent with language facilitate conceptual processing. For example, compatibility between a described action and a required movement speeds comprehension. Conversely, switching between modalities incurs a processing cost in both purely perceptual tasks and conceptual tasks about perceptual properties. Such cross-modal interference suggests coupling between semantic processing and modality-specific perceptual processing. Classic studies demonstrating the action-sentence compatibility effect support this idea(Glenberg & Kaschak, 2002).

Words with action- or perception-related meanings elicit modality-specific activation in motor and sensory cortices. For instance, action words like kick or lick somatotopically activate corresponding motor regions for legs or tongue. Likewise, visual or auditory words activate occipital or temporal sensory areas (Binder & Desai, 2011; Hauk et al., 2004). Such findings reveal a systematic overlap between conceptual and perceptual brain networks, consistent with partially "embodied" semantic representations.

Large datasets have quantified the perceptual and action associations of words. In English, modality exclusivity norms rate the strength of a concept's association with five senses (and action effectors), enabling quantification of a word's "embodiment footprint". Early work by Lynott & Connell (2020) collected ratings for hundreds of concepts across modalities. Recently, the Lancaster Sensorimotor Norms provide 11-dimensional ratings (6 sensory modalities and 5 action effectors) for ~40,000 words. These norms explain differences in concreteness, category structure, and memory advantages for certain concepts. In Chinese, a systematic database of sensory-action ratings for nouns has also been developed (Zhong et al., 2022), offering modality strengths (visual, auditory, tactile, gustatory, olfactory, etc.) and bodily effectors for each word. Such resources allow researchers to characterize a word's embodied profile in multiple modalities.

Overall, the "degree of embodiment" of a concept can be operationalized via multi-modal strength, dominant modality, or modality exclusivity. These indicators correlate with concreteness and also predict imageability, memorability, and even the topology of semantic networks (Barsalou, 2010; Binder et al., 2016; Lynott et al., 2020). In multimodal tasks like image-text retrieval, incorporating sensory-motor features can complement abstract distributional vectors, improving cross-modal alignment and model interpretability. Studies have found that adding modality-specific information (e.g. visual or motor features) to word embeddings enhances performance on cross-modal matching and provides more human-interpretable alignments (Lynott et al., 2020; Shutova et al., 2016).

2.2 Emotional Dimensions and Embodied Cognition

Emotion is another key axis of embodied experience that shapes semantic representation. The classic valence-arousal model (Russell, 1980) describes emotions in a two-dimensional space

(valence: positive-negative, and arousal: high-low activation). These affective dimensions are tightly coupled with attention, memory, and decision-making processes:

Psychological and neural evidence: Emotional valence and arousal modulate cognitive processing speed, memory retention, and selective attention. Positively or negatively valenced words can be processed more quickly depending on context, and high-arousal content tends to be remembered in greater detail (Kensinger, 2009). Neuroimaging and lesion studies reveal distinguishable neural signatures for valence and arousal—for example, the amygdala and ventromedial prefrontal cortex track affective intensity, while regions of prefrontal and cingulate cortex differentiate positive vs. negative valence ((Lindquist et al., 2012). Such findings suggest that emotional dimensions are instantiated in the brain's affective networks, creating an "emotional fingerprint" for semantic stimuli.

Semantic and distributional evidence: A growing body of work indicates that a word's valence and arousal ratings correlate with its position in distributional semantic space and with how it aligns to visual representations. For instance, words that are highly positive or highly negative cluster distinctly in word embedding spaces, and their emotional ratings predict human judgments and memory advantages (Hollis & Westbury, 2016; Recchia & Louwerse, 2015). High-arousal words often have more and stronger associations in semantic networks, reflecting their attention-grabbing nature.

Resources in Chinese: Recent efforts have produced emotion norms for thousands of Chinese words. Xu et al. (2022) report valence and arousal ratings for over 11,000 simplified Chinese words, with analyses of gender differences in ratings. These resources enable researchers to introduce emotion features at the word or instance level in multimodal alignment tasks. For example, one can ask whether positively valenced or high-arousal words align more easily with certain image content, or if emotional congruence between caption and image boosts alignment.

Compared to concrete sensory-motor features, emotion may be more influenced by cultural and subjective factors. However, emotion strongly influences what we attend to and remember (Kensinger, 2009). In an embodied cognition framework, emotion can be seen as an encapsulation of "embodied-social experience," complementing sensory dimensions. Together, sensory and affective features jointly determine a concept's imageability, memorability, and ease of cross-modal association. In other words, a concept rich in perceptual detail and emotional salience is more likely to be vividly visualizable and easily paired with corresponding images, providing a strong signal for image-text alignment.

2.3 Visual Information and Embodiment in Multimodal Alignment

Vision is often highlighted as a dominant embodied modality. Beyond asking whether a concept "has visual attributes," researchers are examining how measurable image statistics (brightness, color entropy, color diversity, contrast, etc.) relate to semantic representations. Several lines of inquiry demonstrate the importance of visual features:

Low-level visual statistics like color and texture carry semantic and affective connotations. Colors can imply category or function (e.g. green for plants), evoke emotions (e.g. red for anger or love), or convey symbolic meanings (Jonauskaite et al., 2020; Palmer et al., 2013). For example, warm colors (reds/yellows) are often associated with positive valence and high arousal, whereas cool colors (blue hues) tend to correlate with calmer, lower-arousal feelings. Similarly, an image's color diversity and complexity can suggest "liveliness" or conceptual richness, potentially affecting interest and memorability. Thus, concrete concepts with strong visual features might also carry consistent emotional tones (e.g., a "sunset" is visually warm and often deemed pleasant).

Modern image-text models (e.g. CLIP) implicitly capture some color and contrast information, but these can still sway alignment. Research shows that certain models behave like "bag-of-words," lacking relational understanding of image content and instead relying on object presence or overall appearance. Visual factors like brightness or saturation can sometimes confound image-text similarity if not accounted for. Using perceptually uniform color spaces (such as JzAzBz) allows more consistent quantification of image attributes like mean brightness or color entropy, which can be related to language features in an interpretable way. For instance, one might find that images with extremely high brightness are harder to align with captions due to reduced contrast, or that captions with highly concrete nouns align

better with images having greater color variability (indicating more objects or details). As noted by Radford et al. (2021) and follow-up analyses, certain visual properties can either facilitate or impede cross-modal matching: a richly colored, high-contrast image may provide more "hooks" for semantic alignment, whereas an overexposed image might be less distinguishable in a joint embedding space.

Visual statistics intersect with sensory and emotional dimensions. A visually striking image (e.g., with high color variance) might align better with descriptive, concrete text, effectively leveraging embodied (visual) information to improve retrieval. At the same time, visual cues also carry emotional weight—color tone can modulate perceived valence or arousal of an image, thereby affecting alignment with text that has emotional connotations. For example, an image dominated by dark, desaturated colors might align well with a negatively valenced caption (a phenomenon related to color-emotion association). Thus, visual statistics serve as both low-level perceptual evidence and high-level semantic/emotional signals. In multimodal learning, incorporating these features can enhance alignment: one study found that adding a simple colorfulness metric improved image-caption retrieval, as it captured an aspect of "visual vividness" not present in text embeddings alone (Palmer et al., 2013; Radford et al., 2021). However, certain visual extremes (e.g., extremely bright images) can reduce alignment quality by washing out distinctive features, an observation in line with human factors in perception. The key is that visual features, in concert with sensory and emotional semantic features, contribute to a concept's overall embodied signature, which in turn influences cross-modal mapping.

3 Methodology

3.1 Lexeme Selection

Textual data were drawn from the Chinese Corpus Internet (CCI 3.0) (Liangdong Wang et al., 2024), a large-scale corpus (~1,000 GB) of digital publications from Mainland China (2001-2023). From this corpus, 184 metaphorical lexemes derived from the six basic color terms (黑 hēi 'black', 白 bái 'white', 红 hóng 'red', 黄 huáng 'yellow', 蓝 lán 'blue', and 绿 lǜ 'green') were identified using the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007).

3.2 Text and Image Data

For each lexeme, up to 100 contextual sentences were extracted from CCI 3.0 and trimmed to a ±100-character window around the target word to capture its immediate context. Parallel visual data were collected from Baidu Images, with the top 100 images per lexeme retained as representative visual exemplars after basic filtering which excluding images with resolution lower than 200x200. The text and image sets are unpaired and serve as multimodal exemplars of the same lexeme rather than item-aligned pairs.

3.3 Multimodal embeddings and alignment

Texts and images are encoded with Chinese-CLIP (ViT-L/14). To reduce sensitivity to outliers and sampling noise, we aggregate the set of text embeddings and the set of image embeddings for each lexeme using robust pooling: (i) a 10% trimmed mean and (ii) the medoid (the exemplar with minimal average cosine distance). Beyond pointwise cosine between pooled vectors, we evaluate set-to-set alignment on the full cross-modal similarity matrix $S = TI^{\top}$ after L2 normalization. We report four alignment metrics used in the Results: trimmed-cosine, spherical-cosine, agreement-weighted cosine, and bi-directional Chamfer (the average of the per-text maxima and the per-image maxima in S).

3.4 Sensory and Emotional Features

Two external rating databases were used to characterize lexemes. The Chinese Noun Sensory-Motor Norms (Zhong & Zhang, 2022), which provide ratings across six sensory modalities (visual, auditory, olfactory, gustatory, tactile, interoceptive) and associated motor effectors. The Simplified Chinese Affective Lexicon (Xu et al., 2022), which provides valence and arousal ratings on 11,310 words, with gender-specific and overall averages. Lexemes were matched to these databases to obtain multidimensional sensory and emotional ratings.

3.5 Visual Features

From the collected images, low-level visual statistics were computed in the JzAzBz perceptual color space, including average luminance, color variability, color entropy, and colorfulness. These measures captured perceptual diversity and distributional properties of the lexemes' visual exemplars.

3.6 Statistical Analyses

We assess associations and predictive power in two steps that are reported in the Results. (1) Correlation. Pearson's r between alignment scores and individual features; correlations use the trimmed-cosine alignment score. (2) Predictive modeling. Ridge regression with 5-fold cross-validation; we report CV-R^2 against two baselines: MeanBaseline (zero point) and RandomBaseline (expected ≈ -1). Models are fit for single-modality and combined feature sets, and we report ΔR^2 for nested contrasts under each alignment metric to quantify incremental value.

4 Result

4.1 Correlation Analysis

feature	r	p
sens_effector_exclusivity	-0.213	0.004
sens_tactile	-0.195	0.008
sens_max_action	-0.133	0.073
sens_auditory	0.130	0.079
sens_gustatory	-0.123	0.096
sens_concreteness	0.106	0.153
sens_head	-0.086	0.246
sens_olfactory	-0.077	0.298
sens_max_sensorimotor	-0.057	0.446
sens_perceptual_mean	-0.052	0.482
sens_max_perceptual	-0.052	0.486
sens_mouth/throat	-0.048	0.516
sens_exclusivity_sensorimotor	-0.043	0.559
sens_visual	0.042	0.568
sens_interoceptive	-0.033	0.654
sens_action_mean	-0.027	0.719
sens_torso	0.011	0.887
sens_modality_exclusivity	-0.009	0.902
sens_leg/foot	0.008	0.914
sens_hand/arm	0.002	0.976

Table 1: Pearson correlations between sensory-motor features and text-image alignment.

Table 1 reports the Pearson correlation coefficients (r) and significance levels (p) between sensory-motor dimensions and text-image alignment. Overall, most features showed weak associations with alignment, but a few dimensions yielded significant or near-significant effects.

The strongest effect was found for effector exclusivity (sens_effector_exclusivity), which correlated negatively with alignment (r = -0.213, p = .004). This suggests that lexemes characterized by greater specificity in their action effectors (e.g., strongly tied to one particular body part) tended to achieve weaker text-image alignment. In other words, highly specialized motor grounding may hinder the integration of visual and linguistic representations.

A second robust result was observed for the tactile dimension (sens_tactile), which also showed a significant negative correlation (r = -0.195, p = .008). Lexemes strongly grounded in tactile experience aligned less well across modalities, likely because tactile sensations are inherently difficult to represent visually.

Several additional features displayed marginal effects. Maximum action ratings (sens_max_action) were weakly negatively correlated with alignment (r = -0.133, p = .073), while auditory strength (sens_auditory) showed a small positive correlation (r = 0.130, p = .079), both trending toward significance. Similarly, gustatory strength (sens_gustatory) trended negatively (r = -0.123, p = .096). Although modest, these findings suggest that auditory and gustatory experiences may exert limited influence on cross-modal integration.

By contrast, most other sensory indices, such as visual (r = 0.042, p = .568), olfactory (r = -0.077, p = .298), and interoceptive (r = -0.033, p = .654), showed correlations close to zero and did not approach significance. Likewise, global embodiment indices including perceptual mean (sens_perceptual_mean) and action mean (sens_action_mean) were nonsignificant, indicating that broad averages of sensory grounding do not strongly predict alignment. These results highlight that specific embodied channels, rather than overall sensory strength, are the key drivers of cross-modal variation.

feature	r	p
emo_Women_Valence_Mean	0.448	0.000
emo_Valence_Mean	0.445	0.000
emo_Men_Valence_Mean	0.444	0.000
emo_Men_Arousal_Mean	0.283	0.000
emo_Arousal_Mean	0.230	0.002
emo_Women_Arousal_Mean	0.195	0.008

Table 2: Pearson correlations between emotional features and text–image alignment.

Table 2 presents the Pearson correlations between emotional dimensions and text-image alignment. In contrast to the sensory domain, the emotional features demonstrated consistently strong and positive relationships with alignment, particularly for valence ratings.

The valence dimension emerged as the most reliable predictor. Regardless of whether ratings were drawn from men, women, or the overall mean, the correlation coefficients were nearly identical ($r \approx 0.45$, $ps < .001$). This indicates that lexemes with more positive affective connotations systematically aligned better across modalities.

The arousal dimension also produced significant positive effects, though the effect sizes were smaller than those for valence. Male arousal ratings showed the strongest association ($r = 0.283$, $p < .001$), followed by the overall mean ($r = 0.230$, $p = .002$) and female arousal ratings ($r = 0.195$, $p = .008$). Collectively, these findings demonstrate that emotional positivity and activation jointly facilitate multimodal alignment, with valence providing the dominant contribution.

feature	r	p
ColorVariabilityBz	0.311	0.000
Colorfulness	0.208	0.005
ColorEntropyBz	0.205	0.005
ColorVariabilityAz	0.170	0.021
HueAngle	-0.157	0.033
ColorEntropyAz	0.148	0.045
AverageColorJz	-0.120	0.105
ColorEntropyJz	0.112	0.130
ColorContrast	0.081	0.275
AverageColorBz	-0.056	0.447
AverageColorAz	0.013	0.862
ColorVariabilityJz	-0.008	0.915

Table 3: Pearson correlations between visual color features and text–image alignment.

Table 3 lists the correlations between image-based visual features and text-image alignment. Compared to the sensory and emotional results, the visual features displayed a more mixed pattern, with some robust positive predictors alongside negative or nonsignificant effects.

The most prominent predictor was color variability along the Bz dimension (ColorVariabilityBz), which correlated moderately and positively with alignment ($r = 0.311$, $p < .001$). Two additional features—colorfulness ($r = 0.208$, $p = .005$) and color entropy in the Bz dimension (ColorEntropyBz, $r = 0.205$, $p = .005$)—also showed significant positive correlations. Together, these results suggest that lexemes whose associated images contain richer and more varied color distributions tend to achieve better cross-modal alignment.

More modest but still significant effects were found for color variability ($r = 0.170$, $p = .021$) and color entropy ($r = 0.148$, $p = .045$) in the Az dimension. These indicate that diversity in the color distribution, even along secondary axes, can enhance semantic-visual consistency.

On the other hand, some features exhibited negative or null associations. Hue angle (HueAngle) was significantly negatively correlated with alignment ($r = -0.157$, $p = .033$), implying that large deviations in hue may disrupt the semantic fit between text and images. Average lightness (AverageColorJz) showed a nonsignificant negative trend ($r = -0.120$, $p = .105$), while other mean color measures (e.g., AverageColorAz, AverageColorBz) and contrast did not contribute meaningfully ($|r| < .1$, ns).

4.2 Regression analysis

We assess generalization with 5-fold Ridge CV-R^2 under four alignment metrics (trimmed cosine, spherical cosine, agreement-weighted cosine, and set-to-set Chamfer). Results are summarized in Table 4. Among single-modality models, Emotion is the only group that achieves positive out-of-fold performance on the cosine family, reaching its best value with trimmed cosine (CV-R^2=0.05). Visual features alone do not surpass the mean predictor on cosine metrics but become informative when alignment is evaluated at the set level: under Chamfer the visual model attains CV-R^2=0.04. Sensory norms fail to generalize on all metrics (CV-$R^2 \leq 0$), consistent with their limited predictive value in this task.

set	cosine_trimmed	cosine_spherical	cosine_agreew	chamfer_bi
MeanBaseline	0.00	0.00	0.00	0.00
Random Baseline	-1.02	-1.10	-0.96	-0.95
sensory	-0.03	-0.05	-0.05	-0.19
emotion	0.05	0.02	-0.01	-0.07
visual	-0.05	-0.05	-0.06	0.04
sensory+ emotion	0.05	0.01	-0.01	-0.18
sensory+ visual	-0.03	-0.04	-0.06	-0.06
emotion +visual	0.00	-0.03	-0.03	0.01
all_three	0.02	-0.01	-0.02	-0.05

Table 4: Cross-validated R^2 values for regression models with different feature sets.

Multi-modality patterns reinforce these observations. Combining Sensory+Emotion does not improve over Emotion on trimmed cosine (0.05 vs. 0.05) and remains near zero or negative elsewhere, indicating that any apparent gains are carried by the emotional dimensions. Emotion+Visual is near zero on cosine metrics but turns positive under Chamfer (0.01). The full model does not outperform the best single or two-way combinations (e.g., −0.05 on Chamfer). Considering nested contrasts clarifies the incremental value: under Chamfer, adding Visual to Emotion yields $\Delta R^2(\mathrm{EV}-\mathrm{E}) = +0.08$, whereas the same addition is non-beneficial on cosine metrics (−0.05/−0.05/−0.02 for trimmed/spherical/agree-weighted). Thus, affective salience is the most reliable cross-modal signal, while perceptual statistics contribute a small but robust additional component specifically in the many-to-many matching regime captured by Chamfer. Negative CV-R^2 values are reported relative to the mean-predictor zero point and indicate poorer-than-mean out-of-fold prediction.

Table 5 reports nested contrasts that isolate the incremental value of each modality. Adding Emotion to Sensory consistently improves performance on all alignment metrics (SE−S: +0.07 trimmed, +0.06 spherical, +0.04 agree-weighted, +0.01 Chamfer), confirming that the gains attributed to the SE model are carried by the emotional dimensions. The reverse contrast is zero or negative (SE−E: 0.00/−0.01/0.00/−0.11), indicating that Sensory contributes no unique information beyond Emotion and can even harm generalization under set-to-set evaluation.

contrast	cosine_trimmed	cosine_spherical	cosine_agreew	chamfer_bi
SE - S	0.07	0.06	0.04	0.01
SE - E	0.00	-0.01	0.00	-0.11
SV - S	0.00	0.01	-0.01	0.13
SV - V	0.02	0.01	0.01	-0.10
EV - E	-0.05	-0.05	-0.02	0.08
EV - V	0.04	0.02	0.03	-0.03
ALL - SE	-0.03	-0.03	-0.02	0.13
ALL - SV	0.05	0.02	0.03	0.01
ALL - EV	0.02	0.01	0.01	-0.06

Table 5: Incremental ΔR² comparisons across feature set combinations.

The behavior of Visual depends on the alignment regime. When alignment is measured at the set level, Visual augments Sensory (SV−S: +0.13 under Chamfer), whereas adding Sensory to Visual degrades performance in the same regime (SV−V: −0.10), again pointing to the limited utility of generalized sensory norms. Critically, the Emotion+Visual contrast shows that Visual provides a small but reliable addition over Emotion only for Chamfer (EV−E: +0.08), while the cosine metrics yield non-beneficial or slightly negative increments (−0.05/−0.05/−0.02). Thus, perceptual statistics contribute additively in the many-to-many matching setting captured by Chamfer, but not in pointwise cosine alignment.

Three-way models mirror these patterns. Relative to SE, adding Visual produces a noticeable improvement under Chamfer (ALL−SE: +0.13) but not under the cosine metrics (−0.03/−0.03/−0.02). Relative to SV, adding Emotion yields small positive increments for the cosine metrics (+0.05/+0.02/+0.03) and only a negligible change for Chamfer (+0.01). In contrast, augmenting the EV model with Sensory is consistently unhelpful (ALL−EV: +0.02/+0.01/+0.01 but −0.06 under Chamfer). Taken together with Table 4, these contrasts substantiate a clear hierarchy: Emotion is the only modality that generalizes on its own; Visual adds limited but reproducible value when alignment is evaluated at the set level; and Sensory norms do not provide unique predictive power under strict out-of-fold testing.

5 Discussion

5.1 Competing Pathways of Embodiment in Chinese Color Metaphors

Our findings reveal a clear asymmetry between sensory–motor and emotional pathways in predicting cross-modal alignment for Chinese color metaphors. Sensory features such as effector exclusivity and tactile grounding showed negative correlations with alignment, and global perceptual indices were nonsignificant. Ridge regression with strict 5-fold cross-validation (with Mean/Random baselines) further indicated that sensory features achieved non-positive CV-R^2, providing no out-of-fold advantage over the mean predictor.

By contrast, emotional ratings exhibited strong and consistent positive correlations ($r \approx .45, p < .001$) and were the only feature group to reach

consistently non-negative generalization across alignment metrics (best under trimmed-cosine, CV- $R^2 \approx 0.05$). Taken together, these results favor affective grounding over purely sensory–motor simulation for metaphorical color lexemes: affective salience affords modest but robust predictive leverage under conservative evaluation, converging with psycholinguistic evidence that affect facilitates lexical processing and memory (Barsalou, 2008; Glenberg & Kaschak, 2002; Pulvermüller, 2005). For items such as *lǜ mào* 绿帽 'betrayal', affective resonance appears to be a more reliable grounding mechanism than narrowly defined sensory-motor associations.

5.2 Generalized Experience vs. Concrete Perception

A second contrast concerns norm-based sensory ratings versus image-derived visual statistics. Whereas sensory norms largely failed to predict alignment, several low-level color statistics—notably color variability and entropy—showed moderate positive correlations (up to $r = .31$). In cross-validated prediction, visual features became informative primarily under the set-to-set Chamfer metric (CV-$R^2 \approx 0.04$), while performance with pooled-vector cosines was near zero. This pattern suggests that concrete perceptual richness in the image sets aligns better with metaphor semantics than generalized sensory norms, particularly when alignment is assessed in a many-to-many manner rather than by a single pooled vector. These observations accord with prior multimodal work linking perceptual complexity and color distributions to semantic and affective outcomes (Kiela et al., 2014; Jonauskaite et al., 2020).

6 Conclusion

We compared sensory–motor norms, emotional ratings, and image-derived visual statistics as predictors of text–image alignment for 184 Chinese color metaphors, using robust pooling and both pooled-vector and set-to-set alignment metrics. Across analyses, emotion—especially valence—was the only feature group that generalized reliably, yielding small but consistently non-negative CV-R^2 under strict 5-fold Ridge (best ≈ 0.05 with trimmed-cosine), and strong positive correlations ($r \approx .45$). Visual statistics (color variability/entropy) correlated positively with alignment and showed modest generalization primarily under Chamfer (≈ 0.04), indicating benefits when alignment is evaluated many-to-many. In contrast, sensory-motor norms did not generalize (CV- $R^2 \leq 0$) and were sometimes negatively related to alignment (e.g., effector exclusivity, tactile), suggesting that generalized sensory ratings are a poor proxy for the perceptual evidence supporting metaphor–image congruence.

Taken together, the results argue for a multidimensional embodiment in which affective salience provides a modest yet robust predictive pathway, complemented by perceptual diversity captured by visual statistics under set-to-set alignment. Methodologically, the study highlights the value of robust pooling and collection-level alignment metrics for multimodal semantics. Limitations include the absence of behavioral validation and the focus on a single metaphor family. Future work should broaden to additional metaphor domains, explore non-linear and interaction-aware models, and incorporate human judgments of lexeme-image congruence to triangulate the corpus-based findings.

References

Barsalou, L. W. (2008). Grounded cognition. Annu.Rev.Psychol., 59(1), 617-645.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. Topics in Cognitive Science, 2(4), 716-724.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. Cognitive Neuropsychology, 33(3-4), 130-174.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. Trends in Cognitive Sciences, 15(11), 527-536.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125(3), 452-465.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9(3), 558-565.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41(2), 301-307.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-

occurrence models of semantics. Psychonomic Bulletin & Review, 23(6), 1744-1756.

Jonauskaite, D., Parraga, C. A., Quiblier, M., & Mohr, C. (2020). Feeling blue or seeing red? Similar patterns of emotion associations with colour patches and colour terms. I-Perception, 11(1), 2041669520902484.

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. Emotion Review, 1(2), 99-113.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. Journal of Experimental Psychology: General, 143(3), 1065.

Liangdong Wang, Bowen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, Tengfei Pan, & Guang Liu. (2024). CCI3.0-HQ: a large-scale Chinese dataset of high quality designed for pre-training large language models. CoRR, abs/2410.18505.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. Behavioral and Brain Sciences, 35(3), 121-143.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3), 1271-1291.

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. Proceedings of the National Academy of Sciences, 107(19), 8877-8882.

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music-color associations are mediated by emotion. Proceedings of the National Academy of Sciences, 110(22), 8836-8841.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6(7), 576-582.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. Paper presented at the International Conference on Machine Learning, 8748-8763.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. Quarterly Journal of Experimental Psychology, 68(8), 1584-1598.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161.

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 160-170.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. Language and Cognition, 1(2), 219-247.

Xu, X., Li, J., & Chen, H. (2022). Valence and arousal ratings for 11,310 simplified Chinese words. Behavior Research Methods, 54(1), 26-41.

Zhong, Y., Wan, M., Ahrens, K., & Huang, C. (2022). Sensorimotor norms for Chinese nouns and their relationship with orthographic and semantic variables. Language, Cognition and Neuroscience, 37(8), 1000-1022.