Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Seeing Through Words: Recovering Visual Information from Chinese Word Use

Yufeng Wu
City University of Hong Kong
Xinyue Ren
City University of Hong Kong
Esther Asare
City University of Hong Kong
Meichun Liu
meichliu@cityu.edu.hk

City University of Hong Kong

Research Article

Keywords: lexical semantics, distributional semantics, visual information recovery, sensorimotor experience, Chinese

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Relations among words may provide access to visual information, as proposed by the Symbol Interdependency Hypothesis. We examined this representational premise by predicting independently constructed visual prototypes from usage-derived representations of 2,571 Chinese lexical items. Text representations were aggregated from up to 500 unmasked occurrences in CCI 3.0, and visual prototypes were constructed from selected web-search images. A ridge mapping predicted each visual target from the text representation of an item held out under nested cross-validation. Full-library mean reciprocal rank was 0.1260, compared with 0.0044 for a part-of-speech (POS) mean baseline. In an exploratory evaluation against twenty same-POS semantic neighbors, mean reciprocal rank was 0.3050, compared with 0.1558 for the POS-mean baseline. We related differences in out-of-fold target specificity to twelve independently collected sensorimotor ratings and textual and visual representation properties. Under mutual adjustment, visual and olfactory associations were positive and the interoceptive association was negative. The eleven specific dimensions added 5.87 percentage points of in-sample explained variance beyond overall embodiment and the measurement covariates. Exploratory interactions indicated a weaker visual association at higher contextual dispersion and a stronger association at higher visual split-half stability. Among the analyzed targets in the sampled Chinese Web environment, lexical representations supported target-specific visual recovery. Its accessibility varied with experiential attributes and the organization of word use and visual representations.

1. Introduction

Patterns of co-occurrence link words to other linguistic expressions used to describe entities, properties, participants, and events. A central issue in lexical semantics is what information about word referents can be inferred from these linguistic relations. The Symbol Interdependency Hypothesis proposes that relations among linguistic symbols encode regularities of the perceptual world, allowing linguistic structure to provide access to perceptual information (Louwerse 2011). Perceptual inference gives this proposal a concrete lexical formulation. Johns and Jones (2012) showed how global linguistic similarity could support the inference of missing perceptual features using words with known experiential representations. We examine the visual dimension of this proposal by asking how well visual information can be recovered from usage-derived lexical representations.

Language-to-vision studies provide direct precedents for this question. Collell et al. (2017) learned a mapping from word embeddings to visual vectors and used the predictions to construct multimodal representations. Günther et al. (2022) predicted visual representations for concrete and abstract words outside the mapping's training vocabulary, evaluating these predictions by asking participants to choose between model-selected and random images. Their study also examined concreteness and the availability of visually represented semantic neighbors as factors associated with prediction. Lexical-level recovery and its variation across words therefore form an established research problem.

The present study examines visual recovery from contextual representations of Chinese wordforms and independently sampled Web-image prototypes. A mapping learned from other words predicts a visual vector for each held-out item. Retrieval from a fixed library then evaluates whether the prediction distinguishes the corresponding target from competing lexical targets. We refer to this target-specific correspondence as predictive text–vision alignment. The linguistic question is how much information about a word's visual target is available from its representation in language.

Visual recovery also varies across lexical items. Previous work has related multimodal representations to concreteness, lexical class, and image dispersion (Hill et al. 2014; Kiela et al. 2014; Köper and Schulte im Walde 2017), while Zhou et al. (2025) examined variability and exemplar aggregation in noun–verb correspondence. These findings motivate examining the experiential attributes of words together with the organization of their linguistic and visual representations. Both characterize variation in the same recovery outcome.

Sensorimotor ratings describe the experiential attributes of the words whose visual targets are being recovered. Distributional representations have been used to recover semantic features and experiential ratings (Grand et al. 2022; Chersoni et al. 2021; Chen et al. 2026), and multimodal models have examined indirect grounding through linguistic relations (Utsumi 2022). Comparing specific dimensions with a separate overall embodiment rating allows us to distinguish general experiential strength from a differentiated experiential profile.

We use the 3,000 Chinese wordforms in the norms of Chen et al. (2026). The image-selection pipeline yielded valid computational visual prototypes for 2,571 wordforms, which formed the predictive sample. Each item had eleven specific sensorimotor ratings and a separately elicited overall embodiment rating. We examine how these attributes relate to differences in visual recovery, first separately and then together, accounting for properties of the sampled text and images. Section 3.1 describes the sample, and Online Resource 1 lists the analyzed items.

The central question is how well visual information can be recovered from lexical usage. We first test held-out retrieval from 2,571 visual candidates, then examine discrimination among semantically similar words of the same part of speech in an exploratory comparison. Experiential ratings and representation properties characterize variation in full-library recovery. Exploratory interactions examine how these associations vary with contextual dispersion and visual split-half stability. Part of speech (POS) is used for descriptive comparisons and statistical adjustment.

2. Research Background

The Symbol Interdependency Hypothesis links linguistic relations to perceptual information. Perceptual-inference models and language-to-vision studies examine how perceptual features or visual representations can be recovered from lexical relations. Research on lexical variation asks how this access differs across words. Together, these traditions motivate studying the degree of visual recovery and its relation to experiential attributes and the organization of word use and images.

2.1 Linguistic relations and perceptual inference

Distributional semantics characterizes word meaning through regularities in linguistic usage. The Symbol Interdependency Hypothesis offers an account of why such regularities can be informative about perception. Because language is used to communicate about the world, relations among words can preserve information about relations among their referents (Louwerse 2011). This premise motivates asking what perceptual information can be recovered from lexical relations. Multimodal distributional semantics integrates linguistic and visual features (Bruni et al. 2014). Perceptual inference instead asks what information can be inferred when a lexical item has a linguistic representation but lacks a corresponding experiential representation. Johns and Jones (2012) used global lexical similarity and known perceptual traces to infer missing feature vectors. Their targets included human-generated semantic features and modality-strength ratings. The distinction between combining information sources and recovering one from another places the present study in the perceptual-inference tradition.

Visual-vector prediction provides a more direct precedent for the present task. Collell et al. (2017) predicted visual vectors from word embeddings and evaluated the resulting multimodal representations on concept-similarity tasks. Günther et al. (2022) learned linear mappings to visual representations derived from image categories, comparing prototype- and exemplar-based approaches and applying the mapping to unseen words. Their behavioral evaluation tested whether the model-selected image fit a word better than a random image. A related line of work maps visual representations to linguistic labels for unseen categories, as in Lazaridou et al. (2014). These approaches connect the modalities through different prediction targets and evaluation criteria.

The present evaluation uses the identity of an observed visual prototype as the retrieval criterion. It asks whether a prediction from a held-out lexical representation contains information specific to that word's target. Collell and Moens (2018) showed that mapped vectors can retain substantial structure from the input modality. Their findings motivate distinguishing successful target retrieval from reconstruction of the visual space. Here, retrieval assesses how well the lexical representation distinguishes the correct visual target within a common candidate library.

2.2 Lexical heterogeneity in multimodal correspondence

Lexical variation in visual recovery already forms part of the direct mapping literature. Günther et al. (2022) examined the roles of concreteness and visually represented semantic neighbors in predicting images for unseen words. Related multimodal work has considered concreteness (Hill et al. 2014), image dispersion (Kiela et al. 2014), and lexical class. Köper and Schulte im Walde (2017) found that visual features contributed differently to compositionality prediction for nominal and verbal compounds. Zhou et al. (2025) examined noun–verb differences in variability and cross-modal alignment, including changes with exemplar aggregation. Together, these studies motivate a lexical account of how recovery varies, with explicit attention to the linguistic and visual material representing each item.

Contextualized representations characterize variation across a word's sampled uses alongside its aggregate lexical vector (Ethayarajh 2019; Bommasani et al. 2020). The aggregated occurrences may differ in their referents, situations, or topical associations. A visual prototype likewise summarizes selected depictions, which may be diverse yet yield a reproducible centroid. Contextual dispersion, visual dispersion, visual split-half stability, and visual separation characterize these properties. Their inclusion in F2 and F3 relates experiential associations to the organization of the linguistic and visual material representing each word.

2.3 Sensorimotor structure and indirect grounding

Multidimensional experience offers a theoretically motivated account of variation in lexical recoverability. Grounded and integrative accounts relate conceptual knowledge to perceptual, motor, affective, and linguistic experience (Barsalou 2008; Andrews et al. 2009; Davis and Yee 2021). This experiential perspective complements the linguistic emphasis of the Symbol Interdependency Hypothesis, which relates linguistic regularities to perceptual information (Louwerse 2011). Experiential dimensions describe differences among words in the visual information accessible through language. Empirically, distributional representations recover object properties and experiential dimensions (Grand et al. 2022; Chersoni et al. 2021), while multimodal distributional models support tests of indirect grounding for abstract concepts (Utsumi 2022). Collectively, these results motivate sensorimotor predictors of text–vision alignment, while leaving open whether the relevant relation is broad or dimension-specific.

The Chinese norms used here distinguish visual, auditory, gustatory, olfactory, tactile, and interoceptive experience; five action-related dimensions; and a separately elicited overall embodiment rating (Chen et al. 2026). Comparable multidimensional norms show that experiential structure may vary across modalities (Lynott et al. 2020; Wingfield and Connell 2023). The F1–F3 sequence therefore tests whether the profile persists under measurement adjustment and whether specific dimensions retain information once their common relation with overall embodiment is controlled.

Human ratings serve different purposes in feature-recovery studies and the present analysis. Chen et al. (2026) predicted sensorimotor ratings from linguistic representations. Here, an independently constructed visual prototype is the prediction target, and the ratings characterize experiential differences among the lexical items whose targets are recovered. This distinction connects the study to feature recovery while keeping visual information recovery as its central outcome.

2.4 Recovering visual information from Chinese lexical usage

The present study examines the representational premise of the Symbol Interdependency Hypothesis through visual recovery from Chinese lexical usage. Held-out retrieval tests whether a word's linguistic representation distinguishes its visual target from those of other words. We then relate differences in recovery to experiential attributes and the organization of the word's sampled uses and depictions. The

experiential analyses distinguish overall embodiment from specific sensory and motor associations. Exploratory interactions examine whether these associations vary with contextual dispersion and visual-target stability. The analyses thus address the degree and lexical distribution of visual recovery within one framework.

3. Methods

The study comprises two linked analyses. First, we construct a text representation and a visual prototype for each lexical item and train a mapping to predict visual representations from textual representations. Each item is evaluated using a mapping trained on other words. Its predicted vector is compared with the observed visual library to obtain target specificity, a score describing how clearly the correct target is distinguished from alternatives. Second, we relate variation in this score to the human sensorimotor ratings and the properties of the text and image collections. F1–F3 estimate progressively adjusted rating profiles, followed by a joint model and an exploratory analysis of contextual and visual moderation. The sections below describe the lexical inventory, representation construction, mapping, and association models.

3.1 Lexical inventory and sensorimotor norms

The lexical inventory supplies the common unit linking human ratings, contextual usage, and visual targets. The starting inventory comprised 3,000 Chinese written wordforms with concept identifiers and POS labels. We used the aligned norms of Chen et al. (2026), which distinguish eleven specific sensorimotor dimensions. Six concern visual, auditory, gustatory, olfactory, tactile, and interoceptive experience, and five concern actions involving the leg and foot, hand and arm, mouth and throat, head, and torso. Overall embodiment was a separate rating rather than an average computed from these eleven dimensions. According to Chen et al. (2026), the norms were collected from 378 native Mandarin speakers. The eleven specific dimensions were rated from 0 to 5, and overall embodiment from 0 to 6. In both scales, 0 indicated no experience and the upper endpoint indicated strong experience.

The image-selection pipeline described in Section 3.2 produced valid visual targets for 2,571 of the 3,000 inventory items; the remaining 429 had no valid visual target. The predictive sample therefore comprised the 2,571 items for which textual representations, visual prototypes, and human ratings could be aligned by concept identifier. It included 1,384 nouns, 711 verbs, 443 adjectives, and 33 items in the remaining POS category. The individual wordform was the unit of analysis, pooling the senses represented in its sampled contexts and images. Each wordform retained its inventory POS label. Continuous variables in the association models were standardized within the relevant analysis sample using population standard deviations. Online Resource 1 provides the complete analyzed inventory with concept identifiers and POS labels.

3.2 Web-derived computational visual prototypes

The visual analysis requires a reproducible target that summarizes recurrent image content for each wordform. The image archive combined Baidu Image Search as the primary source with Bing as a supplemental source. Initial image retrievals were recorded in UTC from 5 to 12 August 2026, followed by targeted Baidu retrievals on 24 and 26 August for low-coverage concepts. Candidate files passed technical filtering, exact SHA-256 deduplication, and perceptual-hash grouping at a maximum Hamming distance of five. Within each concept and near-duplicate group, the earliest search-result rank was retained.

Image embeddings were obtained using DINOv3 ViT-B/16, which provided 768-dimensional, L2-normalized vectors (Siméoni et al. 2025). Average-linkage clustering used cosine distance with a threshold of 0.60, and clusters required at least five images for eligibility. Qwen3.5-4B (Qwen Team 2026) evaluated contact sheets containing up to eight cluster exemplars together with the queried wordform and cluster size. Clusters were eligible for the visual core when classified as visual-core candidates and not assigned a high-confidence negative label for irrelevance, wrong sense, noise, or text- and advertisement-dominated content. Invalid outputs were retried twice; unresolved cases were assigned REVIEW status and were not automatically assigned a negative semantic label.

Eligible images were ordered by a facility-location criterion balancing cluster share, medoid proximity, coverage, and diversity. Adaptive-K allowed the selected image count to vary across words. It considered 10, 15, 20, 25, 30, 40, 50, 75, and 100 images and selected the first size meeting the coverage, prototype-stability, and marginal-gain thresholds for two consecutive steps. Concepts with fewer than ten eligible images retained all eligible images. For larger sets that did not meet the stopping criterion, the procedure retained the maximum candidate size or all available eligible images when fewer were available. The procedure selected 138,206 images for 2,571 lexical items, with a mean of 53.76 and a median of 50, while 429 inventory items had no valid visual target. Each computational visual prototype was the L2-normalized mean of its selected normalized embeddings. Visual dispersion, split-half stability, and average visual separation were entered as measurement properties in F2 and F3.

3.3 Contextual text representations

The linguistic representations were constructed by aggregating representations of each target wordform across its observed contexts of use. Text contexts came from CCI 3.0, an internet-derived Chinese corpus released by the Beijing Academy of Artificial Intelligence (2024). Up to 500 occurrence contexts were selected per item without replacement using surface-exact target spans and a deterministic priority rule with hash and document tie-breakers. Items with fewer available occurrences retained their available contexts. The representation resource contained 1,498,607 occurrences across the 3,000-item inventory and 1,284,107 occurrences across the 2,571 analyzed items. Contextual dispersion was defined as one minus the mean cosine similarity across all distinct pairs of normalized occurrence vectors for a wordform.

Each lexical text vector summarized the contextual representations of the target word itself across its sampled occurrences. The target word remained unmasked during encoding. Chinese RoBERTa-wwm-

ext-large provided 1,024-dimensional contextual representations with a maximum sequence length of 256 tokens (Cui et al. 2020). Target-centered iterative cropping retained the target span. For each occurrence, we averaged the final four encoder layers and pooled the subword tokens overlapping that span. The occurrence vector was L2-normalized, and the wordform representation was the L2-normalized mean of its occurrence vectors. This aggregation of occurrence vectors followed the general approach of deriving static lexical embeddings from contextualized representations (Bommasani et al. 2020).

Text split-half stability was the mean cosine similarity between unit-normalized half-set centroids over 50 deterministic partitions of the occurrence vectors used to construct each word representation. Odd-sized sets differed by one occurrence between halves. The measure used the available context count and no Spearman–Brown correction. Text separation was one minus the mean cosine between a word representation and the other 2,570 word representations. Separation used unit-normalized, uncentered word vectors before fold-specific centering. Together with contextual dispersion, these measures describe within-word variation, split-half agreement, and between-word distinctiveness.

3.4 Nested text-to-vision mapping

The text-to-vision mapping uses ridge regression to transform a 1,024-dimensional lexical text vector into a 768-dimensional visual vector. Each training pair links the text and visual representations of the same word. Ridge regression estimates a linear transformation while penalizing large coefficients. For a held-out word, the trained mapping takes its text vector as input and predicts a location in the visual space. Nested cross-validation separates penalty selection from the evaluation of held-out predictions. Lexical holdout tests whether a learned mapping generalizes to words outside its training set. Ten POS-stratified outer folds were held fixed across conditions. Each item received one prediction from a ridge mapping trained without that item. Within an outer-training set, five-fold inner cross-validation selected the penalty α from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}\}$. Selection maximized mean validation target specificity, with ties resolved in favor of the smallest grid value. The ridge objective minimized the squared Frobenius prediction error plus α times the squared Frobenius norm of the coefficient matrix, with an unpenalized intercept.

Inner penalty selection used centered text. The text mean was estimated from the inner-training items and subtracted from both training and validation vectors. The centered vectors were then L2-normalized. Inner validation used the outer-training visual prototypes as its candidate library, excluding outer-test targets. The mean was then re-estimated on the full outer-training set, and both outer-training and outer-test vectors were centered and L2-normalized before fitting and prediction. The primary condition combined centered text with Adaptive-K targets. Predicted visual vectors were L2-normalized before retrieval.

3.5 Retrieval and target specificity

Retrieval rank and target specificity capture complementary aspects of how a prediction distinguishes the correct target from its alternatives. Each predicted vector was compared using cosine similarity with

all 2,571 visual targets, including its correct target and 2,570 distractors. Ranking the targets by cosine similarity measures how readily the prediction recovers the visual prototype of the correct word. Target specificity was the correct-target cosine minus the mean distractor cosine, divided by the population standard deviation of the 2,570 distractor cosines. It expresses the target's advantage in units of distractor spread. Specificity compares the predicted vector with the target library, whereas visual separation compares the observed prototype with that library. Ranking produced reciprocal rank and cumulative Top-k hit rates, with stable concept order resolving ties.

Two non-item-specific baselines were constructed from outer-training visual targets, using their global mean and their mean within the evaluated item's POS category. The full-library evaluation compared all predictions against the same 2,571 candidate targets. A secondary descriptive evaluation restricted candidates to items of the same POS, including the correct target. Its candidate sets contained 33–1,384 items. We report the two candidate-library conditions separately because their search spaces differ.

An exploratory semantic-neighborhood evaluation compared retrieval within two equally sized candidate sets. Each set contained the correct target and twenty other items of the same POS. In the semantic-neighbor condition, competitors were the twenty nearest contextual text representations by cosine similarity. For each query, the text mean from its outer-training fold was subtracted from the query and candidate text vectors before L2 normalization. Ties followed the fixed concept order. In the random same-POS condition, twenty competitors were sampled uniformly without replacement, with one hundred draws per item. Retrieval metrics were averaged across draws within each item. Both conditions reused the held-out predictions from the primary mapping and the outer-training mean baselines. MRR was the primary measure; Top-1 and Top-5 rates described the proportion of targets retrieved within those ranks.

3.6 Sensorimotor and representation-property models

The association models relate variation in held-out recovery to sensorimotor experience and to the organization of the text and visual representations. The outcome for all association models was the primary-condition out-of-fold target specificity for each of the 2,571 items. All continuous variables, including the outcome, were standardized using full-sample means and population standard deviations. Every model included an intercept, POS indicators with nouns as the reference category, and outer-fold fixed effects. POS adjusted for broad lexical-class differences.

The three model families ask whether an experiential association is present, persists after measurement adjustment, and contributes beyond overall embodiment. F1 comprised twelve separate baseline models. Each model added one human rating to the common POS and fold controls. F2 comprised twelve separate measurement-adjusted models. Each model added one rating together with contextual dispersion, visual dispersion, visual split-half stability, and average visual separation. These four covariates describe the spread and organization of the text and visual representations used in prediction. F3 comprised eleven separate conditional models, one for each specific sensorimotor

dimension. Each model included that dimension, overall embodiment, the four F2 measurement covariates, POS, and fold. Overall embodiment was included as a covariate in every F3 model.

A complementary joint model included all twelve ratings and the four representation properties together, with the same POS and outer-fold controls. We used ordinary least squares to estimate the association of each indicator with standardized target specificity, conditional on the other indicators. Two reduced models provided the comparisons for incremental explained variance. The first contained the four representation properties, POS, and fold. The second additionally included overall embodiment. The full model added all twelve ratings to the first reduced model and the eleven specific ratings to the second.

For the exploratory moderation analysis, we extended the joint model with interactions between each of the twelve ratings and contextual dispersion, and between each rating and visual split-half stability. Both sets of twelve interactions were entered into the same model, retaining all rating and representation-property main effects, POS, and outer fold. The outcome and continuous predictors were standardized using population standard deviations before the interaction products were formed. A functional-form sensitivity analysis added squared terms for all twelve ratings and both moderators to the additive and interaction models. The same twenty-four interactions were retained.

3.7 Inference and uncertainty

Uncertainty was evaluated separately for sensorimotor associations and for the retrieval comparison. All three model families used the HC3 heteroskedasticity-consistent covariance estimator and normal-reference 95% confidence intervals (CIs) (MacKinnon and White 1985). For F1–F3, coefficient inference used 10,000 restricted wild-bootstrap-t draws under $\beta = 0$. Each draw applied Rademacher weights to leverage-adjusted reduced-model residuals, refitted the full model, and recomputed the HC3 t statistic. Two-sided p values used the absolute-t exceedance count with a plus-one correction, and Benjamini–Hochberg (BH) correction was applied separately within F1, F2, and F3 (Benjamini and Hochberg 1995). We report standardized coefficients so that effect magnitudes are comparable across ratings. For F3, ΔR^2 equals the full-model R^2 minus the corresponding model without the specific rating; partial R^2 equals ΔR^2 divided by one minus the reduced-model R^2.

For the joint model, coefficient tests used the normal reference distribution with the HC3 covariance estimator, and confidence intervals were normal-reference 95% HC3 intervals. BH correction was applied across the twelve rating tests. The four representation-property estimates are reported with confidence intervals. An HC3 Wald test with eleven degrees of freedom assessed whether the coefficients of the eleven specific dimensions were jointly zero, conditional on overall embodiment, the representation properties, POS, and fold. In-sample R^2 summarizes fit across the 2,571 items. The nested-model ΔR^2 and block partial R^2 use the same definitions as the F3 increments, with all eleven specific ratings removed together in the reduced model.

For moderation, HC3 Wald tests assessed each twelve-term interaction block conditional on the other block, using a chi-square reference distribution with twelve degrees of freedom. BH correction was

applied across the two block tests and separately across the twenty-four primary interaction coefficients. Dimension-level interpretation required support for both its block and its BH-adjusted coefficient. We report in-sample R^2, the increment obtained by removing each block while retaining the other, and the corresponding partial R^2. Conditional rating slopes were evaluated at the observed 10th, 50th, and 90th moderator percentiles, with the other moderator fixed at its median. Pointwise normal-reference 95% HC3 intervals used the full coefficient covariance. The quadratic sensitivity analysis repeated the two block tests, with BH correction across those two tests.

Retrieval inference used 10,000 target-identity randomizations within POS-by-outer-fold strata. The same permuted identities were applied to the learned mapping and the POS-mean baseline. One-sided p values for positive MRR and mean-specificity advantages used a plus-one correction, and the two tests formed a Holm family. Conditional 95% retrieval intervals were the 2.5th and 97.5th percentiles of 2,000 samples drawn with replacement within the same strata. The fitted predictions and similarity matrix were held fixed while the paired item and candidate indices were resampled.

For the exploratory neighborhood evaluation, 2,000 POS-stratified paired item bootstrap samples provided percentile 95% intervals, holding predictions and candidate sets fixed. The 2,571 lexical items were the resampling units, with random-candidate results first averaged within item. Two-sided paired sign-flip tests with 9,999 randomizations assessed two MRR contrasts. These compared the mapping with the POS-mean baseline on the semantic-neighbor sets and compared the mapping across semantic-neighbor and random same-POS sets. Holm correction was applied across these two tests.

4. Results

Visual recovery from lexical usage is examined at two levels. Retrieval from the full visual library first establishes how well a held-out word’s representation identifies its corresponding target. An exploratory comparison then restricts competition to semantically similar words of the same part of speech, keeping candidate-set size constant against a random same-POS comparison. The remaining analyses characterize differences in the original full-library target-specificity scores. Separate and joint models relate these differences to experiential attributes and representation properties, followed by an exploratory test of whether the experiential associations vary with contextual dispersion and visual-target stability.

4.1 Visual information recovered from lexical usage

Visual-target recovery for held-out lexical items exceeded both non-item-specific baselines. Full-library mean reciprocal rank (MRR) was 0.1260, with a conditional 95% bootstrap interval of 0.1171–0.1354, compared with 0.0032 for the global-mean baseline and 0.0044 for the POS-mean baseline. Relative to the POS mean, the mapping improved MRR by 0.1216 and mean target specificity by 2.1585; both randomization tests had Holm-adjusted p = .0002. Figure 1 shows cumulative recovery overall and by word class.

Among 2,571 candidates, 19.9% of correct targets appeared in the first five positions, 27.7% in the first ten, and 37.8% in the first 25. Median correct-target rank was 68. Figure 1 compares this cumulative pattern across word classes. Same-POS retrieval yielded MRR = 0.1764 and median rank = 29 within smaller candidate sets of 33 to 1,384 items.

Figure 1 shows empirical cumulative recovery overall and by word class. Every word was ranked against the same full library of 2,571 targets; proportions are normalized within each group. The overall curve includes the 33 words in the remaining POS category. Separate curves show nouns, verbs, and adjectives. Steps mark observed changes in cumulative recovery; the rank axis is logarithmic.

Out-of-fold target specificity averaged 2.212, with a standard deviation (SD) of 1.701, and 90.6% of items had a value above zero. Mean specificity was higher for nouns at 2.609 than for verbs at 1.883, adjectives at 1.610, and the small remaining category at 0.768. The distributions overlapped substantially. The next comparison holds both POS and candidate-set size constant while varying the textual similarity of the competing words.

4.2 Visual discrimination among semantically similar words

The exploratory neighborhood comparison examined whether lexical usage distinguished the correct visual target from those of twenty semantically similar words of the same POS. Figure 2 contrasts this condition with twenty randomly selected same-POS competitors. All 2,571 words were evaluated in both conditions, with twenty-one visual candidates per query.

Among semantic neighbors, the mapping achieved MRR = 0.3050, with a conditional 95% interval of 0.2932–0.3162. The POS-mean baseline achieved MRR = 0.1558 on the same candidate sets. The paired advantage was 0.1492, 95% CI [0.1373, 0.1612], Holm-adjusted p = .0002. The correct target ranked first for 13.11% of words and within the first five positions for 49.51%. The uniform random-ranking expectations for twenty-one candidates were 4.76% and 23.81%, respectively.

Semantically similar competitors reduced recovery relative to the equally sized random same-POS sets. Mapping MRR was 0.6222 in the random condition, with a conditional 95% interval of 0.6084–0.6345. The paired semantic-neighbor minus random-condition difference was − 0.3172, 95% CI [− 0.3264, − 0.3072], Holm-adjusted p = .0002. Top-1 and Top-5 rates in the random condition were 49.05% and 78.05%. Figure 2 shows how cumulative target recovery differs across the two competition conditions and the semantic-neighbor baselines.

These comparisons establish target discrimination under semantic competition. The following analyses return to the original full-library target-specificity scores to characterize how recovery varies across lexical items.

4.3 Experiential structure of differences in lexical visual recovery

Differences in full-library recovery were associated with the experiential attributes of words and the organization of their text and visual representations. Figure 3 compares the separate rating profiles before and after measurement adjustment. With POS and outer fold controlled in F1, eleven of twelve ratings had positive coefficients that remained significant after BH correction. Visual experience had the largest observed positive point estimate, overall embodiment was positive, and the interoceptive association was not statistically supported.

Adding contextual dispersion, visual dispersion, visual split-half stability, and average visual separation in F2 changed this profile. Ten specific dimensions retained positive, BH-significant associations, while interoception showed a negative association, $\beta = -0.070$, 95% CI [− 0.106, − 0.035], $q < .001$. Visual experience retained the largest observed positive estimate, $\beta = 0.228$, 95% CI [0.190, 0.266]. Overall embodiment had a smaller estimate that did not meet the adjusted threshold, $\beta = 0.030$, 95% CI [− 0.007, 0.067], $q = .106$.

Figure 3 shows two changes after measurement adjustment. The overall embodiment estimate decreased from 0.138 to 0.030, and the interoceptive estimate changed from 0.005 to − 0.070 after measurement adjustment.

In Fig. 3, the circles represent F1 and squares represent F2; intervals are 95% HC3 confidence intervals for standardized coefficients. Thin lines connect estimates for the same dimension. Filled markers indicate BH-adjusted $q < .05$ and open markers $q \geq .05$ within each family. F1 includes one rating, POS, and outer fold; F2 adds contextual dispersion, visual dispersion, visual split-half stability, and average visual separation.

Visual recovery was associated with each specific experiential dimension beyond overall embodiment in the separate F3 models. All eleven associations met the BH-adjusted threshold. Figure 4 shows the strength and direction of these associations. Visual experience had the largest observed positive coefficient and partial R^2, $\beta = 0.227$, 95% CI [0.189, 0.266], partial $R^2 = 0.0539$. Olfaction had the second-largest partial R^2, $\beta = 0.139$, 95% CI [0.097, 0.181], partial $R^2 = 0.0235$. Interoception retained a negative coefficient, $\beta = -0.075$, 95% CI [− 0.111, − 0.038], partial $R^2 = 0.0063$, while the action-related coefficients remained positive.

Across these models, partial R^2 ranged from 0.0031 for auditory experience to 0.0539 for visual experience. Olfaction, head, and torso had the next largest values, each with a positive coefficient. Figure 4 pairs this conditional information with the direction and uncertainty of each association.

In Fig. 4, rows show separate F3 models in descending order of partial R^2. Horizontal stems show partial R^2; circles indicate positive coefficients and the diamond a negative coefficient. The right-hand column reports standardized coefficients with 95% HC3 confidence intervals. Each model includes one specific dimension, overall embodiment, and the F2 controls. Partial R^2 compares each model with its corresponding reduced model omitting the specific dimension. These model-specific point estimates are nonadditive. All eleven dimensions have BH-adjusted $q < .05$.

The joint model estimated these associations with all twelve ratings and four representation properties entered together. Table 1 reports the mutually adjusted coefficients. Visual experience retained a positive association, β = 0.194, 95% CI [0.149, 0.239], as did olfactory experience, β = 0.157, 95% CI [0.076, 0.238]. Interoceptive experience retained a negative association, β = −0.072, 95% CI [− 0.111, − 0.033]. These three ratings met the BH-adjusted threshold. Intervals for the other nine ratings included zero, including overall embodiment, β = −0.010, 95% CI [− 0.053, 0.032].

The representation properties also showed distinct conditional associations. Contextual dispersion was negatively associated with specificity, β = −0.207, whereas visual dispersion was positively associated with it, β = 0.144. Visual split-half stability had a positive coefficient of 0.280, and average visual separation had a negative coefficient of − 0.276. All four intervals excluded zero. Thus, the contrasting associations of textual and visual dispersion were also present when the experiential ratings were included together.

Together, the ratings increased in-sample R^2 from 0.275 for representation properties, POS, and fold to 0.335 in the joint model, an increment of 0.0595. The reduced model additionally containing overall embodiment had $R^2 = 0.276$. Adding the eleven specific dimensions increased R^2 by 0.0587, or 5.87 percentage points, with block partial $R^2 = 0.0811$ and HC3 Wald $\chi^2(11) = 216.31$, $p < .001$. The specific dimensions thus added information beyond overall embodiment and the representation properties. The next analysis examines whether experiential associations vary with contextual dispersion and visual-target stability.

Table 1
Mutually adjusted experiential and representation-property associations with out-of-fold target specificity.

Indicator	β	95% HC3 CI	BH q
Experiential ratings			
Visual	**0.194**	**[0.149, 0.239]**	**< .001**
Auditory	0.014	[− 0.033, 0.060]	.7576
Gustatory	−0.034	[− 0.116, 0.048]	.7576
Olfactory	**0.157**	**[0.076, 0.238]**	**< .001**
Tactile	−0.045	[− 0.096, 0.006]	.2424
Interoceptive	**−0.072**	**[− 0.111, − 0.033]**	**.0011**
Leg/foot	0.017	[− 0.053, 0.087]	.7576
Hand/arm	0.006	[− 0.053, 0.066]	.8376
Mouth/throat	0.021	[− 0.038, 0.080]	.7576
Head	0.014	[− 0.055, 0.083]	.7576
Torso	0.026	[− 0.072, 0.123]	.7576
Overall embodiment	−0.010	[− 0.053, 0.032]	.7576
Representation properties			
Contextual dispersion	−0.207	[− 0.243, − 0.171]	—
Visual dispersion	0.144	[0.106, 0.182]	—
Visual split-half stability	0.280	[0.249, 0.311]	—
Average visual separation	−0.276	[− 0.314, − 0.239]	—
Model fit and added information			
Joint model: R^2 = .335.			
Eleven specific dimensions beyond overall embodiment and controls:			
ΔR^2 = .0587; block partial R^2 = .0811.			

Note

N = 2,571. The joint model includes all sixteen indicators, POS, and outer fold. The outcome and continuous predictors are standardized. Intervals are normal-reference 95% HC3 confidence intervals. BH correction covers the twelve rating tests; bold rating coefficients indicate q < .05. Dashes identify

representation properties outside this correction family. The model-fit increment is calculated against the reduced model containing overall embodiment, all representation properties, POS, and outer fold.

4.4 Contextual variation and visual stability in lexical recovery

Experiential associations varied with both contextual dispersion and visual split-half stability in the exploratory interaction model. Adding the twenty-four interaction terms increased in-sample R^2 from 0.3347 to 0.3538, an increment of 1.91 percentage points. The contextual-dispersion block was significant conditional on the stability block, HC3 Wald $\chi^2(12) = 40.83$, $q = .000105$, with $\Delta R^2 = 0.0091$ and partial $R^2 = 0.0139$. The stability block was also significant conditional on the dispersion block, $\chi^2(12) = 38.93$, $q = .000108$, with $\Delta R^2 = 0.0082$ and partial $R^2 = 0.0126$.

Table 2 reports all twenty-four interaction estimates. Two met the BH-adjusted coefficient threshold of $q < .05$, both involving visual experience. The visual-rating interaction with contextual dispersion was negative, $\beta = -0.076$, 95% CI [− 0.114, − 0.038], $q = .0011$. The interaction with visual split-half stability was positive, $\beta = 0.084$, 95% CI [0.042, 0.126], $q = .0011$. The remaining twenty-two interactions did not reach the adjusted threshold.

With visual stability fixed at its median, the conditional visual-rating slope decreased from 0.305, 95% CI [0.240, 0.369], at the 10th percentile of contextual dispersion to 0.120, 95% CI [0.051, 0.189], at the 90th percentile. With contextual dispersion fixed at its median, the slope increased from 0.088, 95% CI [0.009, 0.166], at the 10th percentile of visual stability to 0.303, 95% CI [0.237, 0.368], at the 90th percentile. The visual association was therefore weaker across more dispersed contexts and stronger for more stable visual targets within the observed moderator ranges.

Both interaction blocks remained significant after adding quadratic main effects. The contextual-dispersion block yielded $\chi^2(12) = 30.80$, $q = .0021$, and the stability block yielded $\chi^2(12) = 33.28$, $q = .0018$. In this specification, adding the same interaction terms increased in-sample R^2 from 0.3696 to 0.3848, an increment of 1.52 percentage points.

Table 2
Interactions of experiential ratings with contextual dispersion and visual split-half stability.

Rating	Contextual dispersion		Visual stability	
	β [95% CI]	BH q	β [95% CI]	BH q
Visual	**−0.076 [− 0.114, − 0.038]**	**0.0011**	**0.084 [0.042, 0.126]**	**0.0011**
Auditory	−0.022 [− 0.076, 0.031]	0.631	0.031 [− 0.015, 0.077]	0.499
Gustatory	−0.042 [− 0.134, 0.050]	0.631	−0.059 [− 0.156, 0.038]	0.547
Olfactory	0.014 [− 0.086, 0.114]	0.850	0.034 [− 0.051, 0.119]	0.631
Tactile	0.009 [− 0.033, 0.050]	0.773	−0.001 [− 0.052, 0.051]	0.981
Interoceptive	−0.011 [− 0.045, 0.023]	0.658	−0.012 [− 0.047, 0.022]	0.638
Leg/foot	−0.071 [− 0.135, − 0.007]	0.174	−0.026 [− 0.087, 0.035]	0.631
Hand/arm	−0.047 [− 0.103, 0.008]	0.377	−0.067 [− 0.122, − 0.011]	0.151
Mouth/throat	0.047 [− 0.013, 0.107]	0.379	0.030 [− 0.027, 0.088]	0.601
Head	−0.004 [− 0.068, 0.060]	0.945	−0.018 [− 0.079, 0.043]	0.667
Torso	0.079 [− 0.019, 0.177]	0.379	0.087 [− 0.006, 0.180]	0.324
Embodiment	0.015 [− 0.024, 0.054]	0.631	−0.023 [− 0.063, 0.016]	0.547

Note. N = 2,571. The exploratory model estimates all twenty-four interactions together with the twelve ratings, four representation properties, POS, and outer fold. β denotes the coefficient for a product of standardized predictors. Intervals are normal-reference 95% HC3 confidence intervals. BH correction covers all twenty-four interaction tests; bold entries indicate q < .05. Visual stability denotes visual split-half stability, and Embodiment denotes overall embodiment.

5. Discussion

Visual information recovered from word use varied across lexical items and with their experiential and representation properties. This pattern connects a linguistic source of visual information to differences in its accessibility across the lexicon. We interpret the findings in relation to perceptual inference, the experiential attributes of words, and the consistency of their uses and visual depictions.

5.1 Lexical relations provide access to visual information

Distributional relations can carry visual information when words used in related situations also refer to entities, properties, and events with related appearances. Johns and Jones (2012) modeled a comparable route from global lexical similarity to missing perceptual features using known experiential traces. The present findings are consistent with applying this inferential relation to computational visual targets. A mapping learned from other words used a held-out item's linguistic representation to distinguish its corresponding target in the visual library.

Collell et al. (2017) and Günther et al. (2022) provide direct precedents for predicting visual representations of unseen words. The present evaluation asks whether a prediction identifies an independently constructed prototype among a common set of lexical targets. Its advantage over both mean baselines connects this task to a lexical-semantic distinction between information shared across a class and information that differentiates an individual word. The recovery outcome concerns the latter within the evaluated target library.

Semantic-neighborhood competition clarifies what this lexical differentiation involves. Words selected for similarity in linguistic usage were harder to distinguish than equally numerous random words of the same class. Yet the mapped representation retained an advantage over the class mean within these neighborhoods. Linguistic similarity therefore has two roles in the evaluation. It supplies information about related meanings and identifies the alternatives whose visual targets are difficult to separate. Overall recovery and discrimination among similar words capture complementary aspects of correspondence between lexical use and visual content.

The Symbol Interdependency Hypothesis provides a linguistic interpretation of this correspondence. Louwerse (2011) proposes that relations among symbols reflect regularities in situated experience. Recurring descriptions may therefore make information about referents accessible through relations in language. Collell and Moens (2018) distinguish successful prediction from reconstruction of target-space geometry. In the present task, retrieving the correct target identifies information that transfers between the representations while allowing linguistic and visual relations to retain their distinct organization.

Lexical variation specifies where this access is more or less informative. Günther et al. (2022) examined concreteness and visual-neighbor structure as factors in prediction. The present analysis relates recovery to differentiated experiential ratings together with textual and visual representation properties. This extends the characterization of which words support visual recovery and links their experiential attributes to the organization of their uses and depictions.

This approach also distinguishes the information being recovered from the attributes used to characterize its recovery. Grand et al. (2022), Chersoni et al. (2021), and Chen et al. (2026) examined knowledge or features recoverable from linguistic representations, while Xu et al. (2025) identified differences between non-sensorimotor and sensorimotor feature recovery. Here, the target is visual

content derived from images, and human ratings characterize variation in its recovery across words. The analysis therefore addresses the wider question of what information is accessible through distributional relations while retaining a distinct visual retrieval criterion.

Distributional and experiential information contribute complementary perspectives on lexical meaning. Andrews et al. (2009) combined these sources in learning semantic representations, and Davis and Yee (2021) discuss their contributions to semantic memory. The present analysis links the information used to represent a word in language with its visual target and experiential attributes. This combination helps characterize both the availability of visual information and its uneven distribution across words.

5.2 Experiential associations characterize differences in visual recovery

The contribution of the specific dimensions indicates that visual recovery varies with the kinds of experience associated with lexical meaning, beyond overall experiential strength. A global embodiment rating summarizes the strength of bodily experience, whereas specific ratings distinguish relations to perception and action. Multidimensional norms preserve this differentiation (Lynott et al. 2020; Chen et al. 2026), and sensorimotor distance uses the configuration of dimensions to characterize semantic similarity (Wingfield and Connell 2023). A differentiated experiential profile thus provides a more informative description of variation in visual recovery than the global rating alone.

The olfactory association makes the relation between nonvisual experience and visual recovery particularly informative. Indirect-grounding approaches examine how linguistic relations connect experiential information (Utsumi 2022). One possible account is recurrent reference. Descriptions of smell may repeatedly mention entities and events that also recur in the image collection. Another is conventional depiction, in which a visible object regularly stands for an olfactory experience. Recurrent reference predicts correspondence across varied depictions with shared referential content; conventional depiction predicts a closer relation for a restricted depiction convention. Comparing the referents and situations expressed in contexts with those shown in images would distinguish these accounts.

Interoceptive words suggest a contrasting relation between recurring usage and depiction. Internal states may receive consistent linguistic descriptions while being depicted through varied expressions, situations, or associated objects. On this account, the negative conditional association reflects weaker correspondence between recurrent linguistic content and recurrent depiction. Covariance among the ratings and the inventory's semantic composition offer competing explanations. A depiction account predicts closer recovery when contexts and images are matched by sense and situation; a covariance account predicts sensitivity to how correlated ratings and semantic categories are modeled. The distinction concerns how a word's uses relate to its visual target.

The separate and joint profiles distinguish shared experiential organization from mutually adjusted associations. Several ratings may describe overlapping groups of objects, events, and states, producing

broad separate relations with recovery. The joint model identifies a smaller set of conditional associations within that organization. The linguistic interpretation concerns which experiential descriptions are informative about visual recoverability. Comparing recovery across visual, action-related, and internal-state targets would test how their relevance changes with the information being reconstructed.

5.3 Semantic consistency and visual-target stability

The exploratory interactions connect visual recovery to consistency across a word's contexts of use. A linguistic account concerns the recurrence of related referential and experiential content across uses. Syntactically and topically varied sentences may still refer to related situations, while an aggregate vector may also combine uses associated with different visual realizations. Contextual dispersion characterizes variation among encoded occurrences (Ethayarajh 2019; Bommasani et al. 2020). The weaker visual association at higher dispersion motivates testing whether recurring semantic content preserves a clearer relation between a lexical representation and its visual target.

Visual-target stability concerns how reproducibly the image collection represents a word. Split-half agreement measures the consistency of the aggregate target across image subsets. Images may vary in viewpoint, background, and exemplar while retaining an organized category structure, so visual diversity and a stable summary can coexist. The positive conditional associations of visual dispersion and split-half stability are compatible with this account. Research on image dispersion and exemplar aggregation likewise relates correspondence to the organization of image collections (Kiela et al. 2014; Zhou et al. 2025). The stronger visual-rating association at higher split-half stability suggests that experiential distinctions are more clearly expressed when the target is reproducible.

Semantic organization and measurement reliability offer competing explanations of these interactions. Dispersed uses may evoke different situations and correspond less closely to a single visual summary. Lower visual stability may also weaken an association through less reliable measurement of that summary. Contextualized sensorimotor norms connect experiential judgments to particular senses and occurrences (Trott and Bergen 2026). Comparing sense-matched contexts with pooled uses would test the linguistic account, while repeated image subsampling at matched sample sizes would assess the reliability account. These comparisons distinguish what a word's uses share from how reproducibly its sampled images summarize visual content.

Visual recovery thus relates the organization of linguistic usage to the organization of visual depiction. The lexical representation summarizes sampled uses, and the visual prototype summarizes sampled images. Their correspondence can vary with recurring semantic content, experiential associations, and the reproducibility of the visual target. Considering these properties together specifies the conditions under which visual information is accessible from lexical usage.

5.4 Limitations

Several limitations concern the visual targets, the aggregation of word uses, and the interpretation of the association models. The inference concerns correspondence with constructed Web-derived visual targets at the Chinese wordform level. These targets only summarize selected search images and differ from individual mental prototypes. Wordform aggregation pools senses and uses represented in the sampled contexts and images. Because the target word remained unmasked, each text representation combines lexical identity, pretrained lexical knowledge, and information from the sampled contexts; the design does not isolate the unique contribution of contextual variation. The predictive analysis covers the items for which the selection pipeline produced valid targets, while the remaining inventory items characterize the boundary of visual-target availability.

The association models describe conditional lexical covariation rather than causal effects of sensorimotor experience. Text and images were sampled independently from one constructed Web-derived empirical environment, so the resulting profile is anchored to that source configuration. The F3 models estimate each specific rating separately, whereas the joint model estimates mutually adjusted associations among correlated ratings. Collinearity, with a maximum rating variance inflation factor of 8.62, affects the precision and interpretation of the joint coefficients. An interval including zero leaves a range of associations compatible with the data. The block increment quantifies the contribution of the eleven dimensions together but does not partition their shared explanatory information into independent components.

The interaction analyses were exploratory and used fixed out-of-fold specificity scores. Their coefficient intervals condition on those scores, and independent samples would assess the reproducibility of the moderation patterns. Contextual dispersion also reflects contextual, topical, and encoder-dependent variation; semantic stability is an interpretation to be tested rather than a directly validated measure in this study. The interaction models estimate moderation by each representation property and do not test an interaction between semantic and visual stability. Future studies can examine these sources of variation using additional samples, sense-specific contexts, and independently sampled visual targets. The semantic-neighborhood evaluation was also exploratory. Its intervals condition on the fitted predictions and candidate sets, and its competitors were defined by contextual-vector similarity rather than independent human judgments of semantic similarity.

6. Conclusion

This study examined the recovery of visual information from usage-derived lexical representations. Among the analyzed targets in the sampled Chinese Web environment, representations of held-out words supported identification of their independently constructed visual prototypes. An exploratory comparison further showed target discrimination among semantically similar words of the same part of speech. Recovery varied with experiential attributes and properties of the representations. The specific sensorimotor dimensions collectively contributed information beyond overall embodiment, with visual, olfactory, and interoceptive associations retained under mutual adjustment. Exploratory moderation showed a stronger visual association at lower contextual dispersion and greater visual-target stability.

These findings support the representational premise of the Symbol Interdependency Hypothesis that linguistic relations provide access to perceptual information. They extend the empirical account of this premise by showing how visual recovery varies across Chinese lexical items and with the organization of their sampled uses and depictions. Word use thus provides a source of visual information whose accessibility is differentiated across the lexicon.

Declarations

Author Contribution

Y.W. analyzed the data and drafted the manuscript. X.R. collected the data and contributed to drafting the manuscript. E.A. revised the manuscript. M.L. supervised the research and reviewed and revised the manuscript.

Data Availability

Sensorimotor norms are available from Chen et al. (2026). Text representations, Web-derived computational visual prototypes, analysis outputs, and code will be made available in a public repository upon publication. Raw Web images and corpus texts will not be redistributed where redistribution is not permitted.

References

1. Andrews M, Vigliocco G, Vinson D (2009) Integrating experiential and distributional data to learn semantic representations. Psychological Review 116(3):463–498. https://doi.org/10.1037/a0016261
2. Barsalou LW (2008) Grounded cognition. Annual Review of Psychology 59:617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639
3. Beijing Academy of Artificial Intelligence (2024) CCI3-Data [Data set]. Hugging Face. https://huggingface.co/datasets/BAAI/CCI3-Data
4. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
5. Bommasani R, Davis K, Cardie C (2020) Interpreting pretrained contextualized representations via reductions to static embeddings. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics pp 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431
6. Bruni E, Tran NK, Baroni M (2014) Multimodal distributional semantics. Journal of Artificial Intelligence Research 49:1–47. https://doi.org/10.1613/jair.4135

7. Chen J, Parti G, Zhong Y, Huang C-R, Marelli M (2026) Chinese sensorimotor and embodiment norms for 3,000 lexicalized concepts. arXiv preprint, arXiv:2605.22616. https://doi.org/10.48550/arXiv.2605.22616
8. Chersoni E, Santus E, Huang C-R, Lenci A (2021) Decoding word embeddings with brain-based semantic features. Computational Linguistics 47(3):663–698. https://doi.org/10.1162/coli_a_00412
9. Collell G, Moens M-F (2018) Do neural network cross-modal mappings really bridge modalities? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pp 462–468. https://doi.org/10.18653/v1/P18-2074
10. Collell G, Zhang T, Moens M-F (2017) Imagined visual representations as multimodal embeddings. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence pp 4378–4384. https://doi.org/10.1609/aaai.v31i1.11155
11. Cui Y, Che W, Liu T, Qin B, Wang S, Hu G (2020) Revisiting Pre-Trained Models for Chinese Natural Language Processing. In: Findings of the Association for Computational Linguistics: EMNLP 2020 pp 657–668. https://doi.org/10.18653/v1/2020.findings-emnlp.58
12. Davis CP, Yee E (2021) Building semantic memory from embodied and distributional language experience. WIREs Cognitive Science 12(5):e1555. https://doi.org/10.1002/wcs.1555
13. Ethayarajh K (2019) How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proceedings of EMNLP-IJCNLP 2019 pp 55–65. https://doi.org/10.18653/v1/D19-1006
14. Grand G, Blank IA, Pereira F, Fedorenko E (2022) Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nature Human Behaviour 6:975–987. https://doi.org/10.1038/s41562-022-01316-8
15. Günther F, Petilli MA, Vergallito A, Marelli M (2022) Images of the unseen: extrapolating visual representations for abstract and concrete words in a data-driven computational model. Psychological Research 86:2512–2532. https://doi.org/10.1007/s00426-020-01429-7
16. Hill F, Reichart R, Korhonen A (2014) Multi-modal models for concrete and abstract concept meaning. Transactions of the Association for Computational Linguistics 2:285–296. https://doi.org/10.1162/tacl_a_00183
17. Johns BT, Jones MN (2012) Perceptual inference through global lexical similarity. Topics in Cognitive Science 4(1):103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x
18. Kiela D, Hill F, Korhonen A, Clark S (2014) Improving multi-modal representations using image dispersion: Why less is sometimes more. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pp 835–841. https://doi.org/10.3115/v1/P14-2135
19. Köper M, Schulte im Walde S (2017) Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) pp 200–206. https://doi.org/10.18653/v1/W17-1728

20. Lazaridou A, Bruni E, Baroni M (2014) Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pp 1403–1414. https://doi.org/10.3115/v1/P14-1132
21. Louwerse MM (2011) Symbol interdependency in symbolic and embodied cognition. Topics in Cognitive Science 3(2):273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x
22. 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 52:1271–1291. https://doi.org/10.3758/s13428-019-01316-z
23. MacKinnon JG, White H (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29(3):305–325. https://doi.org/10.1016/0304-4076(85)90158-7
24. Qwen Team (2026) Qwen3.5-4B [Model card]. Hugging Face. https://huggingface.co/Qwen/Qwen3.5-4B
25. Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C, Khalidov V, Szafraniec M, Yi S, Ramamonjisoa M, Massa F, Haziza D, Wehrstedt L, Wang J, Darcet T, Moutakanni T, Sentana L, Roberts C, Vedaldi A, Tolan J, Brandt J, Couprie C, Mairal J, Jégou H, Labatut P, Bojanowski P (2025) DINOv3. arXiv preprint, arXiv:2508.10104. https://doi.org/10.48550/arXiv.2508.10104
26. Trott S, Bergen B (2026) Contextualized sensorimotor norms: Multi-dimensional measures of sensorimotor strength for ambiguous English words, in context. Behavior Research Methods 58:264. https://doi.org/10.3758/s13428-026-03134-6
27. Utsumi A (2022) A test of indirect grounding of abstract concepts using multimodal distributional semantics. Frontiers in Psychology 13:906181. https://doi.org/10.3389/fpsyg.2022.906181
28. Wingfield C, Connell L (2023) Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept pairs. Behavior Research Methods 55:3416–3432. https://doi.org/10.3758/s13428-022-01965-7
29. Xu Q, Peng Y, Nastase SA, Chodorow M, Wu M, Li P (2025) Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. Nature Human Behaviour 9:1871–1886. https://doi.org/10.1038/s41562-025-02203-8
30. Zhou Y, Tarr MJ, Yurovsky D (2025) Quantifying the roles of visual, linguistic, and visual-linguistic complexity in noun and verb acquisition. PLOS ONE 20(5):e0321973. https://doi.org/10.1371/journal.pone.0321973

Figures

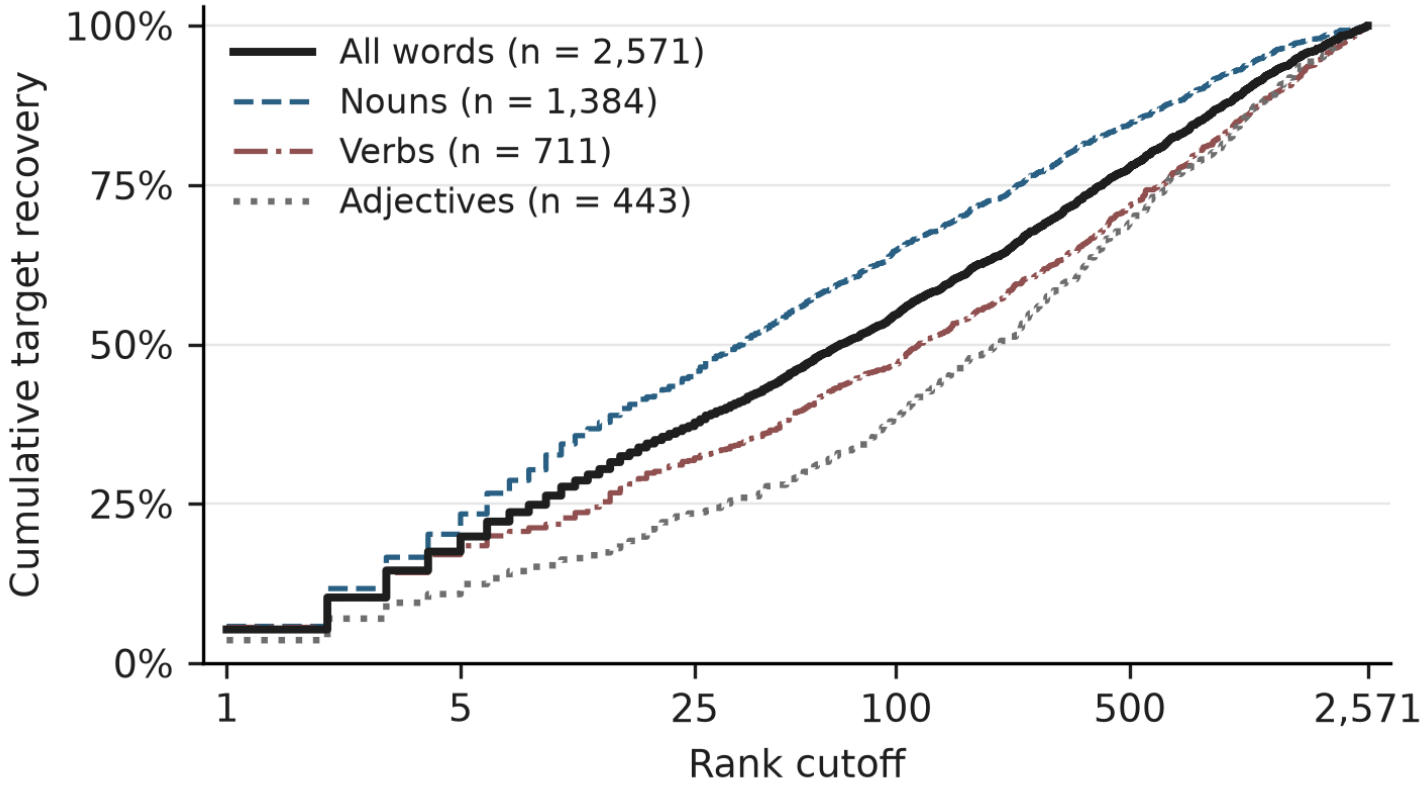


Figure 1

Cumulative visual recovery from lexical usage overall and by word class

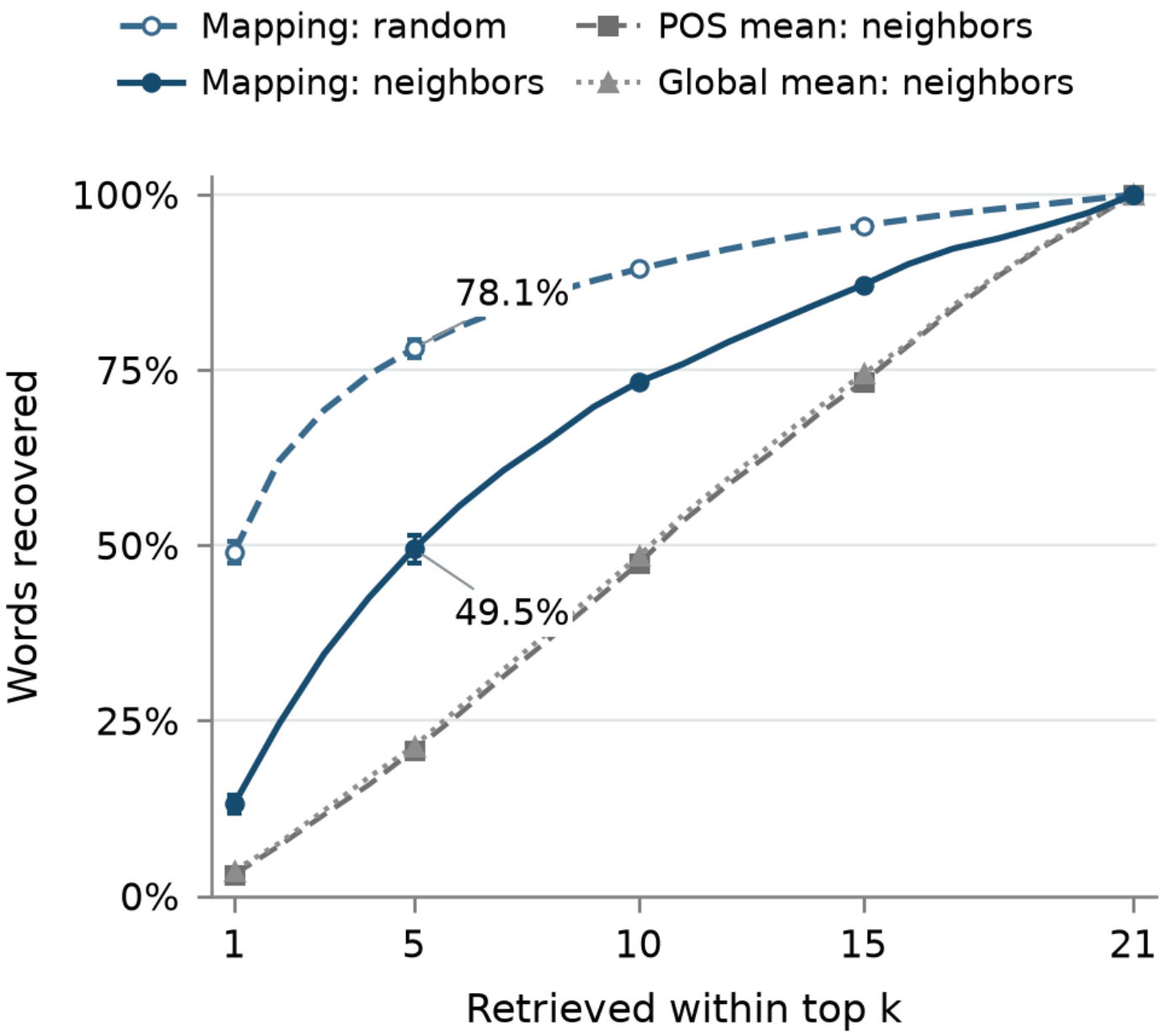


Figure 2

Cumulative visual recovery under semantic-neighbor and random same-POS competition. Each query has twenty-one candidates, including its correct target. Curves show the proportion of 2,571 words retrieved within the first k positions. The mapping is shown under both conditions; POS-mean and global-mean baselines are shown for semantic neighbors. Random-condition rates average one hundred candidate draws within each word. Whiskers at k = 1 and 5 show the mapping's conditional 95% POS-stratified item-bootstrap intervals

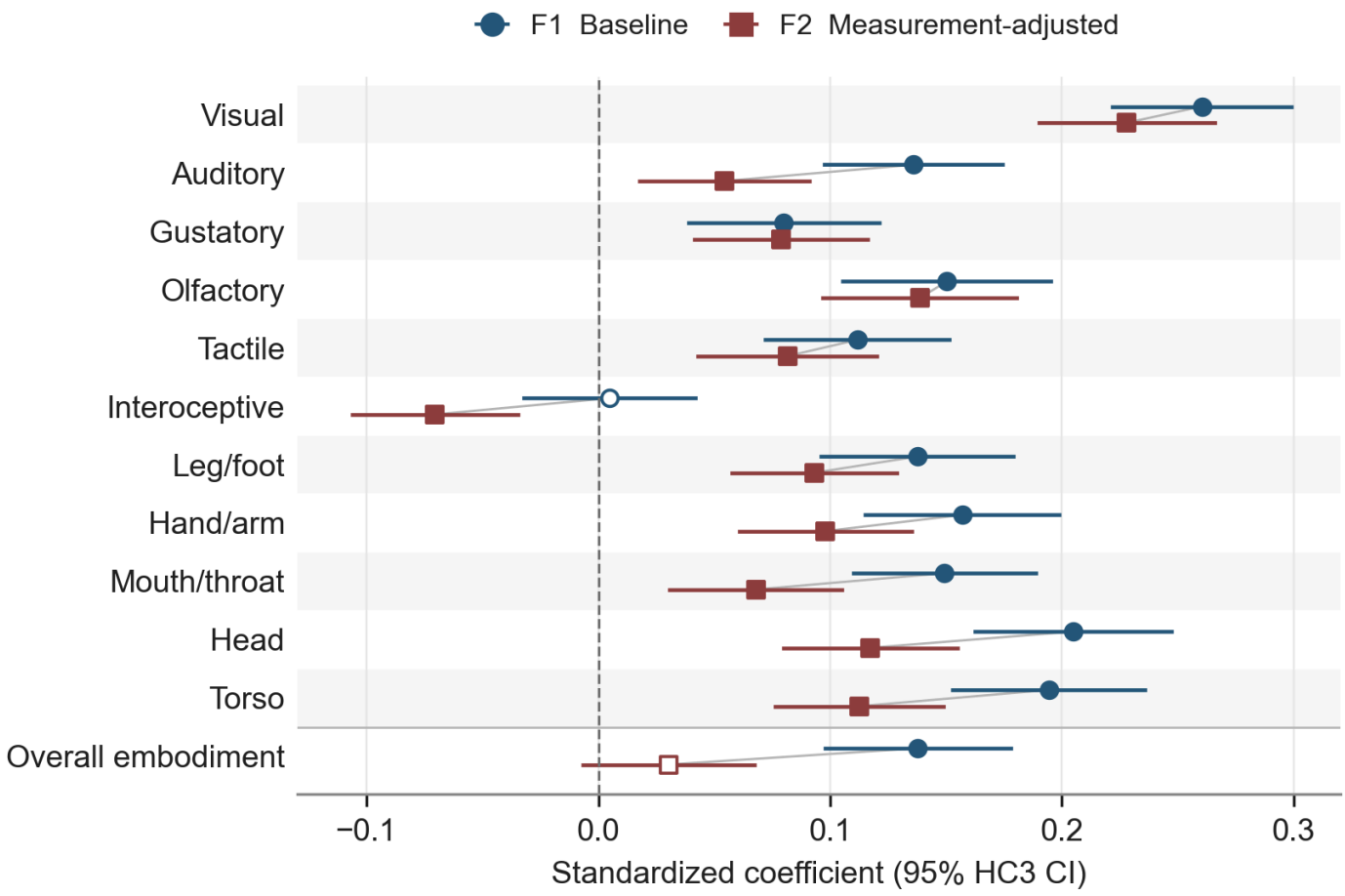


Figure 3

Sensorimotor associations before and after measurement adjustment

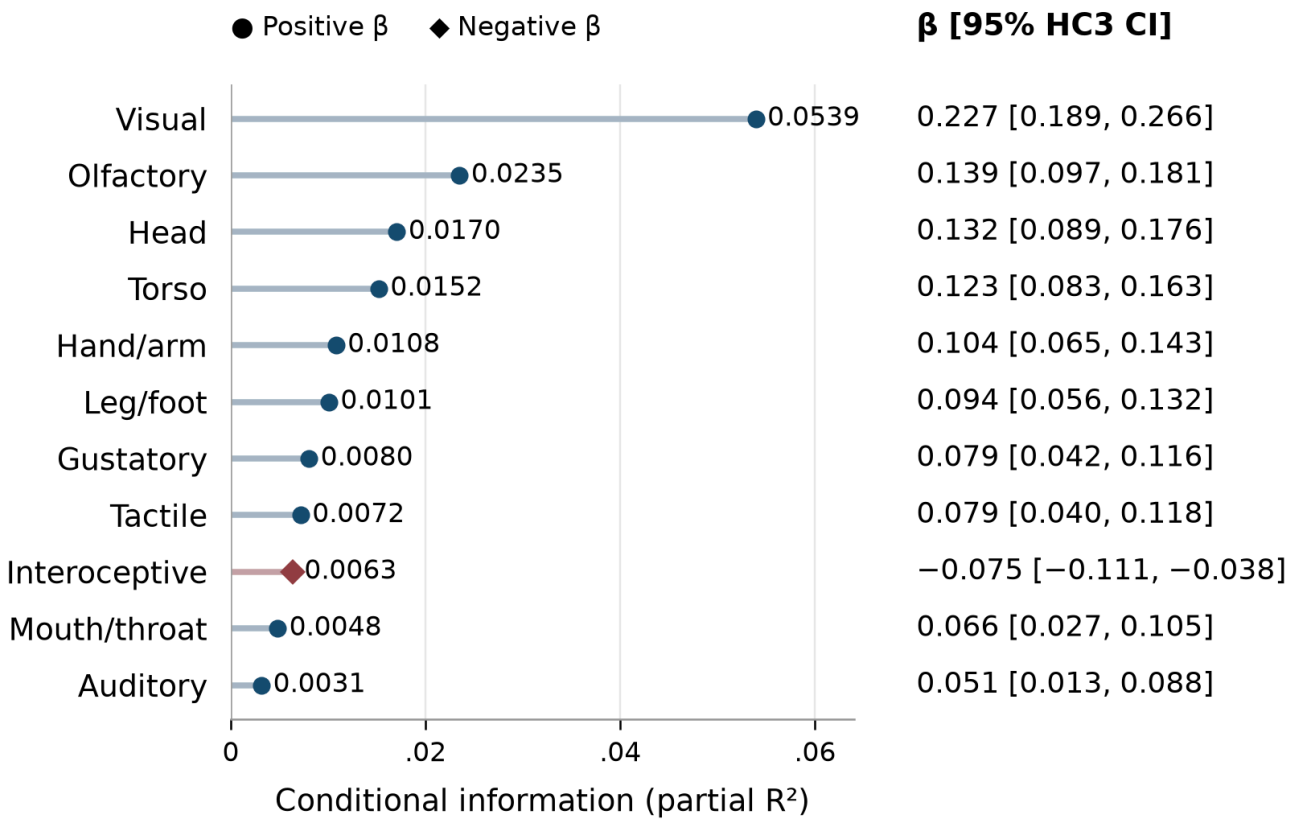


Figure 4

Experiential associations with visual recovery beyond overall embodiment

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.csv