

Title Page

Title: Measuring Meme-Family Structure With Contrastive Sparse Features: Canonical Resemblance and Reddit Engagement

Running head: MEASURING MEME-FAMILY STRUCTURE

Authors and affiliations:

- Eric Gladstone, Independent Researcher, Traverse City, Michigan, USA
- David C. Thompson, Chemical Computing Group, Montreal, Quebec, Canada

Corresponding author: Eric Gladstone, Independent Researcher, Traverse City, Michigan 49684, USA. Phone: +1 206 259 1912. Email: eric.c.gladstone@gmail.com

Co-author contact: David C. Thompson. Email: d.c.thompson.00@gmail.com

ORCID iDs: Eric Gladstone: <https://orcid.org/0000-0003-4393-050X>. David C. Thompson: <https://orcid.org/0000-0001-8050-3852>.

Acknowledgments: Generative AI tools, including ChatGPT (OpenAI) and Claude (Anthropic), were used to assist with coding, data analysis, preparation of figures and tables, formatting, and editorial refinement. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

Declaration of Conflicting Interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical considerations: This study analyzed publicly available Reddit posts and public meme-reference materials. It involved no intervention with human participants and no identifiable private information. Usernames were removed from the analysis data, and raw user content was not redistributed beyond platform terms. A supplementary coding check was carried out by one of the authors, who labelled a sample of public Reddit images against a reference sheet. No other persons took part, and no ethics review was sought.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data Availability Statement: The derived features, analysis outputs, and permitted public post identifiers needed to reproduce the reported analyses are archived on the Open Science Framework (OSF) under DOI 10.17605/OSF.IO/BFWZE, subject to source and platform redistribution terms. For peer review, the materials are available through an anonymized view-only OSF link: https://osf.io/bfwze/?view_only=4633292d4a7c4cf387ab02aa7f8e70b9. The OSF project will be made public on acceptance. Raw user-generated content is not redistributed.

Identifiers with usernames removed and derived representations are shared where permitted. Data are released under CC0 1.0.

Code Availability: All analysis code is version-controlled and archived with the research materials on OSF under DOI 10.17605/OSF.IO/BFWZE. The archived materials link to the corresponding GitHub repository. Code is released under the MIT License. For peer review, the code is included in the anonymized OSF package. The public repository will be released on acceptance.

Author Contributions: Eric Gladstone: Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Visualization, Writing – original draft, and Writing – review and editing. David C. Thompson: Conceptualization, Investigation (supplementary computation), and Writing – review and editing. Both authors share ownership of the core idea and approved the final manuscript. Eric Gladstone is the corresponding and lead author.

Measuring Meme-Family Structure With Contrastive Sparse Features: Canonical Resemblance and Reddit Engagement

Abstract

We introduce a contrastive sparse-feature signature that scores an image's resemblance to a documented meme family. We develop it in a 31-family corpus and validate recovery on 80 held-out families under a protocol fixed before scoring. On 60 randomly sampled held-out families, the signature retained clear recovery advantages over a frequency-based recurrence baseline, sparse centroid and nearest-neighbor rules using the same representation, and surface measures, but no clear contrastive advantage was detected over the whole-image CLIP centroid or a fused SigLIP2–DINOv2 centroid. The development-corpus advantage over the CLIP centroid did not replicate. A supervised classifier recovered membership more accurately. Against hard decoys selected by a separately trained CLIP-family encoder, the signature retained a clear advantage over the CLIP centroid in random held-out families. In 20 look-alike challenge families, the fused centroid outperformed the signature. The signature is a compact set of 15 weighted features without established semantic meanings. In 4,263 Reddit posts, the raw score modestly improved out-of-sample engagement prediction beyond whole-image, surface, timing, and community measures. Family-calibrated fit and classifier scores added near-zero point increments. Much of the increment reflected broad structured resemblance, which we call canonical resemblance. After representation-magnitude and text-length controls, the residual was small and uncertain.

Keywords: computational social science, internet memes, multimodal measurement, sparse autoencoders, social media engagement, Reddit, construct validity

Introduction

Internet memes circulate as multimodal variants that are copied, altered, and recombined across people and communities (Shifman, 2013, 2014; Author & Author, forthcoming). We use meme family to mean a set of variants documented by a meme-reference source as one lineage. Members share some recurring configuration of image, text, or layout while differing, sometimes substantially, in surface form (Figure 1). A family is therefore broader than a template. A template is a reusable base image, whereas a family may span several templates or preserve its identity through a recurring caption structure or relation even when the image changes. Figure 1 illustrates both possibilities in development-corpus families. One retains a caption formula across different base images (Figure 1a), while another retains a reject-and-approve layout across different base characters (Figure 1b). Conversely, a visually similar image can belong to a different documented family (Figure 1c).

This distinction changes the measurement problem. Nearest-example and pairwise-similarity methods are well suited to many template-matching tasks, but documented family membership can depend on structure that recurs across variants rather than on resemblance to a single

exemplar. Whole-image representations capture visual and semantic similarity (Radford et al., 2021), optical character recognition can recover recurring text, and perceptual hashing can identify near-duplicates and template reuse. Our question is different: can a documented family be represented by a single reusable representation, estimated from what recurs among its members and distinguishes them from other families, that assigns new variants a common-form resemblance score? We therefore treat documented meme-family fit as a measurement target distinct from pairwise resemblance.

We approach this problem as one of computational measurement. When model outputs are used as measures of social or cultural constructs, predictive performance alone does not determine what the resulting scores mean (Adcock & Collier, 2001; Grimmer & Stewart, 2013; Jacobs & Wallach, 2021; Kirchmair et al., 2026). We introduce a compact family-level signature estimated contrastively in a sparse feature space and compare it with a frequency-based recurrence baseline, surface measures, whole-image representations, sparse centroids, nearest-neighbor rules, and supervised classification. The supervised one-versus-rest logistic classifier trained on the same sparse representation is the stronger estimator of documented family membership. The contrastive signature serves a different purpose: it represents each family as a positive prototype containing 15 retained features and their weights. The comparison therefore concerns both recovery of documented labels and the properties of the resulting measurement object.

The empirical criterion throughout the measurement analysis is documented family membership. Because Reddit posts lack these labels, we treat the raw signature score as an uncalibrated magnitude of resemblance to the closest documented family. Engagement is analyzed separately as a downstream outcome, asking whether that magnitude contributes information about platform reception beyond the other measured predictors (Table 1).

The study proceeds in two stages. First, we estimate family-level signatures in a development corpus and compare their recovery of documented labels with alternative methods, both in the development families and in held-out families collected under a protocol fixed before scoring. On randomly sampled held-out families, the signature retains clear advantages over the frequency-based recurrence baseline, same-representation sparse centroid and nearest-neighbor rules, and surface measures. No clear contrastive advantage was detected over whole-image dense centroids, and the supervised classifier recovers membership more accurately. Second, we apply the development signatures to newly listed Reddit image posts and ask what information their raw signature scores carry about engagement.

Related Literature and Conceptual Framework

Computational Measurement in Social Science

Computational social science increasingly uses model outputs as measures of social and cultural phenomena that cannot be observed directly. Word embeddings, for example, have been used to measure cultural meaning and social associations (Garg et al., 2018; Kozłowski et al., 2019),

while large-scale digital traces have been used to construct indicators of cultural variation (Obradovich et al., 2022). Such applications create a measurement problem as well as a prediction problem. A model can recover labeled observations accurately while leaving uncertain what its scores represent substantively (Cronbach & Meehl, 1955). Evaluating a computational measure therefore requires evidence about the criterion it recovers, the behavior of its scores across relevant cases, and the limits of the interpretation those scores can support (Adcock & Collier, 2001; Grimmer & Stewart, 2013; Jacobs & Wallach, 2021).

Those questions become more important when different representations recover the same labels with different tradeoffs. Classification performance is one relevant criterion, but a measure may also differ in compactness (how few components represent the construct), inspectability (whether the retained components and their weights can be directly examined), stability (how sensitive it is to reasonable specification choices), and suitability for researcher-directed analysis (whether it produces a reusable quantity for subsequent analysis) (Kirchmair et al., 2026). In our case, the contrastive signature represents a family with 15 weighted sparse features and assigns each candidate a scalar resemblance score. We therefore compare its recovery performance with stronger supervised alternatives rather than treating compactness as a substitute for accuracy. We separately examine sensitivity to feature count and to redrawn evaluation negatives, and we treat cross-family calibration as an unresolved empirical property rather than assuming that scores are directly comparable across families.

A related issue is construct scope. Digital traces can support relatively narrow interpretations without supporting broader ones. Jungherr et al. (2017), for example, argue that political tweeting is more defensibly interpreted as an indicator of attention than of political support. Here, the empirical criterion is documented meme-family membership, which does not establish cultural importance, human recognition, or family structure independent of the reference source.

Computational Analysis of Internet Memes

Computational research on memes has examined their emergence, diffusion, and transformation across online communities (Bauckhage, 2011; Zannettou et al., 2018) and has used multimodal representations for tasks such as hateful-meme detection (Kiela et al., 2020). Existing methods capture several kinds of resemblance. Vision-language encoders measure broad visual and semantic similarity (Radford et al., 2021). Optical character recognition and text similarity can identify recurring captions and phrases. Perceptual hashing detects near-duplicates and repeated templates.

Recent work also shows how effectively modern vision encoders can identify meme templates. Nguyen et al. (2026), for example, compare semi-open-set methods that assign an image to a known template or reject it when no template matches, finding strong performance from encoders including SigLIP and DINOv2. That task differs from ours because a documented meme family can span more than one base template or preserve a recurring relation across changing images. Theisen et al. (2021, 2023), in turn, identify recurring meme genres and

remixed visual motifs without starting from predefined family labels. Our problem is complementary: given a lineage already documented by a reference source, we ask how strongly a new image resembles the structure shared by that lineage.

This distinction motivates comparisons with both pairwise and family-level alternatives. A family measure must aggregate information across members while representing what distinguishes the family from nonmembers. We therefore compare the contrastive signature with a frequency-based recurrence baseline, whole-image and sparse centroids, nearest-neighbor rules, surface measures, and a supervised classifier using the same sparse representation. These comparisons separate the value of the contrastive construction from information already present in the underlying representation.

Research on meme circulation also motivates the downstream application. Work on fringe-web and Reddit meme ecosystems links visual and textual properties of memes to diffusion and popularity (Zannettou et al., 2018; Ling et al., 2021). We ask a narrower question. We test whether resemblance to documented meme forms contributes information about engagement after accounting for whole-image similarity, surface features, posting time, and community.

Sparse Representations and Contrastive Family Signatures

Whole-image embeddings can be decomposed into high-dimensional sparse codes in which only a small subset of units is active for a given observation. This approach builds on sparse coding of natural signals (Olshausen & Field, 1996) and recent work using sparse autoencoders to recover structured features from learned representations (Cunningham et al., 2024; Zaigrajew et al., 2025). Sparse representations do not make individual features automatically interpretable, but they provide a feature basis in which a family can be represented by a small subset of active dimensions rather than by a dense whole-image vector.

Several familiar methods can operate in this space. Centroid and prototype methods summarize the location of class members, including methods that classify observations according to their distance from learned prototypes (Snell et al., 2017). Nearest-neighbor methods retain individual exemplars, while supervised linear classifiers estimate a discriminative boundary. Concept activation vectors provide another related approach by representing a concept as a direction in a learned feature space that can later be inspected or tested (Kim et al., 2018).

The contrastive signature considered here is a particular prototype construction within this broader family of methods. For each documented meme family, it identifies sparse features that are more active among members than among nonmembers, retains the strongest positive differences, and weights those retained features by the size of the contrast. The resulting vector therefore combines within-family recurrence with between-family distinctiveness. Its value depends empirically on three separate questions: whether this construction improves recovery over simpler summaries of the same sparse representation, how it compares with supervised classification, and what its resulting scalar resemblance score captures when applied to new observations.

Family Fit and Platform Engagement

A family-level resemblance score can also be used as a content measure in models of platform outcomes. Prior work has linked computationally measured content characteristics to social-media engagement (Cheng & Li, 2024), but engagement is difficult to predict and depends on factors including timing, social influence, ranking systems, and audience behavior (Cheng et al., 2014; Lindström et al., 2021; Salganik et al., 2006). An observed association with family resemblance could therefore reflect many correlated properties, including topic, emotional content, image quality, posting strategy, audience composition, or broader resemblance to familiar meme forms.

The downstream analysis asks whether the raw signature score contributes out-of-sample information after accounting for whole-image resemblance, surface similarity, posting time, and community. Any association is specific to the sampled Reddit communities and to a platform environment structured by pseudonymous participation, topical moderation, and community voting (Jungherr et al., 2022).

We ask three research questions:

- **RQ1.** How accurately does the contrastive family measure recover documented meme-family membership?
- **RQ2.** Under which adversarial-decoy conditions does the contrastive signature provide information beyond whole-image similarity?
- **RQ3.** Does the raw signature score carry incremental out-of-sample information about engagement among new-listing Reddit image posts after accounting for whole-image, surface, timing, and community variables, and how stable is that association across validation schemes, communities, and sensitivity analyses?

Table 1. Constructs, operational measures, and the interpretations each does and does not support.

Quantity	Operational definition	Criterion or evidence	Interpretation supported	Interpretation not supported
Documented family membership	External meme-reference label	Curated reference galleries	Ground-truth label for recovery evaluation	Latent family reality; human recognition
Raw signature resemblance	Cosine of a candidate's activations, restricted to a family's 15 contrastive features, with the family weight vector	Recovery AUROC versus baselines, in development and held-out families	Resemblance magnitude to a documented family signature	A membership probability comparable across families

Quantity	Operational definition	Criterion or evidence	Interpretation supported	Interpretation not supported
Family-calibrated fit	Raw resemblance standardized by the family's nonmember distribution	—	Relative extremity within a specific family	Cross-family magnitude, which standardization removes
Classifier score	Maximum one-versus-rest predicted probability (as analysed); the maximum decision score gives the same result	Documented labels	Discriminative evidence of documented membership	A calibrated membership probability
Reddit engagement	Within-bin top-quartile (binary) and log-scaled score (continuous)	Grouped cross-validation increment	A modest downstream association with raw resemblance magnitude	Virality, cultural success, causation, or family-specific fit

Note. The contrast d_f uses the training background N_f , which is disjoint from the evaluation negatives and decoys.

Method

Meme-Family Corpus

We compiled images from Know Your Meme (KYM; knowyourmeme.com), a public reference database whose entries document meme lineages and their variants. Development-corpus images were retrieved from entry galleries between 12 and 26 May 2026. Family labels follow the lineages documented by the source rather than visual groupings introduced for this study.

The development families were drawn from a candidate pool of 41 documented meme lineages assembled before data collection to include both image-template families and families defined by recurring text, layout, or relation, with documented variation over time. Gallery variants were retrieved for 32 pool families; two were subsequently merged (Supplement §S2), yielding 31 development families. No gallery variants were retrieved for the remaining nine, which were not included. The development corpus was therefore purposively constructed rather than probability-sampled from the broader population of meme families.

We removed byte-identical files using cryptographic hashes and near-duplicates using perceptual hashing, first within families and then across families. Cross-family duplicates were removed because the same image assigned to multiple family labels would mechanically affect recovery estimates. One pair of labels with near-total overlap was merged. After deduplication, the development corpus contained 680 variants across 31 documented families. Thirty families, comprising 678 variants, had at least five members and entered the recovery analysis. Among

those 30 families, size ranged from 15 to 26 variants, with a median of 24.5. The remaining family contained two variants and was excluded by the size threshold. The Supplement provides the complete family list, retrieval and inclusion procedures, deduplication thresholds, and overlap decisions.

The development-corpus collection procedure requested paginated Know Your Meme gallery routes that were listed under Disallow in the site's robots.txt. We retained the existing development corpus for provenance but did not repeat that retrieval procedure. Held-out-family collection was restricted to routes permitted by robots.txt, which limited retrieval to the first gallery page and produced shallower candidate families. Held-out families were consequently shallower, and we report family-size sensitivity analyses.

Because curated galleries may preferentially contain recognizable or canonical examples, the recovery criterion is documented reference-source membership.

Held-Out-Family Validation

The signature construction, including the choice of $K = 15$ retained features, was developed on the 30 development families. We therefore evaluated the principal recovery comparisons prospectively on held-out families. We fixed and timestamped the sampling frame, eligibility rules, retrieval procedure, comparison set, and interpretive rules before scoring any held-out family (Supplement §S1a).

The sampling frame was the Know Your Meme confirmed-entry sitemap. Candidates were screened in a seeded random order. The random stratum (H) was filled from that order. A separate look-alike challenge stratum (C) was constructed by a fixed token rule from entries related to development families. Eligibility covered Know Your Meme confirmed memes, excluding KYM-gated sensitive content. Retrieval was limited to robots-compliant first-page gallery routes. We applied the development deduplication rules within held-out families and resolved duplicate images across held-out families and against the development corpus.

Because first-page retrieval produced shallower families, eligibility required at least 10 distinct variants after deduplication, a threshold fixed before screening. We also report analyses capping every family at 10 and 15 variants. Screening ended under the pre-specified stopping rule with 60 families in H and 20 in C. Together with the 30 analyzed development families, these comprise 110 families in the recovery analyses. Screening dispositions are reported in the Supplement.

Held-out families were scored with the same $K = 15$ signature procedure and evaluated against the same comparison methods used in development, with the addition of a fused SigLIP2 + DINOv2 whole-image centroid.

Whole-Image and Sparse Representations

Each image was encoded with the pretrained OpenAI CLIP ViT-L/14 model (Radford et al., 2021). The 768-dimensional whole-image embedding was retained without unit normalization when passed to the sparse autoencoder and was unit-normalized for centroid-based similarity measures. We decomposed each embedding into a 6,144-dimensional sparse representation using a pretrained multimodal sparse autoencoder (Zaigrajew et al., 2025; Figure 2a). The Supplement reports the checkpoint, preprocessing steps, activation settings, normalization choices, reconstruction diagnostics, and software versions.

Family-Signature Estimation

For family f , let M_f denote its member set and N_f the nonmember training background. Let z_i denote the sparse activation vector for image i . We compute the feature-wise difference between mean activation among members and mean activation in the background:

$$d_f = \frac{1}{|M_f|} \sum_{i \in M_f} z_i - \frac{1}{|N_f|} \sum_{j \in N_f} z_j.$$

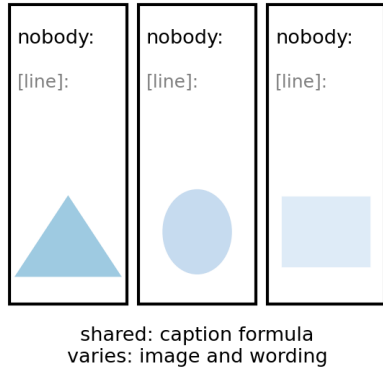
The contrastive signature retains the $K = 15$ features with the largest positive entries in d_f , using the corresponding contrast values as weights. During development, mean recovery AUROC increased from 0.878 at $K = 5$ to 0.899 at $K = 15$ and 0.907 at $K = 50$ (Figure S3a). We selected $K = 15$ on the development families as a compact representation retaining most of the recovery observed at $K = 50$, and used $K = 15$ for all subsequent analyses.

A family signature can therefore be represented either as 15 feature-index and weight pairs or as a 6,144-dimensional vector with 15 positive nonzero entries. For a candidate image c , raw resemblance to family f is the cosine similarity between the candidate's activations on those 15 retained features and the family's 15 weights. Activations on all other sparse features do not enter the score. Figure 2b shows the signature for trade offer, a development family selected by a fixed rule as the family with the median development-corpus recovery AUROC.

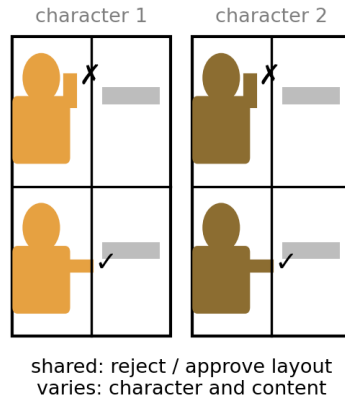
For each candidate, the raw signature score is the highest resemblance across the family bank. The specificity margin is the difference between the highest and second-highest family resemblance scores. These quantities share a numerical form but are not calibrated to a common membership scale across families.

The frequency-based recurrence baseline retains features that are frequently active among family members relative to a background activation threshold and serves as a simple recurrence-based family descriptor rather than a reimplementation of published motif- or template-mining systems. The Supplement provides the full notation, calibration procedures, comparison-method definitions, and sensitivity analyses.

(a) One family:
different images and captions



(b) One family: shared layout,
different base characters



(c) Similar-looking:
different families

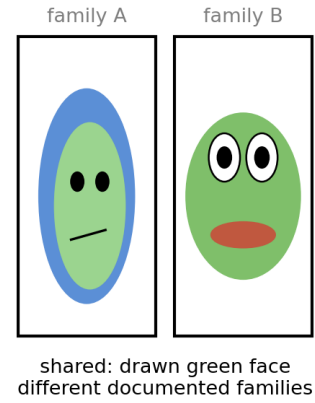


Figure 1. What we mean by a meme family. (a) Variants documented under one Know Your Meme entry share a recurring configuration while differing in image and caption. (b) Within one family, a shared reject/approve layout persists across different base characters; a single base image is one possible carrier of family structure, not the family itself. (c) Visually similar images can belong to different documented families. The measure developed here scores resemblance to the recurring and distinguishing structure of a documented family, not pairwise image similarity. Panels are schematic redraws of development-corpus examples; no held-out-family images are shown.

(a) Estimation pipeline

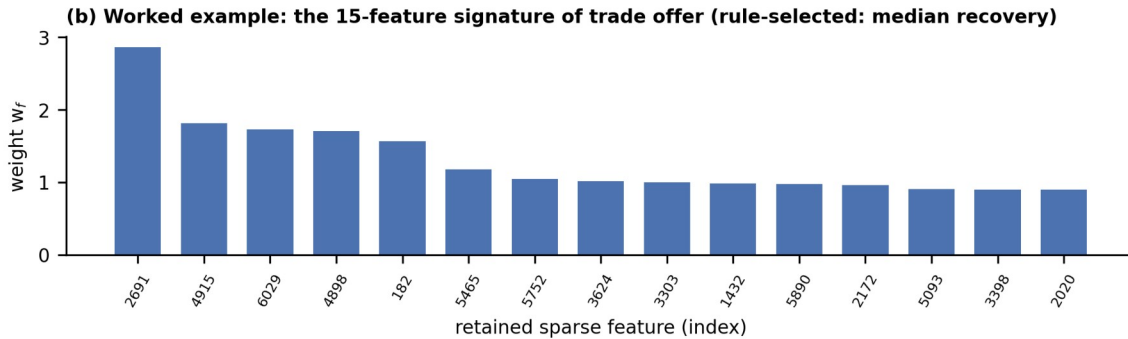
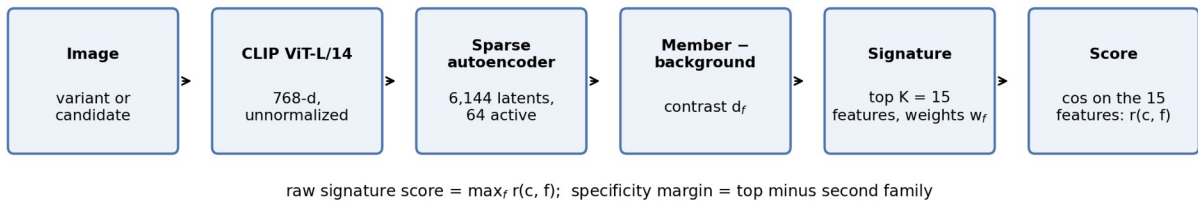


Figure 2. Measurement framework. (a) An image is encoded with OpenAI CLIP ViT-L/14 (768 dimensions, unnormalized for sparse decomposition) and decomposed by the multimodal sparse

autoencoder into a 6,144-dimensional sparse code with 64 active units. Family-member activations are contrasted with the training background, and the 15 largest positive differences and their weights form the family signature. A candidate is scored by cosine similarity between its activations on those 15 features and the signature weights. The maximum resemblance across families is the raw signature score, and the difference between the highest and second-highest scores is the specificity margin. The fixed pipeline is evaluated by leave-one-variant-out recovery in the development and held-out families (RQ1–RQ2) and then applied to Reddit posts (RQ3). (b) The worked example shows the 15 retained features for the trade offer family, ordered by contrastive weight. Sparse-feature indices appear on the horizontal axis. Feature semantics are not assigned.

Comparison Measures

The comparison set was designed to separate the effect of the contrastive construction from the information available in alternative representations of the same images. It included a frequency-based recurrence baseline, the contrastive signature, a whole-image CLIP centroid, a sparse-space centroid, a sparse-space nearest-neighbor rule, a regularized logistic classifier trained on the same sparse features, and surface-similarity measures. The held-out-family evaluation additionally included a pre-specified fused SigLIP2 + DINOv2 whole-image centroid using fixed checkpoints and equal-weight fusion.

Surface measures captured recurring text and image reuse. Text was extracted with the macOS Vision optical-character-recognition framework through the ocrmac library. Text similarity was the cosine similarity between character n-gram TF-IDF vectors using n-grams of length two through four, implemented in scikit-learn (Pedregosa et al., 2011). Image similarity was measured with the discrete-cosine-transform perceptual hash implemented in the imagehash library (Zauner, 2010). We report a fixed composite surface measure defined without selection on evaluation performance and a per-family oracle comparator that takes the best-performing surface measure for each family. The oracle provides a deliberately favorable descriptive benchmark.

Family-Recovery Evaluation

Family recovery was evaluated with leave-one-variant-out estimation. For each held-out family member, all quantities that depended on family membership, including feature selection and signature weights, were recomputed from the remaining $n-1$ members. The held-out image therefore never contributed to the representation used to score it.

For each target family, recovery was evaluated against 200 randomly drawn nonmembers from the other documented families, including the two-variant family excluded from the recovery analysis. These evaluation negatives were not used to estimate any comparison method. Training backgrounds were drawn from a separate pool of remaining nonmembers, with a mean of 435 and a minimum of 428 per family. Adversarial decoys were also excluded from every training

background and evaluated separately. All methods, including the supervised classifier, were fitted within the same leave-one-variant-out loop using the same $n-1$ target-family members. AUROC was calculated within each fold and then averaged. The held-out-family evaluation used the same procedure, except that each target family's 200 evaluation negatives were sampled uniformly from variants of the other families in the 110-family evaluation universe; these negatives were likewise kept separate from every training background.

For a target family, recovery AUROC is the probability that a randomly chosen held-out member receives a higher score than a randomly chosen nonmember. A value of 0.5 indicates chance ranking and 1.0 perfect ranking. Mean recovery is the average family-level AUROC across the relevant family set. Paired advantages are calculated as within-family differences in recovery AUROC between two methods and then averaged across families, with uncertainty estimated by a family-level bootstrap. Discrimination uses the same AUROC definition but replaces random nonmembers with adversarial decoys.

The family is the unit of aggregation throughout these comparisons. We report family-level distributions, paired method differences, and descriptive head-to-head counts. Head-to-head counts descriptively report the number of families in which the contrastive signature exceeds a comparator by more than 0.02. Reported 95% intervals are per-comparison intervals without multiplicity adjustment. Held-out comparisons and interval-interpretation rules were fixed before scoring.

Adversarial Decoys

We constructed two 12-item adversarial-decoy sets for each family. One selected the nearest nonmembers under a separate vision-language encoder trained on different data. The other selected nonmembers using perceptual hashing, providing a non-CLIP selection criterion. True family members were excluded from the candidate pools, and cross-family near-duplicates were removed before selection. Both decoy sets were then scored with the primary representation. These decoys test whether family recovery persists when negatives are deliberately chosen to resemble the target family under criteria different from the contrastive score itself.

Reddit Sample

We collected public image posts from the new listings of seven Reddit communities: r/memes, r/dankmemes, r/meme, r/me_irl, r/shitposting, r/whenthe, and r/historymemes. Sampling from new listings reduces conditioning on realized engagement relative to sampling from top or hot listings, although inclusion still depends on successful posting, survival until collection, subreddit rules, moderation, and interface visibility. Top and hot listings were queried separately to construct a selected-listing contrast.

The rising listing was queried after the new listing but contributed no additional posts once previously observed post identifiers were removed. The denominator sample therefore consists

entirely of posts collected from new listings between April 26 and May 27, 2026. Posts were deduplicated by perceptual hash before analysis.

The final denominator contained 4,263 posts across the seven communities, with community counts ranging from 357 to 749. The selected-listing contrast contained 348 posts. Four additional posts retrieved from hot listings were not included in either analytic sample. For each post, we recorded score, comment count, upvote ratio, creation time, community, and image. Time-invariant content features were computed once per post. The Supplement reports the complete community counts and sample flow.

Engagement Outcomes

We constructed binary and continuous engagement outcomes within community and posting-age bins. The binary outcome identifies posts in the top quartile of engagement within their bin. The continuous outcome uses a log-scaled engagement score. Community and posting age therefore enter both the outcome construction and the predictive models.

For the binary analysis, bin-specific thresholds were recomputed within each training fold so that labels for held-out observations did not depend on the full-sample outcome distribution. Binary and continuous outcomes were analyzed separately because changes in AUROC and R^2 have different scales and interpretations. The Supplement gives the outcome formulas and binning rules.

Predictive Models

The control model included whole-image similarity, surface measures, posting time, and community. The full model added the raw signature score, defined as the maximum raw resemblance across the family bank, and the specificity margin. Binary engagement was modeled with regularized logistic regression and evaluated by AUROC. Continuous engagement was modeled with ridge regression and evaluated by out-of-sample R^2 . Predictor standardization was estimated separately within each training fold. In the decomposition analysis, representation-magnitude and text-length controls comprise the whole-image embedding norm, the sparse activation-vector norm, and OCR text length.

The quantity of interest is the change in out-of-sample performance when the signature variables are added to the control model. The raw signature score is an uncalibrated resemblance magnitude, so the engagement analysis concerns predictive information in resemblance magnitude rather than calibrated membership. Family-calibrated scores are examined separately in the decomposition analysis.

We report the performance increment under four validation schemes, each imposing a different generalization constraint (Table 3). Random post-level folds evaluate new posts drawn from the same sampled population. Perceptual-hash-cluster folds keep detected near-duplicates together. This grouping was specified after the random-fold and leave-one-community-out results were known but before the perceptual-hash-cluster analysis was run. Community-grouped validation

uses leave-one-community-out prediction across the seven communities. Best-fit-family grouping keeps together posts whose highest raw resemblance is to the same family signature (Roberts et al., 2017).

None of the four schemes is designated as primary. Best-fit-family grouping, which yields the largest increment, was selected after the initial results and is treated as post hoc in the decomposition, randomization, and sensitivity analyses (Supplement §S12). Because every family signature is estimated before the Reddit cross-validation stage, best-fit-family grouping tests generalization across groups of posts assigned to different existing signatures. It does not test performance on family signatures absent from estimation.

Perceptual-hash grouping primarily controls near-duplicate dependence because 95.6% of posts form singleton clusters. It therefore does not generally place broader template siblings in the same fold. Best-fit-family grouping reduces dependence among posts assigned to the same family signature, but template siblings can still receive different best-fit assignments. Residual template dependence can remain under all four schemes.

Uncertainty and Sensitivity Analysis

For grouped engagement estimands, uncertainty was evaluated by resampling the relevant grouping units. In the best-fit-family analysis, groups defined by the highest-scoring family signature in the 31-signature bank were sampled with replacement. The grouped cross-validation pipeline was then refitted for each of 2,000 bootstrap iterations following cluster-aware resampling principles (Cameron et al., 2008). If a group was sampled more than once, its copies remained within the same cross-validation group so that duplicated observations could not enter both training and test sets.

These intervals quantify variation across best-fit-family groups. They do not incorporate dependence associated with community, collection period, or residual template structure. We also report the proportion of bootstrap performance increments above zero as a descriptive summary rather than as a hypothesis-test p-value.

Negative controls permuted the raw signature score within community and posting-age bins and, separately, refit the models after randomly shuffling the outcome. Sensitivity analyses removed individual communities, excluded the high-engagement tail, and restricted the analysis to the selected-listing contrast. Table 3 records the sample, grouping unit, number of groups, split rule, generalization target, and remaining dependence for each analysis using the same terminology as the Results.

Results

Family Recovery Across Contrastive, Prototype, and Supervised Methods (RQ1)

In the 30-family development corpus, the contrastive signature recovered documented family membership more accurately than the recurrence baseline, the sparse centroid and nearest-

neighbor rules, the surface measures, and the whole-image CLIP centroid. The supervised logistic classifier was more accurate than the contrastive signature (Table 2, Panel 2A). Because the signature construction and the choice of $K = 15$ were developed on these families, the held-out-family results provide the more informative test of how these comparisons generalize.

Among the 60 randomly sampled held-out families, the contrastive signature retained clear recovery advantages over the recurrence baseline, the sparse centroid and nearest-neighbor rules, and both surface measures, with family-bootstrap intervals excluding zero (Table 2, Panel 2B). No clear contrastive advantage was detected over either whole-image dense centroid: the intervals for the CLIP centroid and the fused SigLIP2 + DINOv2 centroid both include zero. Thus, the development-stage advantage over the CLIP centroid did not replicate in the randomly sampled held-out families. Development and held-out evaluations differ in family depth and in the evaluation-negative universe, so the observed change cannot be attributed uniquely to development status, family size, or negative composition (Supplement §S7a). The same qualitative verdict held when family size was capped at 10 or 15 members and across ten redraws of the evaluation negatives.

The supervised classifier remained the strongest estimator of documented family membership, with higher mean recovery than the contrastive signature in the random held-out families and an interval excluding zero (Table 2, Panel 2B). The supervised classifier was therefore the strongest membership estimator. The contrastive signature instead represents each family with a small positive feature set and supplies the scalar resemblance measure used downstream.

The purposively constructed look-alike challenge stratum exposed a further boundary condition. Among these 20 families, the contrastive advantages over the sparse centroid and nearest-neighbor rule were uncertain, and the fused SigLIP2 + DINOv2 centroid outperformed the contrastive signature, with an interval excluding zero (Table 2, Panel 2B). Because these families were selected by a fixed look-alike rule rather than sampled from the Know Your Meme frame, this result is not a population estimate. It shows that the sparse-baseline advantages become uncertain under some deliberately difficult family configurations and that a strong dense representation can outperform the contrastive construction.

Table 2. Family recovery.

Panel 2A. Development corpus (30 families; leave-one-variant-out protocol). Mean recovery AUROC, the paired contrastive-minus-method difference with its 95% family-bootstrap interval, and the number of families where the contrastive signature exceeds the method by more than 0.02.

Method	Representation	Supervision	Mean AUROC	Contrastive – method [95% CI]	Families > 0.02
Frequency-based recurrence baseline	Sparse	Unsupervised prototype	0.790	+0.109 [0.076, 0.146]	23 of 30
Whole-image CLIP centroid	CLIP	Unsupervised prototype	0.843	+0.056 [0.025, 0.088]	19 of 30
Surface composite (fixed)	pHash / OCR / template	None	0.779	+0.120 [0.061, 0.182]	23 of 30
Oracle surface (per-family maximum)	pHash / OCR / template	Outcome-informed	0.812	+0.087 [0.037, 0.138]	21 of 30
Sparse centroid	Sparse	Unsupervised prototype	0.824	+0.075 [0.051, 0.102]	24 of 30
Sparse nearest neighbor	Sparse	Unsupervised memory	0.813	+0.086 [0.058, 0.114]	25 of 30
Contrastive signature (this study)	Sparse	Weak (contrast)	0.899	—	—
Supervised logistic classifier	Sparse	Supervised	0.963	−0.064 [−0.085, −0.045]	classifier higher by > 0.02 in 21 of 30

Note. All methods were fitted within one common leave-one-variant-out loop with evaluation negatives disjoint from every training background (Method). Under encoder-selected (LAION-CLIP) decoys, the contrastive-minus-CLIP-centroid discrimination difference was +0.181 [0.128, 0.237]. Under perceptual-hash-selected decoys, no clear contrastive advantage was detected (+0.036 [−0.001, 0.074]). The development-stage advantage over the CLIP centroid did not replicate in the random held-out stratum (Panel 2B).

Panel 2B. Held-out-family validation. Random stratum H (60 families sampled at random from Know Your Meme confirmed entries that retained at least 10 distinct first-page variants) and, separately, the purposively constructed look-alike challenge stratum C (20 families). Cells report each method's mean recovery AUROC and its paired contrastive-minus-method difference with a 95% family-bootstrap interval.

Method	H (random) mean	H: contrastive – method [95% CI]	C (purposive challenge) mean	C: contrastive – method [95% CI]
Frequency-based recurrence baseline	0.900	+0.049 [0.032, 0.067]	0.864	+0.063 [0.022, 0.105]
Sparse centroid	0.921	+0.027 [0.015, 0.039]	0.908	+0.019 [−0.010, 0.049]
Sparse nearest neighbor	0.918	+0.030 [0.016, 0.044]	0.901	+0.026 [−0.016, 0.068]
Surface composite (fixed)	0.779	+0.169 [0.109, 0.231]	0.778	+0.148 [0.067, 0.233]
Oracle surface (per-family maximum)	0.825	+0.123 [0.077, 0.170]	0.790	+0.137 [0.078, 0.198]
Whole-image CLIP centroid	0.950	−0.002 [−0.013, 0.008]	0.946	−0.019 [−0.044, 0.003]
Fused SigLIP2 + DINOv2 centroid	0.954	−0.006 [−0.021, 0.010]	0.964	−0.038 [−0.070, −0.011]
Contrastive signature (this study)	0.948	—	0.927	—
Supervised logistic classifier	0.983	logistic – contrastive: +0.035 [0.024, 0.046]	0.981	logistic – contrastive: +0.055 [0.029, 0.088]

Note. C is a separately constructed look-alike challenge stratum and is not generalized to the Know Your Meme frame. Intervals are 95% family-bootstrap intervals. An interval containing zero is interpreted as no clear contrastive advantage. Under encoder-selected (LAION-CLIP) decoys, the contrastive-minus-CLIP-centroid discrimination difference was +0.073 [0.041, 0.108] in H and +0.048 [−0.025, 0.126] in C. Under perceptual-hash-selected decoys, no clear contrastive advantage was detected in H (−0.008 [−0.023, 0.007]).

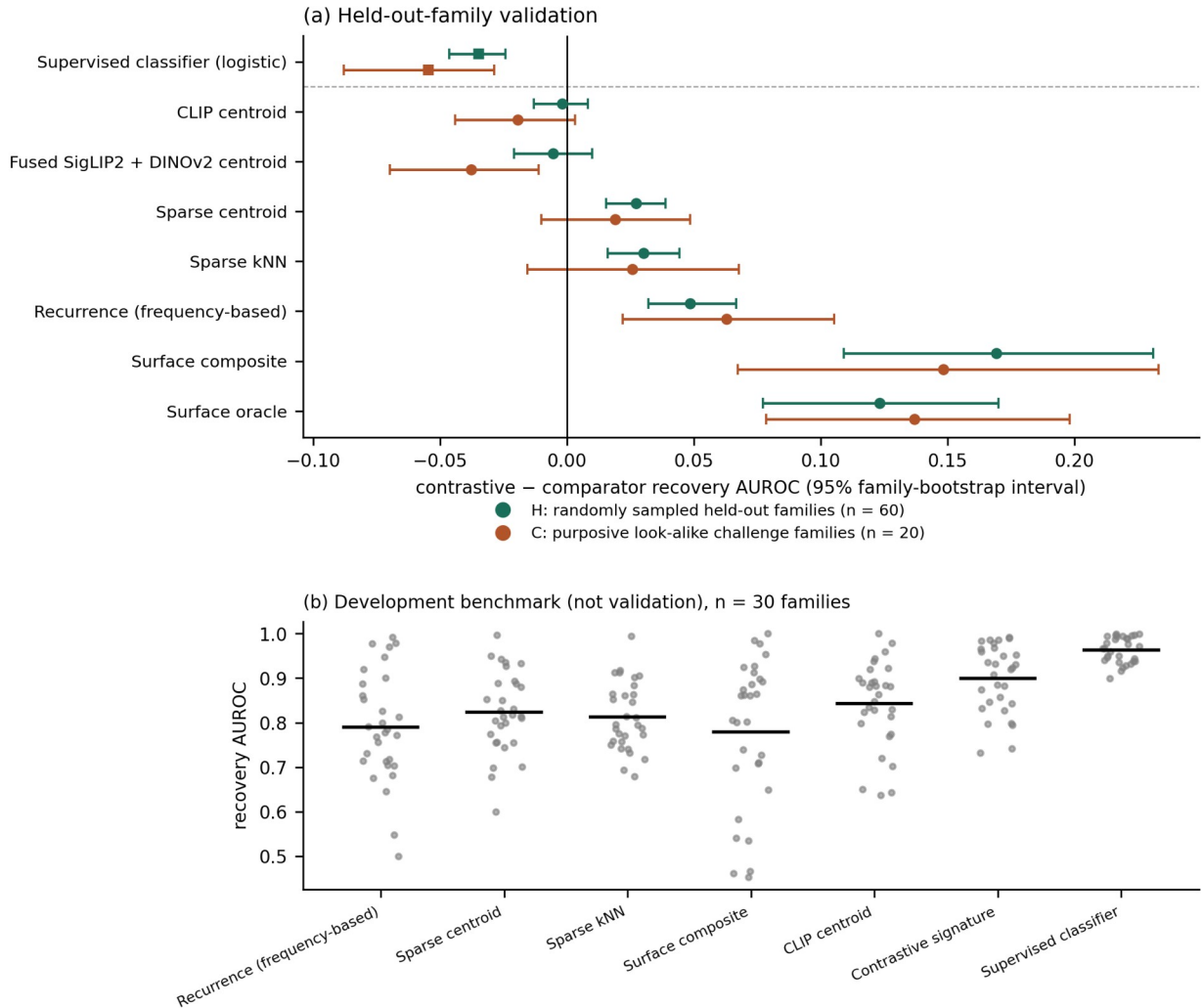


Figure 3. Family recovery. Panel (a) shows paired contrastive-minus-comparator differences in recovery AUROC with 95% family-bootstrap intervals for the randomly sampled held-out families (H) and the separate purposive look-alike challenge stratum (C). The supervised logistic classifier appears above the dashed separator in the same contrastive-minus-comparator orientation, so negative values indicate higher classifier recovery. Other comparators include the surface oracle and fused SigLIP2 + DINOv2 centroid. C denotes the purposive look-alike challenge stratum. Panel (b) shows recovery AUROC by method across the 30 development families, with points representing families and horizontal marks representing method means. The development-stage advantage over the CLIP centroid did not replicate in the random held-out stratum. Additional development diagnostics appear in Figure S13.

Illustrative Cases

Recovery varied substantially across development families (Figures S4 and S13). We selected three development-corpus cases using a rule specified before case inspection: the family with the largest contrastive-minus-CLIP-centroid advantage under encoder-selected decoys (pov you are

x), the family with the smallest such advantage (wojak soyjak), and the family with the largest supervised-classifier advantage over the contrastive signature (distracted boyfriend). For each family, the Supplement reports the retained features and the highest-scoring nonmember under that family's signature. All three examples come from the development corpus.

Performance Under Adversarial Decoys (RQ2)

The contrastive signature showed its clearest advantage over the CLIP centroid when negatives were selected to be difficult under a separate vision-language encoder. Under decoys selected by LAION-CLIP (a CLIP model trained on the public LAION dataset), the contrastive signature discriminated members from decoys better than the CLIP centroid in the development corpus and in the 60 randomly sampled held-out families, with intervals excluding zero (notes to Table 2; Figure 4). Selecting decoys separately within each fold produced the same qualitative result (Supplement §S7a). In the purposive challenge stratum, the corresponding difference was positive but uncertain, with an interval including zero.

The result did not extend to perceptual-hash-selected decoys. No clear contrastive advantage over the CLIP centroid was detected under that selection rule in either the development corpus or the random held-out families. The hard-decoy advantage over the CLIP centroid is therefore conditional on how difficult negatives are defined rather than a general advantage under visual similarity.

Because LAION CLIP and the CLIP centroid belong to the same broad encoder family, this comparison tests whether the contrastive sparse construction separates members from cases that are difficult for CLIP-family whole-image similarity. It does not test visual confusion generally. The supervised classifier also outperformed the contrastive signature under these decoys (Supplement Table S9).

We added the fused SigLIP2 + DINOv2 comparison under encoder-selected decoys after examining the held-out results, so we treat it as exploratory (Supplement §S7a). The contrastive advantage was present in the random held-out families but not in the challenge stratum.

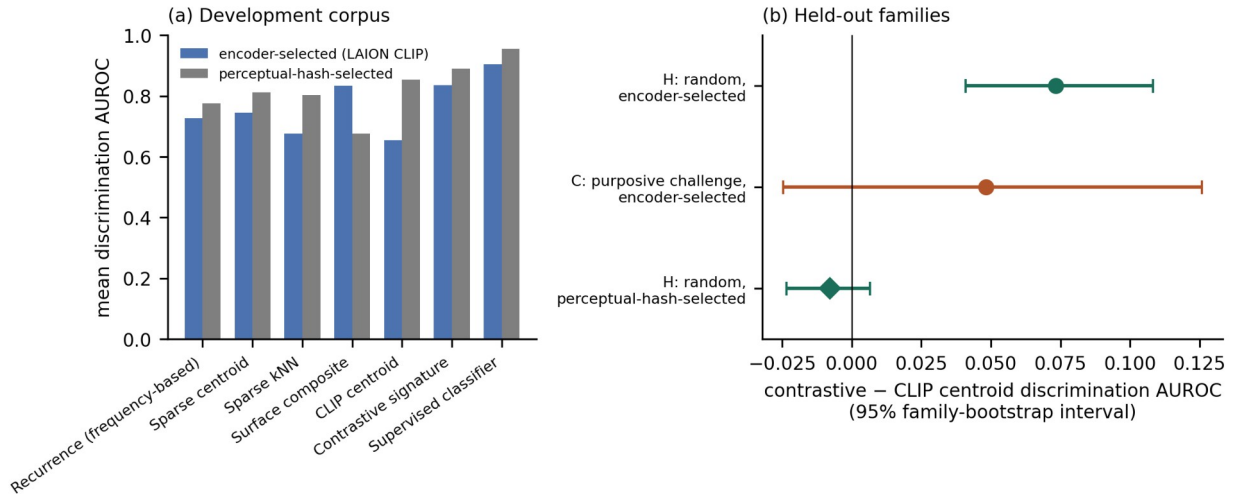


Figure 4. Adversarial discrimination. (a) Mean discrimination AUROC by method under encoder-selected and perceptual-hash-selected decoys in the development corpus. (b) Held-out contrastive-minus-CLIP-centroid discrimination differences with 95% intervals under encoder-selected decoys for the random stratum H and purposive challenge stratum C, and under perceptual-hash-selected decoys for H. The interval excludes zero for encoder-selected decoys in H, but includes zero in C. No clear contrastive advantage was detected under perceptual-hash-selected decoys. Per-family distributions appear in Figure S6.

The Raw Signature Score Carries Incremental Information About Reddit Engagement (RQ3)

Adding the raw signature score and specificity margin to the control model increased the point estimate of out-of-sample predictive performance under all four validation schemes (Table 4). For the continuous outcome, the ΔR^2 estimates ranged from 0.046 to 0.072 depending on the grouping scheme. Intervals were estimated for the perceptual-hash-cluster and best-fit-family schemes. These increments are modest in absolute terms and describe predictive association rather than a causal effect of family resemblance on engagement.

The scale of the increment should be read against the performance of the control model. Under random folds, posting time, community, whole-image similarity, and surface measures together explained 0.107 of the variance in log engagement. When entire communities were held out, the control-model R^2 was -0.018 (Table 4). The larger increments under some grouped schemes therefore partly reflect weaker baseline prediction rather than a larger underlying effect of the signature variables.

For scale, 27.87% of the 427 posts in the top tenth of the raw-signature-score distribution were classified as high engagement, compared with 22.52% in the full 4,263-post sample, a difference of 5.35 percentage points. The full-sample rate is below 25% because this descriptive comparison uses stored labels, whereas the cross-validated models recompute thresholds within training folds. Inferential claims rely on the cross-validated performance increments.

The best-fit family assignments were unevenly distributed across the 31-signature bank. “Nobody / literally no one” was the best-fitting signature for 1,701 of 4,263 posts (39.9%) and for 409 of the 427 posts in the top tenth of the raw-score distribution. Because raw scores are not calibrated across families, the top-decile contrast describes the pooled score distribution rather than a uniform cross-family association.

The random and perceptual-hash-cluster schemes produced similar continuous increments, leave-one-community-out validation a slightly larger one, and best-fit-family grouping the largest (Table 4). Because best-fit-family grouping was selected after the initial results, we treat its larger increment as post hoc. Across the four schemes, the stable pattern is the direction of the point estimate, not a common effect size.

Additional sensitivity analyses tested the generality of this pattern (Figure S9; Supplement Table S17). In the selected-listing contrast ($n = 348$), the performance increment was near zero and its cluster-bootstrap interval included zero. The association observed in the new-listing denominator was therefore not reproduced among these already surfaced posts. The selected-listing contrast uses the stored scores and predictor block for that sample rather than the denominator-sample models reported above, and is based on a smaller, selectively surfaced sample.

The remaining sensitivity analyses did not identify a single community or extreme-engagement tail as the source of the denominator-sample association. Permuting the raw signature score within community and posting-age bins removed the increment, and models fitted to randomly shuffled outcomes performed at chance. Bivariate correlations between raw resemblance and log engagement were positive in all seven communities (Supplement Table S16), although the leave-one-community-out increment was negative when *r/whenthe* was held out (Figure S12). Under best-fit-family grouping with the raw score alone, the continuous increment also remained positive after excluding *r/whenthe* and after removing the top 1% or 5% of posts by engagement (Supplement Table S17).

The cluster-bootstrap intervals reported in the Table 4 note quantify variation across their respective grouping units only. They do not incorporate dependence associated with community, collection period, or residual template structure.

Decomposing the Engagement Association

The engagement result does not by itself identify what information in the signature score carries the association. We therefore compared alternative scores derived from the same sparse representation using all 4,263 posts from the seven communities (Supplement §S16).

Decomposition models add one score at a time, so the raw-signature increment here (0.068) omits the specificity margin included in Table 4 (0.072).

Raw signature resemblance and sparse-centroid resemblance produced similar downstream point increments (difference not tested), whereas family-calibrated fit and classifier membership scores had point estimates near zero, with no interval estimated for those comparisons (Figure

5a; Supplement Table S18). Because the classifier was the strongest estimator of documented membership while the sparse centroid produced a similar downstream increment to the contrastive signature, the engagement association cannot be attributed specifically to correct family identity or to the contrastive construction.

Random-feature signature banks produced negligible increments, and none of the 500 banks reached the observed increment (Monte Carlo $p = 0.002$). Label-permuted signatures, which preserve the sparse representation while breaking the documented family assignments, retained a substantial fraction of the observed association. The observed increment lay above most label-permuted values (Monte Carlo $p = 0.030$ – 0.045 across four seeds), although some permuted banks reached higher values (Figure 5b–c; Supplement Table S19).

Adding representation-magnitude and text-length controls reduced the remaining raw-signature increment to a small and uncertain residual (Figure 5a; Supplement Table S18). Taken together, these results locate much of the association in broader structured resemblance within the sparse representation and in representation magnitude, rather than in calibrated evidence of correct documented-family identity. We use canonical resemblance to documented meme forms as a descriptive label for this broader pattern, not as an independently validated cultural construct.

Table 3. Validation designs and generalization targets for the engagement analysis.

Analysis	Sample	Grouping unit	Groups	Split rule	Generalization question	Remaining dependence
Random 5-fold	Denominator (4,263)	none	—	5-fold shuffle	New posts, same population	template, community, time
Perceptual-hash cluster (Hamming ≤ 10)	Denominator (4,263)	perceptual-hash cluster	4,148	GroupKFold	New near-duplicate groups	near-duplicate control only: 95.6% of posts are in singleton clusters; template siblings not grouped
Community grouped (= leave-one-community-out, 7 groups)	Denominator (4,263)	community	7	GroupKFold (7 groups)	A held-out community	template, time
Best-fit-family grouped	Denominator (4,263)	best-fit family signature	30	GroupKFold	Posts sharing a held-out best-fit signature	residual template
Cluster bootstrap	Denominator (4,263)	best-fit family; pHash cluster; 2,000 iterations	—	grouped refit	Uncertainty across grouping units	community, time, template
Downstream decomposition	Denominator (4,263)	best-fit family	30	GroupKFold	Score construction	residual template

Note. The continuous outcome is $\log_{10}(\text{score} + 1)$; the binary outcome is the top quartile within community-by-posting-age bins. The control predictor block is whole-image similarity, surface measures, posting time, and community; the full block adds the raw signature score and specificity margin. No scheme is treated as primary. The downstream bank contains 31 family signatures, comprising the 30 analysed families plus the excluded two-variant family. Thirty signatures are the best fit for at least one Reddit post, producing the best-fit-family groups shown.

Table 4. Out-of-sample engagement performance across validation designs, ordered by generalization target. Denominator sample $n = 4,263$. AUROC is the binary outcome, and R^2 is the continuous outcome.

Validation scheme	R^2 control	R^2 full	ΔR^2	AUROC control	AUROC full	Δ AUROC
Random 5-fold	0.107	0.153	+0.046	0.513	0.596	+0.083
Perceptual-hash cluster (Hamming ≤ 10)	0.106	0.152	+0.046	0.518	0.601	+0.082

Validation scheme	R^2 control	R^2 full	ΔR^2	AUROC control	AUROC full	Δ AUROC
Community grouped (= leave-one-community-out)	-0.018	0.035	+0.052	0.547	0.617	+0.070
Best-fit-family grouped	0.057	0.129	+0.072	0.525	0.620	+0.094

Note. The raw signature score is an uncalibrated resemblance magnitude, so these increments test whether resemblance magnitude carries predictive information rather than whether calibrated family membership predicts engagement. Cluster-bootstrap 95% intervals for the continuous increment are [0.011, 0.149] under best-fit-family grouping and [0.034, 0.058] under perceptual-hash-cluster grouping. Across repeated cross-validation seeds 1–20, the median continuous increments were +0.046 for random folds, +0.046 for perceptual-hash-cluster folds, and +0.072 for best-fit-family folds; community folds are deterministic. Figure S12 shows the per-fold distributions. Fold-level t-tests are not used because cross-validation folds share training observations and therefore do not provide independent estimates. The repeated-cross-validation medians and per-fold distributions are robustness diagnostics, while the cluster-bootstrap intervals quantify uncertainty across the relevant grouping units. In the selected-listing contrast ($n = 348$), the continuous increment is -0.002 with a cluster-bootstrap 95% interval of $[-0.163, 0.031]$ (Supplement Table S17). Figure 5 and the Supplement report the score-construction decomposition.

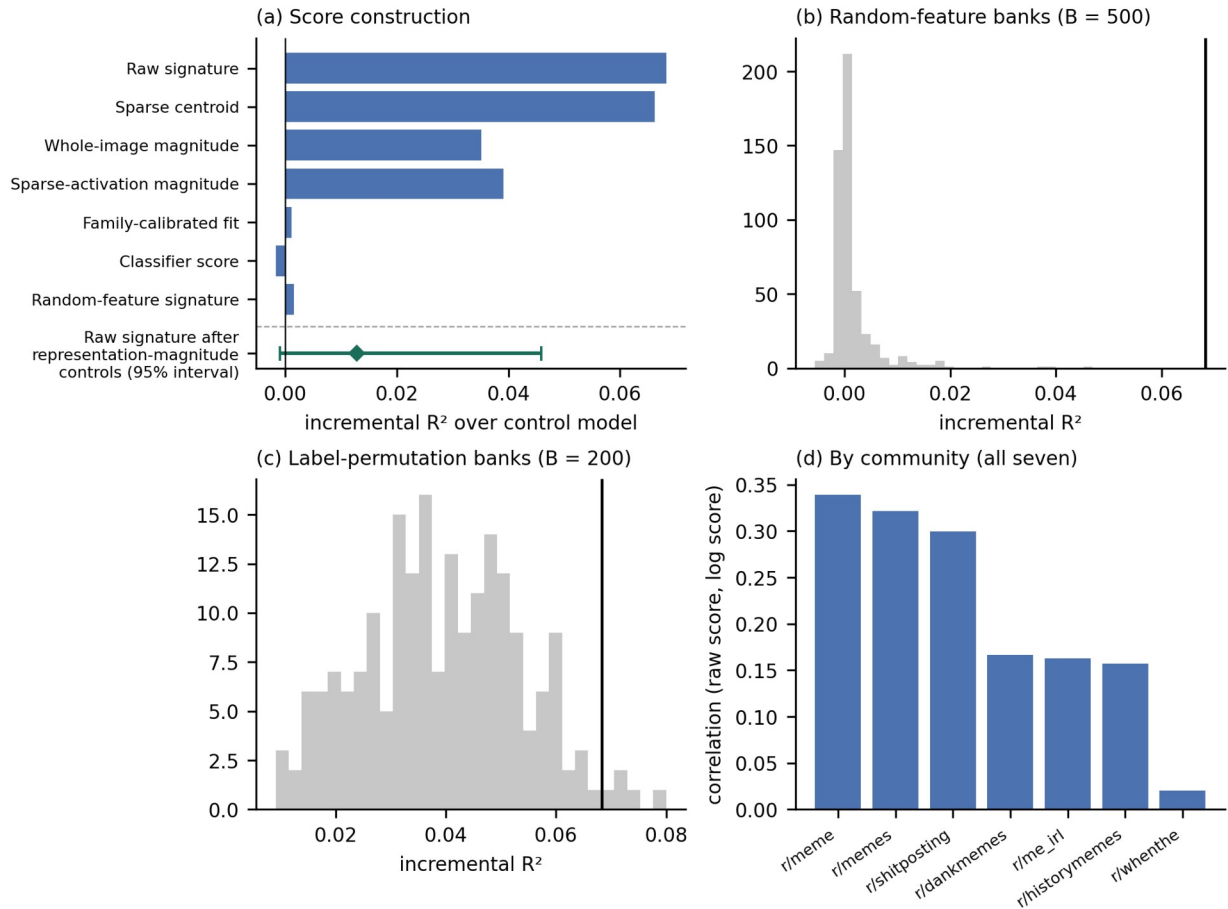


Figure 5. Signal decomposition and randomization controls for the engagement association (all 4,263 posts; family-grouped cross-validation). (a) Incremental R^2 by score construction: raw signature resemblance, sparse-centroid resemblance, whole-image and sparse-activation magnitude proxies, family-calibrated fit, the classifier score, and a random-feature signature maximum. Below the dashed line, the raw-signature increment is shown after representation-magnitude and text-length controls, with its 95% cluster-bootstrap interval. Positive values indicate predictive information beyond the control model. (b) Distribution of the increment across 500 random-feature signature banks, with the observed increment of 0.068 marked (Monte Carlo $p = 0.002$). (c) Distribution across 200 label-permutation signature banks, which preserve the empirical sparse representation while breaking documented family identity, with the observed increment marked (Monte Carlo $p = 0.030$ for the displayed run). (d) Correlation between raw signature resemblance and engagement by community for all seven communities in the denominator sample. The decomposition indicates that much of the downstream association lies in broad sparse-space resemblance and representation magnitude rather than in calibrated family-specific fit, and that much of the raw-signature increment is absorbed by representation-magnitude and text-length controls.

Prospective Engagement Growth

We also asked whether the raw signature score measured at collection predicted subsequent engagement growth after initial engagement was taken into account. Neither of the two completed short-window waves showed detectable incremental prediction. The planned longer-horizon waves could not be collected: beginning 29 May 2026, Reddit returned HTTP 403 responses at the IP level to the unauthenticated public endpoints used for collection, and authenticated access through Reddit's Data API required a registration and approval process that could not be completed during the study window. Prediction over longer horizons therefore remains unevaluated (Supplement §S17).

Discussion

Main Findings

This study evaluates a compact contrastive representation of documented meme families against simpler prototypes, whole-image representations, and supervised classification. The held-out results substantially narrow the claim suggested by the development corpus. Among 60 randomly sampled held-out families, the contrastive signature recovered documented membership more accurately than the frequency-based recurrence baseline, same-representation sparse centroid and nearest-neighbor rules, and surface measures. We detected no clear contrastive advantage over either the CLIP centroid or the fused SigLIP2 + DINOv2 centroid. In the separate look-alike challenge stratum, the fused centroid outperformed the contrastive signature. The supervised classifier was also more accurate than the contrastive signature in both development and random held-out families. The contrastive method is therefore not the strongest general estimator of documented family membership.

Its more specific recovery advantage appears under one adversarial condition. When nonmembers were selected because they were difficult for a separately trained CLIP-family encoder, the contrastive signature discriminated members from decoys more accurately than the CLIP centroid in the random held-out families. The advantage was specific to encoder-selected decoys and did not extend to perceptual-hash-selected decoys. This identifies a setting in which a sparse contrastive representation contributes information beyond CLIP-family whole-image similarity.

The Reddit application concerns a different property of the score. Raw signature resemblance contributed a modest increment to out-of-sample engagement prediction across several validation schemes, but the decomposition does not support attributing that increment specifically to correct family membership or to the contrastive construction. Sparse-centroid resemblance produced a similar point increment, family-calibrated and classifier scores added near-zero point increments, label-permuted signatures retained a substantial fraction of the association, and much of the raw-signature increment was absorbed by representation-magnitude and text-length controls. These

results distinguish documented-family recovery from a broader pattern of structured resemblance in the sparse representation.

Measurement Contribution

The contrastive signature sits between a simple prototype and a supervised classifier. Its 15 positive features are selected because they recur more strongly among members of a documented family than in a nonmember background. The resulting object is compact and explicit: each family is represented by a fixed set of feature indices and weights that can be applied to new observations without refitting on the downstream corpus. This structure does not make the individual sparse features inherently interpretable, but it does make the representation itself compact and explicit in a way that a dense centroid or a full discriminative model is not.

Compactness does not substitute for recovery accuracy. The supervised logistic classifier recovered documented membership more accurately, demonstrating that the underlying sparse representation contains discriminative information that the 15-feature signature does not retain. The held-out comparisons likewise provide no clear evidence that the contrastive construction improves ordinary family recovery over strong whole-image dense centroids. Empirically, the signature provides a reusable sparse family representation that outperforms simpler sparse and surface alternatives, preserves a small explicit feature set, and exceeds the CLIP centroid in random held-out families under hard decoys selected by a separately trained CLIP-family encoder.

The downstream results further show why recovery performance cannot determine a measure's substantive interpretation by itself. The classifier was the strongest family-membership estimator, but its scores added only a near-zero point increment to engagement prediction. Raw sparse-space resemblance did, whether measured with the contrastive signature or the sparse centroid. Stronger label recovery therefore did not translate into a stronger association with the downstream outcome. Computational measures can differ across recovery, calibration, stability, representational form, and downstream behavior, and those properties need not rank methods in the same order (Kirchmair et al., 2026; Törnberg, 2025).

What the Measure Makes Possible

The signature supports three operations. First, it represents each documented family as one reusable object and assigns new images a scalar resemblance score without estimating a new classifier on the application corpus. A whole-image centroid shares this reusable-object property. In the random held-out families, the signature recovered documented membership more accurately than the frequency-based recurrence baseline, sparse centroid and nearest-neighbor rules, and surface measures, but no clear contrastive recovery advantage was detected over either the CLIP or fused dense centroid. The signature's distinctive features are its sparse, inspectable form and its advantage over the CLIP centroid in the random held-out families for hard decoys selected by a separately trained CLIP-family encoder.

Second, the representation allows the downstream association to be decomposed into quantities that behave differently. Raw resemblance, family-calibrated fit, classifier evidence, randomly structured signatures, label-permuted signatures, and representation-magnitude and text-length controls do not produce the same engagement result. This makes it possible to ask whether an observed downstream association is specifically tied to documented family identity or instead reflects broader structure in the representation.

Third, each family signature exposes the 15 retained feature indices and weights. Researchers can compare these features with the family's member images and use them to guide qualitative inspection. This provides representational transparency while leaving individual feature semantics unvalidated. Human recognition of the Reddit family assignments also remains unresolved (Supplement §S17a).

Interpreting the Engagement Association

Among Reddit posts collected from new listings, greater raw signature resemblance was associated with higher engagement after accounting for measured posting time, community, whole-image similarity, and surface features. The incremental predictive gain was modest, observational, and sensitive in magnitude to how the validation problem was defined.

The decomposition indicates that much of the Reddit association reflects broad structured resemblance within the sparse representation and representation magnitude. Evidence tying the association specifically to correct documented-family identity is limited. We refer to this broader empirical pattern as canonical resemblance to documented meme forms.

Limits of the Measure

Because curated reference galleries may disproportionately contain canonical or readily recognizable examples, documented membership is a source-defined recovery criterion rather than an independent latent ground truth. The analysis also uses a single sparse representation, so the results do not establish that the same contrasts would appear with other sparse encoders or feature spaces. Whole-image nearest-neighbor retrieval of the kind evaluated by Nguyen et al. (2026) was not tested.

The held-out analysis addresses development-set overfitting but introduces a sampling difference. Robots-compliant retrieval limited held-out families to the first gallery page, making them shallower than the development families. Family-size sensitivities produce the same qualitative recovery verdict, but development and held-out evaluations also differ in their negative-sampling universe. Development status, family size, and negative composition therefore cannot be fully separated. Claims about new families therefore rest primarily on the 60 randomly sampled held-out families rather than on the stronger development-corpus comparisons. The 20-family look-alike stratum provides a useful boundary condition but is purposive and cannot be generalized to the Know Your Meme frame.

The retained feature indices and weights are inspectable, but their stable human-interpretable meanings remain unvalidated.

Cross-family calibration remains unresolved. Raw scores have a common numerical form, but equal values have not been shown to represent the same membership evidence across families. The downstream decomposition reinforces the importance of this distinction because raw resemblance and family-calibrated fit behave differently. Applications that require cross-family comparisons will need explicit calibration and independent validation.

Limits of the Reddit Application

The Reddit sample covers seven meme-focused communities during approximately one month on one platform. The observed associations may depend on that platform environment, including its users, community boundaries, moderation practices, ranking systems, and voting behavior (Jungherr et al., 2022). Generalization beyond these communities, platform, and period remains untested.

Because the development bank was purposively assembled, the Reddit maximum is conditional on this specific family bank rather than a population-wide measure of meme-family resemblance. Best-fit assignments were concentrated in a small number of families, particularly “Nobody / literally no one”, which accounted for 39.9% of posts and most observations in the highest-score decile. This concentration may partly reflect differences in raw-score scaling across families and further limits cross-family interpretation of the maximum raw score.

The engagement models are observational. Posting time and community capture some features of the posting environment, but the models do not observe all potentially relevant characteristics of posts, authors, audiences, ranking exposure, or contemporaneous events. Residual dependence among related templates can also remain under each validation design. The estimated increments therefore describe predictive association rather than a causal effect of meme resemblance on engagement.

The prospective extension does not resolve that limitation. The completed short-window waves showed no clear incremental prediction after initial engagement was included, and the Reddit access restriction described in the Results prevented both longer follow-up and a larger sample.

Finally, the raw Reddit score remains an uncalibrated resemblance magnitude, and the blinded author-coded check was largely indeterminate. Of 60 unique posts, 0 were assigned to the model-nominated family, 3 to another listed family, 6 to none of the listed families, and 51 were coded uncertain. The sole coder was one of the authors and was not a specialist in meme history or internet culture. Given the predominance of uncertain judgments, the check does not provide a determinate benchmark for the model-nominated family assignments (Supplement §S17a).

Future Validation

A priority for future work is a prospectively designed assessment of family membership by raters with specialist knowledge of meme history and internet culture. Such an evaluation could test directly whether raw signature resemblance corresponds to recognized family membership and where machine and specialist judgments diverge.

Cross-family calibration is a second priority. Raw resemblance and family-calibrated fit behave differently in the Reddit application, while equal raw scores are not yet known to carry the same meaning across families. Future work should evaluate cross-family calibration against an independent criterion.

Replication should also extend to independent family corpora, alternative sparse and dense representations, additional platforms, and a completed prospective engagement design. These comparisons would help distinguish properties of the contrastive construction from properties of the particular corpus, representation, and platform studied here.

Conclusion

The contrastive signature provides a compact representation of documented meme families in a sparse feature space. On randomly sampled held-out families, it recovers documented membership more accurately than the frequency-based recurrence baseline, same-representation sparse centroid and nearest-neighbor rules, and surface measures. It shows no clear general recovery advantage over strong whole-image dense centroids, and supervised classification recovers documented membership more accurately. Under hard decoys selected by a separately trained CLIP-family encoder, however, the contrastive signature retains a specific advantage over the CLIP centroid in the random held-out families.

The downstream application points to a different property of the representation. Raw sparse-space resemblance contributes modest information about Reddit engagement, but that association is not specific to the contrastive construction, and evidence tying it to correct documented-family identity is limited. Sparse-centroid resemblance shares the association, label-permuted signatures retain a substantial portion of it, and the residual after representation-magnitude and text-length controls is small and uncertain.

The contribution is a compact family representation whose recovery behavior, calibration limits, and downstream resemblance signal can be evaluated separately. Distinguishing these quantities is necessary for interpreting computational representations as social-science measures.

Acknowledgments

Generative AI tools, including ChatGPT (OpenAI) and Claude (Anthropic), were used to assist with coding, data analysis, preparation of figures and tables, formatting, and editorial refinement. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

Author Contributions

Author 1 (corresponding author): Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Visualization, Writing – original draft, and Writing – review and editing. Author 2: Conceptualization, Investigation (supplementary computation), and Writing – review and editing. Both authors share ownership of the core idea and approved the final manuscript.

Statements and Declarations

Ethical considerations. This study analyzed publicly available social media posts and public meme-reference materials. It involved no intervention with human participants and no identifiable private information. Usernames were removed from the analysis data, and raw user content was not redistributed beyond platform terms. A supplementary coding check was carried out by one of the authors, who labelled a sample of public Reddit images against a reference sheet. No other persons took part, and no ethics review was sought.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Declaration of conflicting interests. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding. The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability. The derived features, analysis outputs, and public post identifiers needed to reproduce the reported analyses are archived in a repository that will be made public on acceptance, subject to source and platform redistribution terms. For peer review, the materials are available through an anonymized view-only link: https://osf.io/bfwze/?view_only=4633292d4a7c4cf387ab02aa7f8e70b9. The repository DOI and public project page are withheld here for anonymized review. Raw user-generated content is not redistributed. Only identifiers with usernames removed and derived representations are shared. Data are shared under CC0 1.0.

Code availability. All analysis code is version-controlled and archived with a citable DOI and is included in the anonymized review package linked above (https://osf.io/bfwze/?view_only=4633292d4a7c4cf387ab02aa7f8e70b9). Code is released under the MIT License. The public repository will be released on acceptance.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546.
<https://doi.org/10.1017/S0003055401003100>

Author, & Author. (forthcoming). Details withheld to preserve author anonymity for double-anonymized review.

Bauckhage, C. (2011). Insights into internet memes. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), 42–49. <https://doi.org/10.1609/icwsm.v5i1.14097>

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414–427. <https://doi.org/10.1162/rest.90.3.414>

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J., & Leskovec, J. (2014). Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web*, 925–936. <https://doi.org/10.1145/2566486.2567997>

Cheng, Z., & Li, Y. (2024). Like, comment, and share on TikTok: Exploring the effect of sentiment and second-person view on user engagement with TikTok news videos. *Social Science Computer Review*, 42(1), 201–223. <https://doi.org/10.1177/08944393231178603>

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. *International Conference on Learning Representations*.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. <https://doi.org/10.1073/pnas.1720347115>

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 375–385. <https://doi.org/10.1145/3442188.3445901>

Jungherr, A., Posegga, O., & An, J. (2022). Populist supporters on Reddit: A comparison of content and behavioral patterns within publics of supporters of Donald Trump and Hillary Clinton. *Social Science Computer Review*, 40(3), 809–830. <https://doi.org/10.1177/0894439321996130>

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, 35(3), 336–356. <https://doi.org/10.1177/0894439316631043>

- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The Hateful Memes Challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33, 2611–2624.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 80, 2668–2677.
- Kirchmair, T., Koban, K., & Matthes, J. (2026). Profiling digital hate: A multidimensional measurement approach based on the Perspective API. *Social Science Computer Review*. Advance online publication. <https://doi.org/10.1177/08944393261469403>
- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5), 905–949. <https://doi.org/10.1177/0003122419877135>
- Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, 12, Article 1311. <https://doi.org/10.1038/s41467-020-19607-x>
- Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–24. <https://doi.org/10.1145/3449155>
- Nguyen, H., Köller, D. Á., Murgás, L., Pintér, J., Nagy, M., Barnes, K., & Molontay, R. (2026). Meme template identification in the wild: Comparing methods for semi-open-set recognition (Version 2) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2408.08126>
- Obradovich, N., Özak, Ö., Martín, I., Ortuño-Ortín, I., Awad, E., Cebrián, M., Cuevas, R., Desmet, K., Rahwan, I., & Cuevas, Á. (2022). Expanding the measurement of culture with a sample of two billion humans. *Journal of the Royal Society Interface*, 19(190), Article 20220085. <https://doi.org/10.1098/rsif.2022.0085>
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607–609. <https://doi.org/10.1038/381607a0>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models

from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854–856. <https://doi.org/10.1126/science.1121066>

Shifman, L. (2013). Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 18(3), 362–377. <https://doi.org/10.1111/jcc4.12013>

Shifman, L. (2014). *Memes in digital culture*. MIT Press.

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30.

Theisen, W., Brogan, J., Thomas, P. B., Moreira, D., Phoa, P., Weninger, T., & Scheirer, W. (2021). Automatic discovery of political meme genres with diverse appearances. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 714–726. <https://doi.org/10.1609/icwsm.v15i1.18097>

Theisen, W., Cedre, D. G., Carmichael, Z., Moreira, D., Weninger, T., & Scheirer, W. (2023). Motif mining: Finding and summarizing remixed image content. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1319–1328. <https://doi.org/10.1109/WACV56688.2023.00137>

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6), 1181–1195. <https://doi.org/10.1177/08944393241286471>

Zaigrajew, V., Baniecki, H., & Biecek, P. (2025). Interpreting CLIP with hierarchical sparse autoencoders. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 73918–73956.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. *Proceedings of the Internet Measurement Conference*, 188–202. <https://doi.org/10.1145/3278532.3278550>

Zauner, C. (2010). *Implementation and benchmarking of perceptual image hash functions* [Master's thesis, Upper Austria University of Applied Sciences, Hagenberg].

Supplementary Information

Supplement for *Measuring Meme-Family Structure With Contrastive Sparse Features: Canonical Resemblance and Reddit Engagement*

This supplement provides additional details on corpus construction, signature estimation, recovery evaluation in the development and held-out-family samples, sensitivity analyses, and the Reddit engagement application. Family-level and post-level results are provided in the accompanying data files where indicated. Reported statistics are linked to the corresponding analysis outputs in the reproducibility materials.

Supplementary Methods

S1. Meme-Family Corpus Construction

Families were defined by entries in Know Your Meme (KYM; knowyourmeme.com), a public reference database that documents meme lineages and their variants. Development-corpus images were retrieved from entry galleries between 12 and 26 May 2026. Family labels therefore follow the source's documented lineages rather than visual groupings constructed for this study.

Retrieval provenance. The development-corpus collection procedure requested paginated Know Your Meme gallery routes that were listed under Disallow in the site's robots.txt. We retained the already collected development corpus for provenance but did not repeat that retrieval procedure. Subsequent held-out-family collection used only routes permitted by robots.txt, which limited retrieval to the first gallery page and produced shallower candidate families. For that reason, the held-out design does not assume approximately 50 variants per family and includes explicit family-size sensitivity analyses.

After deduplication, the development corpus contained 680 variants across 31 documented families. Thirty families, comprising 678 variants, contained at least five members and entered the recovery analysis. The remaining family contained two variants and was excluded by the size threshold. Among the 30 analyzed families, size ranged from 15 to 26 variants, with a median of 24.5.

The 31-family development corpus came from a 41-family candidate pool assembled before collection to include both image-template families and families whose recurring form extended across changes in image, text, layout, or relation. Gallery variants were retrieved for 32 pool families, two of which were subsequently merged; no gallery variants were retrieved for the remaining nine, which were not included. The resulting corpus is purposive and should not be interpreted as a probability sample of meme families or as a natural family-size distribution. The narrow size range also reflects the bounded retrieval procedure and subsequent deduplication. Ambiguous cases followed the family label assigned by the reference source. Images appearing under more than one family label were handled as described in §S2.

Table S1. Meme-family corpus composition. Final deduplicated sample size and recovery-analysis inclusion for each documented family. Per-family retrieval and exact-, near-, and cross-family removal counts are provided in TableS1_corpus_composition.csv and the reproducibility package.

Family	Final n	In recovery
american chopper argument	18	yes
coffin dance	26	yes
disaster girl	21	yes
distracted boyfriend	20	yes
doge such x much y very z	25	yes
drake hotline bling drakeposting	19	yes
expanding brain galaxy brain	21	yes
hide the pain harold	23	yes
is this a pigeon	25	yes
lolcats i can has cheezburger lolspeak	25	yes
loss is this loss	25	yes
mocking spongebob	25	yes
most interesting man (i don't always x)	2	no
nobody literally no one	26	yes
npc i support the current thing	15	yes
okay boomer	24	yes
pepe the frog	25	yes
pov you are x	25	yes
roll safe	26	yes
skill issue git gud cope seethe mald	25	yes
stonks meme man series	22	yes
surprised pikachu	18	yes
they did surgery on a grape	25	yes
this is fine	26	yes
trade offer	19	yes
two buttons daily struggle	17	yes
wojak soyjak	25	yes
woman yelling at a cat	22	yes

Family	Final n	In recovery
x all the y	23	yes
y all got anymore of	17	yes
yo dawg i heard you like	25	yes

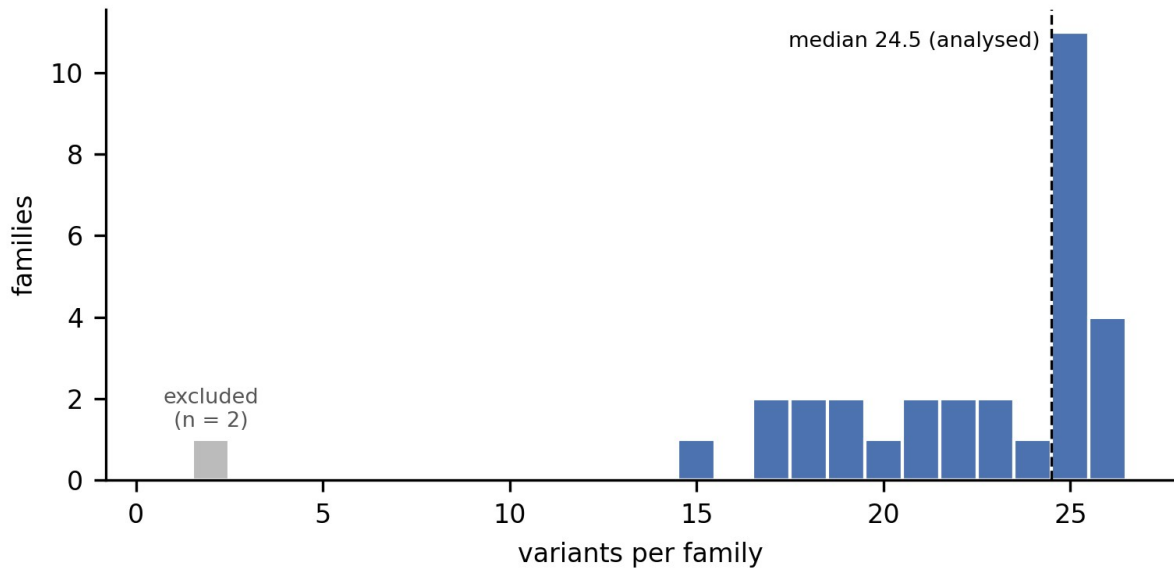


Figure S1. Family-size distribution. Distribution of final family sizes among the 30 analyzed families (15–26 variants; median 24.5), with the excluded two-variant family shown separately. The distribution reflects the bounded retrieval procedure and does not estimate naturally occurring family sizes.

Note. Table S1 and TableS1_corpus_composition.csv report final per-family sample sizes and recovery inclusion for all 31 documented families. Cross-family deduplication reduced the 707-variant within-family-deduplicated corpus to 680 variants (§S2).

S1a. Held-Out-Family Validation Protocol and Dispositions

The held-out-family protocol fixed the sampling frame, eligibility criteria, retrieval and deduplication rules, stopping rule, scoring pipeline, comparison set, and interpretive rules before any held-out family was scored. The protocol was timestamped before scoring. One technical amendment, made before screening resumed, changed only the filename assigned to retrieved images after a file address without an extension caused a processing failure. It did not alter eligibility, screening, image content, deduplication, scoring, or any analysis variable.

- **Frame:** Know Your Meme confirmed-entry sitemap, snapshot 29 September 2026, screened in seeded random order.

- **Strata:** random stratum H, filled from the seeded order, and look-alike challenge stratum C, filled using a fixed token rule from entries related to development families, subject to a per-anchor cap.
- **Eligibility:** Know Your Meme confirmed memes, excluding KYM-gated sensitive content. Gated gallery items were excluded and not decoded.
- **Retrieval:** first gallery page only, using robots-compliant routes.
- **Deduplication:** development-corpus rules applied within families, followed by duplicate resolution across held-out families and against the development corpus.
- **Analyzability:** at least 10 distinct variants after deduplication.
- **Scoring:** the fixed development pipeline with $K = 15$, the same comparison methods, and a fused SigLIP2 + DINOv2 whole-image centroid. For the fused baseline, each embedding was L2-normalized and the two representations were concatenated with equal weight. Checkpoints and fusion weights were fixed without tuning.
- **Evaluation negatives:** for each target family, 200 negatives sampled uniformly from variants of all other families in the 110-family evaluation universe, with adversarial decoys selected from the same universe. The training background comprised the remaining nonmembers, excluding the evaluation negatives and decoys. Because this universe differs from the development evaluation, development and held-out AUROCs are not directly interchangeable (§S7a).

Screening ended under the pre-specified stopping rule after 158 unique entries had been screened, yielding 60 analyzable random-stratum families and 20 analyzable challenge-stratum families.

Table S1a. Screening dispositions.

Disposition	Entries
Analyzable	80
Too small after deduplication	29
Challenge-stratum anchor cap reached	31
Nested development family	8
No gallery	7
Development-corpus entry	3
Screened	158

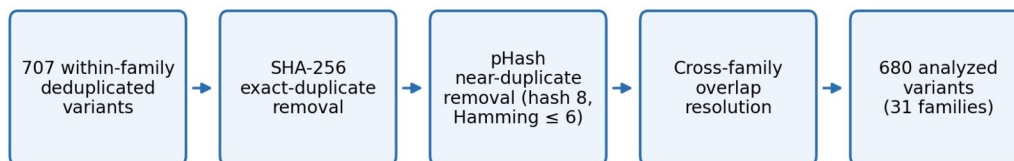
S2. Deduplication and Overlap Handling

Exact duplicates were identified by SHA-256 hashes. Near-duplicates were identified with the discrete-cosine-transform perceptual hash implemented in the imagehash library (Zauner, 2010), using hash size 8 and a Hamming-distance threshold of 6 or less. Near-duplicate clustering was transitive at this threshold and was applied first within families and then across families.

This procedure identified visually equivalent files that differed at the byte level and therefore were not detected by cryptographic hashing. Cross-family deduplication reduced the corpus from 707 to 680 variants and from 32 to 31 family labels. Wojak and Soyjak, which showed near-total cross-family overlap, were merged into one documented family for analysis. One comic-book image appearing under four otherwise unrelated family labels was excluded.

Table S2. Deduplication and overlap handling. Methods, thresholds, and outcomes for exact- and near-duplicate removal within and across families.

Stage	Method	Threshold	Outcome
Within-family exact	SHA-256 hash	identical bytes	incorporated into the 707-variant within-family-deduplicated baseline
Within-family near-duplicate	DCT perceptual hash (size 8)	Hamming ≤ 6	incorporated into the 707-variant within-family-deduplicated baseline
Cross-family exact	SHA-256 hash	identical bytes	23 Wojak/Soyjak duplicates; labels merged into one family
Cross-family contamination	manual review	image assigned to four unrelated labels	one comic image excluded
Net effect			707 $\rightarrow$ 680 variants; 32 $\rightarrow$ 31 family labels



Cross-family step: 23 exact Wojak/Soyjak duplicates merged into one family; 1 image filed under four families excluded as contamination.

Figure S2. Deduplication procedure. Schematic of exact-duplicate, near-duplicate, and cross-family overlap decisions.

S3. Representation and Sparse-Autoencoder Details

Images were encoded with OpenAI CLIP ViT-L/14, producing 768-dimensional whole-image embeddings after standard CLIP preprocessing at 224-pixel input resolution. Embeddings were retained without unit normalization when passed to the sparse autoencoder and unit-normalized for centroid-based whole-image similarity.

Sparse representations were generated with the hierarchical Matryoshka multimodal sparse autoencoder described by Zaigrajew et al. (2025), checkpoint WolodjaZ/MSAE. The representation contains 6,144 latent dimensions with top- k activation at $k = 64$.

As a reproducibility check for the Reddit application, re-encoding all 4,263 Reddit images reproduced the stored raw signature scores to the reported precision: correlation 1.000000 and 100% agreement in the best-fit family.

Analyses used Python 3.9 with NumPy 2.0.2, pandas 2.3.3, scikit-learn 1.6.1, PyTorch with open_clip, imagehash 4.3.2, and ocrmac. Recovery negative sampling used seed 20260512. Engagement cross-validation and bootstrap point estimates used seed 0, with seed sensitivity reported in Table S5a.

Table S3. Computational environment. Models, checkpoints, preprocessing settings, software packages, versions, and random seeds.

Component	Value
Image encoder	OpenAI CLIP ViT-L/14 (768-dim; 224-pixel preprocessing)
Sparse autoencoder	Matryoshka multimodal SAE (Zaigrajew et al., 2025), checkpoint WolodjaZ/MSAE
Latent dimensions	6,144, top-k activation at $k = 64$
Held-out dense baseline	SigLIP2 and DINOv2 image encoders (fixed checkpoints), equal-weight fused centroid
Language	Python 3.9 (macOS / Darwin)
Core libraries	NumPy 2.0.2; pandas 2.3.3; scikit-learn 1.6.1; PyTorch with open_clip; imagehash 4.3.2; ocrmac
Random seeds	recovery negative sampling = 20260512; cross-validation and bootstrap = 0 for reported point estimates (seed sweeps in Table S5a)

S4. Formal Signature Definition

For family f , let M_f denote the member set and N_f the nonmember training background used for estimation. The training background is disjoint from evaluation negatives and adversarial decoys. Let z_i denote the sparse activation vector for image i .

The feature-wise member-minus-background contrast is

$$d_f = \frac{1}{|M_f|} \sum_{i \in M_f} z_i - \frac{1}{|N_f|} \sum_{j \in N_f} z_j.$$

Let T_f denote the $K = 15$ features with the largest positive entries in d_f . Their signature weights are

$$w_f = \max(d_f[T_f], 0), \quad w_f \in \mathbb{R}_+^{15}.$$

For candidate c , let $z_c[T_f]$ be its activation vector on the 15 retained features. Raw resemblance to family f is

$$r(c, f) = \cos(z_c[T_f], w_f).$$

The **raw signature score** is the largest resemblance over the family bank,

$$\max_f r(c, f),$$

and the **specificity margin** is the difference between the largest and second-largest family resemblance scores.

The **family-calibrated fit** standardizes raw resemblance relative to the corresponding family's nonmember distribution:

$$\text{CalibratedFit}(c, f) = \frac{r(c, f) - \mu_{N_f}}{\sigma_{N_f}},$$

where μ_{N_f} and σ_{N_f} are the mean and standard deviation of nonmember resemblance scores for family f .

The recurrence baseline instead retains sparse features that are active among a high proportion of family members and exceed the 95th percentile of the training-background activation distribution.

Three downstream quantities are analyzed separately:

1. **Raw signature resemblance:** uncalibrated cosine resemblance to a family signature.
2. **Family-calibrated fit:** resemblance standardized relative to a family's nonmember score distribution.
3. **Classifier score:** the maximum one-versus-rest predicted probability. Using the maximum decision score produces the same downstream result.

These quantities are evaluated separately because they encode different information and are not interchangeable measures of family membership.

S5. Signature Sensitivity and Stability

Number of Retained Features

Mean recovery AUROC increased gradually over the evaluated values of K and changed little between $K = 25$ and $K = 50$ (Table S4). $K = 15$ was selected during development as a compact

specification that retained most of the recovery observed at the larger settings. It was then held fixed in the held-out-family and downstream analyses.

Negative-Pool Sensitivity

Because signature weights are estimated from member-minus-background contrasts, recovery can depend on the sampled nonmember pool. We therefore repeated the evaluation-negative draw with ten seeds. In the development corpus, mean contrastive recovery ranged from 0.895 to 0.898. The mean contrastive-minus-sparse-centroid difference ranged from +0.073 to +0.077, while the contrastive-minus-CLIP-centroid difference ranged from +0.049 to +0.056. This development-corpus advantage did not replicate in the random held-out families: across the same ten redraws in H, the contrastive-minus-CLIP-centroid difference ranged from −0.0021 to −0.0001.

Table S4. Sensitivity to the number of retained features. Development-corpus mean recovery AUROC under the leave-one-variant-out protocol as a function of K .

K	Mean recovery AUROC
5	0.878
10	0.893
15 (used)	0.899
25	0.906
50	0.907

Table S5. Sensitivity to evaluation-negative sampling. Range across seeds 1–10 of the mean paired method difference.

Comparison	Development corpus	Random held-out stratum H
Contrastive – recurrence	+0.099 to +0.111	+0.044 to +0.049
Contrastive – sparse centroid	+0.073 to +0.077	+0.026 to +0.028
Contrastive – sparse nearest neighbor	+0.083 to +0.087	+0.029 to +0.032
Contrastive – CLIP centroid	+0.049 to +0.056	−0.0021 to −0.0001
Contrastive – fused SigLIP2 + DINOv2 centroid	not part of the original development analysis (§S7a)	−0.0071 to −0.0040

Table S5a. Seed sensitivity across analyses. Main-text point estimates use seed 0 for the engagement analyses and seed 20260512 for recovery negative sampling. Community-grouped folds are deterministic.

Analysis	Seeds	Result
Label-permutation control, engagement decomposition	0–3	Monte Carlo $p = 0.030, 0.035, 0.030, 0.045$
Magnitude-adjusted increment, lower 95% interval bound	0–3	$-0.0010, 0.0002, -0.0014, -0.0017$
Repeated cross-validation, median continuous increment	1–20	random $+0.046$; perceptual-hash cluster $+0.046$; best-fit family $+0.072$
Recovery evaluation-negative redrows	1–10	Table S5

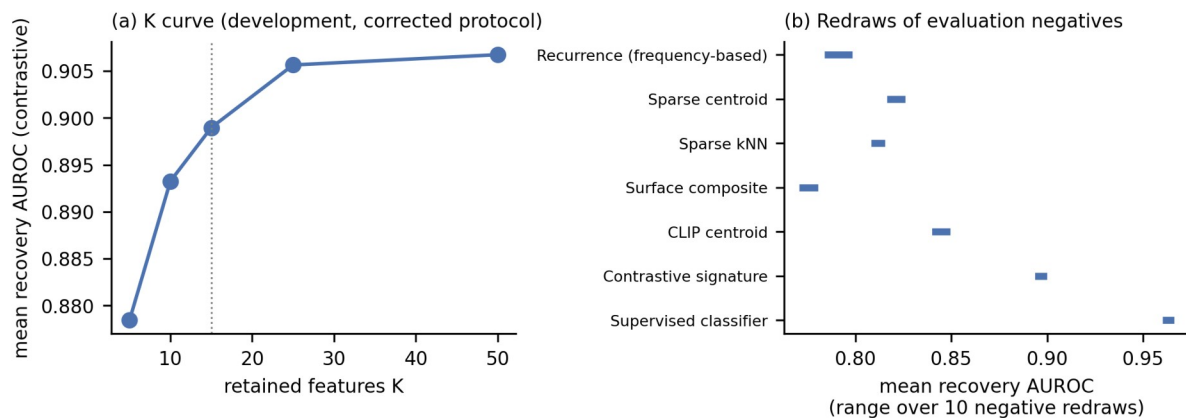


Figure S3. Signature sensitivity. (a) Mean recovery AUROC as a function of the retained-feature count K . (b) Range of each method's mean recovery AUROC across ten evaluation-negative redrows.

Supplementary Recovery Results

S6. Full Method Specifications

All recovery methods were evaluated within a common leave-one-variant-out procedure. For each target family, 200 evaluation negatives were sampled from variants belonging to other families, including the two-variant family excluded from the recovery set. These evaluation negatives were never used to estimate a comparison method. Each method instead used a disjoint training background drawn from the remaining nonmembers, with a mean of 435 and a minimum of 428 observations per target family. Adversarial decoys were likewise excluded from all training backgrounds. Each method, including the supervised classifier, was fitted using the same

$n-1$ target-family members within the same leave-one-variant-out loop. AUROC was computed within each fold before averaging.

The comparison set comprised:

- frequency-based recurrence baseline
- whole-image CLIP centroid
- sparse-space centroid
- sparse-space nearest-neighbor rule
- contrastive signature with $K = 15$
- regularized one-versus-rest logistic classifier using the sparse representation
- character n -gram TF-IDF similarity based on macOS Vision OCR text
- DCT perceptual-hash similarity
- template similarity
- fixed composite surface measure
- per-family oracle surface comparator
- fused SigLIP2 + DINOv2 whole-image centroid in the expanded held-out-family evaluation only, which also re-scored the development families (§S7a)

Hyperparameters. No hyperparameter search was performed for these comparisons. The recovery logistic classifier used L2 regularization with $C = 1.0$ and balanced class weights. Engagement ridge models used $\alpha = 1.0$, and engagement logistic models used $C = 1.0$. Predictors were standardized within each training fold. $K = 15$ was selected on the development corpus and then fixed for all held-out and downstream analyses; the K-curve is reported as a sensitivity (Table S4). The 0.02 descriptive head-to-head margin and the recurrence baseline's 95th-percentile background threshold were also fixed before the held-out analysis.

Recovery protocol. Every method is fitted within one leave-one-variant-out loop. Training backgrounds are disjoint from evaluation negatives and adversarial decoys. The supervised classifier uses the same folds as the other methods, all methods are evaluated against the same negative samples, and AUROC is computed within folds before averaging.

Recurrence thresholds are estimated only from the training background, and surface measures are fitted within each training fold. The fixed composite surface measure has mean recovery 0.779, below the oracle surface comparator's 0.812.

Text similarity uses cosine similarity between character n -gram TF-IDF representations of macOS Vision OCR text (Pedregosa et al., 2011). Perceptual-hash similarity uses the DCT-based method described by Zauner (2010). The oracle surface comparator selects the best-performing surface measure separately for each family and is reported only as a deliberately favorable descriptive benchmark.

The logistic classifier operates on the full 6,144-dimensional sparse representation and is not constrained to a 15-feature prototype. The contrastive signature retains exactly 15 positive

features per family. The difference in representational form should therefore be considered separately from their difference in recovery accuracy.

Table S6. Full method definitions and hyperparameters.

Method	Definition
Frequency-based recurrence baseline	features active in a high proportion of members and above a 95th-percentile training-background threshold
Whole-image CLIP centroid	cosine similarity to the unit-normalized mean CLIP embedding
Sparse-space centroid	cosine similarity to the mean sparse activation vector
Sparse-space nearest neighbor	maximum sparse-space cosine similarity to a family member
Contrastive signature	15 largest positive member-minus-background contrast features
Logistic classifier	one-versus-rest L2 logistic regression on the 6,144 sparse features $C = 1.0$, balanced class weights; maximum predicted probability used in the reported analysis
Text similarity	character n-gram TF-IDF cosine similarity on macOS Vision OCR text (Pedregosa et al., 2011), fitted within each fold
Perceptual-hash similarity	DCT perceptual-hash similarity (Zauner, 2010)
Template similarity	template-match similarity, fitted within each fold
Fixed surface composite	fixed combination of surface measures defined without selection on evaluation performance
Oracle surface comparator	best-performing surface measure for each family; descriptive only
Fused SigLIP2 + DINOv2 centroid (expanded held-out evaluation)	each embedding L2-normalized and concatenated with equal weight; cosine similarity to the normalized mean of the $n-1$ training members

S7. Complete Family-Recovery Results: Development Corpus

Table S7 reports family-level recovery AUROC for all 30 analyzed development families under the leave-one-variant-out protocol. Main-text Table 2, Panel 2A reports the method means, and Table S8 reports paired family-level comparisons.

Across leave-one-family-out checks, mean contrastive recovery ranged from 0.896 to 0.905. Family size was essentially unrelated to the contrastive signature's advantage over the sparse centroid in the development corpus ($r = -0.05$).

Table S7. Family-level recovery matrix, development corpus. Within-family recovery AUROC under the leave-one-variant-out protocol, sorted by contrastive recovery. The same values are provided in TableS7_family_recovery_matrix.csv.

Family	n	Recur	Whole-img	Surface comp.	Sp.cent	Sp.kNN	Contrast	Classif
hide the pain harold	23	0.90	0.81	0.70	0.74	0.78	0.99	0.99
two buttons daily struggle	17	0.98	0.94	0.86	0.93	0.92	0.99	1.00
expanding brain galaxy brain	21	0.92	0.98	0.89	0.95	0.91	0.99	1.00
skill issue git gud cope seethe mald	25	0.86	0.90	0.91	0.88	0.86	0.99	0.99
pepe the frog	25	0.99	1.00	0.54	1.00	0.99	0.98	1.00
lolcats i can has cheezburger lolspeak	25	0.98	0.92	0.53	0.94	0.81	0.98	0.99
x all the y	23	0.95	0.88	0.86	0.89	0.79	0.97	0.99
doge such x much y very z	25	0.97	0.89	0.58	0.79	0.75	0.97	0.99
stonks meme man series	22	0.80	0.92	0.98	0.93	0.91	0.96	0.98
they did surgery on a grape	25	0.77	0.88	0.80	0.85	0.86	0.95	0.99
disaster girl	21	0.68	0.77	0.80	0.83	0.73	0.95	0.97
american chopper argument	18	0.89	0.94	0.89	0.94	0.91	0.94	1.00
coffin dance	26	0.55	0.89	0.45	0.89	0.86	0.93	0.97
y all got anymore of	17	0.85	0.96	0.98	0.89	0.88	0.93	0.97

Family	n	Recur	Whole-img	Surface comp.	Sp.cent	Sp.kNN	Contrast	Classif
okay boomer	24	0.72	0.70	0.90	0.76	0.77	0.92	0.94
trade offer	19	0.79	0.88	1.00	0.85	0.85	0.92	1.00
this is fine	26	0.70	0.83	0.87	0.77	0.72	0.91	0.94
roll safe	26	0.83	0.88	0.86	0.83	0.90	0.89	0.96
woman yelling at a cat	22	0.68	0.65	0.71	0.70	0.69	0.88	0.94
surprised pikachu	18	0.76	0.86	0.74	0.80	0.80	0.87	0.95
drake hotline bling drakeposting	19	0.71	0.72	0.81	0.75	0.76	0.86	0.93
loss is this loss	25	0.73	0.77	0.65	0.75	0.74	0.85	0.95
pov you are x	25	0.79	0.64	0.92	0.81	0.79	0.84	0.90
npc i support the current thing	15	0.65	0.80	0.86	0.70	0.79	0.83	0.92
is this a pigeon	25	0.78	0.85	0.95	0.81	0.74	0.83	0.93
nobody literally no one	26	0.77	0.83	0.93	0.82	0.77	0.80	0.92
wojak soyjak	25	0.81	0.89	0.47	0.80	0.81	0.80	0.98
yo dawg i heard you like	25	0.50	0.64	0.73	0.60	0.68	0.79	0.94
distracted boyfriend	20	0.71	0.83	0.46	0.68	0.76	0.74	0.95
mocking spongebob	25	0.71	0.82	0.71	0.81	0.85	0.73	0.94

Table S8. Paired family-level method comparisons, development corpus. Mean paired family-level differences with 95% family-bootstrap intervals. The final column reports the number of families for which the contrastive-minus-comparator difference exceeds 0.02, except for the classifier row, where the direction is reversed.

Comparison	Mean difference	95% interval	Families > 0.02
Contrastive – recurrence	+0.109	[0.076, 0.146]	23/30
Contrastive – sparse centroid	+0.075	[0.051, 0.102]	24/30
Contrastive – nearest neighbor	+0.086	[0.058, 0.114]	25/30
Contrastive – CLIP centroid	+0.056	[0.025, 0.088]	19/30
Contrastive – surface composite	+0.120	[0.061, 0.182]	23/30
Contrastive – oracle surface	+0.087	[0.037, 0.138]	21/30
Contrastive – classifier	−0.064	[−0.085, −0.045]	classifier higher by > 0.02 in 21/30

Note. The development-corpus contrastive advantage over the CLIP centroid did not replicate in the random held-out families.

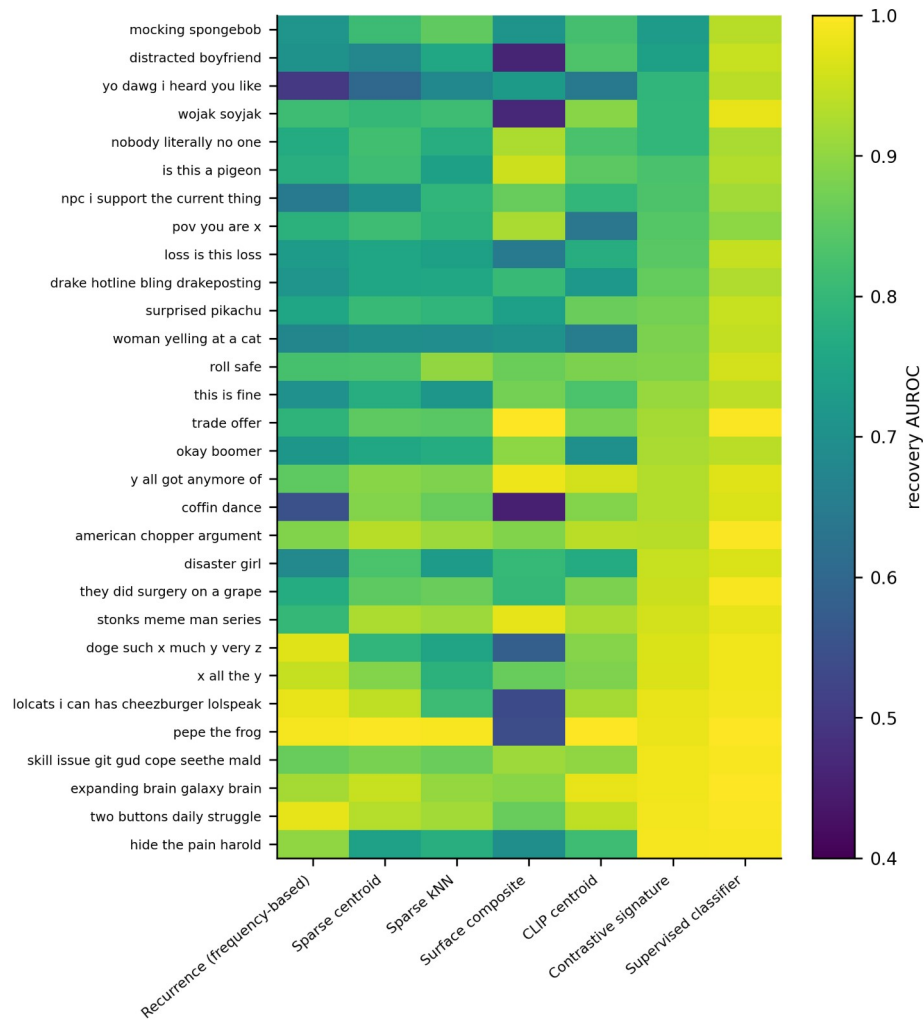


Figure S4. Family-level recovery across methods. Recovery AUROC for all 30 analyzed development families and all comparison methods under the leave-one-variant-out protocol, except the per-family oracle surface comparator (Table 2, Panel 2A).

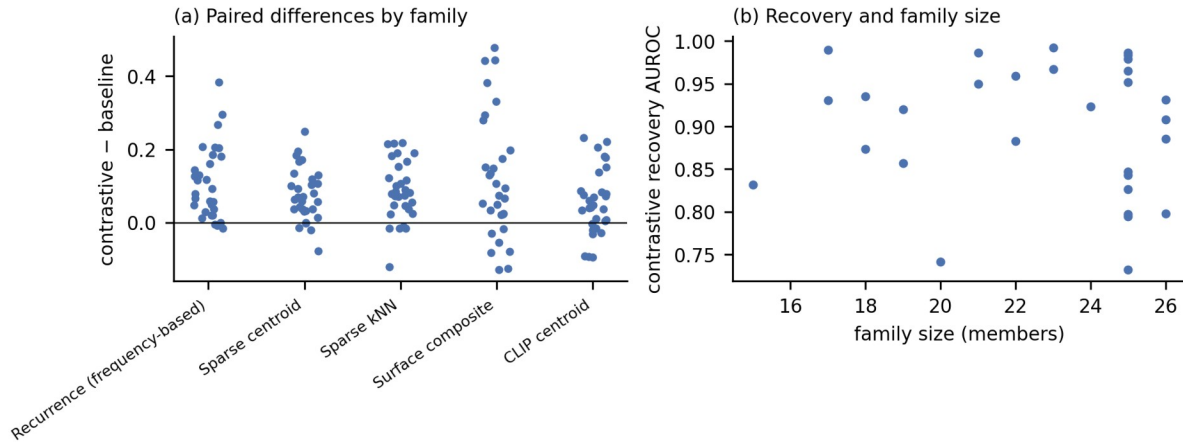


Figure S5. Paired recovery differences and family size. (a) Contrastive-minus-comparator differences by family. (b) Recovery AUROC in relation to family size.

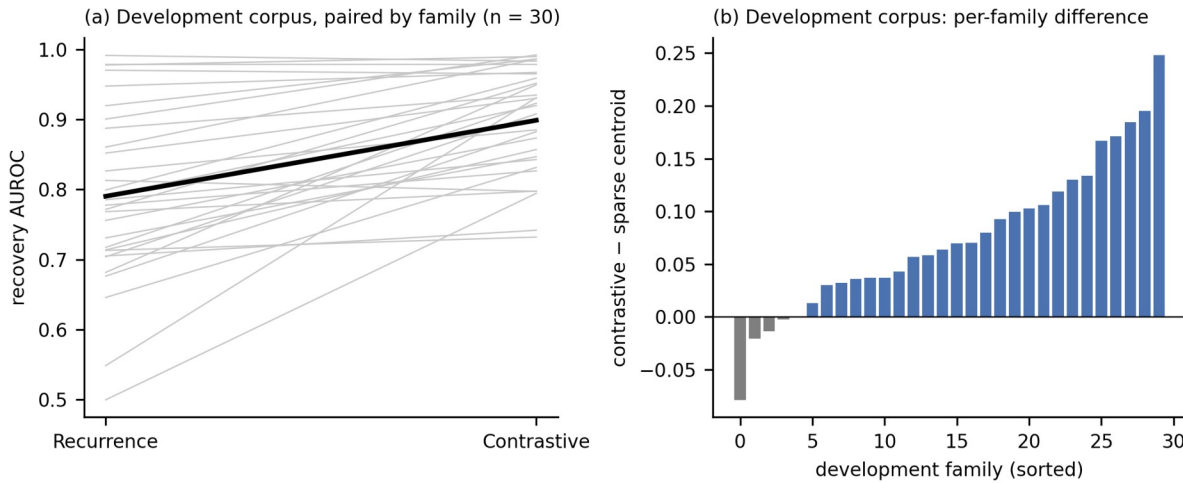


Figure S13. Development-corpus recovery diagnostics. (a) Recurrence and contrastive recovery AUROC paired within each of the 30 development families; the heavy line connects the method means. (b) Sorted family-level contrastive-minus-sparse-centroid differences. These panels describe the development corpus only; main Figure 3a reports the held-out-family comparisons.

S7a. Held-Out-Family Results

Main-text Table 2, Panel 2B reports the principal held-out-family comparisons. This section reports pre-specified sensitivity analyses and additional descriptive comparisons.

Family-size sensitivity. Capping every family, development families included, at 10 or 15 members did not change the qualitative recovery result in random stratum H. The contrastive-minus-CLIP-centroid difference was +0.001 [−0.012, 0.013] at a cap of 10 and −0.002 [−0.013, 0.008] at a cap of 15.

Table S7a. Size-capped sensitivity in random held-out stratum H. Paired contrastive-minus-method recovery AUROC difference with 95% family-bootstrap interval.

Method	Cap 10	Cap 15
Frequency-based recurrence baseline	+0.057 [0.038, 0.077]	+0.051 [0.034, 0.070]
Sparse centroid	+0.030 [0.015, 0.044]	+0.027 [0.016, 0.039]
Sparse nearest neighbor	+0.034 [0.016, 0.054]	+0.033 [0.019, 0.047]
Surface composite	+0.177 [0.113, 0.243]	+0.172 [0.111, 0.234]
Oracle surface	+0.126 [0.077, 0.174]	+0.123 [0.078, 0.170]
Whole-image CLIP centroid	+0.001 [−0.012, 0.013]	−0.002 [−0.013, 0.008]
Fused SigLIP2 + DINOv2 centroid	−0.003 [−0.023, 0.018]	−0.005 [−0.020, 0.010]

Development-to-held-out difference. When development and held-out families were re-evaluated within the same 110-family comparison universe, the development families showed a mean contrastive-minus-CLIP-centroid difference of +0.037 [0.009, 0.066], compared with −0.002 in random held-out stratum H. The +0.037 estimate differs from the +0.056 reported in main-text Table 2, Panel 2A because the evaluation universe differs; the two estimates therefore do not come from otherwise identical runs. In a regression of the family-level contrastive-minus-CLIP difference on stratum and log family size, the held-out-minus-development coefficient was −0.038. Because no held-out family had 20 or more variants, this adjustment cannot separate family size from development status. In the same expanded evaluation, the development families showed no clear contrastive advantage over the fused SigLIP2 + DINOv2 centroid (−0.008 [−0.033, 0.015]); the fused centroid was not part of the original development analysis.

Table S7b. Paired differences by family-size band across all strata. Descriptive only.

Size band	Families	Contrastive – CLIP centroid	Contrastive – fused centroid	Contrastive – sparse centroid
10–14	38	−0.003	−0.008	+0.023
15–19	49	−0.005	−0.019	+0.032
≥ 20	23	+0.042	−0.003	+0.067

Fold-local decoy selection. Selecting LAION-CLIP decoys separately within each fold produced a contrastive-minus-CLIP-centroid difference of +0.072 [0.040, 0.107] in random held-out stratum H, compared with +0.073 [0.041, 0.108] when the decoys were selected against the full member set.

Exploratory fused-centroid comparison. After the held-out-family results had been examined, we additionally compared the contrastive signature with the fused SigLIP2 + DINOv2 centroid

under encoder-selected decoys. The paired contrastive-minus-fused difference was +0.074 [0.034, 0.115] in H, with a fold-local estimate of +0.071 [0.030, 0.113]. In challenge stratum C, the corresponding difference was -0.045 [-0.115, 0.025]. Because this comparison was added after inspection of the held-out results, it is treated as exploratory rather than as part of the pre-specified held-out claim. **Pooled 110-family result (descriptive only).** Pooling all 110 families (the 30 development families, the random held-out families in H, and the purposively selected challenge families in C) gives a contrastive-minus-fused-centroid difference of -0.0121 [-0.0243, -0.0002]. This pooled quantity combines a development sample, a random held-out stratum, and a purposive challenge stratum. It is therefore reported only as a descriptive summary and is not an overall validation estimate. Random stratum H remains the relevant generalization sample.

S8. Adversarial-Decoy Analyses

For each family, adversarial decoys were the 12 nearest nonmembers under two separate selection criteria. True family members were excluded from the candidate pools, and cross-family near-duplicates were removed before selection. Selected decoys were then scored with the primary representation. The main specification selected decoys against the full member set. Selecting them separately within each leave-one-out fold is reported as a sensitivity analysis (below and §S7a).

The encoder-selected condition used a separate CLIP model trained on LAION data to identify visually similar nonmembers. In the development corpus, mean discrimination AUROC was 0.836 for the contrastive signature and 0.654 for the CLIP centroid, a paired difference of +0.181 [0.128, 0.237]. Fold-local decoy selection produced +0.181 [0.127, 0.237]. For comparison, CLIP-centroid discrimination fell to 0.446 when decoys were selected using the primary encoder itself. That condition has direct dependence between the criterion used to choose difficult cases and the representation subsequently scored and is included only as a dependent reference.

The second condition used perceptual hashing as a non-CLIP selector. Mean discrimination AUROC was 0.889 for the contrastive signature and 0.854 for the CLIP centroid. The main-specification contrastive-minus-CLIP interval included zero, +0.036 [-0.001, 0.074], so we do not infer a clear contrastive advantage under this decoy-selection rule. The fold-local interval, [0.0015, 0.076], is reported as a sensitivity and is not used to reverse that interpretation.

Table S9. Adversarial discrimination by method and decoy-selection criterion, development corpus. Mean discrimination AUROC using the 12 nearest nonmembers per family under each selector.

Decoy selector	Contrastive signature	CLIP centroid	Sparse centroid	Sparse nearest neighbor	Fixed surface composite	Recurrence baseline	Logistic classifier
Encoder-selected (separately trained LAION CLIP)	0.836	0.654	0.745	0.676	0.833	0.727	0.904
Perceptual-hash-selected (non-CLIP)	0.889	0.854	0.812	0.802	0.676	0.776	0.956
Primary-encoder-selected (dependent reference)	0.830	0.446	0.719	0.481	0.813	0.691	0.909

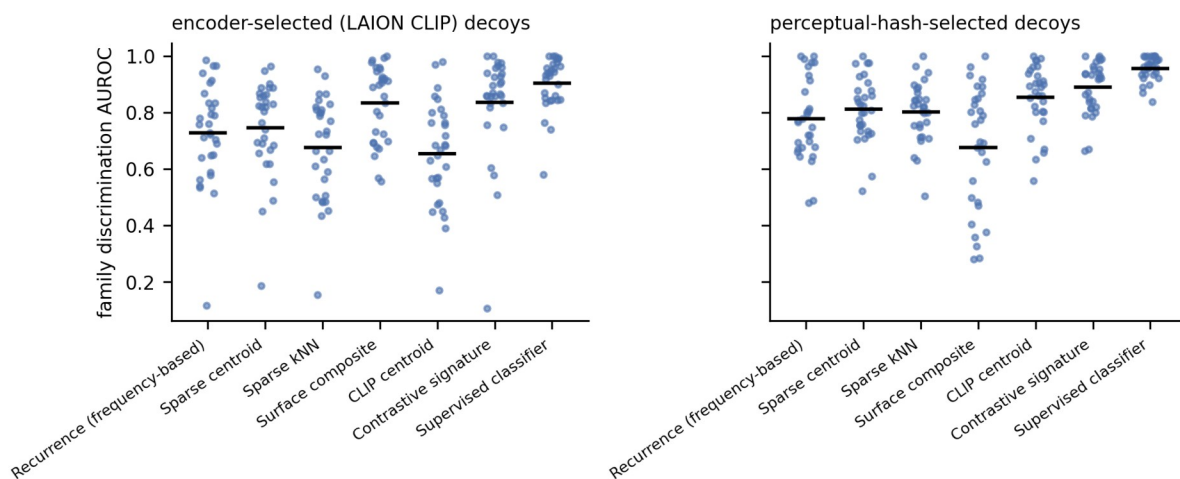


Figure S6. Family-level adversarial discrimination. Per-family discrimination AUROC by method under encoder-selected and perceptual-hash-selected decoys in the development corpus.

S9. Signature Inspection and Illustrative Cases

The family signatures can be inspected through their retained feature indices and weights and by identifying the member and nonmember images that most strongly activate those features. Main-text Figure 2b uses trade offer as the worked example because it was the development family with the median development-corpus contrastive recovery AUROC (family AUROC 0.920; median 0.922). The family was selected by this rule before its signature was inspected.

Additional illustrative cases were selected by rules specified before the cases were viewed and then applied to the development-corpus recovery results: (i) the family with the largest contrastive-minus-CLIP-centroid discrimination difference under encoder-selected decoys, pov you are x (+0.480); (ii) the family with the smallest such difference, wojak soyjak (−0.063); (iii) the family with the largest classifier-minus-contrastive recovery difference, distracted boyfriend (+0.208); and (iv) for each, the single top-ranked encoder-selected decoy. Figure S7 instead displays the highest-scoring nonmember under the family's contrastive signature, drawn from the stored signature export, rather than the encoder-selected decoy named in (iv). All illustrated families come from the development corpus; no held-out family is shown.

These materials are inspection aids rather than semantic validation. The sparse features have not been independently established as human-interpretable concepts, so a retained feature index should not be read as a validated semantic label.

For each example, the accompanying data provide:

- the 15 retained feature indices and weights
- identifiers for the highest-activating member exemplars
- identifiers for the highest-activating nonmember exemplars
- development-family recovery performance
- corresponding classifier performance

An exploratory typology classifies development families as structural, character/style, lexical, or template-based. These assignments were not established through blinded human annotation and are used only for descriptive exploration.

Table S10. Exploratory family-type coding. Descriptive typology only. Per-family assignments are provided in the reproducibility package.

Type	Basis of the family
Structural	recurring spatial or compositional structure
Character/style	recurring character or distinctive visual style
Lexical	recurring text or phrasal template
Template-based	fixed image-macro template

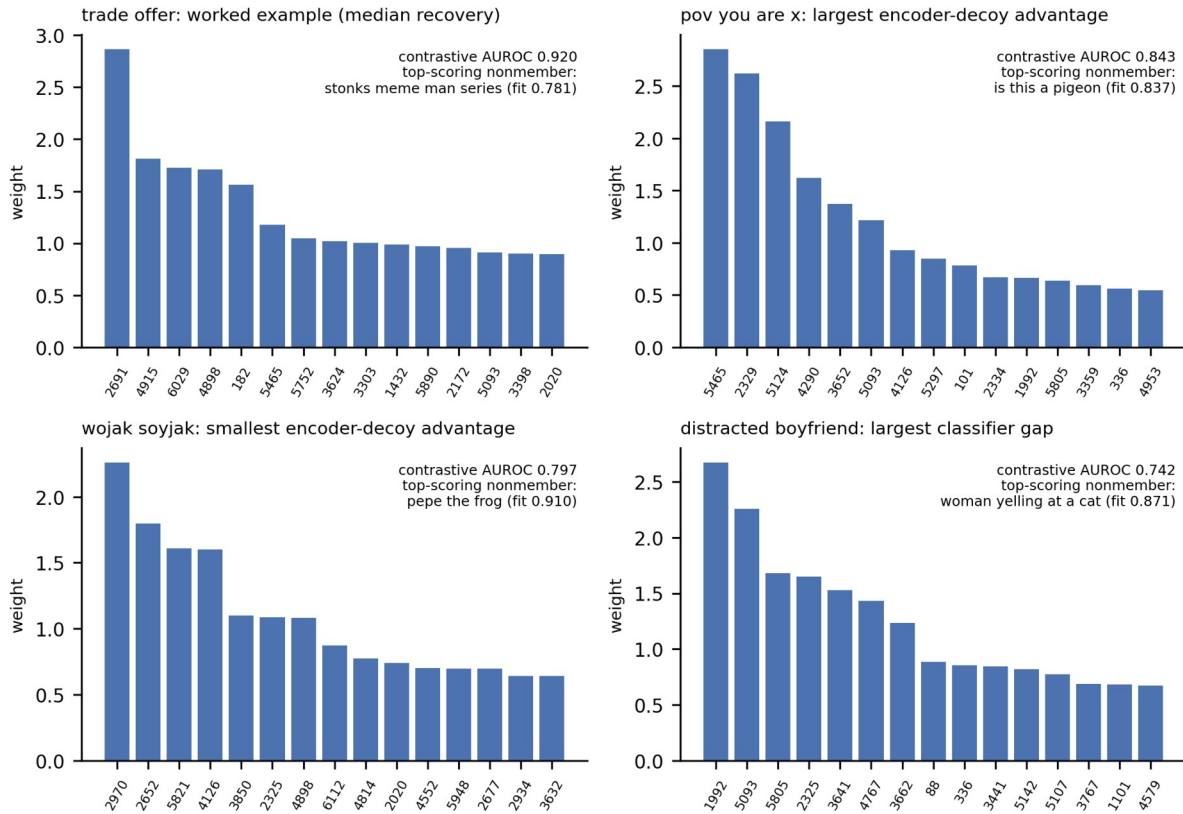


Figure S7. Family-signature inspection examples. The rule-selected worked example and additional cases are shown through the 15 retained contrastive feature weights, the family's contrastive recovery AUROC, and the source family and signature fit of the highest-scoring nonmember under that family's signature. Meme images are not reproduced. Retained features are not assigned semantic labels. Feature indices and identifiers for the highest-activating member and nonmember exemplars are provided in `interpretability_signatures.csv`.

Reddit Supplement

S10. Collection and Sample Flow

Public Reddit image posts were collected from the new listings of seven communities: `r/dankmemes`, `r/memes`, `r/meme`, `r/whenthe`, `r/historymemes`, `r/shitposting`, and `r/me_irl`. Top and hot listings were queried separately to construct the selected-listing contrast. The rising listing was queried after the new listing but contributed no additional posts once previously observed post identifiers were removed.

Collection occurred from April 26 to May 27, 2026. Images were deduplicated by perceptual hash before analysis. The denominator sample contained 4,263 posts: 749 from `r/dankmemes`, 726 from `r/memes`, 711 from `r/meme`, 585 from `r/whenthe`, 573 from `r/historymemes`, 562 from `r/shitposting`, and 357 from `r/me_irl`. The analyzed selected-listing contrast contained 348 posts. Four additional posts retrieved from hot listings were included in neither analytic sample. All

4,263 denominator-sample images were available locally and are included in the full-sample analyses (§S16).

Time-invariant content features were computed once for each post.

Table S11. Sample flow by community. Denominator-sample counts after perceptual-hash deduplication.

Community	Denominator posts
r/dankmemes	749
r/memes	726
r/meme	711
r/whenthe	585
r/historymemes	573
r/shitposting	562
r/me_irl	357
Total	4,263

Note. The analyzed selected-listing contrast contained 348 posts. Four additional hot-listing posts were included in neither analytic sample.

S11. Engagement-Outcome Construction

The primary continuous outcome was

$$\log_{10}(\text{score} + 1).$$

Comment engagement was analyzed separately as

$$\log_{10}(\text{comments} + 1).$$

Posting-age windows were defined within each community using quartiles of post age in days. For the binary outcome, a post was classified as high engagement when its $\log_{10}(\text{score} + 1)$ was at or above the 75th percentile within its community-by-age-window bin. Bins with fewer than four observations received no threshold. Within cross-validation, binary-outcome thresholds were recomputed using only the training fold.

Table S12. Engagement variables and outcome definitions.

Variable	Definition
Continuous score outcome	$\log_{10}(\text{score} + 1)$
Comment engagement	$\log_{10}(\text{comments} + 1)$
Upvote ratio	post upvote ratio

Variable	Definition
Binary high-engagement outcome	1 if $\log_{10}(\text{score} + 1)$ is at or above the 75th percentile within the post's community-by-age-window bin
Age windows	quartiles of post age in days, defined within each community
Threshold rule	bins with fewer than four observations receive no threshold; binary thresholds recomputed using the training fold only

Under best-fit-family-grouped validation, adding the raw signature score and specificity margin produced the outcome-specific increments shown in Table S13.

Table S13. Outcome-specific incremental performance. Continuous-outcome incremental R^2 under best-fit-family-grouped validation.

Outcome	Incremental R^2
$\log_{10}(\text{score} + 1)$	+0.072
$\log_{10}(\text{comments} + 1)$	+0.016
Upvote ratio	+0.076

Note. The $\log_{10}(\text{score} + 1)$ row uses the same model as the "Raw signature score + specificity" row of the ablation analysis in §S12.

The point increment was larger for score and upvote ratio than for comment count.

S12. Model Specifications and Validation Designs

The control model included whole-image similarity, perceptual-hash similarity, template similarity, OCR-based text similarity, log posting age, and community indicators. The full model added the raw signature score and specificity margin.

Continuous outcomes were modeled using ridge regression with $\alpha = 1.0$. The binary outcome used regularized logistic regression with $C = 1.0$. Predictor scaling was estimated separately within each training fold. No data-dependent predictor selection was performed, and the regularization penalties were fixed in advance.

The historical Reddit score bank contained 31 family signatures, including the family excluded from the recovery analysis because it had only two development variants. Rescoring with the matched 30-signature bank produced a raw-score increment of +0.0680, compared with +0.0683 using all 31 signatures.

Best-fit-family groups were highly unequal in size. “Nobody / literally no one” was the best fit for 1,701 of 4,263 posts (39.9%) and “mocking spongebob” for 1,195; together the two largest groups contain 67.9% of posts, and under best-fit-family grouping each forms a fold by itself.

The same signature was the best fit for 409 of the 427 posts in the top decile of raw scores. Because raw scores are not calibrated across families, this concentration may partly reflect differences in raw-score scaling across families.

Four cross-validation schemes were evaluated. Each places a different restriction on how test observations relate to training observations, and none is designated as primary:

1. random five-fold cross-validation
2. perceptual-hash-cluster-grouped cross-validation using Hamming distance ≤ 10
3. community-grouped cross-validation, which with seven communities is leave-one-community-out
4. best-fit-family-grouped cross-validation

Chronology of the validation designs. The original engagement analysis used random folds and a leave-one-community-out classification analysis on 27 May 2026. On 13 July 2026, before the perceptual-hash-cluster analysis was run, an internal analysis plan designated perceptual-hash-cluster grouping as the intended primary design. That analysis could not be executed at the time because the required per-post hash variable was absent. Four schemes were run that day instead (random folds, random folds with within-fold outcome labels, community grouping, and best-fit-family grouping), after which best-fit-family grouping was labeled primary. That designation followed observation of the results, and best-fit-family grouping produced the largest increment among the schemes run. We subsequently executed the previously specified perceptual-hash-cluster design, obtaining $\Delta R^2 = +0.046$ with a 95% interval of [0.034, 0.058]. All four validation schemes are therefore reported without designating one as primary.

Perceptual-hash grouping should be interpreted mainly as a near-duplicate control rather than as comprehensive template grouping. At Hamming distance ≤ 10 , 95.6% of posts belong to singleton clusters, so broader template siblings are generally not forced into the same fold.

Table S14. Validation-design matrix. Grouping unit, generalization question, and remaining dependence for each cross-validation scheme. See also main-text Table 3.

Scheme	Grouping unit	Generalization question	Remaining dependence
Random five-fold	none (shuffled)	new posts from the sampled population	template siblings may share folds
Perceptual-hash cluster (Hamming ≤ 10)	perceptual-hash cluster	new near-duplicate groups	broader template siblings generally not grouped
Community grouped (= leave-one-community-out)	community	community absent from training	within-community template reuse
Best-fit-family grouped	best-fit family signature	posts assigned to a held-out best-fit-signature group	residual template structure across signature groups

Table S15. Control and full-model performance across validation schemes.

Panel A. Absolute model performance.

Scheme	R^2 control	R^2 full	ΔR^2	AUROC control	AUROC full	Δ AUROC
Random 5-fold	0.107	0.153	+0.046	0.513	0.596	+0.083
Perceptual-hash cluster	0.106	0.152	+0.046	0.518	0.601	+0.082
Community grouped (= LOCO)	-0.018	0.035	+0.052	0.547	0.617	+0.070
Best-fit-family grouped	0.057	0.129	+0.072	0.525	0.620	+0.094

The larger increments under some grouped designs partly reflect weaker control-model performance. For the continuous outcome, control-model R^2 was 0.107 under random folds, 0.057 under best-fit-family grouping, and -0.018 under community-grouped validation.

Panel B. Raw-signature and specificity ablation, best-fit-family-grouped validation.

Added predictors	ΔR^2
None	0.000
Raw signature score	+0.068
Specificity only	-0.005
Raw signature score + specificity	+0.072

The incremental association was concentrated in raw signature resemblance. Specificity contributed little additional predictive information.

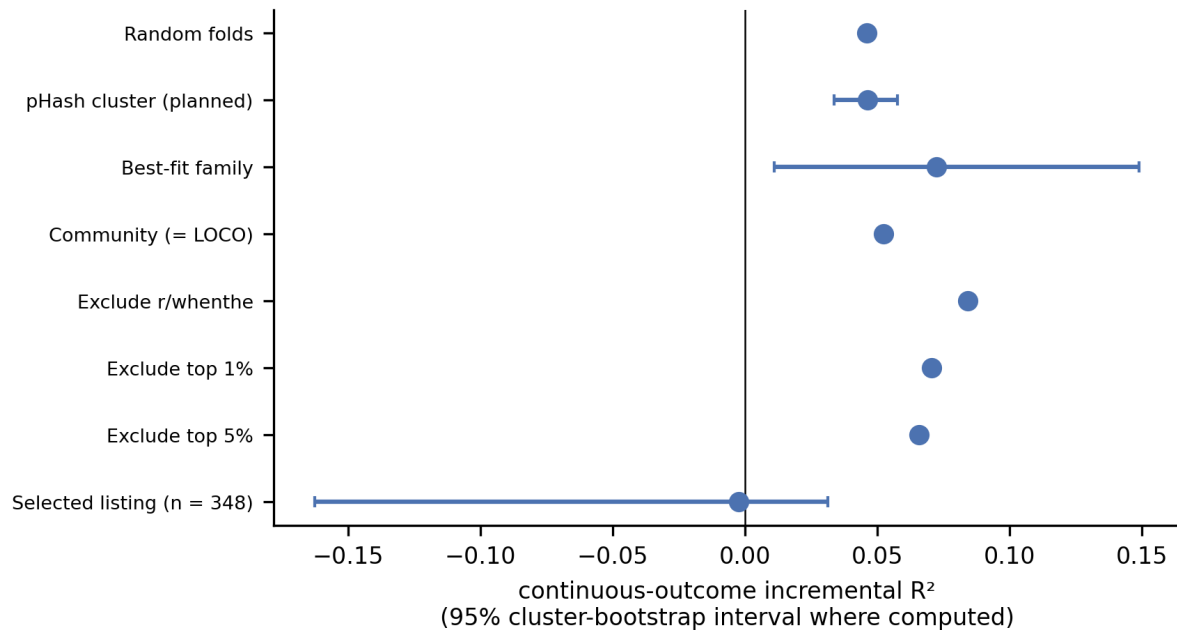


Figure S9. Engagement increment across validation designs and sensitivity analyses.

Continuous-outcome ΔR^2 for each validation scheme, the community- and engagement-tail-exclusion sensitivities, and the selected-listing contrast. Points are increments; horizontal bars show 95% cluster-bootstrap intervals where those intervals were estimated. Scheme rows add the raw signature score and specificity margin; exclusion rows add raw resemblance alone (full-sample reference increment 0.068).

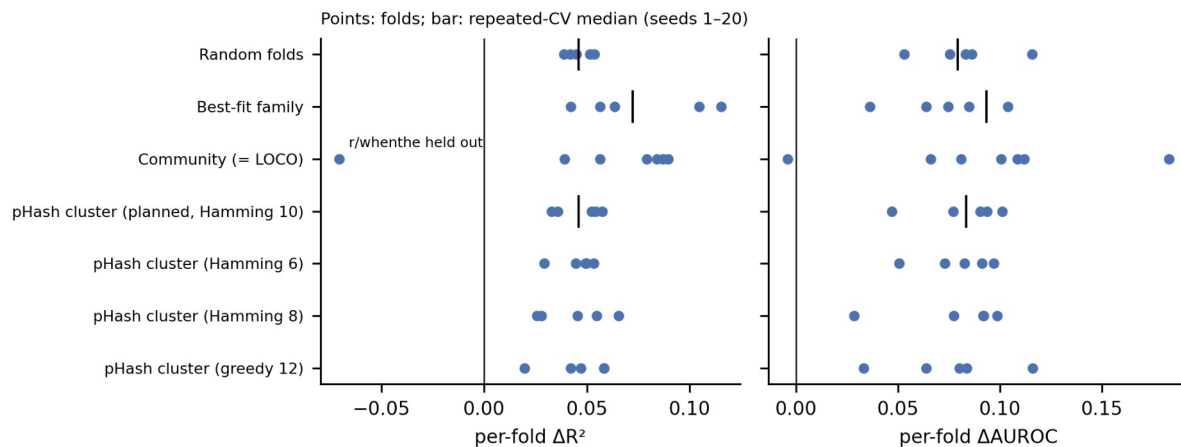


Figure S12. Per-fold engagement increments. Paired fold-level ΔR^2 and $\Delta AUROC$ for each validation scheme from g2_per_fold.csv, with the median across repeated cross-validation seeds 1–20 marked. The additional perceptual-hash-cluster specifications (Hamming thresholds 6 and 8, and greedy-leader clustering at 12) are cluster-definition sensitivities specified in the internal analysis plan before the perceptual-hash-cluster analysis was run. The r/whenthe fold is labeled; in leave-one-community-out validation the increment is negative when r/whenthe is held out.

S13. Bootstrap and Uncertainty

Uncertainty for grouped engagement increments was evaluated with 2,000 cluster-bootstrap iterations over the relevant grouping units (Cameron et al., 2008). Whole groups were sampled with replacement. If a group was drawn more than once, all copies of that draw were assigned to the same cross-validation group so that duplicated observations could not appear in both training and test sets. Intervals use bootstrap percentiles.

For the continuous outcome under best-fit-family grouping, the observed ΔR^2 was 0.072 with a 95% interval of [0.011, 0.149]. Across the 2,000 bootstrap draws, 99.7% of increments were above zero. Under perceptual-hash-cluster grouping, the 95% interval was [0.034, 0.058].

Repeated draws. When a group is drawn more than once, all of its copies share one cluster identity, so copies of the same observations never appear on opposite sides of a train/test split. Under this procedure, the best-fit-family interval is [0.011, 0.149].

These intervals quantify variation across the resampled grouping units. They do not simultaneously incorporate dependence associated with community, collection period, and residual template structure. The proportion of bootstrap increments above zero is reported descriptively and is not treated as a hypothesis-test p-value.

S14. Community and Sensitivity Analyses

The correlation between raw signature resemblance and log engagement varied across communities (Table S16).

Table S16. Community-specific associations. Correlation between raw signature resemblance and log engagement.

Community	Correlation
r/meme	+0.339
r/memes	+0.322
r/shitposting	+0.299
r/dankmemes	+0.167
r/me_irl	+0.163
r/historymemes	+0.157
r/whenthe	+0.020

The point estimate was positive in all seven communities and close to zero in r/whenthe.

Leave-one-community-out validation produced an incremental R^2 of +0.052. Under best-fit-family grouping with raw resemblance added to the control model, the continuous-outcome increment remained positive after excluding r/whenthe (+0.084; $n = 3,678$), after excluding the top 1% of posts by engagement (+0.070), and after excluding the top 5% (+0.066).

In the selected-listing contrast ($n = 348$), the incremental R^2 was -0.002 with a cluster-bootstrap 95% interval of $[-0.163, 0.031]$. This contrast uses the historical score block available for the selected-listing sample, entered as stored raw resemblance and specificity. Its point estimate is near zero, and its interval includes zero. The association observed in the new-listing denominator was therefore not reproduced in this selected-listing sample. Interpretation is limited by the smaller sample and by selection into already surfaced posts, which may restrict the available engagement variation.

Two negative controls were conducted. Permuting raw signature resemblance within community-by-age bins reduced the incremental performance to approximately zero. Models fitted after randomly shuffling the outcome performed at chance.

Table S17. Sensitivity analyses and negative controls. Continuous-outcome incremental R^2 under best-fit-family-grouped validation unless otherwise noted.

Analysis	Incremental R^2
Community grouped (= leave-one-community-out)	+0.052
Exclude r/whenthe	+0.084
Exclude top 1% of posts by engagement	+0.070
Exclude top 5% of posts by engagement	+0.066
Selected-listing contrast ($n = 348$)	-0.002 , 95% $[-0.163, 0.031]$
Negative control: permute score within community-by-age bins	0.000
Negative control: randomly shuffled outcome	+0.000 (full-model AUC 0.489)

Note. The community exclusions, engagement-tail exclusions, and negative controls add raw signature resemblance alone to the control model. The selected-listing contrast uses the historical two-predictor block of stored raw resemblance and specificity. All values in this table were computed on all 4,263 posts (the selected-listing contrast on its 348 posts); the generating script is listed in Table S22. The selected-listing interval uses the cluster bootstrap described in §S13.

Cross-source overlap sensitivity. We assessed overlap between the development meme-family corpus and the 4,263-post Reddit denominator using the same 64-bit DCT perceptual-hash criterion used for corpus deduplication, Hamming distance ≤ 6 . The number of Reddit posts byte-identical to a development-corpus image was 0. Under the perceptual-hash criterion, 14 Reddit posts (0.33% of the denominator) were near-duplicates of 14 development images spanning 9 families.

Excluding exactly these 14 posts produced continuous ΔR^2 estimates of 0.047 under random folds, 0.046 under perceptual-hash-cluster grouping, 0.053 under community grouping, and 0.072 under best-fit-family grouping. The representation-magnitude-adjusted residual was 0.012 with a 95% interval of $[-0.002, 0.044]$ and remained small and uncertain. These estimates show

no material dependence of the reported Reddit result on the identified cross-source near-duplicates.

S15. Signal Decomposition

Signal-decomposition analyses use all 4,263 denominator posts. Unless otherwise noted, estimates use best-fit-family-grouped validation and compare the control model with a model that adds the indicated score (Table S18).

Score definitions. On all 4,263 posts, the magnitude-adjusted increment is 0.013 [−0.001, 0.046], with seed sensitivity in Table S5a. The classifier score is the maximum one-versus-rest predicted probability; the maximum decision score gives the same increment (Table S18). Community-grouped validation is leave-one-community-out and is reported once (Table S15).

Table S18. Signal-decomposition models. Incremental R^2 and correlation with log engagement for each score construction under best-fit-family-grouped validation.

Score construction	ΔR^2	Correlation with log engagement
Raw signature resemblance	+0.068	+0.22
Sparse-centroid resemblance	+0.066	+0.24
Family-calibrated fit	+0.001	−0.01
Classifier score (maximum predicted probability)	−0.002	+0.00
Classifier maximum decision score	−0.002	+0.00
Sparse-activation norm	+0.039	+0.16
Whole-image embedding norm	+0.035	−0.07
Random-feature signature maximum	+0.002	+0.08
Raw signature over generic-magnitude block	+0.013, 95% interval [−0.001, 0.046]	—

Raw signature resemblance and sparse-centroid resemblance produced similar point increments (difference not tested). The full-sample downstream association therefore cannot be attributed specifically to the contrastive construction. Family-calibrated fit and both classifier-score constructions contributed essentially no incremental performance. Sparse-activation and whole-image embedding magnitude each carried smaller positive increments.

Randomization Tests

Table S19. Randomization-test results, all 4,263 posts, seed 0. Null distributions of incremental R^2 and Monte Carlo p -values calculated as $(b + 1)/(B + 1)$ for the observed increment of 0.068.

Null construction	Banks	Null mean	Null 95th percentile	Null maximum	Observed	Monte Carlo p
Random-feature signatures	500	+0.001	+0.010	+0.047	+0.068	0.002
Label-permutation signatures	200	+0.039	+0.062	+0.080	+0.068	0.030

None of the 500 random-feature banks reached the observed increment. Maximizing over an arbitrary collection of sparse feature vectors was therefore insufficient to reproduce the observed downstream association under this randomization.

The label-permutation procedure preserved the empirical sparse representation while breaking the documented family assignments used to estimate the signatures. These permuted signatures retained, on average, 58% of the observed increment. The observed increment exceeded this null, with Monte Carlo $p = 0.030$ – 0.045 across four seeds (Table S5a).

Adding sparse-activation norm, whole-image embedding norm, and text length to the control block reduced the raw-signature increment from 0.068 to 0.013, with a 95% interval of $[-0.001, 0.046]$. The remaining increment is small and uncertain.

Together, these analyses indicate that much of the downstream association resides in broad structured resemblance within the sparse representation and in representation magnitude. Evidence specifically linking the association to correct documented-family identity is limited. We use canonical resemblance to documented meme forms as shorthand for this broader empirical pattern rather than as a separately validated cultural construct.

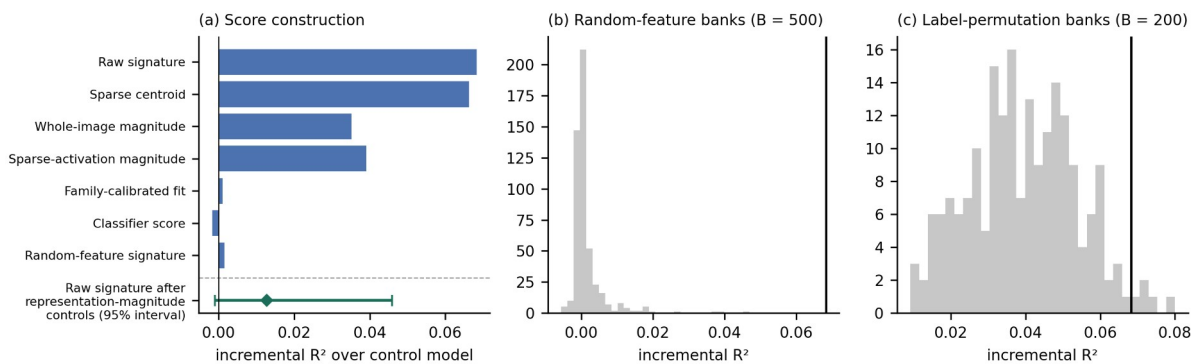


Figure S10. Signal decomposition and randomization controls. (a) Incremental R^2 by score construction. (b) Distribution across 500 random-feature signature banks. (c) Distribution across

200 label-permutation signature banks. The observed raw-signature increment is marked in panels (b) and (c). All panels use the 4,263-post denominator sample.

S16. Image Re-Encoding for the Decomposition Analyses

All 4,263 Reddit images were re-encoded for the decomposition analyses. Re-encoding reproduced the stored raw signature scores to the reported precision, with correlation 1.000000 and 100% agreement in best-fit family. The decomposition analyses therefore use all 4,263 posts from the seven communities (§S15).

S17. Prospective Engagement Analysis

A separate cohort was scored at initial collection and revisited to test whether raw signature resemblance predicted subsequent engagement after initial engagement was included in the model. A total of 1,398 posts were observed at each of three completed waves: initial collection, +4 hours, and +8 hours.

Planned observations at +24 hours, +72 hours, and +7 days could not be collected. Beginning 29 May 2026, Reddit returned HTTP 403 responses at the IP level to the unauthenticated public endpoints used for collection. Authenticated access through Reddit's Data API required a registration and approval process. An access request submitted through Reddit's request form in May 2026 received no response during the study window.

Among posts with complete model data, the incremental continuous-outcome R^2 was -0.002 at +4 hours ($n = 1,366$) and -0.002 at +8 hours ($n = 1,354$). The completed short-window waves therefore provide no clear evidence of incremental prediction beyond initial engagement.

Longer-horizon prospective prediction remains unevaluated.

Table S21. Prospective-wave design and completed analyses. A total of 1,398 posts were observed at each completed wave. Model sample sizes reflect complete-data availability.

Wave	Status	Incremental R^2 after initial engagement (n)
Initial collection (t_0)	completed	baseline
+4 hours	completed	-0.002 (1,366)
+8 hours	completed	-0.002 (1,354)
+24 hours	not collected (HTTP 403 from 29 May 2026)	—
+72 hours	not collected (HTTP 403 from 29 May 2026)	—
+7 days	not collected (HTTP 403 from 29 May 2026)	—

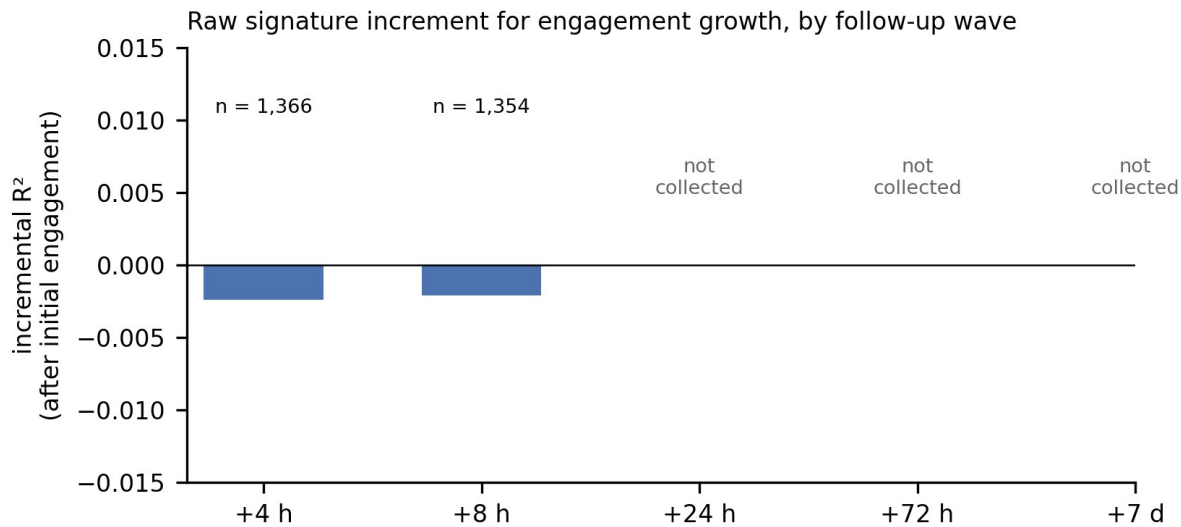


Figure S11. Short-window prospective analysis. Incremental continuous-outcome R^2 associated with raw signature resemblance after adjustment for initial engagement at the completed +4-hour and +8-hour waves. Model n is shown for each wave. The planned +24-hour, +72-hour, and +7-day waves are included for design context but were not collected.

S17a. Blinded Author-Coded Human Check

We conducted a blinded author-coded check to examine whether Reddit posts receiving different raw signature scores could be classified against the documented family bank by a human coder. A stratified random sample of 60 unique posts was drawn, with 10 posts from each of six strata spanning the observed raw-score distribution. Sampling, coding, and analysis rules were fixed before the sample was drawn.

One of the authors coded the posts using only the image and a fixed offline reference sheet containing exemplars from the 31 documented families. For each presentation, the coder selected one listed family, "none," or "uncertain." The raw score, score stratum, model-nominated family, community, title, engagement, and repeat status were hidden during coding. In addition, 10 posts were presented a second time, at least 20 presentations after their first appearance and without indication that they were repeats. Coding labels were fixed before the hidden model information was linked back to the responses.

Among the 60 unique posts, 0 were assigned to the model-nominated family, 3 to another listed family, 6 to none of the listed families, and 51 were coded uncertain. The two presentations of a repeated post received the same response in 10 of 10 cases.

The coder had developed the measure and is not a specialist in meme history or internet culture. Because most judgments were uncertain, the exercise did not yield a determinate human criterion of family membership, and we do not treat it as a validation benchmark. The absence of confirmed model nominations is not evidence that the nominations were incorrect, and the repeat

agreement indicates consistency in the coder's responses, including uncertainty, rather than substantive validity. No additional coders were introduced after the result was observed.

Table S21a. Human-check responses by raw-score stratum. Unique posts per stratum (strata span the observed raw-score distribution from lowest, B1, to highest, B6), by coder response, with the most frequent model-nominated family in each stratum.

Stratum (raw score)	Posts	Model-nominated family	Another listed family	None	Uncertain	Most frequent nomination
B1 (lowest)	10	0	0	2	8	mocking spongebob (4)
B2	10	0	0	1	9	nobody literally no one (6)
B3	10	0	0	1	9	mocking spongebob (7)
B4	10	0	2	1	7	nobody literally no one (9)
B5	10	0	0	0	10	nobody literally no one (7)
B6 (highest)	10	0	1	1	8	nobody literally no one (9)

The pre-specified comparison of the two lowest-score strata (B1–B2) with the two highest (B5–B6) found 0 of 20 versus 0 of 20 confirmed nominations (difference 0.00; Newcombe 95% interval [−0.16, 0.16]). In the highest strata, “nobody literally no one” was the nomination for 7 of 10 posts in B5 and 9 of 10 in B6.

Reproducibility Appendix

S18. Software, Code, and Data Manifest

Analyses were conducted on macOS (Darwin) using Python 3.9. Core software dependencies included NumPy 2.0.2, pandas 2.3.3, scikit-learn 1.6.1, PyTorch, open_clip, imagehash 4.3.2, and ocrmac. Images were encoded using OpenAI CLIP ViT-L/14. Sparse representations used the multimodal sparse autoencoder described by Zaigrajew et al. (2025), checkpoint WolodjaZ/MSAE, with 6,144 latent dimensions. Recovery negative sampling used seed 20260512. Engagement cross-validation and bootstrap point estimates used seed 0, with seed sensitivities reported in Table S5a.

Table S22 maps the principal reported results to their generating scripts and stored outputs; scripts whose results were superseded are listed in the final row.

REPRODUCIBILITY_MANIFEST.csv provides the expanded file-level manifest, including script and output checksums and reproduction commands.

Table S22. Reproducibility manifest.

Result	Script	Output
Development recovery, leave-one-variant-out protocol (Table 2, Panel 2A; Tables S4, S5, S7–S9)	analysis/g3_recovery/	g3_results.json, g3_family_table.csv
Fold-local decoy sensitivity (§S8)	analysis/repair/	decoys_foldlocal.json
Held-out-family validation (Table 2, Panel 2B; §§S1a, S7a)	analysis/heldout/	heldout_results.json, screening_summary.json
Exploratory decoy comparison with the fused centroid (§S7a)	analysis/heldout/exploratory/ x1_hard_decoys_vs_fused.py	exploratory_x1_decoys_vs_fused.js on
Engagement validation schemes (Tables 3, 4, S15; Figure S12)	analysis/g2_template_cluster/	g2_results.json, g2_per_fold.csv
Full-sample decomposition and randomization analyses (Figure 5; Tables S18, S19)	analysis/g1_full_sample/	g1_results.json, identity_check.json
Corrected bootstrap intervals and Monte Carlo p-values	analysis/repair/	bootstrap_samelabel.json, repair_cheap.json
Rule-based case selection (Figure 2b, Figure S7, §S9)	analysis/revision_selections/ select_cases.py	revision_selections.json
Recomputed sensitivity values (Tables S13, S17; Figure S9)	analysis/legacy_values/ recompute_legacy.py	legacy_values.json
Top-decile engagement-rate comparison (RQ3 scale interpretation)	analysis/r1_7_decile/decile_lift.py	decile_lift.json
Signature export used by Figures 2b and S7	analysis/figure_inputs/ export_signatures.py	signatures_submitted_export.json
Best-fit signature concentration (§S12; Results scale anchor)	analysis/concentration/ export_bestfit_concentration.py	bestfit_concentration.json
Development-corpus counts (§S1; Figure S1)	analysis/corpus/ export_corpus_counts.py	corpus_counts.json
Cross-source overlap and exclusion sensitivity (§S14)	analysis/m12_overlap/	m12_overlap.json, m12_exclusion_sensitivity.json
Prospective short-window analysis (§S17; Table S21; Figure S11)	analysis/prospective/ export_prospective.py	prospective.json
Blinded author-coded human check (§S17a)	analysis/human_check/	hc_results.json, hc_summary.json
Figures 1–5 and S1–S13, plotting only	revision/build/figures/ make_figures_r1.py	revision/build/out/figures/

Result	Script	Output
Superseded scripts: RUN_m2m4_full.py, RUN_e2_bootstrap.py, RUN_downstream_embed_score.py (subset filter), RUN_save_nulls.py (raw exceedance p)	—	replaced by the generating sources above

Raw meme images and raw Reddit content are not redistributed where source or platform terms restrict redistribution. The reproducibility package instead provides derived features, analysis outputs, and permitted post identifiers. Access instructions for any restricted review materials are supplied separately with the submission.

S19. Ethics and Data Handling

The study analyzed publicly accessible Reddit posts and publicly available meme-reference materials and involved no intervention with users. Usernames were removed from the analysis data. Post identifiers were retained where required for reproducibility and linkage. Posts that had been deleted or were unavailable when collection occurred were not included in the analyzed sample.

The development-corpus collection procedure requested paginated Know Your Meme gallery routes that were listed under Disallow in the site's robots.txt. The already collected development corpus was retained for provenance, but that retrieval procedure was not repeated. Held-out-family retrieval used only routes permitted by robots.txt and was limited to the first gallery page. The resulting held-out families are therefore shallower than the development families, and the analysis reports explicit family-size sensitivities rather than assuming similar retrieval depth.

Data collection and retention were limited to materials needed for the reported analyses. Redistribution of raw images and raw Reddit content is subject to source and platform terms. Derived datasets remove usernames while retaining permitted identifiers needed for reproducibility. A supplementary coding check was carried out by one of the authors, who labelled a sample of public Reddit images against a reference sheet. No other persons took part, and no ethics review was sought.

Supplementary References

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414–427.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Zauner, C. (2010). *Implementation and benchmarking of perceptual image hash functions* [Master's thesis, Upper Austria University of Applied Sciences].

Zaigrajew, V., Baniecki, H., & Biecek, P. (2025). Interpreting CLIP with hierarchical sparse autoencoders. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 73918–73956.