BEYOND BEST-PROMPT AUROC: TEXTUAL REACHABILITY OF DISTRIBUTION SHIFTS IN VLM

Anonymous authors
Paper under double-blind review

ABSTRACT

Models that connect images and text are increasingly used to describe visual shifts between image collections. A common approach is based on testing many textual descriptions and selecting the one that most strongly separates the two collections, as measured by Prompt AUROC. Yet, strong separation does not necessarily imply a meaningful explanation of the shift. In our experiments, the prompt `necessity`, unrelated to blur, almost perfectly separates original images from blurred versions of them (Prompt AUROC 0.999). Our analysis explains how a large visual shift and a search over many descriptions can produce this result. Under a Gaussian model, we derive that Prompt AUROC depends on both the magnitude of the visual shift and how well text aligns with it. We show that we can largely predict the highest Prompt AUROC from the magnitude of the visual shift and the directional coverage of the candidate text embeddings, independently of the prompts' meanings. We thus introduce *textual reachability*, a label-free measure of the separation achievable through text relative to the best linear separation available from image embeddings. Across 113 diverse shifts, it distinguishes those captured by text from those where arbitrary words achieve similarly high AUROC. It also predicts 1) whether discovered words keep their alignment for other VLM backbones and 2) the performance gains from using text to adapt models to new image collections. Best-prompt AUROC correlates negatively with both.

1 INTRODUCTION

Natural language is increasingly used to describe systematic differences between image collections, such as changes in their content or appearance (Zhu et al., 2022; Dunlap et al., 2024). We refer to these differences as *visual distribution shifts*, or simply *visual shifts*. VisDiff (Dunlap et al., 2024) proposes candidate text descriptions, or *prompts*, and ranks them by how well their image–text similarity scores, obtained from CLIP (Radford et al., 2021), distinguish the two collections. Given a description, the model assigns each image a similarity score. A description distinguishes two collections when images from one consistently receive higher scores than those from the other. The ranking performance is usually measured using the AUROC. We call the highest AUROC found across the candidate descriptions the *best-prompt AUROC*.

Text can also guide adaptation between datasets: 1) by describing how features from one dataset should change to resemble those of another (Dunlap et al., 2023), 2) by specifying the domain in the model's prompts (Ge et al., 2023), or 3) by reducing a classifier's reliance on domain-specific information (Zeng et al., 2025). These models inherently require text to capture (at least partly) the visual shift between the collections. This dependence makes it important to assess whether a candidate description captures the shift. Best-prompt AUROC is a natural candidate for this assessment and it is the criterion commonly used in the literature (Dunlap et al., 2024; Yin et al., 2025; Zhong et al., 2025).

However, a high Prompt AUROC does not necessarily imply that the prompt meaningfully describes the visual shift. Consider original images and blurred versions of the same images: in our experiments, the word `necessity` almost perfectly separates these collections (Prompt AUROC 0.999). Similarly, the word `rages` separates histopathological images from different hospitals in the Camelyon17 dataset with AUROC 0.979. Neither word describes the visual shift being tested (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

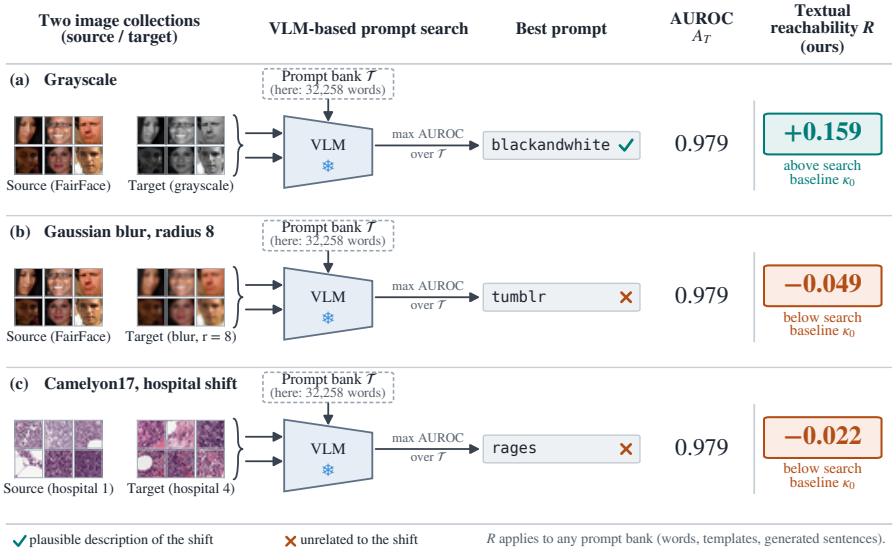


Figure 1: **High best-prompt AUROC does not imply that text captures the visual shift.** With the same CLIP ViT-L/14 and prompt bank $\mathcal{T}$, the best prompt reaches high AUROC in all three cases. For grayscale, the best prompt describes the shift and our textual reachability measure $R > 0$. For Gaussian blur and Camelyon17, the best prompts are unrelated and $R < 0$. Unlike AUROC, R measures how well text actually captures the visual shift.

Our analysis explains how this happens. We measure *visual shift magnitude* as how far apart the image collections' average visual representations are relative to the variation within each collection. Under a Gaussian approximation for the image–text similarity scores, we show that Prompt AUROC depends jointly on visual shift magnitude and how well a text vector aligns with the shift direction in the model's shared image–text embedding space. This relation closely matches empirical AUROC across 22,600 prompt evaluations, explaining 99.5% of the variance in observed AUROC (Appendix N, Figure 4). A large visual shift magnitude can therefore give high AUROC even with a weakly aligned description. Searching many descriptions adds another effect: the candidate text vectors point in many directions, increasing the chance of finding one that partially aligns with the shift even when the corresponding description is unrelated to it.

To solve that, we take a different perspective and ask how much of the available visual separation text can recover. We introduce *textual reachability* to compare the separation achieved through text with the best linear separation available from image embeddings. To account for what the search might find by chance, we randomly reassign images to the two collections and repeat the entire description search. We call the best alignment found in this way the *search baseline*, that is when collection membership carries no information. *Reachability* is the alignment of the best description in excess of this baseline, as a fraction of the largest possible excess. It thus separates shifts that text captures from shifts where a large visual difference gives high AUROC to an arbitrary description. Textual reachability requires no class labels.

Contributions. (i) *We explain why high AUROC can accompany an unrelated text description.* Our analysis separates the effects of visual shift magnitude and text alignment, and closely matches observed scores across 22,600 combinations of shifts and descriptions. (ii) *We quantify the effect of searching many descriptions.* Visual shift magnitude and the geometry of candidate texts predict best-prompt AUROC across 113 shifts, independent of the words' meanings (Pearson correlation 0.902). This effect persists across models and across searches over both words and phrases. (iii) *We introduce and validate textual reachability.* Textual reachability accounts for shift magnitude and for the alignment a search finds by chance, estimated by repeating the search on randomly reassigned images. It predicts whether discovered words retain their alignment for other VLM backbones, and the performance gains from text-guided adaptation of models to new image collections.

2 RELATED WORK

Describing set differences in language. VisDiff (Dunlap et al., 2024) proposes text descriptions with a language model and uses CLIP image–text similarity scores to assess how well they distinguish two image collections. It ranks the descriptions by AUROC to select the best candidates. The same AUROC ranking has since been adopted to compare urban streetscapes (Yin et al., 2025) and architectural styles (Zhong et al., 2025). D5 (Zhong et al., 2023) similarly describes differences between text corpora and also reports AUROC. Yet a large visual shift magnitude can give high best-prompt AUROC even to words poorly describing the shift between the collections. A measure of text alignment must therefore take visual shift magnitude into account.

Language-guided adaptation and generalization. LADS (Dunlap et al., 2023) uses text descriptions (*e.g.*, 'photos of birds' vs. 'paintings of birds') to transform source embeddings toward an unseen domain without target images. DAPL (Ge et al., 2023) learns domain-specific prompts to adapt a classifier from labeled source images to unlabeled target images. CoOp (Zhou et al., 2022a) replaces words surrounding a class name with learned vectors to improve image recognition; CoCoOp (Zhou et al., 2022b) adjusts these vectors for each image to improve generalization to unseen classes. These learned prompts need not correspond to readable descriptions. Later work jointly learns image and text prompts, refines features, or uses CLIP's initial predictions to guide adaptation (Khattak et al., 2023; Yu et al., 2024; Feng et al., 2024). LanCE (Zeng et al., 2025) uses domain descriptions to improve the generalization of classifiers built from named visual concepts. Its domain descriptor orthogonality (DDO) penalty reduces their sensitivity to the domain changes represented by those descriptions. These approaches motivate assessing whether the available text representations provide useful directions for handling a particular visual shift. Their relevance depends on the source–target pair, the image and text encoders, and the prompts considered. We examine this connection through external experiments with LADS and DDO, comparing how reachability and best-prompt AUROC correlate with adaptation and generalization gains.

Predicting transfer, and embedding geometry. Transferability estimation predicts which pretrained representations or training data will be useful for a target task without training every candidate (Nguyen et al., 2020; Huang et al., 2022; Hu et al., 2022; Han et al., 2023). Similarly, we assess how much of a particular visual shift text can recover. CLIP's image and text embeddings occupy different, concentrated regions of their shared space (Liang et al., 2022). We connect the geometry of candidate text directions and visual shift magnitude to best-prompt AUROC, showing how much of this score can be predicted without using the descriptions' meanings.

3 METHOD: TEXTUAL REACHABILITY

Textual reachability measures the separation achievable through text relative to the best linear separation between the two image collections in embedding space. Its construction follows two steps. We first separate visual shift magnitude from text alignment, showing why high Prompt AUROC can mask weak alignment (Sections 3.1, 3.2). We then examine the alignment supplied by searching many descriptions (Section 3.3) and account for it by repeating the search after randomly reassigning collection membership (Section 3.4). These steps explain the contrast in Figure 2(a) and lead to the reachability values evaluated in Section 5. Proofs are in Appendix A.

3.1 SEPARATING SHIFT MAGNITUDE FROM TEXT ALIGNMENT

We begin with a single description and measure what fraction of the best available linear separation between two image collections its text vector recovers. We fix an image encoder f_V and a text encoder f_T, with unit-length embeddings $z(x) = f_V(x)/\|f_V(x)\|$ and $t(q) = f_T(q)/\|f_T(q)\|$ in $\mathbb{R}^p$. A description q, also called a *prompt*, may be a word or phrase. We write $t = t(q)$ for its embedding and define the image score $s_t(x) := t^\top z(x)$.

We observe source and target collections $\mathcal{D}_S, \mathcal{D}_T$, drawn from underlying distributions P_S and P_T. Collection membership is known, but image class labels are not used. For independent images $x^S \sim P_S$ and $x^T \sim P_T$, the *Prompt AUROC* is defined as:

$$\mathrm{AUROC}(t) = \mathbb{P}\left[s_t(x^T) > s_t(x^S)\right] + \tfrac{1}{2}\mathbb{P}\left[s_t(x^T) = s_t(x^S)\right]. \tag{1}$$

It measures how consistently target images receive higher scores than source images, with ties receiving half credit. A value of 1 means target images always score higher, 0.5 corresponds to chance, and values below 0.5 indicate a reversed ordering. We estimate it by averaging over $\mathcal{D}_S \times \mathcal{D}_T$.

Let μ_S, μ_T and Σ_S, Σ_T be the means and covariance matrices of $z(x)$ estimated from $\mathcal{D}_S$ and $\mathcal{D}_T$. We call $\Delta = \mu_T - \mu_S$ the visual *shift vector* and $\Sigma = (\Sigma_S + \Sigma_T)/2$ the average covariance. We assume $\Delta \neq 0$ and that Σ is positive definite.

To measure how separable the collections are from image embeddings alone, we replace the text vector by an arbitrary nonzero direction w and consider the linear score $w^\top z(x)$. We quantify its separation by the standardized mean difference, that is, the gap between the two collections' mean scores divided by their pooled within-collection standard deviation:

$$d'(w) = \frac{w^\top \Delta}{\sqrt{w^\top \Sigma w}}, \qquad d_{\max} = \max_{w \neq 0} d'(w) = \sqrt{\Delta^\top \Sigma^{-1} \Delta}. \tag{2}$$

The maximum is attained at $w^\star = \Sigma^{-1}\Delta$, Fisher's linear discriminant direction, and $d_{\max}$ is the Mahalanobis distance between the collection means, which we call *visual shift magnitude*: the largest standardized separation achievable by any linear score on the image embeddings, without using text.

Furthermore, we define a description's *alignment* as $\kappa(t) = d'(t)/d_{\max}$. An alignment of 0.5 recovers half the maximum separation, whereas a negative sign reverses which collection has higher mean scores. Thus $|\kappa(t)|$ is the recovered fraction when either ordering is allowed.

Alignment has a simple geometric form after whitening the embeddings by the within-collection covariance. Calling $\tilde{z} = \Sigma^{-1/2} z$, the score becomes $t^\top z = (\Sigma^{1/2} t)^\top \tilde{z}$, and the shift becomes $\Sigma^{-1/2}\Delta$. Alignment is then the cosine between these two vectors.

Proposition 1 (Alignment is a whitened cosine)**.** *Let $\Sigma^{1/2}$ denote the positive definite square root of Σ. Then*

$$\kappa(t) = \cos \angle \Big(\Sigma^{1/2} t,\ \Sigma^{-1/2} \Delta \Big) .$$

Hence $|\kappa(t)| \leq 1$, with $\kappa(t) = 1$ exactly when t is a positive multiple of $\Sigma^{-1}\Delta$, and $\kappa(t) = -1$ exactly when it is a negative multiple.

Thus, the best text vector points along $\Sigma^{-1}\Delta$, not along the mean shift Δ. Writing $\Sigma = \sum_i \lambda_i e_i e_i^\top$ gives $\Sigma^{-1}\Delta = \sum_i (e_i^\top \Delta / \lambda_i)\, e_i$: each component of the shift is weighted by the inverse of its variance rather than its size. If target images are brighter on average but brightness also varies widely within each collection, the brightness component of Δ is large but its variance λ_i might be even larger, so it contributes little. Alignment thus rewards descriptions of shifts that are consistent across images, not necessarily the largest or most salient one.

3.2 Why Prompt AUROC does not determine alignment

The distinction between magnitude and alignment helps us formalize the effect illustrated in Figure 3 (Appendix A): increasing visual separation can raise Prompt AUROC while text alignment stays fixed. Under a Gaussian score model, AUROC depends on the product of these two quantities.

Proposition 2 (Gaussian decomposition of Prompt AUROC)**.** *If source and target scores are Gaussian with means $t^\top \mu_S, t^\top \mu_T$ and variances $t^\top \Sigma_S t, t^\top \Sigma_T t$, then*

$$\mathrm{AUROC}(t) = \Phi\left(\frac{d'(t)}{\sqrt{2}} \right) = \Phi\left(\frac{d_{\max} \kappa(t)}{\sqrt{2}} \right), \tag{3}$$

where Φ is the standard normal cumulative distribution function.

Proposition 2 shows why a high Prompt AUROC need not imply strong text alignment: a large visual shift magnitude can compensate for weak alignment. It follows from applying the Gaussian AUROC relation (Green & Swets, 1966) to image–text scores, allowing unequal collection covariances through their average Σ. Section 5.2 tests this relation on real scores (Figure 4, Appendix N).

Corollary 1 (AUROC does not determine alignment)**.** *Under Proposition 2, an AUROC value $\alpha \in (\frac{1}{2}, 1)$ fixes only the product $d_{\max}\kappa = \sqrt{2}\,\Phi^{-1}(\alpha)$.*

Thus, identical AUROC values can hide different levels of text alignment: across shifts, a description with low κ on a large shift and one with high κ on a small shift can reach the same AUROC. Figure 2(a) illustrates this ambiguity on synthetic and real shifts, and Section 5.1 examines the grayscale and blur examples in Table 1. Measuring $\kappa(t)$ resolves the magnitude ambiguity for a given description. For a description selected from many candidates, however, high alignment may partly reflect the size of the candidate set rather than the description itself, which we address next.

3.3 Alignment supplied by the description search

Normalizing by visual magnitude separates alignment from magnitude, but we must still account for selecting the best description among many candidates. Let $\mathcal{T} := \{t_1, \ldots, t_M\}$ contain the candidate text embeddings. Allowing either collection to receive higher scores, define the best alignment and best-prompt AUROC as

$$\kappa_{\max} = \max_{t \in \mathcal{T}} |\kappa(t)|, \qquad A_T = \max_{t \in \mathcal{T}} \max\{\mathrm{AUROC}(t), 1 - \mathrm{AUROC}(t)\}.$$

The candidate text vectors cover different directions in embedding space, so some alignment can arise even without a meaningful relation to the shift. We quantify this geometric contribution by averaging the maximum alignment over uniformly random directions. To obtain a single reference value shared across shifts, we fix a positive definite reference covariance $\bar{\Sigma}$ by pooling within-collection covariances across the registry (Appendix B). Using each pair's Σ would instead give a different coverage value for each pair. Holding $\bar{\Sigma}$ fixed therefore gives one r per encoder and candidate bank, as reported in Table 3; it does not replace Σ in $d_{\max}$, κ or R.

Definition 1 (Mean directional coverage)**.** For a unit vector v in the shared image–text embedding space, define

$$K(v) = \max_{t \in \mathcal{T}} |\cos \angle(\bar{\Sigma}^{1/2} t, v)|, \qquad r = \mathbb{E}_{v \sim \mathrm{Unif}(\mathcal{S}^{p-1})}[K(v)]. \tag{4}$$

Here $\mathcal{S}^{p-1}$ is the unit sphere, and $K(v)$ is the best alignment with v among candidate text vectors transformed by $\bar{\Sigma}^{1/2}$.

Thus, r measures how much alignment the candidate set provides on average with a direction chosen independently of it. It depends on the candidate embeddings and the fixed covariance $\bar{\Sigma}$, but uses no evaluated shift direction. Adding candidates cannot decrease it. It is an average reference level, not a lower bound for every shift. Combining this reference alignment with visual magnitude gives a prediction of best-prompt AUROC through the decomposition established in Proposition 2.

Proposition 3 (Predicting best-prompt AUROC from search geometry)**.** *Under the Gaussian score model, $A_T = \Phi(d_{\max}\kappa_{\max}/\sqrt{2})$, where $\kappa_{\max}$ is the best alignment among the candidates defined above. Replacing it by the mean directional coverage r defines the* geometric AUROC prediction

$$A_{\mathrm{geom}} = \Phi\left(\frac{d_{\max} r}{\sqrt{2}}\right). \tag{5}$$

If, at fixed $d_{\max}$ and $\Sigma = \bar{\Sigma}$, the shift direction varies so that $v_\Delta = \Sigma^{-1/2}\Delta / \|\Sigma^{-1/2}\Delta\|$ is uniform on the sphere independently of $\mathcal{T}$, then $\mathbb{E}_{v_\Delta}[A_T] \le A_{\mathrm{geom}}$ by Jensen's inequality.

The expectation above is over random shift directions at fixed magnitude. For individual shifts, we evaluate A_{geom} as a prediction: once r is fixed, it depends on the shift only through $d_{\max}$, without requiring its text alignment $\kappa_{\max}$. Despite this simplification, it tracks observed best-prompt AUROC across 113 shifts at Pearson correlation 0.902, with median absolute error 0.081 (Figure 2(b)). This agreement shows that visual magnitude and average directional coverage can largely anticipate the discrimination achieved by prompt search (i.e., best-prompt AUROC), without identifying a meaningful description of the shift. Corollary 2 (Appendix A) bounds the prediction error in terms of the departure of $\kappa_{\max}$ from r under the Gaussian score model. This shared geometric reference motivates accounting for search when interpreting alignment. To calibrate the alignment measured for a particular pair of collections, we next construct a baseline from that pair's images.

3.4 Textual reachability calibrated for search

We obtain this baseline by randomly reassigning the images to the two collections and repeating the description search. Each permutation π preserves collection sizes. We recompute the mean

difference Δ_π and average covariance Σ_π from the reassigned collections, following Section 3.1. The alignment of a candidate t under this reassignment is $\kappa_\pi(t) = \cos\angle(\Sigma_\pi^{1/2}t, \Sigma_\pi^{-1/2}\Delta_\pi)$. We define the *search baseline* as the average best alignment over these permutations:

$$\kappa_0 = \mathbb{E}_\pi \left[\max_{t \in \mathcal{T}} |\kappa_\pi(t)| \right]. \tag{6}$$

We compare the observed best alignment $\kappa_{\max}$ with κ_0 to measure how much more alignment the search recovers than under random collection assignments. We define *textual reachability*, for $\kappa_0 < 1$, as

$$R = \frac{\kappa_{\max} - \kappa_0}{1 - \kappa_0}. \tag{7}$$

In equation 7, subtracting κ_0 removes the average alignment obtained under permutation, while dividing by $1 - \kappa_0$ scales this excess relative to the remaining gap to perfect alignment. Thus, $R = 0$ when the search matches the permutation mean, and $R = 1$ when a candidate recovers the maximum linear separation. Negative values fall below the permutation mean, and $R \leq 1$.

Proposition 4 (Invariance to magnitude at fixed Σ and κ_0)**.** *Fix Σ, $\mathcal{T}$ and the permutation mean κ_0 defined in equation 6. Replacing Δ by $c\Delta$, for $c > 0$, leaves $\kappa(t)$, $\kappa_{\max}$ and R unchanged, while $d_{\max}$ scales by c. Under the Gaussian score model, A_T increases with c when $\kappa_{\max} > 0$.*

Real image edits can also change Σ and κ_0; Section 5.1 therefore tests stability across corruption severity empirically. Reachability depends on the chosen encoders and descriptions, while the words' meanings require a separate check. Section 5.3 tests its practical relevance through cross-encoder transfer and adaptation gains (Table 2). For estimation, we use covariance shrinkage and evaluate selected prompts on held-out images. Appendix B details these estimates and the split-half adjustment for sampling noise; the bounds above apply to the unadjusted definitions.

4 Experimental setup

We evaluate whether textual reachability ranks visual shifts consistently with the relevance of the selected descriptions, whether visual magnitude and directional coverage predict best-prompt AUROC, and whether R predicts cross-encoder transfer and adaptation gains. The following shift registry, candidate descriptions and protocols support these evaluations.

Registry. Our *shift registry* contains 113 source–target comparisons from 4 groups. The 27 paired FairFace transformations (Kärkkäinen & Joo, 2021) cover colour, geometry, image quality and two synthetic spatial fingerprints. Source and target contain the same underlying images. Each severity level of blur, JPEG compression, noise and resampling counts as a shift, allowing us to vary severity while keeping the operation fixed. The 56 GenImage comparisons (Zhu et al., 2023) comprise 14 real-to-generated shifts (seven generators in raw and canonically re-encoded form) and 42 ordered comparisons between generator domains. We include all 12 ordered shifts among four Camelyon17 centres (Bándi et al., 2019; Koh et al., 2021), with balanced tumour and normal patches, and 18 natural domain pairs: 6 from DomainNet (Peng et al., 2019) and 12 from OfficeHome (Venkateswara et al., 2017). Geometric prediction, cross-encoder transfer and zero-shot adaptation use all 113 shifts. Implementation details are in Appendix K; pair lists are in Appendices D and E.

Reference descriptions. For paired transformations and natural domain pairs, the applied operations and dataset media types provide reference descriptions, fixed before prompt search. We use these references to semantically assess whether selected prompts describe the visual change and compare this assessment with R and best-prompt AUROC.

Encoders and search. Prompt discovery uses OpenAI CLIP ViT-L/14 (Radford et al., 2021) and exhaustive search over 32,258 readable single-token words. To test dependence on the candidate bank, we also search 39,995 templated expressions and 20,000 LLM-proposed descriptions. Appendix F examines multi-token PEZ (Wen et al., 2023) and continuous prompts. Cross-encoder transfer evaluates the selected strings without re-optimization using OpenCLIP ViT-B/32 and ViT-H/14 pretrained on LAION-2B (Schuhmann et al., 2022; Cherti et al., 2023), and SigLIP ViT-B/16 trained on WebLI (Zhai et al., 2023). None is used for prompt search, tuning or model selection.

Text-guided adaptation. Domain-conditioned zero-shot prediction rewrites the class prompt and is compared with plain zero-shot classification on all 113 shifts. Head adaptation compares LADS-style feature translation (Dunlap et al., 2023) and DDO-style projection (Zeng et al., 2025) with a source-only head. It uses a separate set of 36 semantic domain shifts from DomainNet, OfficeHome, PACS (Li et al., 2017), Office-31 (Saenko et al., 2010) and VisDA (Peng et al., 2017), specified before the runs (Appendix I).

Protocol. Search, prompt selection and final evaluation use disjoint image splits. We use 199 permutations for the search baseline κ_0 and 49 for the visual probe. Transfer and adaptation use Spearman correlations; adaptation correlations control for the unadapted accuracy drop. Confidence intervals use bootstrap resampling over shifts, and adaptation outcomes average over three seeds.

5 Results

5.1 High Prompt AUROC can coexist with low textual reachability

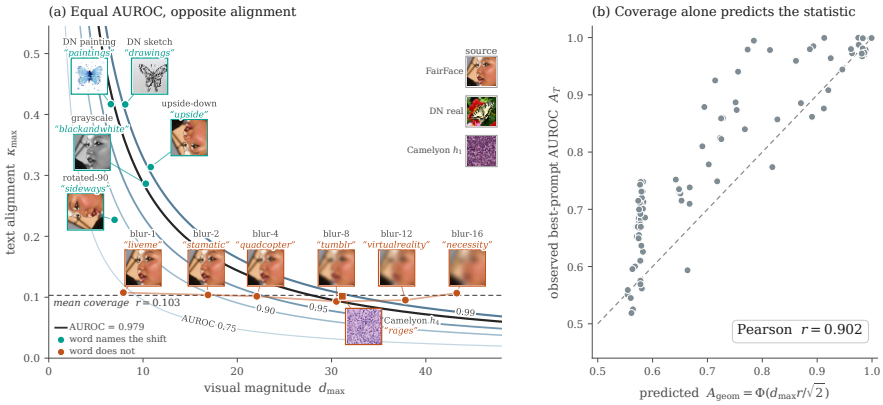


Figure 2: **Best-prompt AUROC confounds visual magnitude and text alignment.** **(a)** Registered shifts plotted in the $(d_{\max}, \kappa_{\max})$ plane. Iso-AUROC curves are hyperbolas: $d_{\max}\kappa_{\max} = \sqrt{2}\Phi^{-1}(\alpha)$; the dotted line marks the mean directional coverage $r = 0.103$ of the 32,258-word search. Here, a Gaussian blur-based visual shift on FairFace varies in shift magnitude $d_{\max}$ when varying the blur radius, but its alignment $\kappa_{\max}$ remains close to r. As a result the prompt search returns unrelated descriptions. Instead, other nameable shifts with similar AUROC but higher alignment are correctly identified. DN stands for DomainNet. **(b)** Across the 113 pooled shifts, $d_{\max}$ and r alone largely predict best-prompt AUROC, without using a shift-specific description.

We first assess whether ranking visual shifts by textual reachability agrees with the relevance of their selected descriptions. Among the examples in Table 1, meaningful descriptions have positive R and rank above unrelated ones with negative R, whereas high Prompt AUROC occurs in both cases.

Grayscale and Gaussian blur at radius 8 both attain best-prompt AUROC $A_T = 0.979$ after rounding, yet their alignments are 0.286 and 0.093, and their reachabilities are $+0.159$ and -0.049, respectively. The grayscale prompt, `blackandwhite`, describes the transformation, whereas the blur prompt, `tumblr`, does not. Figure 2(a) illustrates how different combinations of magnitude and alignment yield the same AUROC. Similarly, JPEG compression at quality 10 attains $A_T = 0.876$ with the unrelated prompt `stamatic`, yet $R = -0.064$.

We then vary corruption severity while keeping the transformation type fixed to test the magnitude effect illustrated in Figure 3. Along the Gaussian-blur ladder, $d_{\max}$ increases from 7.9 to 43.3 and A_T rises from 0.749 to 0.999, while R remains between -0.056 and -0.030. Along the JPEG ladder, $d_{\max}$ ranges from 2.9 to 24.6 and A_T from 0.626 to 0.980, while R remains between -0.064 and $+0.006$. Thus, stronger visual separation can drive AUROC close to 1 without the descriptions capturing a greater fraction of the shift. Reachability reflects this distinction by remaining near zero.

Table 1: **High best-prompt AUROC A_T can accompany unrelated descriptions; negative reachability R flags their weak alignment.** Selected paired FairFace transformations, natural-domain pairs and hospital-acquisition shifts, ordered by R within each block. Teal marks prompts plausibly related to the transformation or domain difference, rust unrelated ones. Negative R means that the best text alignment falls below the search baseline κ_0. Complete tables appear in Appendices D and E.

Shift	$d_{\max}$	A_T	$\kappa_{\max}$	R	prompt
Synthetic transformations					
Upside-down	10.82	0.995	0.314	**+0.194**	✓ upside
Grayscale	10.30	**0.979**	0.286	**+0.159**	✓ blackandwhite
Rotated-90	6.98	0.879	0.227	**+0.092**	✓ sideways
Visible watermark	18.66	0.999	0.186	**+0.038**	✓ sample
Sinusoidal fingerprint	5.95	0.738	0.122	**−0.040**	× newprofilepic
Gaussian blur r=4	22.07	**0.944**	0.102	**−0.046**	× quadcopter
Gaussian blur r=8	30.51	**0.979**	0.093	**−0.049**	× tumblr
JPEG q=10	18.60	0.876	0.086	**−0.064**	× stamatic
Real domain pairs					
DN real to painting	6.58	0.958	0.417	**+0.322**	✓ paintings
DN real to sketch	8.11	0.989	0.416	**+0.316**	✓ drawings
OH real-world to clipart	8.50	0.961	0.307	**+0.180**	✓ cartoon
OH product to art	9.56	0.941	0.231	**+0.091**	✓ artworks
C17 h_1 to h_4	31.14	**0.979**	0.102	**−0.022**	× rages
C17 h_3 to h_4	28.84	**0.967**	0.095	**−0.042**	× greenwood

Dataset abbreviations: DN: DomainNet; OH: OfficeHome; C17: Camelyon17.

A controlled experiment isolates the effect of magnitude directly in embedding space. We fix a ten-dimensional subspace defined by text contrasts and add shifts with no component in that subspace. Increasing their magnitude raises A_T from 0.555 to 0.974, while R remains negative (Appendix C). Supplying correct descriptions for the image transformations also yields strong alignment on positive controls but median alignment below κ_0 on corruptions (Appendix M). Low reachability therefore cannot be explained solely by omission of the correct descriptions from the candidate bank.

5.2 Testing the Geometric Prediction Across Shifts and Search Banks

Previous examples show that high AUROC A_T can coexist with low reachability R. We now ask how much of best-prompt AUROC can be predicted from shift magnitude and search geometry alone, as proposed in Section 3.3. We check the Gaussian AUROC relation, then evaluate the geometric prediction across shifts, and finally examine its dependence on the search bank and encoder.

Across 22,600 (shift, prompt) pairs, predicting AUROC as $\Phi(d_{\max}\kappa/\sqrt{2})$ using Equation 3 explains 99.5% of the variance in observed AUROC, without fitting a nonlinear mapping. Predictions closely follow the identity line (Appendix N, Figure 4), supporting the Gaussian approximation.

Table 3 shows that the evaluated candidate banks have modest directional coverage, with r between 0.103 and 0.148, below that of matched-size sets of uniformly random directions. Nevertheless, even $r = 0.103$ yields a geometric AUROC prediction of 0.90 when $d_{\max} = 17.6$. Broad coverage is therefore unnecessary for predicting high best-prompt AUROC on sufficiently large shifts.

With $r = 0.103$ fixed across all 113 shifts, A_{geom} achieves Pearson correlation 0.902 with observed best-prompt AUROC and median absolute error 0.081 (Figure 2(b)). Only $d_{\max}$ varies between shifts in this prediction; neither their directions nor their descriptions are used. Appendix L examines how estimating $\bar{\Sigma}$ from different image collections changes r and the resulting predictions.

The error of the geometric AUROC prediction depends on both magnitude and the difference between $\kappa_{\max}$ and r (Corollary 2, Appendix A). Under the Gaussian score model, this error vanishes at large magnitude for fixed positive $\kappa_{\max}$ and r. The Camelyon17 results illustrate this effect: at $d_{\max} \in [5, 9]$, A_{geom} underestimates observed AUROC by 0.085 to 0.134, whereas the six pairs at

Table 2: **Reachability R is positively and best-prompt AUROC negatively associated with every external criterion.** Reported results are Spearman correlations with bootstrap 95% CIs within brackets. Adaptation rows give partial correlations controlling for the unadapted accuracy drop.

Criterion	R	best-prompt AUROC
Cross-encoder transfer of the discovered string ($n = 113$)		
OpenCLIP ViT-B/32	$+0.658$ $[+0.511, +0.766]$	-0.234 $[-0.422, -0.040]$
OpenCLIP ViT-H/14	$+0.664$ $[+0.517, +0.768]$	-0.197 $[-0.371, -0.015]$
SigLIP ViT-B/16	$+0.453$ $[+0.267, +0.607]$	-0.252 $[-0.459, -0.092]$
Text-guided adaptation, recovery fraction (partial, severity controlled)		
zero-shot, full registry ($n = 113$)	$+0.548$ $[+0.114, +0.931]$	-0.353 $[-0.510, -0.172]$
head adaptation, semantic ($n = 36$)	$+0.470$ $[+0.147, +0.695]$	-0.545 $[-0.760, -0.206]$

$d_{\max} \in [28, 32]$ are predicted to within 0.016. These high-magnitude pairs attain AUROC between 0.967 and 0.981 with unrelated prompts such as `rages` and `greenwood` (Appendix E).

The geometric prediction also applies to the two sentence banks: their respective coverage values give Pearson correlations of 0.913 and 0.908 with best-sentence AUROC. Thus, modest directional coverage combined with large visual magnitude is relevant across the evaluated banks and encoders.

5.3 Cross-Encoder Transfer and Adaptation Gains

Having shown how visual shift magnitude and search geometry affect best-prompt AUROC, we ask if textual reachability R yields a more useful measure. In Table 2, we compare R and best-prompt AUROC on three outcomes external to their construction: alignment retained by discovered strings in other encoders, gains from domain-conditioned zero-shot prediction and from head adaptation.

Transfer. Each discovered string is re-encoded without re-optimization, with Δ, Σ, and $d_{\max}$ recomputed per transfer model (Appendix H). R correlates positively with retained alignment across all three encoders ($\rho \in [0.45, 0.66]$), whereas best-prompt AUROC correlates negatively. Thus, higher R identifies shifts whose prompts retain more alignment outside the discovery encoder.

Zero-shot adaptation. Across the full registry, R correlates positively with the fraction of the accuracy drop recovered ($\rho = 0.548$), while best-prompt AUROC correlates negatively ($\rho = -0.353$), controlling for the unadapted drop. The same sign contrast holds for raw accuracy gains.

Head adaptation. On the 36 semantic domain shifts, the corresponding correlations are $+0.470$ for R and -0.545 for best-prompt AUROC. Their paired difference on the same bootstrap resamples is $+1.015$ (95% CI $[+0.443, +1.399]$). These evaluations support R as a diagnostic of transferable alignment and adaptation benefit for the tested encoders, candidate banks, and interventions.

6 Limitations and Conclusion

Textual reachability is specific to the evaluated encoder and candidate bank, but remains stable under richer searches: four-token prompts preserve the ranking of shifts and do not recover shifts that no single word reaches (Appendices F and G). Exact invariance to magnitude requires fixed Σ and κ_0 (Proposition 4); stability under real image edits is assessed empirically.

Our results show that high best-prompt AUROC can result from large visual shift magnitude combined with modest text alignment (Figure 2). Textual reachability distinguishes shifts with similarly high AUROC but substantially different recovered alignment (Table 1). It is also positively associated with cross-encoder transfer and text-guided adaptation gains, whereas best-prompt AUROC is negatively associated with these outcomes (Table 2). We therefore recommend reporting R alongside best-prompt AUROC when text is used to describe or correct a visual shift, together with a separate check that the selected descriptions explain it.

REPRODUCIBILITY STATEMENT

The assumptions and definitions underlying textual reachability are stated in Section 3, with proofs in Appendix A. Section 4 and Appendix K describe the datasets, preprocessing, pretrained encoders, candidate descriptions, evaluation splits, and implementation details. Appendix B specifies the estimators and permutation calibration, while Appendix K describes uncertainty estimation and repeated runs. The protocols for transfer across encoders and text-guided adaptation, including baselines and evaluation criteria, are detailed in Appendices H and I. Appendix N documents the empirical check of the Gaussian approximation. All experiments use public image datasets and pretrained vision–language encoders.

AI ASSISTANCE DISCLOSURE

Large language models were used solely to improve the clarity, grammar, and readability of the authors' original writing. They were not used for research ideation, literature discovery, methodological design, experiment execution, data generation or analysis, mathematical reasoning, result interpretation, or the drafting of substantive scientific content. All AI-assisted edits were reviewed and approved by the authors, who take full responsibility for the final content of the paper.

REFERENCES

P. Bándi, O. Geessink, Q. Manson, M. van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

L. Dunlap, C. Mohri, D. Guillory, H. Zhang, T. Darrell, J. E. Gonzalez, A. Raghunathan, and A. Rohrbach. Using language to extend to unseen domains. In *International Conference on Learning Representations*, 2023.

L. Dunlap, Y. Zhang, X. Wang, R. Zhong, T. Darrell, J. Steinhardt, J. E. Gonzalez, and S. Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24199–24208, 2024.

R. Feng, T. Yu, X. Jin, X. Yu, L. Xiao, and Z. Chen. Rethinking domain adaptation and generalization in the era of CLIP. *arXiv preprint arXiv:2407.15173*, 2024.

C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Z. Han, Z. Zhang, F. Wang, R. He, W. Su, X. Xi, and Y. Yin. Discriminability and transferability estimation: a Bayesian source importance estimation approach for multi-source-free domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7811–7820, 2023.

J. Hu, H. Zhong, F. Yang, S. Gong, G. Wu, and J. Yan. Learning unbiased transferability for domain adaptation by uncertainty modeling. In *Proceedings of the European Conference on Computer Vision*, pp. 223–241, 2022.

L.-K. Huang, J. Huang, Y. Rong, Q. Yang, and Y. Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9201–9225, 2022.

K. Kärkkäinen and J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. MaPLe: multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. WILDS: a benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664, 2021.

D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017.

V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

C. V. Nguyen, T. Hassner, M. Seeger, and C. Archambeau. LEEP: a new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7294–7305, 2020.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pp. 213–226. Springer, 2010.

C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein. Hard prompts made easy: gradient-based discrete optimization for prompt tuning and discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51008–51025, 2023.

Jun Yin, Jing Zhong, Peilin Li, Ruolin Pan, Pengyu Zeng, Miao Zhang, and Shuai Lu. Urbansense: A framework for quantitative analysis of urban streetscapes leveraging vision large language models. *arXiv preprint arXiv:2506.10342*, 2025.

X. Yu, S. Yoo, and Y. Lin. CLIPCEIL: domain generalization through CLIP via channel refinement and image-text alignment. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4267–4294, 2024.

Z. Zeng, Y. Su, J. Sun, T. Wen, H. Zhang, Z. Wang, B. Chen, H. Liu, and J. Ma. Explaining domain shifts in language: concept erasing for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9517–9526, 2025.

X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Jing Zhong, Jun Yin, Peilin Li, Pengyu Zeng, Miao Zang, Ran Luo, and Shuai Lu. Archilense: A framework for quantitative analysis of architectural styles based on vision large language models. *arXiv preprint arXiv:2506.07739*, 2025.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a.

K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022b.

M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. GenImage: a million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782, 2023.

Zhiying Zhu, Weixin Liang, and James Zou. GSCLIP: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*, 2022.

A PROOFS

Throughout, $\Sigma \succ 0$ and $\Delta \neq 0$.

Lemma 1 (Optimal linear separation)**.** $\max_{w \neq 0} d'(w) = \sqrt{\Delta^\top \Sigma^{-1} \Delta}$*, attained at* $w^\star = \Sigma^{-1}\Delta$ *up to positive scaling.*

Proof. d' is invariant to positive scaling of w, so restrict to $w^\top \Sigma w = 1$. Substituting $u = \Sigma^{1/2} w$ and $v = \Sigma^{-1/2}\Delta$ turns the problem into $\max_{\|u\|=1} u^\top v = \|v\|$, attained at $u = v/\|v\|$, that is $w \propto \Sigma^{-1}\Delta$, with value $\|v\|^2 = \Delta^\top \Sigma^{-1} \Delta$ under the square root. This is Fisher's linear discriminant; we include it for self-containedness. □

Proof of Proposition 1. Put $u = \Sigma^{1/2} t$ and $v = \Sigma^{-1/2}\Delta$. Then $t^\top \Delta = u^\top v$ and $t^\top \Sigma t = \|u\|^2$, while $d_{\max} = \|v\|$ by Lemma 1. Hence $\kappa(t) = u^\top v/(\|u\|\|v\|)$, a cosine, so $|\kappa| \leq 1$ by Cauchy–Schwarz with equality if and only if $u \parallel v$, that is $t \propto \Sigma^{-1}\Delta$. □

Proof of Proposition 2. $\mathrm{AUROC}(t) = \Pr(t^\top z^T > t^\top z^S)$. The difference of two independent Gaussians is Gaussian with mean $t^\top \Delta$ and variance $t^\top \Sigma_S t + t^\top \Sigma_T t = 2\, t^\top \Sigma t$, using $\Sigma = \frac{1}{2}(\Sigma_S + \Sigma_T)$. Therefore $\mathrm{AUROC}(t) = \Phi\big(t^\top \Delta / \sqrt{2\, t^\top \Sigma t}\big) = \Phi(d'(t)/\sqrt{2})$, and $d'(t) = d_{\max}\kappa(t)$ gives the second equality. Defining Σ as the average of the two covariances is what makes the identity exact without assuming $\Sigma_S = \Sigma_T$. □

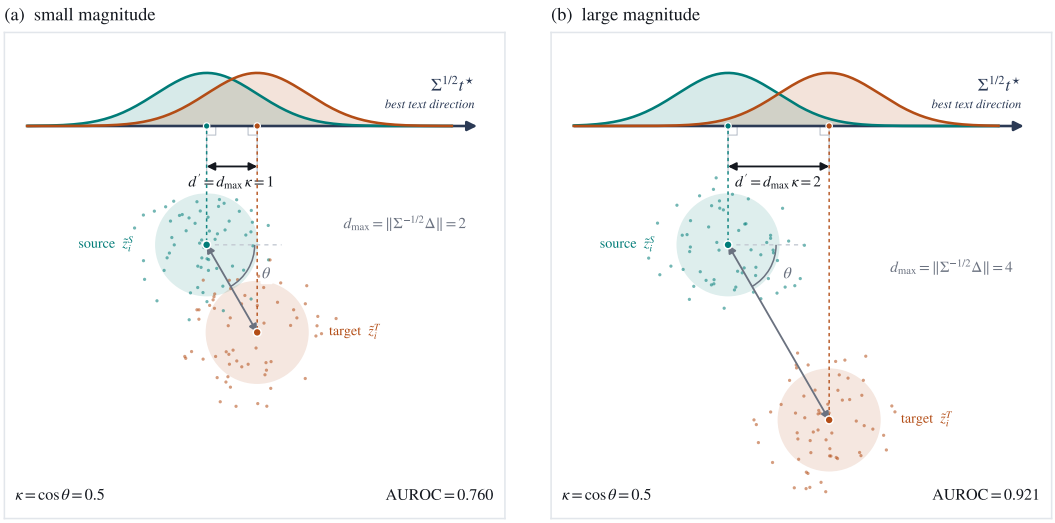


Figure 3: **Same text alignment, different visual shift magnitude.** Colored dots represent image embeddings from two collections; the curves above show their score distributions along the text direction. In this Gaussian illustration, shift magnitude doubles from (a) to (b), while text alignment stays fixed ($\kappa = 0.5$). The scores become easier to separate, raising AUROC from 0.760 to 0.921 without improving text alignment.

Proof of Corollary 1. Φ is strictly increasing, so $\mathrm{AUROC} = \alpha$ if and only if $d_{\max}\kappa = \sqrt{2}\,\Phi^{-1}(\alpha)$. For fixed α this is a hyperbola in $(d_{\max}, \kappa)$, and for each $\kappa \in (0, 1]$ it has the unique solution $d_{\max} = \sqrt{2}\,\Phi^{-1}(\alpha)/\kappa$. □

Proof of Proposition 4. By Proposition 1, $\kappa(t) = \cos\angle(\Sigma^{1/2}t,\ \Sigma^{-1/2}\Delta)$, and the cosine of an angle is invariant to positive rescaling of either argument, so $\Delta \mapsto c\Delta$ with Σ held fixed leaves $\kappa(t)$ unchanged for every $t \in \mathcal{T}$, hence leaves $\kappa_{\max}$ unchanged. The search baseline κ_0 is held fixed by the proposition's assumption. This is not implied by permutation: changing the image collections can change the permutation distribution and hence κ_0. R in equation 7 is a function of those two quantities and is therefore unchanged. By contrast $d_{\max} = \|\Sigma^{-1/2}\Delta\|$ scales by c, so $A_T = \Phi\big(c\, d_{\max}\kappa_{\max}/\sqrt{2}\big)$ is strictly increasing in c whenever $\kappa_{\max} \neq 0$. □

Proof of Proposition 3. We proceed in six steps. Steps 1 to 3 derive exact identities under the Gaussian score model. Step 4 defines an auxiliary random-direction model; step 5 bounds its expected AUROC by the geometric prediction, and step 6 inverts that prediction.

Step 1 (whitened coordinates). Whitening image embeddings uses $z \mapsto \Sigma^{-1/2}z$. The corresponding score direction transforms as $t \mapsto \Sigma^{1/2}t$, preserving the inner product. Define, for a text direction $t \neq 0$ and the shift vector,

$$u_t \;=\; \frac{\Sigma^{1/2}t}{\|\Sigma^{1/2}t\|}, \qquad v \;=\; \frac{\Sigma^{-1/2}\Delta}{\|\Sigma^{-1/2}\Delta\|}, \qquad u_t, v \in \mathcal{S}^{p-1}.$$

Both are well defined because $\Sigma \succ 0$ and $\Delta \neq 0$. By Proposition 1,

$$\kappa(t) \;=\; \cos\angle\Big(\Sigma^{1/2}t,\ \Sigma^{-1/2}\Delta\Big) \;=\; u_t^\top v.$$

Two normalizations are involved and should not be confused: the text embeddings in $\mathcal{T}$ are unit vectors in the original geometry, but $\Sigma^{1/2}t$ is not a unit vector, which is why u_t is renormalized here. Nothing is lost, since κ is invariant to positive rescaling of t.

Step 2 (the maximum over the searched set is exactly $K_\Sigma(v)$). Taking the maximum of $|\kappa|$ over $\mathcal{T}$ and defining the pair-specific coverage function K_Σ gives

$$\max_{t\in\mathcal{T}} |\kappa(t)| \;=\; \max_{t\in\mathcal{T}} \big|u_t^\top v\big| \;=\; \max_{t\in\mathcal{T}} \Big|\cos\angle\Big(\Sigma^{1/2}t,\ v\Big)\Big| \;=\; K_\Sigma(v).$$

This is an identity, not an approximation, and it holds for every $\mathcal{T}$ and every Δ. For a fixed pair geometry, the search interacts with the shift direction only through the single unit vector v.

Step 3 (best-prompt AUROC). Under Proposition 2, $\mathrm{AUROC}(t) = \Phi(d_{\max}\kappa(t)/\sqrt{2})$. Using $1 - \Phi(x) = \Phi(-x)$ and the definition of A_T,

$$A_T = \max_{t\in\mathcal{T}} \max\{\mathrm{AUROC}(t), 1 - \mathrm{AUROC}(t)\} = \Phi\left(\frac{d_{\max}K_\Sigma(v)}{\sqrt{2}}\right). \tag{8}$$

This allows either score orientation without assuming that $-t$ is the embedding of a readable prompt in the bank. Given the shift and searched set, the Gaussian model determines best-prompt AUROC from the magnitude $d_{\max}$ and the directional coverage $K_\Sigma(v)$.

Step 4 (an auxiliary random-orientation model). For the observed pair, the direction $v = \Sigma^{-1/2}\Delta/\|\Sigma^{-1/2}\Delta\|$ is fixed. To interpret average coverage, we introduce a separate reference model. Holding Σ, $\mathcal{T}$ and a magnitude $d > 0$ fixed, draw $V \sim \mathrm{Unif}(\mathcal{S}^{p-1})$ and set

$$\Delta_V = d\,\Sigma^{1/2}V.$$

These constructed shifts have constant magnitude $\|\Sigma^{-1/2}\Delta_V\| = d$ and whitened direction V. Writing $\kappa_V(t)$ for their text alignment gives

$$\mathbb{E}_V\left[\max_{t\in\mathcal{T}} |\kappa_V(t)|\right] = \mathbb{E}_V[K_\Sigma(V)] =: r_\Sigma.$$

At $\Sigma = \bar{\Sigma}$, this expectation equals the reference coverage r. The expectation concerns this auxiliary construction; it makes no randomness assumption about the observed pair.

Step 5 (mean coverage and the geometric prediction). Under the Gaussian score model, $A_{\mathcal{T}}(V) = \Phi(dK_\Sigma(V)/\sqrt{2})$. Since $K_\Sigma(V) \geq 0$ and Φ is concave on $[0, \infty)$,

$$\mathbb{E}_V[A_{\mathcal{T}}(V)] \leq \Phi\left(\frac{d\,r_\Sigma}{\sqrt{2}}\right).$$

At $\Sigma = \bar{\Sigma}$, the right-hand side is $\Phi(dr/\sqrt{2})$. For an observed pair, using the shared r defines $A_{\text{geom}} = \Phi(d_{\max} r/\sqrt{2})$. When $\Sigma \neq \bar{\Sigma}$, this is a prediction whose accuracy is evaluated empirically in Section 5.2.

Step 6 (the threshold). Φ is a strictly increasing bijection onto $(0, 1)$, so for $\alpha \in (\frac{1}{2}, 1)$,

$$A_{\text{geom}} \geq \alpha \iff \frac{d_{\max} r}{\sqrt{2}} \geq \Phi^{-1}(\alpha) \iff d_{\max} \geq \frac{\sqrt{2}\,\Phi^{-1}(\alpha)}{r},$$

which is the stated threshold, and completes the proof. □

Corollary 2 (Sensitivity of the Geometric AUROC Prediction)**.** *Under the Gaussian score model, for any shift and nonempty searched set,*

$$|A_T - A_{\text{geom}}| \;\leq\; \frac{d_{\max}}{\sqrt{2}}\,\varphi\!\left(\frac{d_{\max}\min(\kappa_{\max}, r)}{\sqrt{2}}\right)|\kappa_{\max} - r|,$$

where φ is the standard normal density. If $m = \min(\kappa_{\max}, r) > 0$ is held fixed, the magnitude-dependent factor is maximal at $d_{\max} = \sqrt{2}/m$ and tends to zero as $d_{\max} \to 0$ and $d_{\max} \to \infty$. Thus the bound is largest at intermediate magnitude. If $\kappa_{\max} = 0 < r$, the bound still holds, but the actual discrepancy tends to $1/2$ at large magnitude.

Proof of Corollary 2. Write $a = d_{\max}\kappa_{\max}/\sqrt{2}$ and $b = d_{\max} r/\sqrt{2}$, both nonnegative, so that $A_T = \Phi(a)$ by equation 8 and $A_{\text{geom}} = \Phi(b)$. The mean value theorem gives $|\Phi(a) - \Phi(b)| \leq |a - b| \max_{\xi \in [a \wedge b,\, a \vee b]} \varphi(\xi)$, and φ is decreasing on $[0, \infty)$, so that maximum is $\varphi(a \wedge b) = \varphi\big(d_{\max}\min(\kappa_{\max}, r)/\sqrt{2}\big)$; substituting $|a - b| = d_{\max}|\kappa_{\max} - r|/\sqrt{2}$ gives the bound. For the shape of the prefactor put $m = \min(\kappa_{\max}, r) > 0$ and $x = d_{\max} m/\sqrt{2}$, so that $\frac{d_{\max}}{\sqrt{2}}\varphi\big(d_{\max} m/\sqrt{2}\big) = x\varphi(x)/m$. Since $\frac{d}{dx}\, x\varphi(x) = (1 - x^2)\varphi(x)$, the map $x \mapsto x\varphi(x)$ increases on $[0, 1]$, decreases on $[1, \infty)$ and vanishes at both ends, so the prefactor is maximal at $x = 1$, that is at $d_{\max} = \sqrt{2}/m$, and tends to zero as $d_{\max} \to 0$ and as $d_{\max} \to \infty$. □

The corollary bounds, rather than equates, the discrepancy between the observed statistic and the fixed-r prediction by the product of $|\kappa_{\max} - r|$ and a magnitude-dependent sensitivity; when $\kappa_{\max} \geq r$, the latter peaks at $d_{\max} = \sqrt{2}/r$ and decays on either side. R uses the analogous pair-specific excess over κ_0. The twelve Camelyon17 pairs of Appendix E exhibit both factors, and Section 5.2 reads them that way.

Remark 1 (What the assumption of step 4 does and does not require). The uniformity assumption is placed on the *whitened* direction v, not on the shift vector Δ. This distinction is not cosmetic. CLIP's embedding geometry is strongly anisotropic, so a Δ drawn uniformly in the original space maps to a v that is far from uniform, and conversely. The proposition should therefore be read as a statement about the geometry in which separation is actually measured, which is the whitened one, and this is the same geometry in which κ is a cosine by Proposition 1.

A shift related to the candidate descriptions can have $K_\Sigma(v)$ above the average directional reference, but the proposition does not guarantee this for any individual shift. Textual reachability uses the distinct, pair-specific permutation baseline κ_0; it is not a test of uniformity of v. The shared r is used only for the geometric AUROC prediction. Its empirical agreement with best-prompt AUROC supports the usefulness of that prediction, without establishing that real shift directions are uniformly distributed or that a selected description is meaningful.

Remark 2 (Why the null must re-run the search). Let $S(\pi) = \max_{t \in \mathcal{T}} |\kappa_\pi(t)|$ and let $t^\star$ be selected on the unpermuted data. For every permutation, $S(\pi) \geq |\kappa_\pi(t^\star)|$, hence

$$\mathbb{E}_\pi[S(\pi)] \geq \mathbb{E}_\pi[|\kappa_\pi(t^\star)|].$$

Evaluating only $t^\star$ therefore omits the advantage of selecting a candidate under each reassignment. The absolute alignment of a fixed candidate generally has a positive mean; it need not be zero. Nor must the permutation mean equal r, whose reference distribution is different.

B ESTIMATOR DETAILS

Attenuation. $\kappa_{\max}$ is a cosine with the estimated direction $\widehat{\Delta}$, and a cosine with a noisy direction is attenuated exactly as a correlation is attenuated by measurement error. We estimate the reliability $\widehat{\varrho}$ of the estimated displacement direction by split-half agreement of two independently estimated directions and apply $\tilde{\kappa} = \kappa/\sqrt{\widehat{\varrho}}$ as a sampling-noise adjustment before forming equation 7, following Spearman (1904). Reliabilities are high across the registry, with a median of 0.98, and lowest for the smallest-magnitude cross-generator pairs, where the correction matters most. For a common positive $\widehat{\varrho}$ within a shift, the adjustment is monotone in κ and therefore preserves within-shift rankings. The exact cosine bound applies to unadjusted alignment; dividing by estimated reliability can produce a value above one. The correction is an estimator adjustment, not an additional population identity.

Reference geometry. For each encoder, $\bar{\Sigma}$ is the registry-pooled within-collection covariance, fixed before evaluating individual shifts; the corresponding r is computed once for each candidate bank and used unchanged throughout. Within-collection centering separates this reference from the covariance of one concatenated, uncentered image pool, which also includes between-collection mean differences. The reference is needed only to compare candidate-bank coverage and form A_{geom}. Pair-specific Σ and permuted Σ_π remain the covariances used to construct alignment and its search baseline.

Table 3: **Modest directional coverage can yield high predicted AUROC for large visual shifts.** p is the embedding dimension, M the number of candidate descriptions, and r their mean directional coverage i.e. the average best alignment with a uniformly sampled direction, using the reference covariance $\bar{\Sigma}$. Uniform reports the corresponding coverage of M independent directions sampled uniformly from the unit sphere. $d_{90} = \sqrt{2}\,\Phi^{-1}(0.90)/r$ is the visual shift magnitude at which the geometric prediction A_{geom} reaches 0.90.

Set	Encoder	p	M	r	Uniform	d_{90}
tokens	CLIP ViT-L/14 (OpenAI)	768	32,258	0.103	0.154	17.6
tokens	OpenCLIP ViT-B/32 (LAION-2B)	512	32,258	0.148	0.188	12.3
tokens	OpenCLIP ViT-H/14 (LAION-2B)	1024	32,258	0.112	0.134	16.2
tokens	SigLIP ViT-B/16 (WebLI)	768	32,258	0.111	0.154	16.3
sentences	CLIP ViT-L/14 (templated)	768	39,995	0.115	0.156	15.7
sentences	CLIP ViT-L/14 (LLM bank)	768	20,000	0.119	0.150	15.3

An AUROC-scale variant. The same idea can be expressed on the scale practitioners already read, as $\Delta_A = A_T - A_{\text{geom}}$: observed best-prompt AUROC minus the semantics-free prediction of equation 5. It tracks R closely on the paired transformations and is useful when a reader wants the correction stated in AUROC units. We carry R in the main text because its unadjusted definition explicitly scales alignment above the search baseline by the remaining gap to perfect alignment. Although Δ_A is also bounded, it does not have this interpretation.

Permutation null. 199 permutations for text statistics and 49 for the visual probe, group sizes preserved, the full enumeration repeated inside each. Image–text scores can be cached, but the reassigned group means, covariances and candidate alignments must be recomputed for every permutation. The maximizing prompt can change under reassignment.

C CONTROLS

Planted shifts with known alignment. We construct shifts whose magnitude and alignment are controlled independently. A rank-10 text-contrast subspace is fixed before evaluation; paired targets displace every source embedding along a direction whose squared projection into that subspace is known by construction, scaled by the data's own standard deviation along the planted direction. Table 4 gives the zero-alignment rows: as magnitude grows, best-prompt AUROC climbs from 0.555 to 0.974 while R stays at or below zero throughout. At the largest separation roughly one word in

three hundred of the entire vocabulary exceeds AUROC 0.90 on a shift orthogonal to the fixed rank-10 text-contrast subspace. Orthogonality to this subspace does not imply orthogonality to every text embedding. These targets are constructed directly in embedding space and need not correspond to actual images; the paired image transformations of Appendix D test the same mechanism using actual images.

Table 4: Controlled shifts orthogonal to the fixed rank-10 text-contrast subspace, with planted subspace energy equal to zero. "Median word" is the AUROC of the median word in the vocabulary; "> 0.9" the fraction of all $32{,}258$ words exceeding AUROC 0.90.

snr	$d_{\max}$	A_T	median word	> 0.9	R
1.4	4.60	0.555	0.512	0.0%	-0.149
2.4	7.89	0.593	0.518	0.0%	-0.143
6.0	19.70	0.708	0.535	0.0%	-0.130
12.0	39.28	0.847	0.556	0.0%	-0.109
25.0	79.29	0.974	0.625	0.3%	-0.078

Natural positive control. On FairFace with a source balanced by age and a target containing a larger proportion of older subjects, the two collections have the same proportions in the other recorded demographic strata, and no image appears in both collections. The search returns `elderly` with $\kappa_{\max} = 0.315$ against a null of $\kappa_0 = 0.169$, giving $R = +0.176$. The next four words are `retirement`, `retired`, `elder` and `ageing`; `elderly` is selected in all five splits. This control tests whether the selected word reflects an age-mixture change. The statistic R is unchanged by swapping the same source and target collections, since $\kappa_{\max}$ uses absolute alignment. Directional prompt retrieval is therefore distinct from the two-sided reachability statistic defined in Section 3.3.

Negative control. Balanced-to-balanced, same dataset, disjoint images: the visual probe reaches $A_V = 0.501$ with $p = 0.560$. There is no detectable shift, and we report reachability as undefined rather than zero, because its alignment input $\kappa = d'(t)/d_{\max}$ normalizes by a magnitude that is not distinguishable from zero.

D Full transformation taxonomy

Tables 5 and 6 give all 27 paired FairFace transformations and all 18 natural domain pairs, ordered by R. The rows shown in Table 1 are drawn from here without reordering.

All natural-domain pairs have positive reachability, with $R \in [0.091, 0.322]$ and $\kappa_{\max} \in [0.231, 0.417]$. Their reachability range overlaps that of the paired transformations; for example, upside-down images attain $R = 0.194$. We display these families separately to distinguish natural changes in style or medium from controlled transformations of the same images. Both families belong to the main registry.

E Camelyon17 acquisition shifts in full

Hospital acquisition is the sharpest illustration of Proposition 3 on real data, because magnitude varies by a factor of six across pairs of the same kind. Table 7 gives all twelve ordered pairs.

Every pair involving centre h_4 has $d_{\max} > 28$ and is separated above AUROC 0.96, and every one of them is predicted to within 0.016 by magnitude and mean directional coverage alone. These high AUROC values accompany prompts that do not describe the acquisition difference. Their proximity to the geometric prediction illustrates why high discrimination alone does not establish a meaningful explanation.

Table 5: All 27 paired transformations, ordered by textual reachability. The block spans $d_{\max}$ from 2.35 to 43.28 and A_T from 0.600 to 1.000.

Shift	$d_{\max}$	A_T	$\kappa_{\max}$	R	prompt
Upside-down	10.82	0.995	0.314	$+0.194$	`upside`
Grayscale	10.30	0.979	0.286	$+0.159$	`blackandwhite`
Rotation 90°	6.98	0.879	0.227	$+0.092$	`sideways`
Sepia	17.08	0.985	0.208	$+0.071$	`blackandwhite`
Visible watermark	18.66	0.999	0.186	$+0.038$	`sample`
Posterized	26.91	1.000	0.165	$+0.020$	`colorized`
JPEG q=20	16.61	0.980	0.152	$+0.006$	`stamatic`
Channel mix	2.35	0.600	0.163	$+0.003$	`avengersendgame`
Gaussian noise σ=60	29.19	0.999	0.143	-0.001	`pixel`
JPEG q=90	2.86	0.626	0.158	-0.003	`stamatic`
Resample 0.10	19.74	0.964	0.133	-0.013	`greenwood`
JPEG q=5	24.62	0.973	0.129	-0.015	`pixel`
Gaussian noise σ=40	24.31	0.980	0.127	-0.017	`photoshopped`
Checkerboard fingerprint	10.07	0.840	0.139	-0.018	`newprofilepic`
Resample 0.06	27.09	0.987	0.115	-0.028	`photographing`
Gaussian blur r=16	43.28	0.999	0.107	-0.030	`necessity`
JPEG q=50	12.97	0.857	0.118	-0.036	`supernatural`
Resample 0.25	15.49	0.885	0.112	-0.037	`via`
Sinusoidal fingerprint	5.95	0.738	0.122	-0.040	`newprofilepic`
Gaussian blur r=12	37.83	0.999	0.095	-0.044	`virtualreality`
Gaussian blur r=2	16.89	0.862	0.104	-0.045	`stamatic`
Gaussian blur r=4	22.07	0.944	0.102	-0.046	`quadcopter`
Gaussian blur r=8	30.51	0.979	0.093	-0.049	`tumblr`
Brightened	12.48	0.774	0.094	-0.054	`tages`
Gaussian noise σ=20	19.33	0.908	0.096	-0.054	`filtered`
Gaussian blur r=1	7.91	0.749	0.108	-0.056	`liveme`
JPEG q=10	18.60	0.876	0.086	-0.064	`stamatic`

F SEARCH CAPACITY BEYOND ONE TOKEN

Exhaustive one-token search is an interpretability instrument, not the capacity limit of the text tower. We therefore enlarge the searched set in two ways and recompute R for each, with κ_0 obtained by rerunning *the same* search on label-permuted shift directions (same optimiser, restarts and iterations; 19 permutations). PEZ (Wen et al., 2023) optimises $M \in \{1, 2, 4\}$ tokens projected onto the vocabulary at every step; continuous slots optimise $M = 4$ unprojected token embeddings. Both maximise $|\kappa|$ directly, so that $\kappa_{\max}$ remains a maximum of the same quantity over a larger set. All searches use the pipeline's Σ for each shift, and the exhaustive row reproduces Table 1 exactly. At M=1, PEZ recovers a median 84% of the exhaustive $\kappa_{\max}$, so PEZ rows are lower bounds.

More tokens mostly improve coverage. Enlarging the set raises $\kappa_{\max}$ from 0.19 to 0.25, but its coverage and hence its permutation mean rise almost as fast (κ_0 from 0.15 to 0.20), so median R increases only from 0.045 to 0.077. The ordering of shifts is preserved (Spearman 0.88, 0.84, 0.77 against one-token R), and cross-generator shifts remain the most reachable family at every M.

Unreachable shifts stay unreachable. For the 30 shifts with one-token $R \leq 0$, median R remains at or below zero at every M (-0.030, -0.038, -0.046, -0.011). The Gaussian-blur ladder (median $R = -0.011$ at M=4) and the Camelyon17 hospital pairs (0.001) are not recovered by longer prompts.

Continuous prompts reach every direction, including meaningless ones. Continuous slots attain $\kappa_{\max} = 0.993$ on every shift, but equally on label-permuted directions ($\kappa_0 = 0.994$; coverage $r = 0.93$); only 4% of shifts exceed their own null and R is undefined. The text tower can thus express essentially any direction, so continuous capacity measures what the encoder can express,

Table 6: All 18 natural domain pairs, ordered by textual reachability. Every natural pair has positive R, although its range overlaps that of the paired transformations. DN denotes DomainNet, OH OfficeHome.

Shift	$d_{\max}$	A_T	$\kappa_{\max}$	R	prompt
DN real to painting	6.58	0.958	0.417	$+0.322$	`paintings`
DN real to sketch	8.11	0.989	0.416	$+0.316$	`drawings`
DN sketch to real	8.34	0.984	0.415	$+0.312$	`drawings`
DN painting to real	6.76	0.954	0.408	$+0.308$	`painting`
OH real-world to art	4.15	0.832	0.313	$+0.202$	`deviantart`
OH art to real-world	4.38	0.776	0.301	$+0.184$	`fanart`
OH clipart to real-world	8.51	0.930	0.306	$+0.180$	`fanart`
OH real-world to clipart	8.50	0.961	0.307	$+0.180$	`cartoon`
OH clipart to art	8.79	0.950	0.306	$+0.177$	`logo`
OH art to clipart	8.61	0.961	0.304	$+0.175$	`logo`
DN painting to sketch	6.94	0.908	0.297	$+0.173$	`sketches`
DN sketch to painting	6.93	0.933	0.295	$+0.169$	`diagram`
OH clipart to product	11.68	0.979	0.288	$+0.156$	`illustrating`
OH product to clipart	11.89	0.982	0.277	$+0.141$	`webcomic`
OH real-world to product	5.62	0.817	0.240	$+0.109$	`refurbishment`
OH art to product	9.23	0.932	0.244	$+0.108$	`artworks`
OH product to real-world	5.58	0.844	0.238	$+0.107$	`blogpost`
OH product to art	9.56	0.941	0.231	$+0.091$	`artworks`

Table 7: All twelve Camelyon17 acquisition shifts, ordered by magnitude, with the semantics-free prediction A_{geom} of equation 5 and its residual. For the six highest-magnitude pairs the prediction slightly overestimates the observed AUROC, and the selected prompts are not plausible descriptions of a scanner or a staining protocol.

Pair	$d_{\max}$	A_T	A_{geom}	residual	$\kappa_{\max}$	R	prompt
$h_0 \to h_3$	5.22	0.733	0.648	$+0.085$	0.171	$+0.037$	`orange`
$h_3 \to h_0$	5.22	0.735	0.648	$+0.087$	0.158	$+0.006$	`haram`
$h_0 \to h_1$	8.14	0.825	0.723	$+0.101$	0.191	$+0.037$	`bobble`
$h_3 \to h_1$	8.21	0.859	0.725	$+0.134$	0.196	$+0.061$	`trose`
$h_1 \to h_0$	8.23	0.823	0.725	$+0.097$	0.175	$+0.035$	`bobble`
$h_1 \to h_3$	8.34	0.859	0.728	$+0.131$	0.191	$+0.058$	`trose`
$h_4 \to h_3$	28.57	0.972	0.981	-0.009	0.098	-0.031	`rages`
$h_3 \to h_4$	28.84	0.967	0.982	-0.015	0.095	-0.042	`greenwood`
$h_4 \to h_0$	29.14	0.981	0.983	-0.002	0.099	-0.023	`rages`
$h_0 \to h_4$	30.01	0.969	0.986	-0.016	0.101	-0.021	`rages`
$h_4 \to h_1$	30.24	0.974	0.986	-0.013	0.108	-0.020	`rages`
$h_1 \to h_4$	31.14	0.979	0.988	-0.009	0.102	-0.022	`rages`

not whether language describes the shift. This is why R is defined over a vocabulary. Corollary 1 and Proposition 3 hold for any fixed searched set, and Table 8 shows the continuous limit of that statement: with $r \to 1$, AUROC becomes a function of $d_{\max}$ alone.

Consequence for reading R. R is defined relative to a searched set. Enlarging a nested bank cannot decrease unadjusted $\kappa_{\max}$, but it also raises κ_0, so R need not increase. We therefore report R with the set it was computed on and that set's coverage (Tables 3 and 8); richer searches must repeat their own selection under permutation.

G TOKEN SEARCH VERSUS SENTENCE SEARCH

The methods examined in this paper search over natural-language sentences, not isolated tokens, so we additionally evaluate prompt discrimination and reachability on two sentence banks: a determin-

Table 8: Search capacity beyond one token (95 shifts outside the natural block). r: mean coverage of each search over random directions; κ_0: the same search under permuted labels; sig.: share of shifts whose $\kappa_{\max}$ exceeds the 95th percentile of its own permutation draws.

search	M	r	med. $\kappa_{\max}$	med. κ_0	med. R	Spearman vs. $R_{M=1}$
exhaustive tokens	1	0.103	0.191	0.149	0.045	—
PEZ	1	0.092	0.155	0.133	0.025	0.88
PEZ	2	0.122	0.214	0.169	0.062	0.84
PEZ	4	0.157	0.255	0.200	0.077	0.77
continuous slots	4	0.926	0.993	0.994	undef.	—

istic bank of $39{,}995$ templated expressions, and a bank of $20{,}000$ unique LLM-proposed descriptions generated from a single prompt covering image quality, colour, geometry, medium, acquisition device, medical imaging, generation artefacts, weather, lighting and artistic style. Mean directional coverage values are in Table 3.

Per-shift agreement between the two search modes is high: Spearman 0.90 for the templated bank and 0.86 for the LLM bank over the reported 113-shift analysis sample, and recomputing equation 5 with the corresponding sentence-bank coverage predicts observed best-sentence AUROC at Pearson 0.913 and 0.908. Sentence search recovers slightly more alignment than token search, with a median gain in κ of $+0.020$ across families, and the thresholds move accordingly, from $d_{\max} \geq 17.6$ for AUROC 0.90 under token search to ≥ 15.7 under the templated bank. The mechanism and the taxonomy are unchanged; what changes is the amount of alignment a richer search obtains, as expected when a bank offers better directions. The banks are not nested, so candidate count alone does not guarantee this improvement.

H Cross-encoder transfer in full

A prompt discovered on ViT-L/14 is re-encoded, as the same string, by each transfer model, with $\Delta^{(m)}, \Sigma^{(m)}, d_{\max}^{(m)}$ recomputed from that model's own image tower. There is no re-optimization and no tuning on the transfer models. Differing tokenizers are a feature of the design: we transfer the word, not the token id.

Table 9: Cross-encoder transfer, per encoder. "Transfer excess" is the alignment the re-encoded string attains on the target encoder in excess of that encoder's own coverage reference. Entries are Spearman correlations with discovery-encoder R or best-prompt AUROC, not alignment values.

Transfer encoder	R	AUROC
OpenCLIP ViT-B/32 (LAION-2B)	$+0.658$ $[+0.511, +0.766]$	-0.234 $[-0.422, -0.040]$
OpenCLIP ViT-H/14 (LAION-2B)	$+0.664$ $[+0.517, +0.768]$	-0.197 $[-0.371, -0.015]$
SigLIP ViT-B/16 (WebLI)	$+0.453$ $[+0.267, +0.607]$	-0.252 $[-0.459, -0.092]$

The transfer models use different architectures and pretraining corpora; their dimensions are given in Table 3. SigLIP also uses a sigmoid training objective. Positive correlations occur across these tested models, including SigLIP, where the association is weaker. These comparisons do not isolate which architectural, tokenizer or training-data differences account for that change.

I Text-guided adaptation: outcome and scope

Which outcome. For a shift with source accuracy a_S, unadapted target accuracy a_T and post-adaptation accuracy a_M, the two natural outcomes are the raw gain $a_M - a_T$ and the recovery fraction $(a_M - a_T)/(a_S - a_T)$. Severe shifts both separate more easily and leave more room to recover, so all reported associations are partial correlations controlling for the unadapted drop $a_S - a_T$. The recovery fraction is the quantity of scientific interest, since it asks what share of the damage language repairs, and the raw gain is the more directly measured one. Table 2 reports recovery fractions; the raw-gain sign comparison is summarized in Section 5.3. Recovery is undefined when

$a_S = a_T$ and represents recovery of a loss only when $a_S > a_T$. The analysis sample must therefore be distinguished from registry size.

Table 10: Seed-to-seed reliability of the adaptation outcomes across 89 shifts with a mean of 3.9 seeds each. ICC is the single-measurement intraclass correlation. The reported reliability of the replicate mean belongs to this repeated-run study; it must not be interpreted as the reliability of a three-seed mean in the main analysis.

Outcome	ICC (single)	reliability of replicate mean
zero-shot, raw gain	0.937	0.983
zero-shot, recovery	0.245	0.557
head adaptation, recovery	0.286	0.608

Scope for head adaptation. Our LADS-style feature translation and DDO-style projection experiments are inspired by Dunlap et al. (2023) and Zeng et al. (2025); they are adaptations of those mechanisms to our evaluation, rather than claims to reproduce the original methods on identical benchmarks. We evaluate a separate set of 36 semantic domain shifts from DomainNet, OfficeHome, PACS, Office-31 and VisDA. Corruption ladders, generator comparisons, hospital acquisition and synthetic fingerprints are outside this evaluation's declared scope. The primary specification is partial Spearman correlation between reachability and recovery fraction, controlling for the unadapted accuracy drop, on the descriptor-text arm; the discovered-prompt arm is secondary. At $n = 36$, a Fisher-z approximation gives a detectable correlation of about 0.46 at 80% power for a two-sided 5% test with one controlled variable. This is an approximate planning calculation, not a guarantee for partial Spearman correlation or dependent shifts.

Table 11: Head adaptation on semantic domain shifts, descriptor arm, recovery fraction, controlling for the unadapted drop, $n = 36$. Reachability is positively associated with recovery; best-prompt AUROC is negatively associated. Brackets are bootstrap 95% confidence intervals.

R	best-prompt AUROC
$+0.470$ [$+0.147$, $+0.695$]	-0.545 [-0.760, -0.206]

J SEARCH BASELINES FOR DESCRIPTOR MATCHING

Not every description-selection method uses AUROC. Zeng et al. (2025) maximize CLIP matching between a generated description and a class-level mean-embedding difference. We examine a related global criterion: the maximum raw-space cosine between a word embedding and the collection-level mean difference Δ. This controlled diagnostic uses our vocabulary; it is not a replication of their class-conditioned captioning procedure.

For fixed $\mathcal{T}$, define

$$C_{\max} = \max_{t \in \mathcal{T}} \frac{t^\top \Delta}{\|t\| \|\Delta\|}, \qquad C_0 = \mathbb{E}_\pi \left[\max_{t \in \mathcal{T}} \frac{t^\top \Delta_\pi}{\|t\| \|\Delta_\pi\|} \right].$$

Unlike AUROC, this cosine is invariant to positive scaling of Δ. It still benefits from searching many candidates. We repeat this matching search under permutation to obtain its own baseline C_0, rather than subtracting the whitened-alignment baseline κ_0.

The reported analysis contains 113 shifts and uses the same 32,258-word bank as the token-search experiments.

The median observed match is 0.065, compared with a median permutation baseline of 0.057; the reported median paired excess is 0.008. The bank and its search baseline therefore matter when interpreting the absolute matching score. Enlarging a nested bank can increase both the observed maximum and its baseline, so an unadjusted threshold does not automatically transfer between banks.

The reported excesses are $+0.056$ for grayscale and $+0.030$ for upside-down images, compared with $+0.005$ for blur at radius 16 and $+0.009$ for Camelyon17 $h_1 \to h_4$. These examples agree

Table 12: Global descriptor matching on the reported 113-shift analysis sample. Observed is the largest raw-space cosine to Δ; the search baseline repeats this criterion under permutation. Summaries describe point estimates and do not establish statistical indistinguishability.

Statistic	median	minimum	maximum
observed best match	0.065	0.043	0.128
permutation baseline	0.057	0.053	0.067
excess over baseline	0.008	83 of 113 shifts within 0.02	

with the qualitative distinction in Section 5.1; point estimates alone do not establish significance or guarantee semantic correctness.

Raw-space matching and separation alignment use different geometries. Isotropic Σ makes their signed cosines equal for every direction; with anisotropic Σ, equality may hold in special directions but does not hold in general. Raw cosine is already bounded. The normalization in R additionally expresses alignment above its search baseline relative to the remaining gap to perfect separation. Neither criterion by itself certifies that a description is meaningful.

K Registry, datasets and implementation

Registry. As described in Section 4, the main registry contains 113 source–target shifts in four regimes. The first comprises 27 paired image transformations of FairFace (Kärkkäinen & Joo, 2021), covering colour, geometry, compression, noise, resolution, and two synthetic spatial fingerprints. Blur, JPEG compression, noise, and resampling are evaluated as severity ladders, with each severity level counted as a separate shift. The second comprises 56 GenImage shifts (Zhu et al., 2023) across seven generators: 14 real-to-generated comparisons, corresponding to raw and canonically re-encoded versions for each generator, and 42 ordered cross-generator comparisons using the raw domains. The canonical encoding uses a 256px resolution and JPEG quality 95 to examine the contribution of encoding artefacts. The third comprises all 12 ordered acquisition shifts among four Camelyon17 centres (Bándi et al., 2019; Koh et al., 2021), using balanced tumour and normal patches. The fourth comprises 18 ordered natural-domain pairs: 6 from DomainNet (Peng et al., 2019) and 12 from OfficeHome (Venkateswara et al., 2017), with 20 shared classes within each dataset and class-stratified subsampling. FairFace also supplies the matched demographic-mixture controls described in Appendix C.

Analysis scope. The main registry contains 113 shifts. Main-text analyses report their evaluation sizes separately. The supplementary reference-covariance sensitivity uses a separate 163-shift sample. These are analysis counts, not alternative totals for the main registry. The natural-domain pairs are included in these analyses and are also reported separately in the family-level tables. Head adaptation uses a separate evaluation sample of 36 semantic domain shifts spanning DomainNet, OfficeHome, PACS (Li et al., 2017), Office-31 (Saenko et al., 2010), and VisDA (Peng et al., 2017). Table 2 reports the sample size for each external evaluation, and Appendix I details the adaptation protocols and scope.

Encoders. Prompt search uses only OpenAI CLIP ViT-L/14 (Radford et al., 2021). Transfer is evaluated on OpenCLIP ViT-B/32 and ViT-H/14 pretrained on LAION-2B (Schuhmann et al., 2022; Cherti et al., 2023) and on SigLIP ViT-B/16 trained with a sigmoid loss on WebLI (Zhai et al., 2023). No search, tuning or model selection is performed on any transfer encoder.

Vocabulary and sentence banks. 45,307 readable token ids reduced to 32,258 by requiring that tokenization of the decoded string round-trips to the original id. Sentence banks: 39,995 templated expressions, deterministic, and 20,000 unique LLM-proposed descriptions from a single generation prompt, released with the code.

Estimation. Embeddings are ℓ_2-normalized. Covariances use Ledoit–Wolf shrinkage. Confidence intervals are bootstrap over shifts with 4000 resamples unless stated otherwise. Predictive-validity

analyses average over three seeds. Search, model selection and final evaluation use disjoint splits. Every heavy step is content-hash cached. All experiments run on a single 32 GB GPU.

L SENSITIVITY TO THE REFERENCE COVARIANCE

Status and scope. The headline figures, counts and thresholds retain the registry-pooled reference, for which $r = 0.103$. The purpose of this check is to assess the transport of the geometric prediction across reference pools; it does not assume that the numerical value of r is covariance-invariant. The sensitivity sweep was run on an enlarged registry of 163 shifts (56 semantic, 107 other); its correlations are therefore not directly comparable with the $n = 113$ figure quoted in the main text, and the two are kept in separate blocks below.

One coverage value per reference pool. For a reference pool $\mathcal{C}$ with covariance $\bar{\Sigma}_{\mathcal{C}}$, keep the encoder and candidate bank $\mathcal{T}$ fixed and define

$$u_{\mathcal{C},t} = \frac{\bar{\Sigma}_{\mathcal{C}}^{1/2} t}{\|\bar{\Sigma}_{\mathcal{C}}^{1/2} t\|}, \qquad r_{\mathcal{C}} = \mathbb{E}_{v \sim \mathrm{Unif}(\mathcal{S}^{p-1})} \left[\max_{t \in \mathcal{T}} |u_{\mathcal{C},t}^{\top} v| \right].$$

Within each reference-pool condition, this single coverage value is held fixed across evaluated shifts and substituted into $A_{\mathrm{geom}} = \Phi(d_{\max} r_{\mathcal{C}} / \sqrt{2})$. Here the subscript identifies the sensitivity condition; it introduces no shift-specific fitted coverage value. Each pool is a set of source-side images used only to estimate $\bar{\Sigma}_{\mathcal{C}}$; the coverage is computed from it and frozen before any shift is evaluated. We verified that the reported coverage values are produced by the normalised-direction formula above: recomputing r directly from $u_{\mathcal{C},t}$ agrees with the implementation to 1.5×10^{-8}.

Table 13: Reference-covariance sensitivity, with the reported analysis samples separated. Panel (a) reproduces the main-text result. Panel (b) compares reference pools on one common supplementary sample; its groups contain 107 other shifts and 56 semantic-domain shifts. Comparisons between panels do not isolate the effect of the reference pool. MC is the spread of r over eight Monte Carlo seeds; Err. is median absolute AUROC-prediction error.

(a) Main-text analysis, reported $n = 113$

Reference	n	r	Pearson	Err.
Registry-pooled	113	0.1030	0.902	0.081

(b) Supplementary sensitivity sample, $n = 163$

				Other ($n = 107$)		All ($n = 163$)		Semantic
Reference pool	n_{cal}	r	MC	Pearson	Err.	Pearson	Err.	Err.
Camelyon17 source	3200	0.1090	±0.0005	0.913	0.070	0.827	0.086	0.165
All sources pooled	4000	0.1136	±0.0003	0.915	0.067	0.831	0.081	0.156
FairFace source	801	0.1174	±0.0002	0.916	0.063	0.835	0.076	0.149
DomainNet real	2990	0.1185	±0.0002	0.917	0.062	0.836	0.073	0.147
GenImage real	4000	0.1250	±0.0002	0.919	0.056	0.842	0.071	0.135

Within the supplementary sample, Pearson correlation on the 107 other shifts ranges from 0.913 to 0.919, with median absolute error from 0.070 to 0.056. On all 163 shifts, the corresponding ranges are 0.827 to 0.842 and 0.086 to 0.071. The semantic-domain group has larger median errors, between 0.135 and 0.165. These results cannot be compared directly with the main-text 0.902 correlation to infer an improvement from changing the reference pool, because the evaluated samples differ.

Across the five alternative pools the coverage ranges from 0.109 to 0.125, a relative spread of about 15%, and all five exceed the registry-pooled value of 0.103. The Monte Carlo uncertainty is at least an order of magnitude smaller than the between-pool spread, so what varies is the reference covariance, not the sampling of directions. The FairFace pool is limited to 801 distinct images because the 27 paired transformations share their source images; its covariance therefore leans more heavily on the shrinkage term than the others, and we report it with that caveat rather than excluding it.

The ranking is invariant to the choice of pool. Substituting a different coverage value rescales $d_{\max}$ inside a monotone $\Phi(\cdot)$, so the predicted *ordering* of shifts cannot change. In the supplementary analysis, the Spearman correlation between the geometric prediction and observed best-prompt AUROC is 0.877 for all five pools, identical to twelve decimal places. Only the absolute calibration moves, and it moves little: Pearson 0.827 to 0.842, median absolute error 0.086 to 0.071. The reference pool must accompany any quoted numerical value of r and any threshold derived from it; it does not affect which shifts the mechanism identifies as separable on magnitude alone.

Semantic-domain shifts have a positive median residual. For each reference pool, the reported median signed residual on the 56 semantic-domain shifts equals the median absolute residual, between 0.135 and 0.165. This supports a positive median departure from the prediction. Equality of these medians does not establish that every, or nearly every, individual residual is positive. Mean directional coverage is an average reference, not a floor that semantic shifts are guaranteed to exceed.

Thresholds depend on the chosen coverage. For each reference condition, the magnitude threshold is

$$d_\alpha(r) = \frac{\sqrt{2}\,\Phi^{-1}(\alpha)}{r}.$$

Table 14 evaluates this expression using the displayed coverage values. Exceedance counts use the same supplementary sample of 163 shifts in every row, including the registry-pooled reference. They are not counts for the main registry of 113 shifts.

Table 14: Magnitude thresholds recomputed from the displayed coverage values. The last two columns count shifts, out of the 163 of the supplementary sensitivity sample, whose geometric prediction reaches the indicated AUROC. All rows use this same denominator. These are thresholds for the prediction, not guarantees on observed AUROC.

Reference pool	r	$d_{0.75}$	$d_{0.90}$	$d_{0.95}$	$d_{0.99}$	$N_{0.75}$	$N_{0.90}$
Registry-pooled	0.1030	9.3	17.6	22.6	31.9	72	43
Camelyon17 source	0.1090	8.8	16.6	21.3	30.2	74	47
All sources pooled	0.1136	8.4	16.0	20.5	29.0	77	50
FairFace source	0.1174	8.1	15.4	19.8	28.0	82	51
DomainNet real	0.1185	8.0	15.3	19.6	27.8	83	51
GenImage real	0.1250	7.6	14.5	18.6	26.3	85	52

Every alternative pool yields larger mean directional coverage than the registry-pooled reference and therefore lower prediction thresholds. On this fixed supplementary sample, the corresponding exceedance counts cannot decrease. This is a property of the geometric prediction and does not establish a lower bound for measured Prompt AUROC.

Changing the reference pool changes a geometric prediction, not the measured AUROC of a prompt. The check does not extend the Jensen statement of Proposition 3 to pairs with a different covariance.

M CORRECT DESCRIPTIONS VALIDATE THE REACHABILITY TAXONOMY

Validation with known descriptions. Explicitly supplied correct descriptions reproduce the distinction identified by R: strong alignment for positive controls and low alignment for corruption and fingerprint shifts. On 21 paired transformations, we compare a canonical description of the applied operation, specified before evaluating the scores, and a selected paraphrase with the word returned by the existing search. This directly tests whether low reachability reflects omission of the correct description from the candidate bank, complementing the sentence-bank experiment in Appendix G.

Correct descriptions recover the positive controls. Canonical descriptions attain $\kappa = 0.243$ for grayscale, 0.304 for upside-down images, 0.226 for sepia and 0.219 for watermarking, against baselines around 0.10 for these examples. Their strong alignment recovers the positive-control side of the taxonomy: these visual changes are accessible through descriptions of the applied operations in the frozen encoder.

Table 15: **The correct-description control supports the reachability taxonomy.** The low-alignment corruption pattern persists when correct descriptions are supplied, alongside successful positive controls (values in the text). Exceedances compare point estimates with the reported bank null.

Candidate	Median κ	Above bank baseline
All 21 *transformations*		
Discovered word	0.098	$18/21$
Canonical description	0.060	$9/21$
Selected paraphrase	0.052	$10/21$
Bank null	0.075	—
15 *corruption and fingerprint variants*		
Canonical description	$\mathbf{0.054}$	$\mathbf{4/15}$
Bank null	0.074	—

Low reachability survives the correct-description control. For the 15 variants involving blur, JPEG compression, noise, resampling and fingerprints, median canonical alignment is 0.054, below the median bank baseline of 0.074; only 4 exceed their baseline, leaving 11 at or below it. Thus the tested canonical descriptions remain weakly aligned with most of these corruption and fingerprint shifts. These comparisons are to the reported bank baseline, not to a separately calibrated null for a single fixed description. Across all 21 transformations, canonical descriptions and selected paraphrases also remain below the bank null at the median (Table 15). These results support the distinction between positive controls and the tested corruptions. They show that supplying these particular correct descriptions does not remove the low-alignment pattern; they do not establish an upper bound over all possible descriptions.

Search finds effective operational directions. As a complementary result, the discovered word achieves greater alignment than the canonical description on 15 **of** 21 **transformations**. The search thus finds text directions that capture more of the standardized visual displacement than the supplied literal description in most tested cases. This supports interpreting the discovered prompt as an operational direction, with descriptive fidelity assessed separately.

N Empirical check of the Gaussian AUROC relation

Proposition 2 connects shift magnitude and text alignment to AUROC under a Gaussian model for image–text scores. We test how accurately this relation describes real scores, which need not follow Gaussian distributions.

For each of 22,600 (shift, prompt) pairs, we compare the prediction

$$A_{\mathrm{pred}}(t) = \Phi\left(\frac{d_{\max}\kappa(t)}{\sqrt{2}}\right) = \Phi\left(\frac{d'(t)}{\sqrt{2}}\right)$$

with empirical AUROC, computed by ranking images using their matching scores $t^\top z(x)$. The prediction uses the mean difference and covariance defined in Section 3.1; no nonlinear mapping is fitted to the observed AUROC values. Both quantities use the same source–target ordering, with signed alignment $\kappa(t)$. Although $d_{\max}\kappa(t) = d'(t)$ holds by construction, its connection to AUROC through Φ is a prediction that can be checked against the data.

Figure 4 shows close agreement with the identity line. The predictions explain 99.5% of the variance in observed AUROC ($R^2 = 0.995$), supporting the decomposition as an accurate approximation across the evaluated pairs. This check assesses the AUROC relation; it does not establish that every underlying score distribution is Gaussian.

O Reproducibility

Proof sketches appear in Section 3 and full proofs in Appendix A. The registry, the analysis bundle containing per-shift statistics, permutation nulls and bootstrap draws, the two sentence banks, the

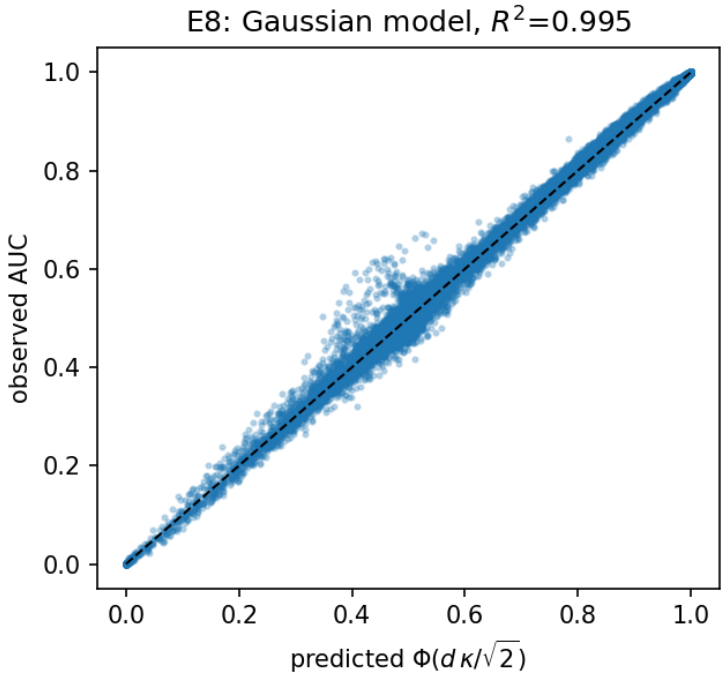


Figure 4: **The Gaussian relation closely predicts Prompt AUROC.** Each point is one of 22,600 (shift, prompt) pairs. Predicted and observed AUROC closely follow the dashed identity line, with $R^2 = 0.995$ and no fitted nonlinear mapping. The horizontal-axis symbol d denotes $d_{\max}$.

blinded semantic-audit items, the estimator definitions and split protocol, and the notebooks that produce every table and figure in this paper are released with the code. All encoders and datasets are public. Criteria for the predictive-validity analyses, including the scope declaration and primary specification for head adaptation in Appendix I, were fixed before the corresponding runs and are reproduced in the released bundle.