SEEING IS NOT ADDRESSING: AUDITING LINGUISTIC ACCESS TO FROZEN VISUAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual distinctions are often finer than those reflected in linguistic conceptualization. Vision–language models exhibit a similar asymmetry: a distinction can remain discriminable in frozen image geometry while being weakly addressable through the native text interface. We study this gap by separating *visual discriminability* from *linguistic addressability* in text-to-image retrieval. Using FactorAtlas, a fully crossed testbed of 23,040 images spanning shape, hue, pattern, and nuisance variation, we compare both readouts on held-out images of the same distinctions. We then derive image-side contrasts that separate each value from its alternatives for matched visual grounding, and test whether this reduces the native-text access gap across factors and models. Direction-specific and visual-absence controls tie these gains to the relevant visual contrast; the gains persist after global alignment and extend to compositional retrieval and natural images. Together, these results show that visual discriminability and linguistic addressability need not coincide, and that matched visual grounding can probe and reduce the resulting access gap.

1 INTRODUCTION

Visual representations often preserve finer distinctions than are reflected in linguistic conceptualization (Liao et al., 2024). Such distinctions become addressable through language when relevant words are linked to their visual referents, and these associations are learned through repeated co-occurrence (Yu & Smith, 2007; Smith & Yu, 2008). Contrastively trained vision–language models have a similar learning structure, acquiring cross-modal correspondences from large-scale paired image–text data (Radford et al., 2021; Zhai et al., 2023). The paired text, however, need not describe every visual distinction present in the image. Accordingly, in the resulting shared embedding space, some of these distinctions can remain **discriminable in the image representation while being only weakly addressable** through the model's text interface.

In **text-to-image retrieval**, this asymmetry motivates separating two aspects of performance: whether image geometry supports the relevant visual distinction, and how well the native text query can access it. Related work shows that representations within each modality retain structured information (Koishigarina et al., 2026), with compositional and ordinal structure in image embeddings (Berasi et al., 2025; Sonthalia et al., 2026). Meanwhile, cross-modal alignment and mismatch have been studied at broader representational and compositional scales (Liang et al., 2022; Kamath et al., 2023; Koishigarina et al., 2026). We instead take **each named visual distinction as the unit of analysis**, asking how the image-side structure underlying its discriminability relates to native linguistic access.

We operationalize this question using **FactorAtlas**, a fully crossed testbed of 23,040 images spanning 8 shapes, 12 hues, 10 patterns, and 24 nuisance realizations. For each named factor value, we compare two readouts under held-out conditions: ***visual discriminability***—how reliably each value can be distinguished from its alternatives in frozen image geometry—and ***linguistic addressability***—how well the corresponding text query retrieves images exhibiting that value. We then derive a target-versus-rest visual contrast for each value and use its direction to ground the native text query. We call this ***matched visual grounding*** and test whether it improves distinction-specific linguistic access.

Across our experiments, visual discriminability and linguistic addressability diverge, while matched visual grounding yields held-out access gains across factors and models. These gains are direction-specific and weaken as the underlying visual contrast is attenuated. Broad alignment reduces part of the mismatch but leaves residual value-specific access gaps that matched grounding further reduces.

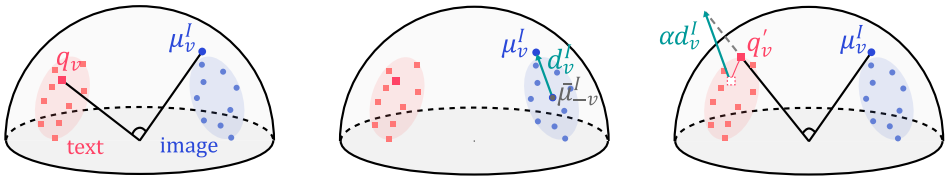
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

These access gains, in turn, support exact compositional retrieval under held-out conditions, while the same value-specific intervention can be applied selectively and generalizes to natural images. Together, these findings show that linguistic access can lag behind visual structure already present in frozen representations, suggesting a way to expose and use under-accessed visual distinctions.

2 From Visual Distinctions to Linguistic Access

2.1 Visual Discriminability and Linguistic Addressability

Our analysis separates two questions for the same named visual property: whether its distinction remains discriminable in frozen image geometry and how well the native text query can access it. We term these *visual discriminability* and *linguistic addressability*, measured from image-derived evidence and the native text query, respectively. When they diverge, we use the held-out improvement obtained by grounding the native text query with the corresponding image-side structure, which we call *matched access gain*, as an interventional readout of how much the access gap can be reduced.



Shared Embedding Space | Value-specific Visual Contrast | Matched Visual Grounding

Figure 1: **From visual discriminability to linguistic access.** A visual distinction may remain discriminable in frozen image geometry yet weakly accessed by its native text query; matched visual grounding connects the two through a value-specific visual contrast.

2.2 Matched Visual Grounding via Value-Specific Visual Contrasts

We instantiate this concept by representing each visual value v with an image-side direction that distinguishes it from the other values in the same factor vocabulary. We use this direction to ground the corresponding native text query, a procedure we call *matched visual grounding* (Fig. 1).

Let $I(x) \in \mathbb{R}^d$ denote the frozen image embedding. For a factor with value set $\mathcal{V}$ and value $v \in \mathcal{V}$, let C_v denote a visual support set for v. We define the target and average-rest image prototypes as

$$\mu_v^I = \mathrm{norm}\left(\frac{1}{|C_v|}\sum_{x\in C_v}\mathrm{norm}(I(x))\right), \qquad \bar{\mu}_{-v}^I = \frac{1}{|\mathcal{V}|-1}\sum_{u\neq v}\mu_u^I.$$

We leave $\bar{\mu}_{-v}^I$ unnormalized so that its score under a unit query equals the average score of the remaining value prototypes. Their difference defines the matched visual contrast and its unit direction:

$$\Delta_v^I = \mu_v^I - \bar{\mu}_{-v}^I, \qquad d_v^I = \frac{\Delta_v^I}{\|\Delta_v^I\|}.$$

Given the native text query q_v, we ground it along this direction:

$$q_v'(\alpha) = \mathrm{norm}\big(q_v + \alpha d_v^I\big), \qquad \alpha \geq 0. \tag{1}$$

The parameter α controls the grounding strength. Comparing retrieval with $q_v'(\alpha)$ against the native query q_v yields the *matched access gain* introduced above.

2.3 Geometric Guarantee

We derive a guarantee in terms of the target-versus-average-rest similarity margin.

$$M_v(q) = q^\top \Delta_v^I, \qquad c = q_v^\top d_v^I.$$

Proposition 1. Matched similarity-margin guarantee

For every $\alpha \geq 0$ for which $q_v'(\alpha)$ is defined,

$$M_v(q_v'(\alpha)) \geq M_v(q_v).$$

The grounded similarity margin has the closed form

$$M_v(q'_v(\alpha)) = \|\Delta_v^I\| \frac{c+\alpha}{\sqrt{1+\alpha^2+2\alpha c}}. \tag{2}$$

For $|c| < 1$, $M_v(q'_v(\alpha))$ increases strictly with α. When the initial margin is negative ($-1 < c < 0$), it crosses zero at

$$\alpha^\star = -c = -q_v^\top d_v^I.$$

The guarantee is local to the matched target-versus-average-rest margin; whether it translates into held-out retrieval gains and broader retrieval behavior is evaluated empirically below. The full proof, including endpoint cases, is given in Appendix A.

3 Diagnosing and Reducing Distinction-Level Access Gaps

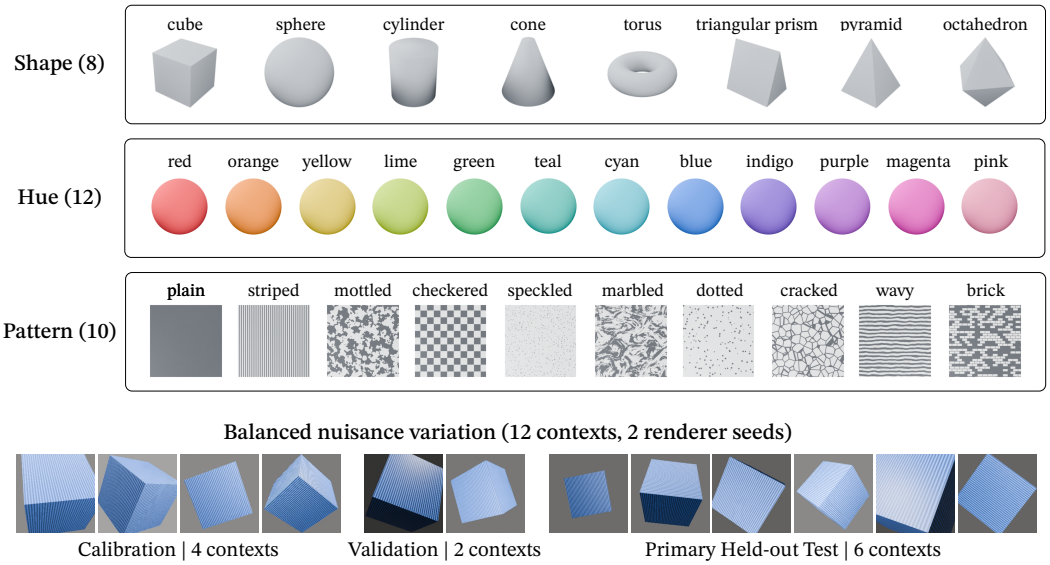


Figure 2: **FactorAtlas and its primary held-out protocol.** The testbed factorially combines 8 shapes, 12 hues, and 10 surface patterns under balanced nuisance variation. Disjoint nuisance contexts are used for calibration, validation, and held-out testing, while the semantic factors remain fully crossed and balanced.

3.1 FactorAtlas and Evaluation Protocol

FactorAtlas is a procedurally rendered visual testbed designed to compare visual discriminability and linguistic addressability under controlled variation. It contains 23,040 images spanning 8 shapes, 12 hues, 10 surface patterns, and 24 nuisance realizations. Each realization is jointly determined by a context and renderer seed, varying object pose, viewpoint, illumination, background, and material appearance while preserving the underlying semantic factors. This lets the same semantic distinctions recur across varied visual conditions.

For each semantic factor, we evaluate all of its values jointly. When one factor is the target, the other factors are fully crossed and balanced, so every target value appears equally often with every combination of the remaining factors. This rules out simple cross-factor frequency shortcuts, preventing the readout from relying on correlations between the target value and the other factors.

Our primary protocol holds out nuisance contexts, using four for calibration, two for validation, and six for testing, with both renderer seeds of each context assigned to the same split. For each factor value, the native text query is formed by averaging and renormalizing the embeddings of three fixed prompt templates specific to its factor. Calibration defines the visual support sets in Section 2.2, while validation selects the factor-level grounding strength α. All choices are fixed before evaluation on held-out contexts, so the readouts are tested under nuisance conditions not used to construct or tune them. Full construction and evaluation details are provided in Appendix B.

3.2 Discriminability, Addressability, and Matched Access

Discriminability versus addressability. We first compare *visual discriminability* and *linguistic addressability* on the same held-out gallery. For each factor value, the former is measured with an image-derived query built from calibration images, whereas the latter uses the corresponding native text query. Thus, the two readouts differ only in how the query is obtained.

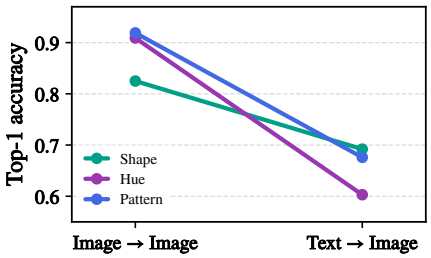


Figure 3: **Factor-value top-1 readout gap.** Visual discriminability exceeds native linguistic addressability across all factors.

As shown in Fig. 3, image-based queries identify the correct value more reliably than native text queries across shape, hue, and pattern. This reveals the distinction-level asymmetry motivating our audit, as the relevant visual distinctions remain readily discriminable in image geometry even when their linguistic handles provide weaker access.

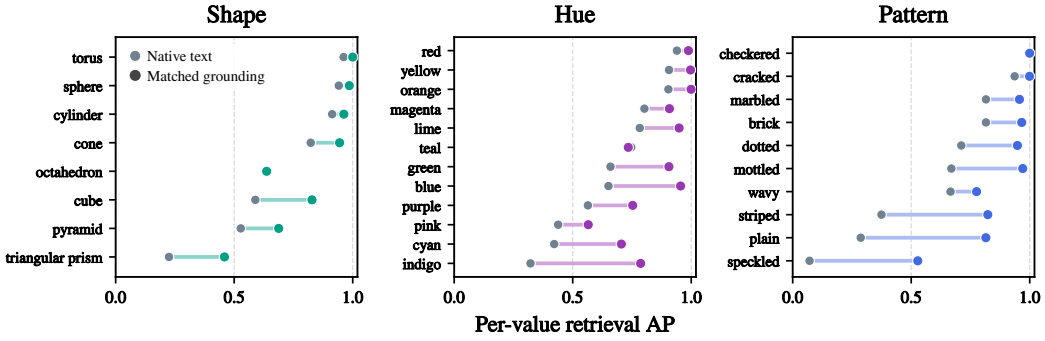


Figure 4: **Matched access gain across individual factor values.** Native text retrieval and matched visual grounding are compared for each value of shape, hue, and pattern on held-out images.

Matched access gain across factors. We next ask whether grounding the text query with value-specific image-side information can reduce this access gap. Using the matched visual grounding procedure in Section 2.2, we estimate each visual direction from calibration images and select the factor-level strength α on validation before evaluating the held-out gallery.

Figure 4 shows the result for every value of shape, hue, and pattern. Matched grounding improves average precision for 29 of the 30 values across all three factor vocabularies. Although the gains vary in magnitude across values, their broadly positive pattern shows that matched image-side information can substantially improve linguistic access on held-out images.

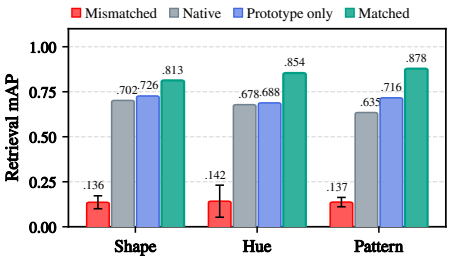


Figure 5: **Visual contrast matters.** Matched contrast outperforms prototype-only and mismatched-direction controls.

Specificity of the visual contrast. The observed access gain raises a natural question: does it depend on the matched target-versus-rest visual contrast, or is moving the text query toward the target visual prototype sufficient? We test the latter with *prototype-only* grounding,

$$q_v^+(\lambda) = \mathrm{norm}\left((1-\lambda)q_v + \lambda\mu_v^I\right),$$

with λ selected on validation and fixed before test.

On SigLIP2 Base, prototype-only grounding yields modest gains over the native query but remains consistently below matched grounding (Fig. 5). Moreover, assigning each target a nonmatching visual contrast sharply degrades retrieval, even when averaged over all nonmatching assignments. Together, these controls isolate what drives the access gain. Rather than reflecting generic attraction toward target images, the improvement depends specifically on the matched visual direction that separates the target from its alternatives. This interpretation is also consistent with the geometric guarantee, which applies to movement along the matched target-versus-rest visual contrast.

Prompt and lexical controls. The native query already averages three fixed prompt embeddings, reducing its dependence on any single wording. We nevertheless test whether the observed access gain depends on this ensemble or on particular prompt formulations. In the pattern setting, native mAP ranges from 0.501 to 0.629 across five fixed templates, whereas matched grounding improves performance for every template and remains between 0.872 and 0.881 (Table 1). Thus, the access gain is not tied to the three-template ensemble or to any one tested formulation.

Table 1: **Prompt robustness in the pattern setting.** Matched grounding improves across all tested query forms.

Query form	Native	Matched
`{v}`	.629	.876
`a {v} surface`	.549	.874
`an object with a {v} pattern`	.577	.881
`a surface that is {v}`	.528	.874
`a textured {v} surface`	.501	.872

We also ask whether lexical substitution alone can close the access gap. For each pattern, we form an equal-weight ensemble of the canonical term and two predetermined lexical alternatives. This lexical ensemble reaches 0.512 mAP, compared with 0.635 for the original native ensemble, and therefore does not reproduce the improvement from matched grounding. Together, these controls indicate that the observed access gains are not explained by prompt formulation or lexical alternatives. The prompt-robustness result also holds for hue and shape, with full results reported in Appendix C.3.

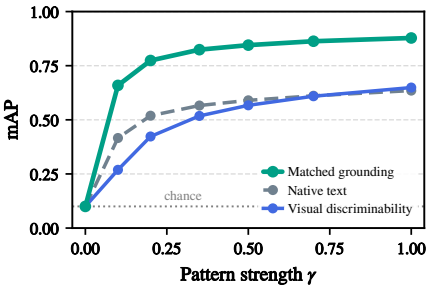


Figure 6: **Matched access gain under attenuated visual evidence.** Matched access gain remains substantial as the pattern signal weakens and disappears when the visual distinction is removed.

Dependence on visual evidence. We further extend the analysis by varying the strength of the underlying visual distinction for the pattern factor, where the visual signal can be directly controlled. Each patterned image is progressively blended with its nuisance-matched plain counterpart, producing seven signal strengths from $\gamma = 1.0$ to $\gamma = 0.0$. At each strength, visual directions are re-estimated from calibration images and the grounding strength α is selected on validation.

As shown in Fig. 6, as the visual signal weakens, the available image-side contrast also decreases, and retrieval performance under matched grounding declines accordingly. Even at $\gamma = 0.1$, however, retrieval with matched-grounded queries reaches 0.659 mAP, compared with 0.416 for native-text retrieval, while the estimated visual directions retain a mean cosine of 0.825 with their full-signal counterparts. At $\gamma = 0$, all pattern variants collapse to nuisance-matched plain images, and grounding provides no gain over chance. These results indicate that matched access gains depend on the presence of the distinction-specific visual contrast itself. Full results are provided in Appendix C.4.

3.3 Robustness Across Models and Held-Out Structure

Table 2: **Matched access gain across factors, backbones, and held-out protocols.** Each cell reports native → **matched** full-gallery mAP and factor-value top-1 accuracy. Per-backbone rows use the context holdout; the lower block reports seven-backbone macro results across additional protocols.

	Shape		Hue		Pattern	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
Context holdout, per backbone						
CLIP family						
CLIP ViT-B/16	.519 → **.723**	.543 → **.690**	.649 → **.809**	.583 → **.836**	.361 → **.830**	.425 → **.844**
EVA02-B/16	.563 → **.804**	.645 → **.767**	.586 → **.796**	.656 → **.841**	.515 → **.843**	.544 → **.884**
FG-CLIP Base	.599 → **.740**	.585 → **.706**	.678 → **.892**	.720 → **.901**	.555 → **.848**	.496 → **.882**
SigLIP family						
SigLIP SO400M	.730 → **.850**	.741 → **.835**	.667 → **.825**	.627 → **.856**	.663 → **.891**	.718 → **.911**
SigLIP2 Base	.702 → **.813**	.692 → **.789**	.678 → **.854**	.603 → **.882**	.635 → **.878**	.676 → **.911**
SigLIP2 Large	.744 → **.837**	.770 → **.830**	.590 → **.736**	.573 → **.756**	.620 → **.875**	.653 → **.927**
SigLIP2 SO400M	.764 → **.851**	.774 → **.838**	.488 → **.592**	.432 → **.620**	.621 → **.867**	.676 → **.914**
Seven-backbone macro, per protocol						
Context holdout	.660 → **.803**	.678 → **.779**	.619 → **.786**	.599 → **.813**	.567 → **.862**	.598 → **.896**
Unseen combinations	.667 → **.814**	.690 → **.798**	.618 → **.801**	.592 → **.839**	.594 → **.876**	.619 → **.906**
24 nuisance partitions	.663 → **.809**	.687 → **.791**	.621 → **.788**	.596 → **.818**	.570 → **.862**	.603 → **.898**

Across backbones and factors. We broaden the context-holdout audit to cover shape, hue, and pattern across seven frozen vision–language backbones. Matched visual grounding improves both mAP and factor-value top-1 for all 21 backbone–factor combinations. Though the gains vary in magnitude across models and factors, the improvement is consistent in every case.

Unseen semantic combinations. We also test a different kind of holdout based on semantic combinations. Here, individual factor values remain present in calibration, validation, and test, but some combinations of the two non-target factors appear in only one split. For example, when pattern is the target, evaluation uses hue–shape combinations that were not seen when the pattern directions were built or tuned. By design, all nuisance contexts and renderer seeds appear in every split, so the evaluation isolates unseen semantic combinations rather than unseen nuisance conditions. Matched visual grounding improves both metrics for all 21 backbone–factor combinations under this protocol. The access gains therefore do not depend on memorizing particular semantic combinations and generalize to new combinations of otherwise familiar factor values.

Robustness to nuisance partitions. The context holdout used in our main analysis shows broad gains across backbone–factor combinations, but a single nuisance split could still be unusually favorable. We therefore repeat the audit across 24 balanced partitions of the joint context–renderer realizations, assigning eight groups to calibration, four to validation, and twelve to test. Visual directions are re-estimated separately for every partition. Matched visual grounding improves both mAP and factor-value top-1 across all backbone–factor–partition configurations, with little variation across partitions. This consistency indicates that the observed access gains are robust to how the available nuisance realizations are divided among calibration, validation, and test.

Taken together, these results show that matched access gains are not confined to a particular backbone–factor setting or held-out protocol. Complete backbone-, protocol-, and value-resolved results are reported in Appendix D.

3.4 Global Alignment and Residual Access Gaps

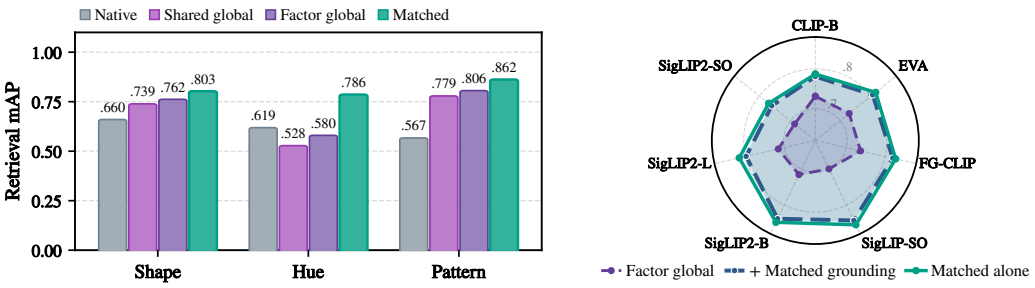


Figure 7: **Global alignment leaves residual access gaps.** (**Left**) Factor-level macro-mAP under the context holdout. (**Right**) Factor-global alignment before and after matched grounding across seven backbones, with native-space matched grounding shown for reference.

We close this analysis by asking whether the access gaps identified for individual visual distinctions persist after a broader correction to the text–image representation geometry. Such broad approaches are common in prior work on fine-grained cross-modal alignment, where correspondence is improved through shared transformations applied across many distinctions rather than distinction-specific corrections. We test this using full-rank linear maps on the text representation, similar in spirit to prior work on linear cross-modal alignment (Koishigarina et al., 2026; Moayeri et al., 2023).

For a native query q, the aligned query is

$$q^G = \mathrm{norm}(Aq), \tag{3}$$

where $A \in \mathbb{R}^{d\times d}$ is a bias-free, identity-initialized map and the pretrained encoders remain fixed. We learn either a separate map A_f for each factor or a single map A_{shared} across all three factors. Both map types are trained on calibration image–caption pairs with factor-replacement hard negatives and selected on validation before test evaluation. We use these maps as expressive baselines for broad cross-modal alignment.

Global alignment improves overall retrieval but leaves substantial residual access gaps across factors. Under the context holdout, seven-backbone macro-mAP rises from 0.616 natively to 0.682 with shared-global alignment and 0.716 with factor-global alignment. Both maps improve pattern and shape mAP but reduce hue mAP, whereas matched visual grounding achieves the highest mAP on all three factors (Fig. 7, left).

We next test whether the value-specific visual structure remains useful even after broad alignment. To examine this, we fix each global map and use matched visual directions estimated from calibration images to ground the aligned queries, with the grounding strength reselected on validation. Factor-global alignment provides the stronger broad correction, yet matched visual grounding further raises macro-mAP from 0.716 to 0.806 and yields additional access gains across all seven backbones (Fig. 7, right). The same overall trend holds across additional global alignment baselines and broader held-out evaluations (Appendix E). Thus, even after a broad correction to text–image alignment, matched grounding that targets the value-specific visual contrast can further reduce the remaining access gaps.

4 Consequences and Generalization of Improved Access

Having established gaps between visual discriminability and linguistic addressability, together with consistent gains from matched visual grounding, we next turn to the broader consequences of improved access and its generalization beyond the controlled setting.

4.1 From Factor Access to Compositional Retrieval

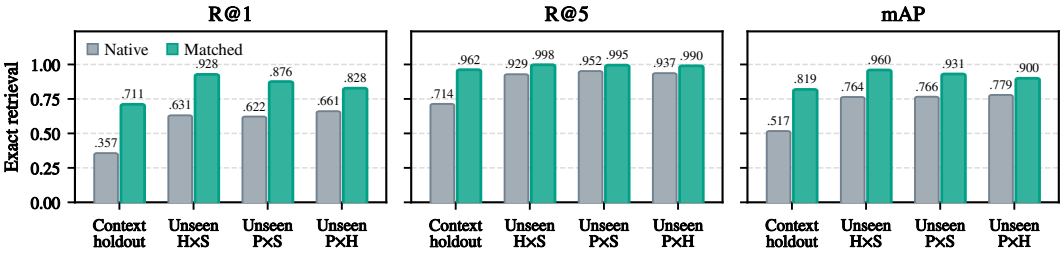


Figure 8: **Factor-level access gains support exact compositional retrieval.** Native and matched queries are compared under the context holdout and three unseen-combination protocols. Matched visual grounding improves exact R@1, R@5, and mAP across all held-out conditions. H×S, P×S, and P×H denote held-out hue–shape, pattern–shape, and pattern–hue combinations.

Matched grounding is local to the visual contrast associated with each factor value, so improved access to an individual factor value does not by itself establish that the resulting queries remain compatible when several distinctions must be satisfied jointly. We therefore evaluate exact composition, where a query specifies one pattern, one hue, and one shape and the correct image must match all three values simultaneously. The exact target must outrank near-miss images that differ in one or more factors.

Following prior work showing the utility of factorized inference for compositional retrieval (Alshehri et al., 2026), we score each requested factor value within its own vocabulary before combining the resulting factor-level evidence. Native and matched queries use the same composition rule, so the comparison isolates the effect of improved factor-level access rather than differences in the composition operator. Specifically, for each factor f, $p_f(v \mid x)$ denotes the softmax-normalized similarity of image x to value v within that factor vocabulary, using a shared temperature τ.

For each factor f, we define

$$p_f(v \mid x) = \frac{\exp(\langle q_{f,v}, x\rangle/\tau)}{\sum_{u\in V_f} \exp(\langle q_{f,u}, x\rangle/\tau)}, \qquad S(y,x) = \sum_{f\in\{\text{shape,hue,pattern}\}} w_f \log p_f(y_f \mid x).$$

Each factor-specific grounding strength α_f is chosen on factor-level validation and then fixed. The shared temperature τ and factor weights w_f are chosen on compositional validation to balance how factor-level evidence is combined before test evaluation.

In this setting, matched visual grounding substantially improves exact retrieval across all held-out conditions and metrics (Fig. 8). Under the context holdout, R@1 rises from 0.357 to 0.711, R@5 from 0.714 to 0.962, and mAP from 0.517 to 0.819. The gains also persist under unseen-combination protocols, where individual factor values remain familiar but their test-time combinations are new. The factor-level access gains therefore carry over to exact compositional retrieval, enabling new multi-factor combinations to be retrieved jointly rather than only improving isolated factor readouts.

4.2 Validation-Guided Selective Grounding

Because access gains vary across individual factor values, the value-level audit can also guide where grounding should be applied. Some native queries are already effective, while others show clear validation gains from matched grounding. We therefore retain grounding only for values whose validation AP gain remains positive under bootstrap resampling of nuisance groups; otherwise, we keep the native query. A single grounding strength is selected per factor on validation, and all decisions are fixed before test evaluation.

Table 3: **Validation-guided selective grounding.** Test mAP across the context and unseen-combination protocols.

Policy	Test mAP
Native	.6797
Ground every value	.8220
Validation-guided	**.8245**

On SigLIP2 Base, across the context-holdout and unseen hue–shape-combination protocols, grounding every value raises mean test mAP from 0.6797 to 0.8220, while the validation-guided policy reaches 0.8245. Treating each value-specific visual distinction separately enables more targeted grounding, applying it only where validation supports a gain while yielding a small additional improvement.

4.3 Matched Access Gains Beyond FactorAtlas

Table 4: **Natural-image transfer settings.** We evaluate texture, garment pattern and length, and material distinctions using dataset-specific splits that separate evaluation examples from those used to estimate and tune the grounding.

Dataset	Target distinction	Held-out setting
DTD (Cimpoi et al., 2014)	Texture	Official partitions
Fashionpedia (Jia et al., 2020)	Pattern / garment length	Category-held-out / image-disjoint
COCO-Facet (Li et al., 2026)	Material	Image-disjoint
UT-Zappos (Yu & Grauman, 2014)	Material	Product-disjoint

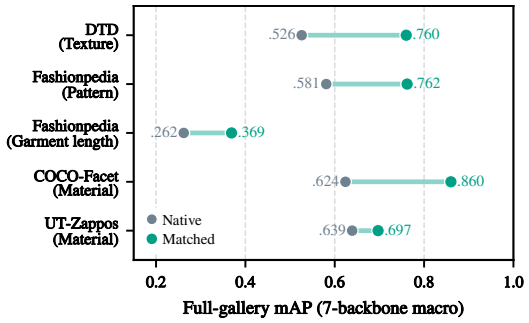


Figure 9: **Matched Access Gain beyond FactorAtlas.** Seven-backbone macro mAP improves across natural texture, garment, and material distinctions.

We finally evaluate matched visual grounding beyond the controlled FactorAtlas setting using natural images. To do so, we draw on datasets that support fine-grained vocabularies for visual factors such as texture, garment pattern and length, and material, and construct controlled text-to-image retrieval settings with dataset-specific splits (Table 4). Despite substantial within-dataset variation in how target values are visually realized, matched grounding yields consistent access gains across the evaluated settings, with mAP improving from 0.262 to 0.369 for Fashionpedia garment length and from 0.624 to 0.860 for COCO-Facet material when averaged across seven backbones (Fig. 9). These results show that the diagnosis and intervention developed in the controlled setting also carry over to natural images.

We further examine whether visual contrasts derived in FactorAtlas can themselves transfer to natural images. For the four pattern values shared with Fashionpedia, directly applying the corresponding contrasts and grounding strengths raises mAP from 0.583 to 0.719 across seven backbones, without target-domain fitting or validation. This suggests that visual structure derived for a well-defined distinction in a controlled setting can remain useful when the same distinction recurs in natural images. Full results and protocols are reported in Appendix H.

5 Relation to Prior Work

Fine-grained and compositional vision–language retrieval. Vision–language models such as CLIP and SigLIP enable retrieval and zero-shot recognition through cross-modal similarity (Radford et al., 2021; Zhai et al., 2023). Yet benchmarks such as Winoground, VL-CheckList, SugarCrepe, and COLA reveal limitations of native vision–language matching on fine-grained attributes, relations, binding, and composition (Thrush et al., 2022; Zhao et al., 2022; Hsieh et al., 2023; Ray et al., 2023). Recent work has further examined these challenges in multi-condition retrieval, attribute binding, and factorized inference (Chow et al., 2026; Lu et al., 2026; Zhang et al., 2025; Alshehri et al., 2026). Complementing these lines of work, we ask whether weak retrieval for a named visual distinction reflects limited image-side discriminability or limited access through the native text query.

Representation, hidden structure, and linguistic access. Shared VLM embedding spaces exhibit meaningful semantic structure across image and text representations (Papadimitriou et al., 2025). On the image side in particular, frozen embeddings preserve compositional organization and ordinal directions identifiable from visual examples (Berasi et al., 2025; Sonthalia et al., 2026). Yet information preserved within a modality may remain weakly accessible through native image–text similarity when image and text representations are not well aligned (Koishigarina et al., 2026). Building on this gap, we take each named visual distinction as the unit of analysis, enabling a finer-grained audit of the relation between image-side structure and linguistic access.

Cross-modal alignment and distinction-specific access. Prior work studies broad cross-modal mismatch (Liang et al., 2022) and approaches for improving image–text correspondence through learned mappings, structure-preserving alignment, or stronger modality-specific representations (Moayeri et al., 2023; Eslami & De Melo, 2024; Gröger et al., 2026; Gong et al., 2025; Huang et al., 2026). Among these approaches, LABCLIP improves attribute–object correspondence through a learned transformation shared across text embeddings (Koishigarina et al., 2026). Support-based adaptation provides another related direction, using labeled visual examples to improve downstream performance with frozen vision–language models (Zhang et al., 2021). In our approach, we focus on distinction-specific access, deriving a matched target-versus-rest contrast for each visual value and using it to adjust the native text query on held-out data. This value-specific intervention tests whether the corresponding image-side structure can reduce the access gap, including after global alignment.

6 Conclusion

A visual distinction can remain discriminable in frozen image geometry while being only weakly addressed by its native text query, so retrieval failure need not imply that the corresponding distinction is absent from the visual representation. Matched visual grounding provides a held-out intervention: when the relevant image-side contrast is present, using that contrast can improve linguistic access and reduce the observed access gap.

This interpretation is supported by direction-specific and visual-evidence controls, robustness across models and held-out conditions, and residual gains after global alignment. Improved access to individual factor values supports exact compositional retrieval and validation-guided selective grounding, while the same matched-grounding logic extends to natural-image distinctions. More broadly, strong aggregate cross-modal alignment does not by itself ensure reliable linguistic access to every distinction that remains discriminable in image geometry.

These conclusions are conditioned on visual discriminability. Access gains vary across settings, the geometric guarantee is local, and the global maps cover only one family of broader alignment corrections. Within this scope, what frozen image geometry can reliably distinguish and what the native language interface can reliably address should be treated as separate empirical questions. When a distinction remains visually discriminable but linguistically under-addressed, matched visual grounding provides a held-out intervention that tests this diagnosis and can improve linguistic access. Extending this perspective beyond predefined distinctions and labeled visual support could enable vision–language systems to more flexibly expose and use visual structure that their representations already support, even when that structure is not well captured by the native language interface.

AI USE STATEMENT

We used generative AI tools to assist with language editing, manuscript organization, and literature discovery during the preparation of this paper. AI tools were also used to suggest alternative phrasings and presentation choices for figures, tables, and experimental descriptions. All scientific claims, numerical results, experimental settings, and citations were reviewed and verified by the authors against the underlying experiments and source materials. Generative AI was not used to generate experimental data or to determine the reported results. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies representation and retrieval behavior in frozen vision–language models using procedurally generated images and existing public image datasets. We do not conduct human-subject experiments or collect new personal data. Natural-image experiments use previously released datasets and annotations, which we cite and use for research purposes according to their established protocols. We are not aware of additional ethical risks specific to the proposed audit beyond the broader limitations and biases of the pretrained models and datasets on which it is applied.

REPRODUCIBILITY STATEMENT

We provide the definitions and evaluation protocol for the audit and matched visual grounding in the main paper. The appendix contains the full proof of the geometric guarantee (Appendix A), backbone-, protocol-, and value-resolved FactorAtlas results (Appendix D), prompt and lexical controls (Appendix C.3), and complete natural-image protocols and backbone-resolved results (Appendix H). These sections document the calibration–validation–test splits, evaluation metrics, intervention selection, and dataset construction needed to reproduce the reported experiments.

REFERENCES

Sultan Alshehri, Zhantao Yang, Han Zhang, and Marios Savvides. Similarity is not logic: Factored inference for dual-encoder vision-language models. *arXiv preprint arXiv:2607.23052*, 2026.

Davide Berasi, Matteo Farina, Massimiliano Mancini, Elisa Ricci, and Nicola Strisciuglio. Not only text: Exploring compositionality of visual representations in vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24917–24927. IEEE, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Wei Chow, Yuan Gao, Linfeng Li, Xian Wang, Qi Xu, Hang Song, Lingdong Kong, Ran Zhou, Yi Zeng, Yidong Cai, et al. Merit: Multilingual semantic retrieval with interleaved multi-condition query. *Advances in Neural Information Processing Systems*, 38:74806–74867, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Sedigheh Eslami and Gerard De Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

Federico Girella, Davide Talon, Ziyue Liu, Zanxi Ruan, Yiming Wang, and Marco Cristani. Lots of fashion! multi-conditioning for image generation via sketch-text pairing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19711–19720. IEEE, 2025.

Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. *arXiv preprint arXiv:2506.02557*, 2025.

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbic. With limited data for multimodal alignment, let the structure guide you. *Advances in Neural Information Processing Systems*, 38:151747–151776, 2026.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer cross-modality representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5131–5139, 2026.

Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European conference on computer vision*, pp. 316–332. Springer, 2020.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4933–4944, 2023.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations*, volume 2026, pp. 96106–96127, 2026.

Siting Li, Xiang Gao, and Simon Du. Highlighting what matters: Promptable embeddings for attribute-focused image retrieval. *Advances in Neural Information Processing Systems*, 38:115077–115110, 2026.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Chenxi Liao, Masataka Sawayama, and Bei Xiao. Probing the link between vision and language in material perception using psychophysics and unsupervised learning. *PLOS Computational Biology*, 20(10):e1012481, 2024.

Xuan Lu, Kangle Li, Haohang Huang, Rui Meng, Wenjun Zeng, and Xiaoyu Shen. Beyond global similarity: Towards fine-grained, multi-condition multimodal retrieval. *arXiv preprint arXiv:2603.01082*, 2026.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, pp. 25037–25060. PMLR, 2023.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. *arXiv preprint arXiv:2504.11695*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36:46433–46445, 2023.

Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. *Advances in Neural Information Processing Systems*, 38:66169–66203, 2026.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Chen Yu and Linda B Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420, 2007.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Qi Zhang, Yuxu Chen, Lei Deng, and Lili Shen. Abe-clip: Training-free attribute binding enhancement for compositional image-text matching. *arXiv preprint arXiv:2512.17178*, 2025.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

A Full Proof of the Geometric Guarantee

Let $\|q\| = \|d\| = 1$, $\Delta = \|\Delta\| d$, and define

$$q_\alpha = \frac{q + \alpha d}{\|q + \alpha d\|}$$

whenever $q + \alpha d \neq 0$. For any unit query r, define the matched score gap

$$M(r) = r^\top \Delta.$$

Proposition. For every $\alpha \geq 0$ for which q_α is defined,

$$M(q_\alpha) \geq M(q).$$

Proof. Let $c = q^\top d \in [-1, 1]$. Since

$$\|q + \alpha d\|^2 = 1 + 2\alpha c + \alpha^2,$$

we obtain

$$M(q_\alpha) = \|\Delta\| \frac{c + \alpha}{\sqrt{1 + 2\alpha c + \alpha^2}}.$$

For $|c| < 1$,

$$\frac{d}{d\alpha} M(q_\alpha) = \|\Delta\| \frac{1 - c^2}{(1 + 2\alpha c + \alpha^2)^{3/2}} > 0,$$

so the matched score gap increases strictly with α.

For $c = 1$, we have $q = d$, hence $q_\alpha = d$ for every $\alpha \geq 0$, and therefore

$$M(q_\alpha) = M(q) = \|\Delta\|.$$

For $c = -1$, $q = -d$ and

$$q_\alpha = \begin{cases} -d, & 0 \leq \alpha < 1, \\ \text{undefined}, & \alpha = 1, \\ d, & \alpha > 1, \end{cases}$$

so

$$M(q_\alpha) = \begin{cases} -\|\Delta\|, & 0 \leq \alpha < 1, \\ +\|\Delta\|, & \alpha > 1. \end{cases}$$

Thus $M(q_\alpha) \geq M(q)$ for every defined $\alpha \geq 0$. □

Negative initial gap. If $-1 < c < 0$, then $M(q) < 0$, and the grounded score gap crosses zero at

$$\alpha^\star = -c = -q^\top d.$$

Hence

$$M(q_{\alpha^\star}) = 0, \qquad M(q_\alpha) > 0 \quad \text{for } \alpha > \alpha^\star.$$

The singular endpoint $c = -1$ is excluded from this continuous crossing statement.

Scope. The guarantee concerns only the matched calibration target-versus-average-rest gap. It does not imply monotonic improvement in held-out retrieval, preservation of unrelated factors, or successful composition; these are evaluated empirically.

B FactorAtlas Construction and Evaluation Protocol

FactorAtlas is the complete product of 10 patterns, 12 hues, 8 shapes, 12 nuisance contexts, and two renderer replicas, totaling 23,040 images. Contexts 0–3 are calibration (7,680 images), 4–5 validation (3,840), and 6–11 test (11,520); both replicas of a context remain in the same split. Contexts vary pose, viewpoint, illumination, background, roughness, and material appearance. Images are rendered at 224×224 using Cycles. Camera elevation is sampled in $[0.14, 0.98]$ radians, camera distance in $[3.65, 6.35]$, lens in $[42, 66]$, object pitch and roll in $[-0.72, 0.72]$, yaw in $[0, 2\pi]$, world strength in $[0.08, 0.72]$, and material roughness in $[0.25, 0.90]$. Each scene uses three randomized area lights. Calibration constructs prototypes and visual contrasts; validation selects all hyperparameters; test is evaluated after those choices are fixed.

The native query is the normalized mean of three prompts: `{v}`, `a {v} surface`, and `an object with a {v} pattern` for pattern; `{v}`, `a {v} object`, and `a {v} surface` for hue; and `{v}`, `a {v}`, and `an object shaped like a {v}` for shape. $\alpha \in \{0, .05, \ldots, 8\}$ is selected by validation macro-mAP, breaking ties by factor-value top-1 and then smaller α. Macro-mAP averages per-value full-gallery AP; factor-value top-1 is an image-wise K-way readout and is distinct from retrieval R@1.

All encoders are frozen. We use official preprocessing for CLIP ViT-B/16 (Radford et al., 2021), EVA02-B/16 (Sun et al., 2023), FG-CLIP Base (Xie et al., 2025), SigLIP SO400M (Zhai et al., 2023), SigLIP2 Base, SigLIP2 Large, and SigLIP2 SO400M (Tschannen et al., 2025). Exact checkpoints are respectively `openai/clip-vit-base-patch16`, OpenCLIP `EVA02-B-16` with `merged2b_s8b_b131k`, `qihoo360/fg-clip-base`, `google/siglip-so400m-patch14-384`, `google/siglip2-base-patch16-224`, `google/siglip2-large-patch16-256`, and `google/siglip2-so400m-patch16-384`.

C Full Diagnostic Controls

C.1 Intervention-Free Separation

We begin with the comparison that does not modify the text query. For each factor value, the image-prototype query is the normalized centroid estimated only from labeled calibration images, whereas the native query is produced by the frozen text encoder from the prompt ensemble specified in Appendix B. Both queries retrieve the same held-out test gallery. The comparison therefore asks whether a distinction remains discriminable in calibration-derived image geometry even when the frozen text interface addresses it less reliably; it is not a comparison between two trained classifiers.

Table 5: **Intervention-free factor-value top-1.**

	Pattern	Hue	Shape
SigLIP2 Base			
Image prototype	.919	.909	.825
Native text	.676	.603	.692
Seven-backbone macro			
Image prototype	.907	.835	.815
Native text	.598	.599	.678

Table 5 reports the image-wise factor-value top-1 readout. The image prototype exceeds the native text query for pattern, hue, and shape on SigLIP2 Base, and the same ordering remains after averaging the seven backbones. The gap is largest for pattern in the seven-backbone macro (.907 versus .598), but is also present for hue and shape. This is the intervention-free evidence that visual discriminability can exceed native linguistic addressability; matched grounding has not yet been applied in this table.

C.2 Specificity of the Visual Distinction

The main results raise a stronger question: is the target-versus-rest visual distinction itself important, or is it sufficient to pull the text query toward images of the correct value? We test the latter possibility

with target-prototype attraction. For each value v, we interpolate the native text query toward its calibration-derived target image centroid:

$$q_v^+(\lambda) = \text{norm}\left((1-\lambda)q_v + \lambda\mu_v^I\right).$$

This baseline uses the correct target-image support but does not contrast the target value against the remaining values in its factor vocabulary. By comparison, target-versus-rest grounding uses the direction

$$d_v^I = \text{norm}\left(\mu_v^I - \bar{\mu}_{-v}^I\right),$$

which explicitly represents the visual distinction between the target value and its alternatives.

For each backbone and factor, a single λ shared by all values is selected from $\{0, 0.0125, \ldots, 1\}$ using validation mAP and then fixed before test evaluation. The grounding strength α is selected independently on the same validation split. Both methods use the same calibration support, native queries, held-out gallery, and evaluation metric.

Table 6: **Target attraction versus target-versus-rest visual contrast.** Each factor cell reports native $\rightarrow$ target-prototype attraction $\rightarrow$ **target-versus-rest grounding** mAP under the primary context holdout. The final column reports the validation-selected attraction strengths for pattern, hue, and shape.

Backbone	Pattern	Hue	Shape	$\lambda_{\text{P/H/S}}$
CLIP ViT-B/16	.361 → .568 → **.830**	.649 → .656 → **.809**	.519 → .584 → **.723**	.713/.163/.500
EVA02-B/16	.515 → .643 → **.843**	.586 → .595 → **.796**	.563 → .649 → **.804**	.525/.150/.488
FG-CLIP Base	.555 → .689 → **.848**	.678 → .708 → **.892**	.599 → .653 → **.740**	.700/.363/.500
SigLIP SO400M	.663 → .745 → **.891**	.667 → .683 → **.825**	.730 → .765 → **.850**	.613/.175/.538
SigLIP2 Base	.635 → .716 → **.878**	.678 → .688 → **.854**	.702 → .726 → **.813**	.425/.100/.400
SigLIP2 Large	.620 → .732 → **.875**	.590 → .597 → **.736**	.744 → .756 → **.837**	.450/.113/.213
SigLIP2 SO400M	.621 → .726 → **.867**	.488 → .497 → **.592**	.764 → .772 → **.851**	.463/.100/.138
Seven-backbone macro	.567 → .689 → **.862**	.619 → .632 → **.786**	.660 → .701 → **.803**	–

Target-prototype attraction improves over native retrieval in all 21 backbone–factor cases, showing that target attraction alone can yield gains. However, target-versus-rest grounding performs better than prototype attraction in all 21 cases. Across factors and backbones, macro-mAP increases from .616 under native retrieval to .674 with prototype attraction and .817 with target-versus-rest grounding. Thus, target attraction alone does not explain the observed access gain; explicitly isolating the visual distinction between each value and its alternatives yields substantially larger held-out gains. This held-out result is separate from, but consistent with, the local geometric guarantee for movement along the matched target-versus-average-rest calibration gap.

C.3 Prompt and Lexical Controls

We test whether matched access gain can be explained by the wording of the native text query. The detailed tables below use the frozen SigLIP2 Base backbone and the primary context holdout. Visual directions are estimated from calibration contexts 0–3, grounding strength is selected on validation contexts 4–5, and all results are reported on test contexts 6–11.

C.3.1 Fixed Prompt Templates

For each factor, we define five grammatical templates that express the same factor value in different ways. Every template is applied uniformly to the complete factor vocabulary; no template is selected separately for individual values. Table 7 lists the complete template bank.

For each factor–template combination, we compare the native query with its matched-grounded counterpart. A single factor-level α is selected on validation separately for each template and fixed before test evaluation. Table 8 reports all 15 test endpoints. Matched grounding improves over the corresponding native query in every case.

The validation-selected strengths for T1–T5 are $[.95, 1.00, .90, 1.05, 1.40]$ for pattern, $[.80, .80, .85, .65, 1.05]$ for hue, and $[2.40, 2.70, 3.00, 2.95, 3.55]$ for shape. The result is therefore not tied to any of the tested query forms. Across all seven audited backbones, matched

Table 7: **Fixed factor-specific prompt templates.** The placeholder `{v}` is replaced by every value in the corresponding factor vocabulary.

ID	Pattern	Hue	Shape
T1	`{v}`	`{v}`	`{v}`
T2	`a {v} surface`	`a {v} object`	`a {v}`
T3	`an object with a {v} pattern`	`an object that is {v}`	`an object shaped like a {v}`
T4	`a surface that is {v}`	`a {v} colored object`	`a geometric {v}`
T5	`a textured {v} surface`	`a {v} surface`	`a three dimensional {v} object`

Table 8: **Prompt robustness on SigLIP2 Base.** Each entry reports native → matched test mAP.

Template	Pattern	Hue	Shape
T1	.629 → .876	.657 → .846	.704 → .813
T2	.549 → .874	.636 → .853	.696 → .813
T3	.577 → .881	.646 → .853	.685 → .813
T4	.528 → .874	.656 → .862	.642 → .813
T5	.501 → .872	.585 → .846	.659 → .813
Improved	5/5	5/5	5/5

grounding improves all $7 \times 3 \times 5 = 105$ backbone–factor–template endpoints under the same per-template evaluation.

C.3.2 Pattern Lexical Substitution

We separately test whether lexical substitution alone can close the observed pattern access gap. For each pattern, we construct an equal-weight ensemble from three predetermined lexical forms using the template `a TERM surface`. No term is selected using validation or test performance. This experiment does not apply matched grounding to the lexical variants; it tests whether replacing the words alone can reproduce the observed access gain.

Table 9: **Pattern lexical-substitution control.** The canonical native ensemble is compared with an equal-weight ensemble of three fixed lexical forms.

Value	Additional lexical forms	Canonical AP	Lexical AP
Plain	solid, unpatterned	.287	.109
Striped	banded, lined	.375	.367
Wavy	undulating, wave patterned	.666	.558
Checkered	checkerboard, chequered	.999	.976
Dotted	polka dotted, dot patterned	.711	.490
Cracked	fissured, crackle patterned	.936	.744
Mottled	blotchy, variegated	.669	.345
Speckled	flecked, peppered	.071	.084
Marbled	veined, marble patterned	.815	.687
Brick	brickwork, brick patterned	.815	.758
Macro mAP		.635	.512
Label top-1		.676	.495

Lexical substitution decreases aggregate mAP from 0.635 to 0.512 and decreases AP for nine of the ten pattern values. Together, the two controls show that matched access gain persists across the tested factor-specific prompt formulations, while lexical substitution alone does not reproduce it.

C.4 Visual-Evidence Dependence

The preceding controls establish direction specificity, but labeled calibration support could still appear to produce gains independently of a repeatable visual contrast. To test dependence on that signal, each patterned image is blended with a nuisance-matched plain rendering of the same scene. Thus pose, viewpoint, illumination, shape, hue, and background are held fixed while the pattern evidence

is progressively attenuated. A strength of $\gamma = 1$ is the original patterned image, and $\gamma = 0$ is the matched plain image. Visual directions are reconstructed from calibration and α is reselected on validation independently at every strength, so intermediate points do not reuse the full-signal direction.

Table 10: **Pattern attenuation on SigLIP2 Base.** Performance is mAP; cosine is measured against the full-signal directions.

γ	Image query	Native	Matched	Cosine
1.00	.649	.635	.878	1.000
.70	.609	.611	.863	.995
.50	.567	.590	.845	.981
.35	.518	.566	.824	.962
.20	.423	.519	.774	.920
.10	.269	.416	.659	.825
.00	.100	.100	.100	-.001

As the signal weakens, image-side discriminability and the absolute level of retrieval achieved by matched grounding decline rather than failing abruptly. At $\gamma = .10$, the image-side direction is weaker but remains aligned with its full-signal counterpart (mean cosine .825), and matched mAP remains .659 compared with native .416. At $\gamma = 0$, images with different pattern labels are identical within each nuisance-matched group, all three readouts reach the ten-way chance value of .100, and the matched access gain disappears. The three readouts use different queries and therefore are not upper bounds on one another; in particular, matched mAP may exceed image-prototype mAP. The relevant result is their shared dependence on a repeatable visual contrast: weak but stable evidence can support matched access gains, whereas labels alone cannot do so after that contrast is removed.

C.5 Finite Grounding and the Pure Visual Endpoint

Our main analysis treats matched visual grounding as an intervention on the native text query. The native query q_v is moved along the value-specific visual contrast d_v^I, with a factor-level grounding strength selected on validation. A possible limiting interpretation is that the resulting gains arise simply because the grounded query approaches an image-derived visual readout, effectively discarding the native text query.

We therefore characterize the relation between finite matched grounding and this limiting endpoint. This is not a separate adaptation method or an additional leaderboard comparison. Rather, it is an endpoint analysis of the same intervention family defined in Eq. (1).

The intervention path. Recall that matched grounding is defined as

$$q'_v(\alpha) = \mathrm{norm}\big(q_v + \alpha d_v^I\big), \qquad \alpha \geq 0. \tag{4}$$

The two endpoints are

$$q'_v(0) = q_v \tag{5}$$

and

$$\lim_{\alpha \to \infty} q'_v(\alpha) = d_v^I. \tag{6}$$

Thus, the native text query, the validation-selected finite grounded query, and the pure visual direction lie on the same intervention path:

$$\underbrace{q_v}_{\alpha=0\text{: native text}} \longrightarrow \underbrace{q'_v(\widehat{\alpha})}_{\text{validation-selected finite grounding}} \longrightarrow \underbrace{d_v^I}_{\alpha\to\infty\text{: pure visual endpoint}} . \tag{7}$$

The pure visual direction d_v^I is therefore the limiting endpoint of the same query family rather than an unrelated baseline.

Geometric location of a finite grounded query. The grounding strength also has a simple geometric interpretation. Let q and d be unit vectors with $q \neq -d$, and let $\alpha \geq 0$. Define

$$q_\alpha = \frac{q + \alpha d}{\|q + \alpha d\|} \tag{8}$$

and let

$$c = q^\top d. \tag{9}$$

Then

$$q_\alpha^\top q = \frac{1 + \alpha c}{\sqrt{1 + \alpha^2 + 2\alpha c}}, \tag{10}$$

while

$$q_\alpha^\top d = \frac{c + \alpha}{\sqrt{1 + \alpha^2 + 2\alpha c}}. \tag{11}$$

Subtracting gives

$$q_\alpha^\top q - q_\alpha^\top d = \frac{(1 - \alpha)(1 - c)}{\sqrt{1 + \alpha^2 + 2\alpha c}}. \tag{12}$$

Proposition. Let q and d be distinct, non-antipodal unit vectors and let

$$q_\alpha = \mathrm{norm}(q + \alpha d). \tag{13}$$

Then q_α is closer in cosine similarity to q than to d whenever $0 \leq \alpha < 1$. It is equally close to the two directions at $\alpha = 1$. It is closer in cosine similarity to d than to q whenever $\alpha > 1$.

Proof. Because $q \neq d$, we have $c = q^\top d < 1$. Therefore, $1 - c > 0$. The denominator in Eq. 12 is positive whenever q_α is defined. Hence, the sign of $q_\alpha^\top q - q_\alpha^\top d$ is determined entirely by $1 - \alpha$. The three cases follow immediately. □

This proposition describes only the geometric position of the grounded query along the intervention path. It does not assign a fraction of retrieval performance to the text or visual component. Accordingly, α should not be interpreted as an attribution weight.

Empirical behavior under limited calibration support. We compare finite grounding with the corresponding pure visual endpoint as the amount of calibration support used to estimate the visual contrast varies. For each factor value, we subsample $n \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ calibration images. Each budget uses 12 balanced resamples; visual directions are re-estimated and α is independently selected on the unchanged validation split for every resample before test evaluation.

As shown in Fig. 10, finite grounding and direction-only retrieval differ substantially at small support, particularly for hue and pattern, and the gap generally narrows as support increases. From one to 256 images per value, direction-only mAP increases from .524 to .811 for shape, .187 to .813 for hue, and .372 to .867 for pattern; the corresponding finite-grounding trajectories are .718 to .812, .679 to .841, and .672 to .876.

The selected α also changes with support. At one image per value, mean α is .10, .03, and .06 for shape, hue, and pattern; at 256 images per value, it reaches 2.78, .53, and .74. Thus, validation selects only a small movement from the native query at very limited support, with factor-dependent movement along the intervention path as additional calibration evidence becomes available.

For completeness, the resampled sweep stops at 256 images per value so that every plotted point follows the same 12-resample protocol. Using the complete calibration support, finite/direction-only mAP is .813/.812 for shape, .854/.833 for hue, and .878/.870 for pattern, with selected α values 2.90, .65, and .70, respectively.

Scope. This analysis characterizes the intervention path rather than attributing retrieval performance to separate linguistic and visual components. Because both the estimated visual direction and the validation-selected α vary with calibration support, α should not be interpreted as a mixture weight or as the fraction of performance contributed by either component. The comparison instead shows that finite grounding and the pure visual endpoint need not behave interchangeably across support regimes.

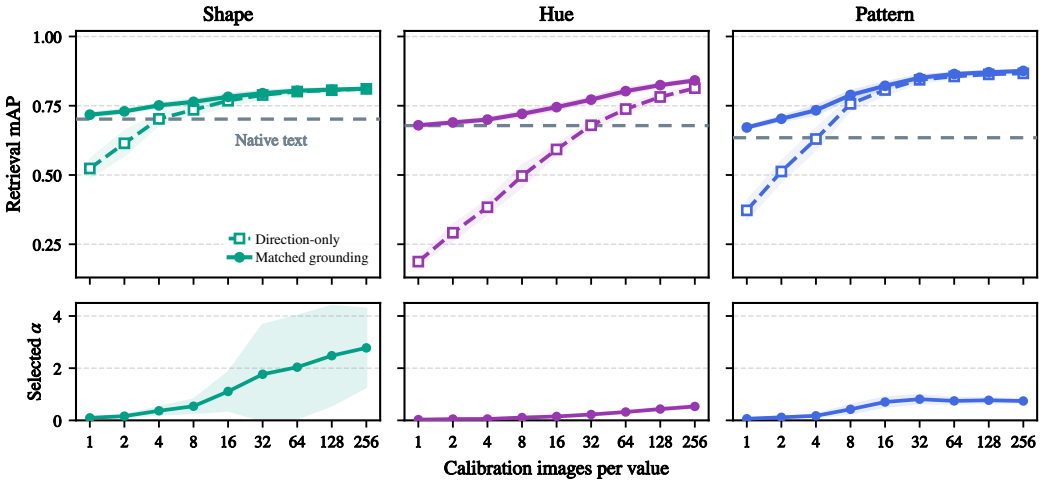


Figure 10: **Finite grounding and the pure visual endpoint under limited calibration support.** Top: test mAP for direction-only retrieval and validation-selected finite grounding. Bottom: corresponding selected grounding strength α. Points and shading show mean ± 1 standard deviation over 12 balanced calibration resamples; the dashed line denotes native-text retrieval.

D Backbone-, Protocol-, and Value-Resolved Readouts

This appendix expands the breadth results summarized in Table 2. Factor-level mAP is the macro average of one-vs-rest AP over the complete factor vocabulary. Factor-value top-1 is image-wise classification accuracy obtained by assigning each image to its highest-scoring value query. It is distinct from the exact-composition retrieval R@1 reported later.

D.1 Unseen Factor Combinations

The unseen-combination evaluation withholds combinations of the two non-target factors while keeping every individual value represented in calibration. Pattern is evaluated on unseen hue–shape cells, hue on unseen pattern–shape cells, and shape on unseen pattern–hue cells. Visual directions and grounding strengths are re-estimated independently for each backbone and held-out protocol.

Table 11: **Per-backbone matched access gains on unseen factor combinations.** Each entry reports native $\rightarrow$ **matched**.

	Pattern		Hue		Shape	
Backbone	mAP	Top-1	mAP	Top-1	mAP	Top-1
CLIP ViT-B/16	.371 → **.838**	.434 → **.856**	.650 → **.820**	.579 → **.856**	.520 → **.729**	.549 → **.707**
EVA02-B/16	.533 → **.856**	.562 → **.888**	.578 → **.816**	.645 → **.841**	.579 → **.817**	.660 → **.787**
FG-CLIP Base	.587 → **.864**	.529 → **.910**	.698 → **.906**	.710 → **.950**	.606 → **.746**	.602 → **.736**
SigLIP SO400M	.691 → **.908**	.739 → **.921**	.662 → **.838**	.613 → **.908**	.735 → **.862**	.756 → **.847**
SigLIP2 Base	.673 → **.895**	.709 → **.924**	.672 → **.872**	.601 → **.910**	.710 → **.832**	.695 → **.810**
SigLIP2 Large	.651 → **.888**	.669 → **.924**	.576 → **.745**	.570 → **.766**	.753 → **.853**	.788 → **.846**
SigLIP2 SO400M	.649 → **.884**	.689 → **.922**	.491 → **.613**	.426 → **.641**	.769 → **.861**	.782 → **.848**
Macro	.594 → **.876**	.619 → **.906**	.618 → **.801**	.592 → **.839**	.667 → **.814**	.690 → **.798**

Matched grounding improves both metrics for all 21 target-aligned backbone–factor evaluations. This result tests generalization to unseen combinations of observed semantic values rather than transfer to unseen nuisance contexts.

D.2 BALANCED NUISANCE PARTITIONS

We construct 24 deterministic balanced partitions of the joint context–renderer realizations. Each partition assigns eight groups to calibration, four to validation, and twelve to testing. Visual directions are estimated from calibration data and the factor-level grounding strength is selected on validation independently within every fold.

Table 12: **Matched access gains across 24 balanced nuisance partitions.** Values are mean ± standard deviation across folds after averaging the seven backbones within each fold.

Factor	mAP: native → matched	Top-1: native → matched	Positive endpoints
Pattern	.5701±.0014 → **.8618±.0005**	.6031±.0019 → **.8977±.0008**	168/168
Hue	.6207±.0026 → **.7877±.0032**	.5961±.0019 → **.8182±.0040**	168/168
Shape	.6630±.0026 → **.8093±.0030**	.6871±.0031 → **.7905±.0049**	168/168

Each factor contributes 7 backbones × 24 folds = 168 endpoints. Matched grounding improves both mAP and top-1 at every endpoint. The small fold-level variation indicates robustness to split choice within the FactorAtlas nuisance generator; it should not be interpreted as transfer to an unrelated renderer distribution.

D.3 VALUE-RESOLVED ACCESS GAINS

Aggregate factor scores can hide whether access gains are shared broadly across the vocabulary or dominated by a few values. We therefore recompute AP separately for every one of the 10 pattern, 12 hue, and 8 shape values, and then average each value over the seven backbones. Native and matched entries use the same primary context holdout; only the query changes.

Factor-level improvement need not be uniform over individual values. Table 13 reports per-value AP averaged across the seven backbones under the primary context holdout.

Table 13: **Value-resolved AP under the primary context holdout.** Each entry reports the seven-backbone macro native → matched AP and its change. Access gains are broad but value-dependent; *teal* is the only value with a negative macro change in this protocol.

Pattern			Hue			Shape		
Value	Native → Matched	Δ	Value	Native → Matched	Δ	Value	Native → Matched	Δ
plain	.227 → .770	+.542	red	.825 → .898	+.072	cube	.551 → .823	+.271
striped	.372 → .854	+.482	orange	.789 → .923	+.133	sphere	.895 → .978	+.083
wavy	.592 → .763	+.171	yellow	.767 → .886	+.119	cylinder	.832 → .908	+.076
checkered	.975 → .998	+.023	lime	.669 → .878	+.209	cone	.753 → .915	+.161
dotted	.643 → .884	+.241	green	.623 → .842	+.218	torus	.885 → .999	+.114
cracked	.865 → .996	+.131	teal	.606 → .594	-.011	pyramid	.514 → .649	+.135
mottled	.517 → .920	+.403	cyan	.445 → .619	+.174	triangular prism	.261 → .466	+.205
speckled	.077 → .547	+.469	blue	.630 → .897	+.267	octahedron	.589 → .683	+.094
marbled	.726 → .931	+.205	indigo	.379 → .758	+.379			
brick	.677 → .953	+.276	purple	.577 → .729	+.152			
			magenta	.695 → .852	+.157			
			pink	.427 → .558	+.131			

The table shows that access gains are distributed across the vocabularies rather than arising from a single value. Large gains occur for values such as *plain*, *striped*, *speckled*, *indigo*, and *cube*, while values already near ceiling, such as *checkered* and *torus*, leave less room for improvement. *Teal* is the sole negative seven-backbone macro endpoint in this protocol. We therefore interpret the result as broad factor-level access gains, not as a claim that every individual value must improve.

E Global Alignment and Alternative Adaptation

E.1 Full-Rank Learned Alignment

The first global comparison asks whether the observed access gap can be removed by learning one expressive transformation over a whole vocabulary, rather than by grounding values individually. This baseline receives direct factor supervision and is retrained for every held-out protocol, making it a target-aware broad correction rather than a zero-shot transfer map.

The global baseline applies $q^G = \mathrm{norm}(Aq)$ with a bias-free full-rank map initialized to identity. Encoders remain frozen. We train one shared map or one map per factor from calibration image–caption pairs, using cyclic factor replacements as hard negatives. Training alternates `a {pattern} {hue} {shape} object` and `a {hue} {shape} with a {pattern} surface`. We use AdamW for 30 epochs, batch size 256, weight decay 10^{-5}, learning rates $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$, and seed 20260820. Checkpoints are chosen by validation macro-mAP and top-1. Each held-out protocol is trained separately. Before training, caption-cache rows are asserted against image identifiers and metadata hashes to verify ordering.

Table 14: **Primary-context global comparison.** Seven-backbone macro-mAP.

Factor	Native	Shared	Factor	Matched
Pattern	.567	.779	.806	**.862**
Hue	.619	.528	.580	**.786**
Shape	.660	.739	.762	**.803**
Macro	.616	.682	.716	**.817**

Table 14 compares four query representations on the same primary-context test galleries. Shared alignment uses one map for all three factors, factor alignment uses a separately trained map for each factor, and matched grounding uses the value-specific image contrast without learning a full-rank map. The global maps substantially improve pattern and shape. Their effect is not uniform, however, and both global maps reduce hue mAP relative to native retrieval in this split. Matched grounding has the highest macro result and the highest result for each factor. The point is not that global alignment fails, but that a broad correction does not uniformly eliminate value-specific access gaps.

E.2 Additional Whole-Vocabulary Linear Maps

We also compare closed-form maps trained from whole-vocabulary centroid correspondences. A factor-specific map uses the K native text queries and K calibration image centroids of one factor; a shared map uses all 30 pattern, hue, and shape correspondences. Validation selects among ridge $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$, truncated low-rank ridge with ranks $\{1, 2, 4, K-1\}$ (or 29 for the shared map), and row-span orthogonal Procrustes. Each candidate is blended with the identity using weights $\{0, .125, \ldots, 1\}$. Selection uses validation macro-mAP, then top-1, then the smaller map blend.

Table 15: **Additional whole-vocabulary maps on SigLIP2 Base.** Values are primary-context test mAP. Validation selects Procrustes for all map scopes in this split.

Method	Pattern	Hue	Shape
Native	.635	.678	.702
Factor-specific map	.772	.727	.789
Shared map	.775	.712	.768
Matched grounding	**.878**	**.854**	**.813**

These closed-form baselines test whether the conclusion depends on the optimization details of the full-rank learned map. Although validation chooses Procrustes in this particular split, the candidate family also includes ridge and low-rank solutions and is selected without test data. Both factor-specific and shared maps improve over native retrieval, but neither reaches matched grounding on

any of the three factors in Table 15. This provides a second, independently fitted whole-vocabulary comparison rather than treating a single global learner as definitive.

E.3 Post-Global Residual Access Gains

The direct comparison above still leaves open whether global alignment and matched grounding address the same part of the mismatch. We therefore apply them sequentially. For post-global evaluation, the selected global map is first frozen. Image directions are then recomputed from calibration data, and a new factor-level α is selected on validation. The global map is never updated using the residual result, and test data are used only for the final evaluation. A positive gain therefore means that value-matched image-side structure remains useful for improving access after the broad transformation has already been applied.

Table 16: **Residual access gains after global alignment.** Macro-mAP averages context and three unseen-combination protocols.

	Shared global			Factor global		
Backbone	Global	+Matched	Improved	Global	+Matched	Improved
CLIP B/16	.680	.743	11/12	.735	.765	10/12
EVA02-B/16	.725	.794	10/12	.738	.791	10/12
FG-CLIP Base	.708	.785	10/12	.753	.797	8/12
SigLIP SO400M	.764	.823	9/12	.775	.827	9/12
SigLIP2 Base	.740	.813	11/12	.753	.816	11/12
SigLIP2 Large	.758	.794	7/12	.762	.795	7/12
SigLIP2 SO400M	.699	.747	9/12	.717	.752	8/12
Macro	.725	**.785**	67/84	.748	**.792**	63/84

The table averages four held-out conditions, with all three target factors evaluated under each condition, giving 12 endpoints per backbone. Adding matched grounding raises the shared-global macro from .725 to .785 and the factor-global macro from .748 to .792. The gain is positive for 67/84 and 63/84 individual endpoints, respectively, so it is strong but not universal. In the primary context holdout alone, adding matched grounding yields positive gains for all 21 factor–backbone endpoints: shared rises from .682 to .802 and factor global from .716 to .806. These residual gains show that the tested broad maps leave residual value-specific access gaps that matched visual grounding can further reduce.

E.4 Full-Support Cache Adaptation

We additionally test whether comparable held-out gains can be obtained by directly retrieving through the labeled calibration support. We implement a training-free Tip-Adapter-style cache (Zhang et al., 2021) using all calibration images, while keeping the image and text encoders frozen. This is a full-support cache control rather than a canonical few-shot reproduction of Tip-Adapter or its fine-tuned variant.

Let K contain the normalized calibration image embeddings and let L contain class-count-normalized factor-wise one-hot labels. For a normalized test-image embedding z, the cache score is

$$s_{\mathrm{cache}}(z) = \exp\big(\beta(zK^\top - 1)\big)\, L.$$

The final score combines the native text similarity with the cache score:

$$s_{\mathrm{Tip}}(z) = s_{\mathrm{native}}(z) + \eta s_{\mathrm{cache}}(z).$$

The cache contains all 7,680 calibration images, corresponding to 768 examples per pattern value, 640 per hue value, and 960 per shape value. The affinity sharpness β and mixture strength η are selected using validation mAP, with factor-value top-1 used as a tie-breaker. We also evaluate the cache-only and nearest-support limits to ensure that the selected result is not caused by a truncated hyperparameter range. All choices are fixed before evaluation on the primary-context test gallery.

The full-support cache yields substantial gains over native retrieval for every backbone and factor. Target-versus-rest grounding nevertheless achieves higher full-gallery mAP in all 21 backbone–factor

Table 17: **Full-support cache adaptation across backbones and factors.** Each factor cell reports native → Tip-Adapter-style cache → **target-versus-rest grounding** full-gallery mAP under the primary context holdout. All hyperparameters are selected on validation and fixed before test evaluation.

Backbone	Pattern	Hue	Shape
CLIP ViT-B/16	.361 → .800 → **.830**	.649 → .703 → **.809**	.519 → .677 → **.723**
EVA02-B/16	.515 → .797 → **.843**	.586 → .644 → **.796**	.563 → .740 → **.804**
FG-CLIP Base	.555 → .827 → **.848**	.678 → .767 → **.892**	.599 → .688 → **.740**
SigLIP SO400M	.663 → .847 → **.891**	.667 → .753 → **.825**	.730 → .827 → **.850**
SigLIP2 Base	.635 → .841 → **.878**	.678 → .723 → **.854**	.702 → .760 → **.813**
SigLIP2 Large	.620 → .822 → **.875**	.590 → .657 → **.736**	.744 → .786 → **.837**
SigLIP2 SO400M	.621 → .821 → **.867**	.488 → .560 → **.592**	.764 → .815 → **.851**
Seven-backbone macro	.567 → .822 → **.862**	.619 → .687 → **.786**	.660 → .756 → **.803**

cases. Averaged across all factors and backbones, mAP increases from .616 under native retrieval to .755 with the Tip-Adapter-style cache and .817 with target-versus-rest grounding. Thus, labeled support can itself yield substantial gains through cache-based adaptation, but the matched target-versus-rest intervention yields larger full-gallery gains without retaining the calibration set as an inference-time cache.

F Exact Composition Details

F.1 Common Factor-Posterior Readout

Factor-level mAP evaluates one semantic axis at a time and does not show that the exact shape–hue–pattern target outranks near-miss images that differ in one or more factors. We therefore evaluate exact compositions using the same readout for native and matched queries, isolating the effect of improved factor-level access.

For each context–renderer-seed pair in the evaluation split, we form a separate retrieval gallery. Under the context holdout, each gallery contains all 960 shape–hue–pattern triples exactly once. For unseen-pair protocols, evaluation is restricted to triples in the corresponding held-out factor-pair cells. We average over all valid query–gallery pairs, yielding $6 \times 2 \times 960 = 11{,}520$ instances for the context holdout and $12 \times 2 \times 120 = 2{,}880$ for each unseen-pair protocol.

We use

$$\log p_f(v \mid x) = \frac{\langle q_{f,v}, x \rangle}{\tau} - \log \sum_{u \in \mathcal{V}_f} \exp\left(\frac{\langle q_{f,u}, x \rangle}{\tau}\right), \quad S(y, x) = \sum_f w_f \log p_f(y_f \mid x). \quad (14)$$

Factor-level α_f values are fixed first. We select $\tau \in \{.005, .01, .02, .05, .1, .2, .5\}$ and positive simplex weights on a 0.1 grid using validation exact MRR followed by R@1. Native and matched use the same gallery and composition operator.

Table 18: **Exact composition.** Seven-backbone macro results.

	Native			Matched		
Condition	R@1	R@5	mAP	R@1	R@5	mAP
Context	.357	.714	.517	**.711**	**.962**	**.819**
Unseen H×S	.631	.929	.764	**.928**	**.998**	**.960**
Unseen P×S	.622	.952	.766	**.876**	**.995**	**.931**
Unseen P×H	.661	.937	.779	**.828**	**.990**	**.900**

Because each evaluated query has a single exact positive, AP equals the reciprocal rank of that target, so the reported mAP is numerically identical to mean reciprocal rank. Under the context holdout, matched grounding raises exact R@1 from .357 to .711 and R@5 from .714 to .962. Gains persist

for unseen factor-pair combinations: in unseen H×S, for example, pattern queries are evaluated on hue–shape cells withheld from calibration, with P×S and P×H defined analogously. Thus, stronger factor-level access carries over to exact retrieval of new multi-factor combinations.

G Selectivity and Sample Efficiency

G.1 Validation-Guided Selective Grounding

Matched grounding need not be applied to every value. Some native queries are already effective, while others show a clear validation gain after grounding. This experiment asks whether held-out audit evidence can select between the native and grounded query for each value without consulting test performance.

We first select one factor-wide grounding strength α on validation. For each value, we then estimate the uncertainty of its validation AP gain using 2,000 bootstrap resamples over context–renderer-replica groups. Grounding is retained only when the lower bound of the resulting 95% confidence interval is positive; otherwise, the native query is kept. All decisions are fixed before test evaluation. Here, the unseen H×S protocol uses the same held-out hue–shape cells when evaluating all three target factors; it is distinct from the target-specific unseen-combination macro reported in the main breadth analysis.

Table 19: **Selective grounding on SigLIP2 Base.**

Protocol	Factor	α	Ground/keep	Native	All	Selective	Improved/declined
Context	Pattern	.70	10/0	.635	.878	.878	10/0
	Hue	.65	11/1	.678	.854	.855	11/0
	Shape	2.90	7/1	.702	.813	.813	7/0
Unseen H×S	Pattern	.70	9/1	.673	.895	.895	9/0
	Hue	.10	7/5	.671	.698	.697	6/1
	Shape	7.35	7/1	.721	.793	.809	7/0
Macro/total			51/9	.6797	.8220	**.8245**	50/1

Across the two protocols and three factors, the gate grounds 51 of 60 values and retains the native query for nine. Relative to native mAP of .6797, grounding all values reaches .8220 and the validation-guided policy reaches .8245. Of the 51 grounded decisions, 50 improve on test and one declines. The small aggregate difference from grounding everything is not the primary point; the result demonstrates that held-out audit evidence can localize where grounding is useful without inspecting test outcomes. We treat this as a consequence of the audit rather than an additional central method claim.

G.2 labeled Support-Size Scaling

The main experiments use all available calibration images, but the visual contrast may be estimated from substantially less labeled support. We vary the number of calibration images per value while keeping the validation and test sets unchanged. This experiment measures both matched access gain and the agreement between each estimated direction and its full-support counterpart.

For each value, we sample $n \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ calibration images without replacement and re-estimate its centroid and target-versus-average-rest direction. Sampling is balanced across values, and each finite support budget is evaluated using 12 independent calibration resamples. For every resample, the grounding strength α is selected again on the unchanged validation split. The test set is never sampled or used for selection. One-shot support therefore uses one image per value, corresponding to 10, 12, or 8 total calibration images for pattern, hue, or shape, respectively.

Table 20 reports results for SigLIP2 Base, averaged across the three factors and two held-out protocols, yielding six factor–protocol endpoints. The first protocol is the primary context holdout. The second applies the same held-out hue–shape cells when evaluating all three target factors. Direction cosine measures the cosine similarity between each support-limited direction and the corresponding direction estimated from all calibration images.

Table 20: **Matched access gains with increasing labeled support.** Finite-budget results average 12 calibration resamples.

Images/value	1	2	4	8	16	32	64	128	256	Full
Test mAP	.695	.711	.730	.755	.774	.795	.806	.812	.819	.822
Gain	+.015	+.032	+.051	+.076	+.095	+.115	+.127	+.133	+.139	+.142
Direction cosine	.466	.594	.719	.825	.897	.946	.972	.986	.994	1.000

Table 21: **Factor-resolved support-size subset.** Entries are test mAP.

Factor	1	8	32	256	Full
Pattern	.688	.802	.860	.884	.887
Hue	.670	.698	.731	.769	.776
Shape	.728	.767	.793	.805	.803

The first table aggregates the six factor–protocol endpoints, whereas Table 21 separates the three factors after averaging the two protocols. Access gains and direction quality increase smoothly with support size. Thirty-two images per value achieve about 81% of the full-support gain, and 64 achieve about 89%. Shape is already close to its full-support mAP at moderate budgets, whereas pattern benefits more strongly from additional support. The positive one-shot mean does not imply a stable one-shot direction: its mean cosine to the full-support direction is only $.466$. Thus the sweep shows that substantial access gains can be obtained with limited labeled support at the aggregate level, while also showing why the full calibration set is preferable for a stable geometric estimate.

H Natural-Image Protocols and Full Results

H.1 Shared Evaluation Protocol

All natural-image experiments use frozen, normalized image and text embeddings. For each target value, we estimate normalized calibration centroids and construct a normalized target-versus-average-rest visual direction. A single factor-level grounding strength is selected on validation by full-gallery macro-mAP and then fixed before test evaluation. Test images are never used to select prompts, directions, grounding strengths, thresholds, or checkpoints.

Because the natural-image datasets differ in their annotations and grouping structure, we construct dataset-specific calibration, validation, and test partitions while preserving this shared evaluation contract. The primary metric is full-gallery macro-mAP.

H.2 Backbone-Resolved Results

Table 22 reports every backbone-level result behind the seven-backbone macro results in the main paper. The columns intentionally retain separate datasets and targets rather than pooling incompatible natural protocols: DTD tests texture classes, Fashionpedia tests garment pattern and length, COCO-Facet tests localized material, and UT-Zappos tests footwear material under product-disjoint splits. Each cell reports native retrieval followed by matched visual grounding; the final row averages backbones within each setting, not examples across datasets.

Matched visual grounding yields positive gains for every evaluated backbone on DTD, Fashionpedia pattern, Fashionpedia length, and COCO-Facet material. It yields positive gains for six of seven backbones on UT-Zappos. The remaining EVA02-B/16 endpoint changes from 0.6653 to 0.6649 and is therefore unchanged at the three-decimal precision used in the paper. Because the 96-product UT-Zappos cohort also exhibits substantial variation across its five product-disjoint rotations, we treat it as a boundary result rather than part of the stable-transfer count.

Table 22: **Backbone-resolved natural-image transfer.** Each cell reports native → **matched** full-gallery macro-mAP. Fashionpedia pattern uses the garment-category holdout, and COCO-Facet material uses the natural localized crop. Matched grounding improves all 28 backbone–setting endpoints in the four stable transfer settings; UT-Zappos is reported separately as a higher-variance boundary.

Backbone	DTD	Fashionpedia		COCO-Facet	UT-Zappos
	Texture	Pattern	Length	Material	Material
CLIP ViT-B/16	.359 → **.668**	.452 → **.618**	.178 → **.317**	.604 → **.850**	.522 → **.617**
EVA02-B/16	.444 → **.731**	.567 → **.768**	.220 → **.361**	.668 → **.819**	.665 → .665
FG-CLIP Base	.500 → **.751**	.566 → **.789**	.241 → **.352**	.601 → **.899**	.631 → **.676**
SigLIP SO400M	.621 → **.803**	.631 → **.799**	.297 → **.396**	.615 → **.869**	.677 → **.751**
SigLIP2 Base	.564 → **.771**	.620 → **.789**	.285 → **.370**	.627 → **.874**	.678 → **.717**
SigLIP2 Large	.588 → **.797**	.620 → **.778**	.308 → **.403**	.628 → **.860**	.662 → **.726**
SigLIP2 SO400M	.607 → **.798**	.614 → **.793**	.302 → **.387**	.623 → **.851**	.642 → **.724**
Seven-backbone macro	.526 → **.760**	.581 → **.762**	.262 → **.369**	.624 → **.860**	.639 → **.697**

H.3 Dataset-Specific Protocols

DTD texture. We use all 5,640 DTD images and all 47 official texture classes. Each of the ten official partitions provides 40 calibration, 40 validation, and 40 test images per class. The official class token is used as the fixed native query. Directions and grounding strengths are re-estimated independently within each partition, and the table reports performance averaged over the ten partitions.

Fashionpedia garment pattern. We use localized garment crops and fine-grained attributes from Fashionpedia. Each annotated garment box is placed on an aspect-ratio-preserving square canvas and resized to the model input resolution. The pattern vocabulary contains six values: *plain*, *floral*, *stripe*, *check*, *abstract*, and *dot*.

The resulting cohort contains 2,400 crops, balanced over the Cartesian product of six patterns and four garment categories. We perform four leave-one-category-out evaluations. For each run, one garment category supplies the entire test gallery and is excluded from both direction estimation and grounding-strength selection. This tests whether pattern directions estimated from other garment categories transfer to a previously unseen garment category.

Fashionpedia garment length. We separately evaluate an eight-value garment-length vocabulary: *mini*, *micro*, *above-the-knee*, *knee*, *midi*, *below-the-knee*, *maxi*, and *floor*. The cohort contains 2,847 localized dress crops.

Five source-image-disjoint calibration–validation rotations are constructed from the Fashionpedia training set. The untouched Fashionpedia validation gallery is used as test in every rotation. Consequently, variation across rotations measures robustness to the calibration and validation samples rather than uncertainty from different test galleries.

During cohort construction, local English descriptions from LOTS/Sketchy (Girella et al., 2025) were linked to the Fashionpedia examples. However, the evaluated images, garment boxes, category labels, and visual attributes come from Fashionpedia, and retrieval uses fixed factor prompts rather than the LOTS sketch-generation task. We therefore refer to the evaluated dataset as Fashionpedia while recording both sources in the data provenance.

COCO-Facet material. We join official COCO-Facet material labels with COCO-Stuff localization (Caesar et al., 2018). We retain images containing exactly one requested material label and exclude images containing a segment from another requested material class.

For each image, the natural crop is the tight bounding box of the largest individual target-material segment. It is not the union of spatially disconnected segments. We apply fixed inclusion criteria requiring the target mask to occupy at least 1% of the image, fill at least 25% of its bounding box,

and have a box covering at most 50% of the full image. We additionally require at least 24 images per material class.

The resulting cohort contains 306 images: 172 *metal*, 29 *stone*, and 105 *wood*. Five deterministic rotations are image-disjoint across calibration, validation, and test. The main-paper result uses the unmodified natural crop.

UT-Zappos material. We evaluate four footwear materials: *Leather*, *Suede*, *Patent Leather*, and *Synthetic*. The cohort is restricted to a fixed footwear stratum and balances material counts within retained toe-style–heel-height cells.

We retain one deterministic image per product, yielding 96 products with 24 products per material. Products never cross calibration, validation, and test roles. Results are averaged over five product-disjoint rotations.

H.4 Cross-Dataset Reuse of Shared Visual Distinctions

The matched natural-image experiments above estimate visual contrasts and select grounding strength separately within each target dataset. We additionally ask a stricter question: when two domains share the same visual distinction, can image-side directions estimated in one domain remain useful in the other without any target-side fitting or validation? We test this using the four pattern values shared by FactorAtlas and Fashionpedia.

For each backbone, we construct FactorAtlas pattern directions from the complete ten-value vocabulary using the four calibration contexts and reuse the corresponding grounding strength selected in FactorAtlas. No Fashionpedia image is used to estimate a direction or select the grounding strength. We map FactorAtlas *plain*, *striped*, *checkered*, and *dotted* to Fashionpedia *plain*, *stripe*, *check*, and *dot*, respectively.

We evaluate all four Fashionpedia category-held-out partitions. Each held-out gallery contains 600 images spanning all six Fashionpedia pattern labels, so the two unmapped labels remain natural negatives rather than being removed from the gallery. Across seven backbones, directly reusing the FactorAtlas directions improves macro-mAP from .583 to .719 (Table 23). The aggregate gain is positive for all seven backbones and for all 28 backbone–partition evaluations. At the individual backbone–value level, 20 of 28 cases improve, indicating that cross-dataset reuse is consistent in aggregate but remains value-dependent. These results suggest that when the same visual distinction is shared across domains, its image-side direction can remain useful beyond the data from which it was estimated.

Table 23: **Cross-dataset reuse of shared pattern distinctions.** For each backbone, FactorAtlas-derived pattern directions and its fixed FactorAtlas grounding strength are reused without Fashionpedia fitting or validation. Values are four-value macro-mAP averaged over the four Fashionpedia category-held-out partitions.

Backbone	Native	Reused direction	Δ
CLIP ViT-B/16	.414	**.490**	+.075
EVA02-B/16	.575	**.720**	+.145
FG-CLIP Base	.568	**.800**	+.232
SigLIP SO400M	.640	**.767**	+.127
SigLIP2 Base	.631	**.795**	+.164
SigLIP2 Large	.628	**.750**	+.121
SigLIP2 SO400M	.623	**.711**	+.088
Seven-backbone macro	.583	**.719**	+.136