From Embeddings to Interpretable Representations: Orthogonal Projection onto Researcher-Specified Concepts

Yohei Nishimura
Wisconsin School of Business
ynishimura@wisc.edu

Abstract

Pretrained vision–language embeddings are rich but uninterpretable. Their coordinates mean nothing individually, which keeps them out of downstream models whose variables must be measurements. We propose a training-free, deterministic projection that reduces such embeddings to a small set of orthogonal scores on concepts the researcher specifies in natural language, plus a compressed residual that retains what the concepts miss. It requires no labeled images. We show that the representation is invariant to the rotations of the embedding space that contrastive pretraining leaves undetermined, so the projection is exactly reproducible from prompt lists and a public encoder. An empirical example with a five-trait brand-personality lexicon illustrates what the axes select on real images and shows that the axes are stable under prompt resampling and insensitive to the orthogonalization order.

1. Introduction

Images are now a common input to quantitative analysis. Recommender systems rank them, firms choose advertising creative based on them, and a growing empirical literature in marketing, economics, and the social sciences seeks to use them as variables in downstream models [8, 10, 15, 21, 30]. The standard first step is to embed each image with a pretrained vision–language encoder such as CLIP [27]. The embedding preserves rich semantic content and requires no task-specific training. The second step is where the difficulty lies. An embedding is a point in a space of more than a thousand dimensions whose individual coordinates mean nothing, and a researcher who wants to ask an interpretable question of the data cannot ask it of raw coordinates. What downstream analysis needs is a low-dimensional representation whose coordinates are *measurements*.

What properties must such a representation have? We argue for four, formalized in Section 3.1. It should be interpretable without labeled training data (R1). It should be deterministic, so that results and counterfactuals are exactly reproducible (R2). Its coordinates should mean the same thing for every entity type, so that cross-entity comparisons and interactions are well-defined (R3). And it should retain what the named concepts miss instead of discarding it (R4). Existing approaches satisfy some requirements and violate others. Supervised deep learning [21] and generative pipelines built on variational autoencoders [4, 17] are the standard methods for image-based measurement in marketing and business research. Both require labeled data or produce latent spaces that depend on the random draws of a training run, violating R1 and R2. Concept bottleneck models need concept labels or discard what the concepts miss [18]. Interpretable basis changes such as POLAR discard the complement of the concept subspace [22], violating R4. Nullspace projections keep the complement and discard the concepts [28].

We propose a construction that satisfies all four requirements and involves no training (Figure 1). The researcher names each concept with a handful of positive and negative text prompts. The normalized difference of the prompt centroids, computed with the same model's text encoder, defines a semantic direction in embedding space. Gram–Schmidt orthonormalization turns the directions into axes that do not double-count shared variation. Principal components of the orthogonal complement, fitted once on a reference corpus, compress everything the concepts do not capture. The result is a fixed affine map with orthonormal rows. Its first K_{sem} coordinates score the named concepts on a cosine scale, and its remaining K_{res} coordinates summarize the residual. The method rests on a single empirical premise, stated as Assumption 1.

Because the map is simple, its measurement properties can be stated as theorems. The representation depends on the encoder only through inner products, so it is invariant to the orthogonal transformations that contrastive pretraining leaves unidentified (Proposition 3). The projection is therefore fully specified by the prompt lists, the concept ordering, the reference corpus, and the name of a public encoder. Any researcher can reproduce the same numbers without access to a training run. Neither a VAE latent space nor a proprietary supervised score offers this. The map never inflates the distance between two images (Remark 2).

We illustrate the construction with an empirical example

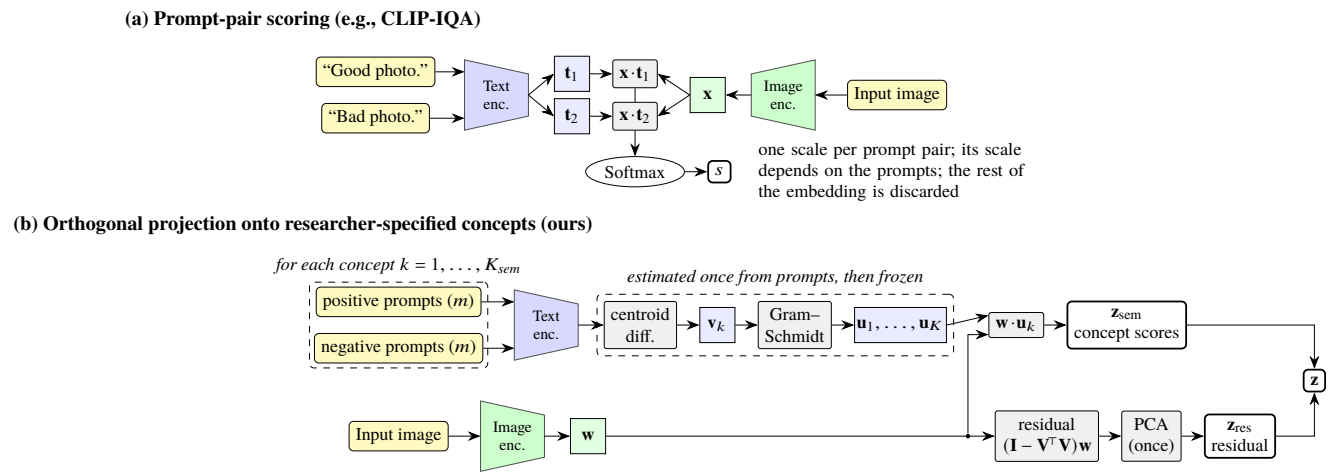

Figure 1: Prompt-based scoring versus the proposed projection. (a) A prompt pair yields one score per image through a softmax over two similarities; the score's scale depends on the prompts, and the rest of the embedding is discarded. (b) For each concept, the difference of positive and negative prompt centroids gives a direction $\mathbf{v}_k$; Gram–Schmidt turns the K_{sem} directions into orthonormal axes. An image embedding $\mathbf{w}$ is projected onto the axes (cosine-scale concept scores) and its residual is compressed by a PCA fitted once. Both encoders are frozen; nothing is trained.

(Section 4). We instantiate the axes with a five-trait brand-personality lexicon [1] and a public encoder, and show what the axes select on real sponsored-post images. The example also checks robustness. Resampling the prompts moves each axis by a mean cosine of no less than 0.84 with nine positive and nine negative prompts per concept, and the Gram–Schmidt ordering has a negligible effect. Across all 120 orderings, no axis rotates below a cosine of 0.91 from its canonical-order counterpart. Every numerical check uses text encodings alone.

Our contributions are as follows.

- **Method.** A training-free, closed-form construction of interpretable low-dimensional representations of vision–language embeddings. Zero-shot concept axes from prompt differences, enforced orthonormality, and systematic retention of the residual subspace.
- **Theory.** The representation is invariant to the rotations of the embedding space that pretraining leaves undetermined, and it never inflates distances.
- **Empirical example.** An instantiation with a five-trait brand-personality lexicon that illustrates the axes on real images and checks their stability under prompt resampling and reordering, from text encodings alone.

2. Related Work

2.1. Measuring Images in Business and Social Science

The mainstream approach to extracting perceptual attributes from images in marketing research is supervised. Deep networks are trained on human-labeled examples to score brand attributes portrayed in social media photos [21]. Generative pipelines built on variational autoencoders measure and manipulate product aesthetics [4, 17]. Both families are effective, but their measurements depend on labeled data and on a particular training run. The supervised score cannot be reproduced without the labels. A VAE's latent coordinates are identified at most up to rotation, and even when a convention such as the posterior mean makes each measurement deterministic, the encoder itself is the outcome of a stochastic training run, so rebuilding it yields a different map. A more recent approach scores images with text prompts directly, exploiting the shared embedding space of contrastive encoders. Prompt pairs recover continuous perceptual attributes such as image quality and aesthetic look-and-feel [32]. Our method develops this zero-shot route into a full measurement method. Multiple orthogonal axes instead of one or two scales, a retained residual instead of a single score, and a proof that the representation does not depend on the rotations that pretraining leaves undetermined (Proposition 3).

2.2. Semantic Projection and Linear Structure in Embeddings

Our axis construction builds on semantic projection in word embeddings. Linear regularities in embedding spaces were documented early [23]. Kozlowski et al. [19] built cultural dimensions (affluence, gender, status) as differences of antonym vectors and projected words onto them. Grand et al. [13] showed that such projections recover human ratings of object features across many semantic scales. This linear

structure extends to multimodal spaces: embedding arithmetic and multimodal neurons in CLIP [12], compositional linear structures in vision–language models [31], and formal accounts of linearly encoded concepts in language models [26]. Contrastive encoders also exhibit a modality gap, a roughly constant offset between text and image embeddings [20]. Our construction handles it by differencing two text centroids, which cancels the shared offset. Relative to this literature, we extend semantic projection to the cross-modal setting and add two elements it lacks: enforced orthonormality of the axis set and systematic retention of the orthogonal complement. Assumption 1 is precisely this literature's hypothesis, and our validation experiments test it in the target domain.

2.3. Interpretable Low-Dimensional Representations

A smaller literature builds interpretability into the coordinate system itself. POLAR maps pretrained word embeddings onto antonym-pair directions, following the semantic differential. SensePOLAR extends this to contextual embeddings [11, 22]. Both discard whatever lies outside the polar subspace, and neither enforces orthogonality. Densifier learns an orthogonal transformation that concentrates task-relevant semantics into a few dimensions, and DensRay is its analytical variant [9, 29]. Both need labeled lexicons to find the axes. Concept whitening aligns orthogonal axes of a network layer with concepts, but it needs concept example sets and retraining [6]. Nullspace projection makes the opposite choice from POLAR: embedding debiasing estimates a concept subspace and *removes* it [3, 28]. We apply the same projection with the opposite aim and keep both the concept subspace and a compressed complement. In factor-analysis terms, the concept axes are confirmatory and the residual components are exploratory. To our knowledge, no existing method combines zero-shot prompt-derived axes, enforced orthonormality, and a retained complement in one deterministic map.

2.4. Concept-Based Interpretability

A large literature explains vision models through concepts. Concept activation vectors estimate concept directions from labeled example sets [16]. Concept bottleneck models route predictions through supervised concept layers [18]. Post-hoc and label-free variants replace the supervision with embeddings of concept words from CLIP-type models [25, 33, 34]. Concepts can be transferred across models by aligning to CLIP space [24], and SpLiCE decomposes CLIP embeddings into sparse combinations over a large concept dictionary [2].

The substantive difference from this literature is one of purpose. Concept-based interpretability explains or constructs classifiers, so stochastic training, per-model concept sets, and lossy bottlenecks are acceptable. We construct variables for downstream quantitative models, where determinism, cross-entity consistency, invariance to the encoder's undetermined rotations, and exact decomposition are required. The methods above neither need nor provide these properties.

3. Method

3.1. Setup and Requirements

We assume access to a pretrained joint text–image encoder consisting of a text encoder $\phi_{\text{txt}} : \mathcal{T} \to \mathbb{R}^D$ and an image encoder $\phi_{\text{img}} : \mathcal{X} \to \mathbb{R}^D$ trained with a contrastive objective, so that inner products across modalities encode semantic similarity [7, 27]. Both encoders output ℓ_2-normalized embeddings; we write $\mathbf{w} = \phi_{\text{img}}(x) \in \mathbb{R}^D$ for the embedding of an image x and use CLIP-family models as the running example, although nothing in the construction is specific to them.

The researcher specifies a set of K_{sem} *concepts* $C = \{c_1, \ldots, c_{K_{\text{sem}}}\}$: the semantic attributes along which images are to be measured. These may be the five brand-personality traits of Aaker [1], or equally aesthetic, emotional, or topical attributes. The goal is a map $f : \mathbb{R}^D \to \mathbb{R}^K$ with $K \ll D$ whose output $\mathbf{z} = f(\mathbf{w}) = [\mathbf{z}_{\text{sem}}; \mathbf{z}_{\text{res}}]$ splits into a semantic block $\mathbf{z}_{\text{sem}} \in \mathbb{R}^{K_{\text{sem}}}$, one coordinate per concept, and a residual block $\mathbf{z}_{\text{res}} \in \mathbb{R}^{K_{\text{res}}}$ summarizing what the concepts do not capture. Four requirements govern the construction:

R1 Zero-shot interpretability. Coordinate k of $\mathbf{z}_{\text{sem}}$ measures concept c_k, without labeled training examples or fine-tuning.

R2 Determinism. Two things must be deterministic. First, the measurement: given the projection, the same input must yield the same representation. Second, the projection: rebuilding it from the same prompts, reference corpus, and encoder must return the same map. No random initialization, minibatch order, or sampling step may enter its construction.

R3 Cross-entity measurement consistency. A single f applies to every embedding in the application, whatever entity it describes. This ensures that coordinate k has the same meaning everywhere and cross-entity comparisons and interactions are well-defined (Section 3.5).

R4 Information retention. For a fixed total number of dimensions K, the map discards as little of $\mathbf{w}$ as possible. The part of $\mathbf{w}$ orthogonal to the concept directions is compressed into $\mathbf{z}_{\text{res}}$.

The construction in Section 3.2–Section 3.5 satisfies all four: R1 and R2 by construction (Section 3.2 and Section 3.4), R3 by applying one map to every entity (Section 3.5), and R4 by retaining the complement (Section 3.4).

The construction below is built entirely from affine operations including centroid differences, orthogonal projections, and a fixed affine map. For these operations to constitute measurement, the embedding geometry must have one property. Semantic content must be encoded in directions, so that moving along a direction varies an attribute and inner products recover its degree. We state this as the method's single substantive premise.

Assumption 1 (Approximate linear encoding)**.** *For each concept $c_k \in C$ there exists a direction $\mathbf{d}_k \in \mathbb{R}^D$ such that, over the image population of interest, the inner product $\mathbf{d}_k^\top \mathbf{w}$ is approximately monotone in the degree to which the image expresses c_k.*

Assumption 1 is the linear representation hypothesis specialized to our setting. It is not implied by the contrastive training objective, so it cannot be taken for granted. It is, however, supported by consistent evidence: additive semantic regularities in word embeddings [13, 19, 23], embedding arithmetic and multimodal neurons in CLIP-type models [12], compositional linear structure in vision–language spaces [31], and formal accounts of linearly encoded concepts in large language models [26]. We therefore treat Assumption 1 as an empirical claim. Everything downstream of it is exact linear algebra (Section 3.4), so the method can fail only if Assumption 1 fails.

3.2. Concept Axes from Prompt Differences

For each concept c_k the researcher writes m positive prompts $\{p^+_{k,1}, \ldots, p^+_{k,m}\}$ exemplifying the concept and m negative prompts $\{p^-_{k,1}, \ldots, p^-_{k,m}\}$ describing its antithesis. The raw direction for c_k is the normalized difference of the prompt centroids:

$$\tilde{\mathbf{v}}_k = \frac{1}{m}\sum_{s=1}^{m} \phi_{\text{txt}}\left(p^+_{k,s}\right) - \frac{1}{m}\sum_{s=1}^{m} \phi_{\text{txt}}\left(p^-_{k,s}\right), \qquad \mathbf{v}_k = \frac{\tilde{\mathbf{v}}_k}{\|\tilde{\mathbf{v}}_k\|_2}. \tag{1}$$

This construction adapts *semantic projection* from word embeddings to the cross-modal setting. In word embeddings, projections onto antonym-based difference directions recover interpretable scales that correlate with human ratings [13, 19]. Because text and images share the contrastively trained embedding space, $\mathbf{v}_k^\top \mathbf{w}$ is a well-defined cross-modal similarity [27]. Prompt-pair scoring of this kind is known to recover continuous perceptual attributes of images [32].

Two design rules make Eq. 1 a measurement rather than an arbitrary coordinate. First, positive and negative prompts share a *neutral anchor* phrase (e.g., "a product," "a photograph") and differ only in the attribute description. Subtracting the negative centroid then cancels the anchor's contribution and isolates the direction that separates the concept from its antithesis [12]. Second, averaging over m positive and m negative paraphrases reduces noise from idiosyncratic phrasing, so $\mathbf{v}_k$ approaches the abstract direction of the concept (the empirical example in Section 4.3 quantifies the remaining sensitivity).

3.3. Orthonormalization

The raw directions $\mathbf{v}_1, \ldots, \mathbf{v}_{K_{\text{sem}}}$ are in general correlated, because related concepts share vocabulary and visual connotations. Correlated axes double-count shared variation and make coordinate-wise readings ambiguous. With a modified Gram–Schmidt sweep,

$$\tilde{\mathbf{u}}_k = \mathbf{v}_k - \sum_{l=1}^{k-1} \left(\mathbf{v}_k^\top \mathbf{u}_l\right) \mathbf{u}_l, \qquad \mathbf{u}_k = \frac{\tilde{\mathbf{u}}_k}{\|\tilde{\mathbf{u}}_k\|_2}, \tag{2}$$

we obtain $\mathbf{V}_{\text{ortho}} = [\mathbf{u}_1; \ldots; \mathbf{u}_{K_{\text{sem}}}] \in \mathbb{R}^{K_{\text{sem}} \times D}$ with $\mathbf{V}_{\text{ortho}} \mathbf{V}_{\text{ortho}}^\top = \mathbf{I}_{K_{\text{sem}}}$.

Gram–Schmidt is order-dependent: axis k is defined as concept c_k with its components along concepts $c_1, \ldots, c_{k-1}$ removed. When the concept set has a natural hierarchy, this is a feature. The researcher orders concepts by theoretical primacy, and each later axis carries an explicit partialling-out interpretation. The choice of ordering is also checkable at negligible cost, by recomputing the axes under every ordering and measuring how much they move. Section 4.4 measures the ordering's actual effect in our empirical example.

3.4. Unified Projection with Residual Retention

Next, we retain the orthogonal complement in compressed form. Define the residual of an embedding as its projection onto the complement of the concept subspace,

$$\mathbf{r} = \left(\mathbf{I}_D - \mathbf{V}_{\text{ortho}}^\top \mathbf{V}_{\text{ortho}}\right) \mathbf{w} \in \mathbb{R}^D, \tag{3}$$

and let $\mathbf{W}_{\text{PCA}} \in \mathbb{R}^{K_{\text{res}} \times D}$ collect the top K_{res} principal components of the residuals $\{\mathbf{r}_n\}$ of a *reference corpus* of images from the application domain, with mean $\bar{\mathbf{r}}$. Both are estimated once and then frozen. The full map is the affine projection

$$f(\mathbf{w}) = \mathbf{P}\mathbf{w} + \mathbf{b}, \qquad \mathbf{P} = \begin{pmatrix} \mathbf{V}_{\text{ortho}} \\ \mathbf{W}_{\text{PCA}} \left(\mathbf{I}_D - \mathbf{V}_{\text{ortho}}^\top \mathbf{V}_{\text{ortho}}\right) \end{pmatrix} \in \mathbb{R}^{K \times D}, \tag{4}$$

where $\mathbf{b} = [\mathbf{0}; -\mathbf{W}_{\text{PCA}}\bar{\mathbf{r}}]$ absorbs the PCA centering and $K = K_{\text{sem}} + K_{\text{res}}$. The construction involves no gradient-based training: its cost is $2mK_{\text{sem}}$ text encodings and one PCA on the reference corpus. Figure 2 summarizes the construction and its two-stage use.

The map inherits the following properties directly from its construction.

Proposition 1 (Orthonormal measurement system)**.** *The rows of $\mathbf{P}$ are orthonormal in $\mathbb{R}^D$.*

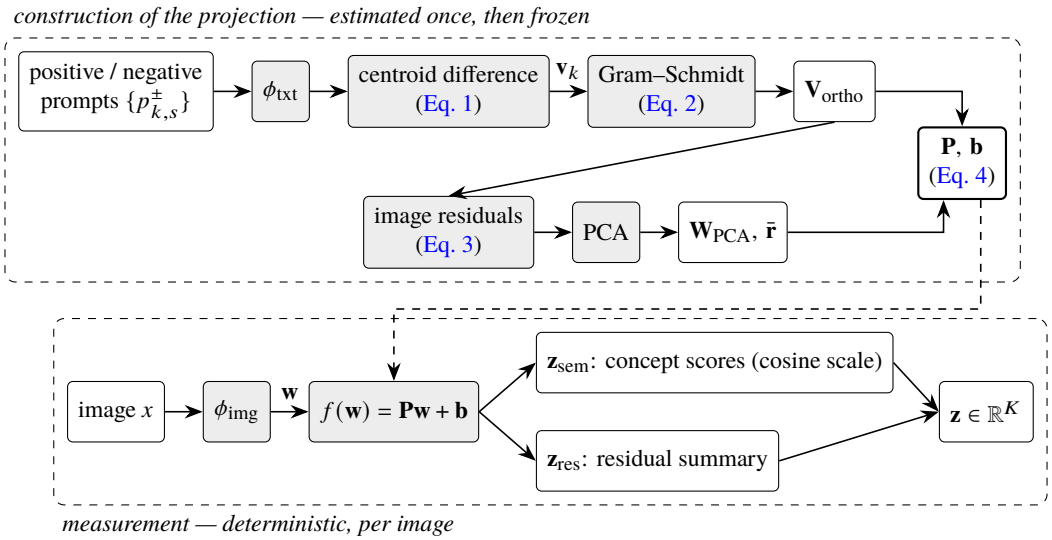

Figure 2: The proposed orthogonal projection. Top: the projection is built once and frozen as the affine map $(\mathbf{P}, \mathbf{b})$. Prompt-difference directions are orthonormalized into $\mathbf{V}_{\text{ortho}}$, and principal components of image residuals yield $\mathbf{W}_{\text{PCA}}$. Bottom: every image embedding is scored by the same map, producing interpretable concept coordinates $\mathbf{z}_{\text{sem}}$ and a compressed residual $\mathbf{z}_{\text{res}}$.

Proof. The first K_{sem} rows are orthonormal by Eq. 2. The principal components are orthonormal eigenvectors of the residual covariance; because the residuals lie in the orthogonal complement of the span of $\mathbf{V}_{\text{ortho}}$, so do the eigenvectors, hence the two blocks are mutually orthogonal. □

Proposition 2 (Exact decomposition)**.** *Every embedding splits exactly as* $\mathbf{w} = \mathbf{V}_{ortho}^{\top}\mathbf{z}_{sem} + \mathbf{r}$ *with* $\|\mathbf{w}\|_2^2 = \|\mathbf{z}_{sem}\|_2^2 + \|\mathbf{r}\|_2^2$*, and* $\mathbf{z}_{res}$ *is the variance-optimal* K_{res}*-dimensional linear summary of* $\mathbf{r}$ *over the reference corpus.*

Proof. The identity follows from Eq. 3 and orthonormality of $\mathbf{V}_{\text{ortho}}$; the Pythagorean split from the orthogonality of the two terms; the optimality statement is the standard PCA property applied to the residuals. □

Proposition 3 (Invariance to undetermined rotations of the encoder)**.** *The representation depends on the encoders only through inner products.* $f(\mathbf{w})$ *is a function of the Gram matrix of the image embedding, the prompt embeddings, and the reference-corpus embeddings. Consequently, if every embedding is transformed by a common orthogonal matrix* $\mathbf{Q} \in \mathbb{R}^{D\times D}$*, the representation is unchanged: the construction applied in the rotated space satisfies* $f'(\mathbf{Q}\mathbf{w}) = f(\mathbf{w})$*, up to the sign convention chosen for the principal components.*

Proof. Each semantic coordinate is a linear combination of inner products between $\mathbf{w}$ and prompt embeddings, with coefficients (centroid weights, Gram–Schmidt coefficients, normalizations) that are functions of prompt–prompt inner products alone. The residual block is a kernel-PCA computation: residual inner products, the residual covariance spectrum, and projections onto its eigenvectors are functions of the same Gram matrix. An orthogonal $\mathbf{Q}$ preserves every inner product, hence every coordinate. Eigenvectors are unique up to sign when the leading eigenvalues are simple, which holds generically; fixing a sign convention removes the remaining ambiguity. □

Proposition 3 matters because a contrastive objective constrains only inner products, so pretraining identifies the encoder at most up to an orthogonal transformation of $\mathbb{R}^D$. The measurement is thus a function of exactly the part of the encoder that *is* identified. This indeterminacy is a known problem. It was documented for word embeddings by Carrington et al. [5], and independently trained vision–language encoders have recently been shown to be related by a single orthogonal map [14]. Proposition 3 shows that the projection is unaffected by it. The practical consequence is replicability. The projection is fully specified by the prompt lists, the concept ordering, the reference corpus, and the name of a public encoder. Any researcher reproduces the same numbers without access to a training run. Learned measurement models lack this property. The latent coordinates of a generative encoder are identified at most up to rotation and vary with initialization in practice. A supervised score depends on proprietary labels and on the realized training run.

Remark 1 (Cosine interpretation). Since $\|\mathbf{u}_k\|_2 = \|\mathbf{w}\|_2 = 1$, each semantic coordinate $z_{\text{sem},k} = \mathbf{u}_k^{\top}\mathbf{w} \in [-1, 1]$ is the

cosine similarity between the image and the k-th orthogonalized concept axis, giving all concept scores a common, bounded scale.

Remark 2 (Distances are never inflated). Because $\mathbf{P}$ has orthonormal rows, $\|f(\mathbf{w}) - f(\mathbf{w}')\|_2 = \|\mathbf{P}(\mathbf{w} - \mathbf{w}')\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2$, with equality if and only if $\mathbf{w} - \mathbf{w}'$ lies in the span of the rows of $\mathbf{P}$. Two images are therefore never farther apart in the representation than in the embedding space. A learned nonlinear encoder has no such guarantee.

Remark 3 (Determinism). f is a fixed affine map whose construction involves no training, so R2 holds in both senses, conditional on a fixed public encoder. Any downstream result can be reproduced exactly from the prompt lists, the concept ordering, the reference corpus, and the encoder.

3.5. Cross-Entity Consistency

Applications rarely involve a single population of images. In a two-sided market, for instance, one may need representations for (i) each agent i on side A, (ii) each agent j on side B, and (iii) the outcome image produced by a matched pair (i, j). The method handles all of them with the same frozen f: entity-level embeddings (e.g., the mean CLIP embedding of an agent's image history) and outcome-level embeddings are projected identically, $\mathbf{z}_i = f(\mathbf{w}_i)$, $\mathbf{z}_j = f(\mathbf{w}_j)$, $\mathbf{z}_{ij} = f(\mathbf{w}_{ij})$.

Because coordinate k measures concept c_k on every entity type, quantities such as the coordinate-wise interaction $z_{i,k}\, z_{j,k}$ (shared expression of concept k across the two sides) or the distance $\|\mathbf{z}_i - \mathbf{z}_j\|$ are meaningful measurements. This is the property that makes the representation usable inside downstream quantitative models. A separately estimated encoder per entity type, or any stochastic encoder, would break the alignment between coordinates that these terms require.

4. Empirical Example: Brand-Personality Axes

This section works through one empirical example. We instantiate the method with a five-trait brand-personality lexicon, show what the resulting axes select on real images, and then check how sensitive the axes are to the choice of prompts and to the Gram–Schmidt ordering. The numerical checks use text encodings alone. Each prompt is encoded once, every replicate is arithmetic on the cached embeddings, and the full battery runs in seconds on a single GPU.

4.1. Setup

We instantiate the concept lexicon with the five brand-personality traits of Aaker [1] (sincerity, excitement, competence, sophistication, ruggedness), using $m = 9$ positive and 9 negative prompts per trait. The full lists appear in Appendix A. Prompts share category-neutral anchors ("a product," "a brand item") and describe visual style (lighting, palette, texture, composition) rather than product content. The encoder is OpenCLIP ViT-bigG/14 with public weights [7], a CLIP model whose image and text encoders share a 1,280-dimensional space ($D = 1{,}280$); every number in this section is therefore reproducible from the prompt lists and the encoder name alone (Proposition 3). Unless noted, axes are Gram–Schmidt orthonormalized in the order listed above, $B = 2{,}000$ resampling replicates are used, and every cosine is measured against the axes built from the full prompt lists.

4.2. What the Axes Select

We applied the projection to a proprietary corpus of sponsored Instagram posts from a micro-influencer marketing platform and scored every image on the five axes. Figure 3 shows, for each axis, two images drawn from the fifty highest-scoring images (left column) and two drawn from the fifty lowest-scoring images (right column). Scores at these extremes are cosines of roughly ±0.25 to ±0.32 (Remark 1). The selections read as the intended style contrasts. High sincerity scores go to warm linen-and-wood flat lays, low scores to saturated neon installations. High excitement scores go to vivid streamers and umbrella canopies, low scores to muted still lifes of perfume and cutlery. High ruggedness scores go to off-road gear and campfires, low scores to pastel florals and tea sets. Competence and sophistication behave likewise, with polished product arrays and gilded accessories at the top and cluttered shelves or dim snapshots at the bottom.

4.3. Stability under Prompt Resampling

The prompts are one of many phrasings the researcher could have written. To measure how much the axes depend on the particular phrasings, we resample the prompts with replacement, separately within the positive set and within the negative set of each concept, recompute Eq. 1–Eq. 2 on each replicate, and record the cosine between the resampled axis and the full-sample axis. Table 1 reports these cosines. Their means range from 0.84 (competence) to 0.91 (sincerity), with 5% quantiles between 0.76 and 0.85. A stricter check splits each prompt list into two disjoint halves (four to five positive and four to five negative prompts per concept) and compares the axes built from each half. These agree at cosines of 0.37–0.60, identifying competence as the least stable trait. Repeating the resampling with $m = 3$, 5, and 7 prompts per set shows that stability rises with the number of prompts: the mean cosine increases from 0.67–0.78 at $m = 3$ to 0.84–0.91 at $m = 9$ without saturating. The design rule "add prompts, especially for unstable concepts" is thus supported by these numbers. Finally, deleting any single prompt leaves its axis within a cosine of 0.98 of the original, so no single prompt determines an axis.

4.4. Sensitivity to the Orthogonalization Order

The main theoretical concern of Section 3.3, order dependence of Gram–Schmidt, has a negligible effect for this

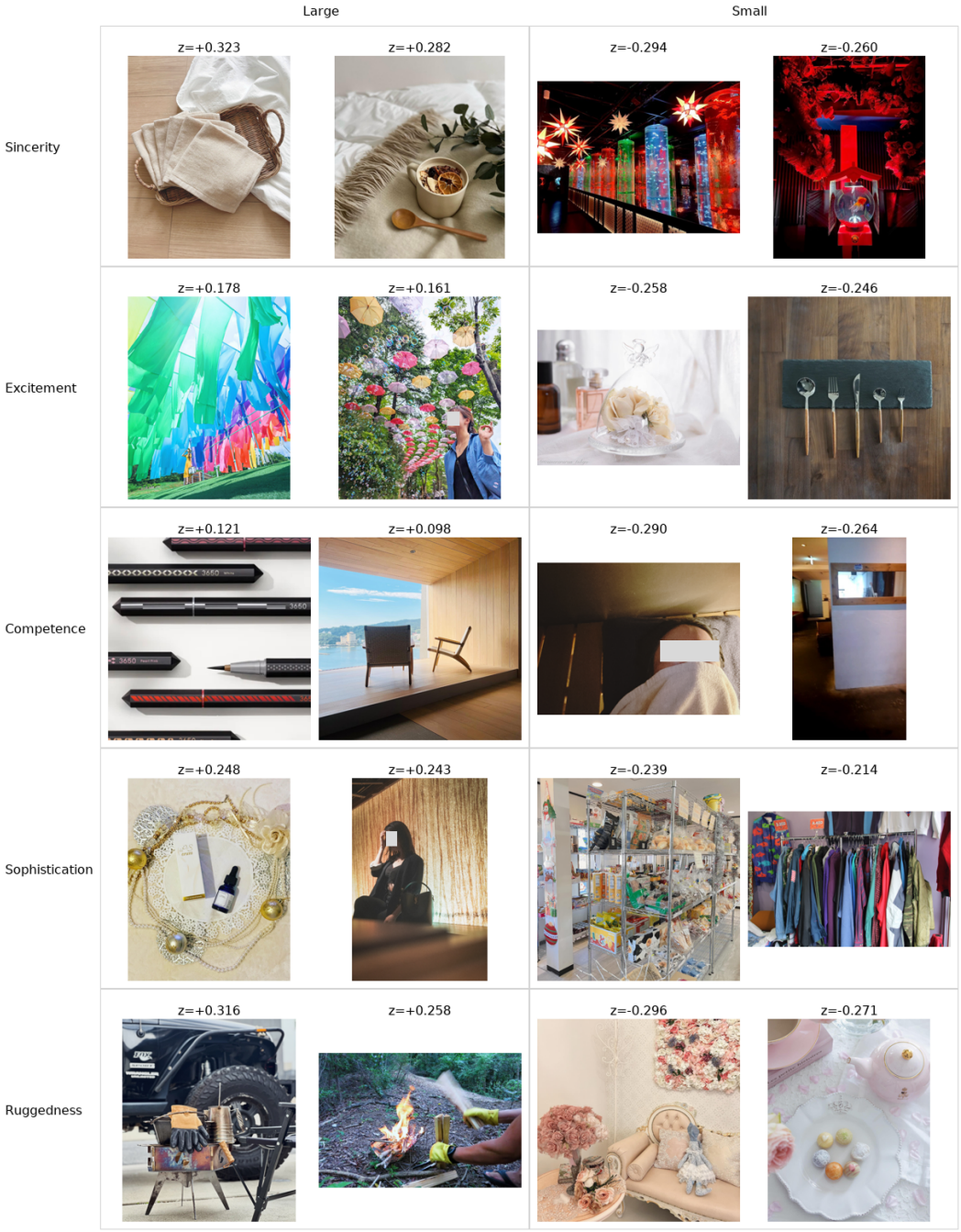

Figure 3: Highest- and lowest-scoring images on each orthogonalized brand-personality axis. Rows: the five axes of Brand Personality. Left column: two images drawn from the fifty highest scores on the axis. Right column: two drawn from the fifty lowest. The score $z_{\text{sem},k}$ is printed above each image. Images are sponsored Instagram posts from a proprietary corpus; faces are masked.

Table 1: Axis stability in the brand-personality example ($m = 9$, $B = 2{,}000$). Resampling: mean (5% quantile) cosine between resampled and full-sample axes. Split-half: mean cosine between axes from disjoint prompt halves. Ordering: worst case (mean) over all $5! = 120$ Gram–Schmidt orderings, against the canonical-order axes. LOO: worst case over single-prompt deletions.

Axis	Resampling	Split-half	Ordering	LOO
Sincerity	0.91 (0.85)	0.60	0.91 (0.95)	0.99
Excitement	0.89 (0.82)	0.53	0.92 (0.96)	0.99
Competence	0.84 (0.76)	0.37	0.97 (0.98)	0.98
Sophistication	0.86 (0.78)	0.44	0.95 (0.98)	0.99
Ruggedness	0.88 (0.82)	0.52	0.96 (0.98)	0.99

lexicon. The raw prompt-difference directions are already close to orthogonal. The largest pairwise cosine is -0.37 (sincerity–excitement), and all others are below 0.20 in magnitude, so orthogonalization is a modest correction. Recomputing the axes under all 120 possible orderings rotates no axis below a cosine of 0.91 with its canonical-order counterpart, and the mean cosine across orderings is 0.95 or higher for every axis (Table 1). This does not mean that ordering never matters; a lexicon with strongly overlapping concepts would behave differently. It means the concern is checkable. The entire ordering sweep costs a few seconds of arithmetic on cached text embeddings, and we recommend reporting it alongside any application of the method.

5. Conclusion

We proposed a deterministic, training-free construction that turns the embeddings of a pretrained vision–language encoder into interpretable measurements: zero-shot concept axes from prompt differences, enforced orthonormality, and a compressed residual that retains what the concepts miss. The construction rests on a single empirical premise, approximately linear encoding of the chosen concepts. Everything downstream of it is exact linear algebra. This simplicity is what makes the measurement properties provable. The representation is invariant to the rotations of the embedding space that contrastive pretraining leaves undetermined, so the projection is reproducible from prompt lists and a public encoder alone.

There are some limitations. If Assumption 1 fails for a concept, its coordinate no longer measures that concept. Only external validation against human judgments can detect this; our empirical example does not include human validation, and we leave it to future work.

While the empirical example found the effect of the Gram–Schmidt ordering negligible for the brand-personality lexicon, that finding is lexicon-specific. The ordering sweep should be rerun for any new concept set.

Two extensions appear most valuable. First, the construction applies verbatim to any contrastive joint-embedding encoder, so text, audio, and video measurements (e.g., via CLAP-style encoders) require only new prompts. Second, comparing axes built from different public encoders would make the choice of encoder itself a robustness check on the measurement.

References

[1] Jennifer L Aaker. Dimensions of brand personality. *Journal of marketing research*, 34(3):347–356, 1997.

[2] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[4] Alex Burnap, John R Hauser, and Artem Timoshenko. Product aesthetic design: A machine learning augmentation. *Marketing Science*, 42(6):1029–1056, 2023.

[5] Rachel Carrington, Karthik Bharath, and Simon Preston. Invariance and identifiability issues for word embeddings. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829. IEEE, 2023.

[8] Giovanni Compiani, Ilya Morozov, and Stephan Seiler. Demand estimation with text and image data. *The RAND Journal of Economics*, 2026.

[9] Philipp Dufter and Hinrich Schütze. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191, 2019.

[10] Daria Dzyabura, Siham El Kihal, John R Hauser, and Marat Ibragimov. Leveraging the power of images in managing product return rates. *Marketing Science*, 42(6):1125–1142, 2023.

[11] Jan Engler, Sandipan Sikdar, Marlene Lutz, and Markus Strohmaier. Sensepolar: Word sense aware interpretability for pre-trained contextual word embeddings. In *Findings of*

the Association for Computational Linguistics: EMNLP 2022, pages 4607–4619, 2022.

[12] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

[13] Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987, 2022.

[14] Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas K Garg. Canonicalizing multimodal contrastive representation learning. In *ICLR 2026 Workshop on Representational Alignment (Re {\textasciicircum} 4-Align)*, 2026.

[15] Justin Huang, Rupali Kaul, and Sridhar Narayanan. Novelty in content creation: Experimental results using deep learning–based image recognition on a large social network. *Marketing Science*, 45(2):335–358, 2025.

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. Pmlr, 2020.

[19] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5): 905–949, 2019.

[20] 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.

[21] Liu Liu, Daria Dzyabura, and Natalie Mizik. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4):669–686, 2020.

[22] Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of the web conference 2020*, pages 1548–1558, 2020.

[23] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

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

[25] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

[26] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[27] 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*, pages 8748–8763. PmLR, 2021.

[28] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7237–7256, 2020.

[29] Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, 2016.

[30] Ankit Sisodia, Alex Burnap, and Vineet Kumar. Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *Journal of Marketing Research*, 62(3): 405–428, 2025.

[31] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15349–15358. IEEE, 2023.

[32] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[33] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19187–19197. IEEE, 2023.

[34] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

A. Prompt Lists

This appendix reproduces the complete prompt lists of the brand-personality instantiation used in Section 4. Together with the encoder name (OpenCLIP ViT-bigG/14, `laion2b_s39b_b160k`), the concept ordering as listed, and the reference corpus, these prompts fully specify the projection (Proposition 3).

Sincerity

Positive prompts:

1. A product displayed in warm soft natural sunlight on a rustic wooden surface with linen cloth and dried wildflowers
2. A person with a genuine candid smile holding a brand item in a cozy sunlit home with warm earthy tones
3. A wholesome flat lay of a product on natural kraft paper with handwritten notes and soft diffused daylight
4. A brand item arranged on a weathered farmhouse table with fresh greenery and gentle morning golden light
5. A cheerful candid photo of a person sharing a product with friends outdoors in warm amber afternoon light
6. A product nestled in natural materials like cotton, wood, and dried plants with soft warm color palette
7. A behind-the-scenes style photo of a product in an everyday setting with honest imperfect natural lighting
8. A simple product presentation on a hand-thrown ceramic plate with soft beige and cream tones
9. A heartfelt lifestyle photo of a person using a product at home with cozy textures and warm muted pastels

Negative prompts:

1. A product photo with heavy digital retouching and artificial neon lighting with exaggerated color saturation
2. A brand item on cold polished black marble with dramatic harsh shadows and metallic gold accents
3. A product displayed under cool sterile clinical white lighting on a glossy reflective surface
4. A product shot with extreme high-contrast bold graphic elements and vibrant clashing neon colors
5. A glamorous editorial of a product with heavy stylized makeup model and opulent dark backdrop
6. A product photo with gritty industrial textures and rough weathered metal surfaces in harsh light
7. A product placed in a chaotic high-energy scene with motion blur and intense saturated color grading
8. A brand item in a sleek futuristic setting with cool blue LED lighting and glass surfaces
9. A product on a crowded flashy display with neon signs and artificial nightlife atmosphere

Excitement

Positive prompts:

1. A product displayed with bold neon pink and electric blue lighting on a vivid pop-art patterned background
2. A dynamic action shot of a person with a brand item in motion with vibrant color streaks and energy
3. A creative product flat lay with liquid splashes and vivid saturated colors on a bright geometric surface
4. A product in a trendy urban scene with colorful graffiti murals and bold diagonal composition
5. A brand item at a lively event with confetti, colorful lights, and energetic blurred crowd in background
6. A daring low-angle shot of a product against a dramatic vivid sunset sky with bold silhouettes
7. A product displayed with holographic iridescent textures and futuristic neon geometric patterns
8. A high-energy photo of a person using a product outdoors with dynamic movement and vivid warm saturation
9. A trendy overhead shot of a product surrounded by bold colorful props with strong diagonal lines and contrast

Negative prompts:

1. A product centered on a plain white background with flat even studio lighting and no styling
2. A conventional product photo on a neutral beige surface with conservative symmetrical composition
3. A quiet product arrangement with muted brown and cream tones on a library shelf in soft diffused light
4. A product on a standard display with flat fluorescent lighting and no visual dynamism
5. A simple product on a plain wooden surface with minimal styling and subdued earth-tone palette
6. A calm still-life product photo with a single item on smooth gray concrete in serene soft light
7. A product in a traditional setting with sepia tones and vintage antique props and soft vignette
8. A brand item in a peaceful pastoral scene with gentle pastel greens and quiet muted atmosphere
9. A product photo with deliberate negative space and restrained monochrome palette on matte surface

Competence

Positive prompts:

1. A product photographed with precise directional lighting and crisp sharp details on a clean white surface
2. A professional flat lay of brand items arranged in a perfect geometric grid with meticulous alignment
3. A product in a sleek modern workspace with organized accessories and clean neutral tones

4. A polished product presentation with clinical precision lighting showing fine details in sharp focus
5. A well-dressed professional person presenting a brand item in a modern minimalist office setting
6. A product displayed alongside structured data visualizations and organized layout in clean design
7. A sharp macro close-up of product craftsmanship details with controlled studio lighting on neutral background
8. A systematic before-and-after style product display with clear measured results and clean typography
9. A premium product on a glass surface with structured geometric props and cool precise lighting

Negative prompts:

1. A blurry out-of-focus product photo with uneven harsh shadows and cluttered messy background
2. A poorly lit product image with incorrect white balance taken casually in a dim room
3. A product in a chaotic disorganized scene with random scattered items and tangled clutter
4. A low-effort product selfie in a messy mirror with harsh unflattering flash and poor framing
5. A product on a wrinkled bedsheet with haphazard random objects and no intentional composition
6. A faded washed-out product photo with dull flat colors and poor exposure in ambient light
7. A product with handmade cardboard signage and uneven hand-drawn lettering on cluttered surface
8. A product photo with distracting busy background and no clear focal point or visual hierarchy
9. A hasty product snapshot with motion blur and tilted horizon on a messy countertop

Sophistication

Positive prompts:

1. A product photographed with dramatic chiaroscuro lighting on polished dark marble with golden bokeh accents
2. An elegant editorial of a brand item with a model in silk under artistic moody dramatic side lighting
3. A product arranged with velvet fabric and gold leaf details with soft warm candlelight ambiance
4. A premium brand item displayed in a grand interior with chandelier reflections and marble surfaces
5. A product on dark velvet with crystal and fresh roses in soft golden hour light and shallow depth of field
6. A refined product presentation with warm golden tones and rich luxurious textures in a classically styled setting
7. A designer brand item on Italian marble with European city bokeh lights at twilight in the background
8. A sophisticated flat lay with the product surrounded by dark satin, pearls, and muted gold tones
9. A product in an art-gallery-like setting with dramatic spotlight and dark moody negative space

Negative prompts:

1. A product in bright plastic packaging on a cluttered store shelf with fluorescent overhead lighting
2. A brand item in a plain everyday setting with basic furniture and flat unflattering daylight
3. A product on a plastic tray with disposable items and paper wrappers in a casual environment
4. A product on a simple clothesline in a suburban backyard with ordinary household background
5. A brand item in a discount aisle with harsh lighting and large bold clearance sale signage
6. A product on a basic countertop under bright convenience-store-style white fluorescent light
7. A product with rough outdoor textures like mud, gravel, and weathered surfaces in harsh sunlight
8. A brand item in a colorful cluttered children's playroom with bright primary-color plastic surfaces
9. A casual low-effort product photo with no props or styling on a plain laminate table

Ruggedness

Positive prompts:

1. A product displayed on rough weathered rock with dramatic mountain landscape and golden hour sidelight
2. A brand item on muddy terrain with heavy-duty gear and a dusty off-road vehicle in the background
3. A product on a person's wrist or hand against a rugged cliff face with raw natural stone textures
4. A product beside a campfire with pine forest, smoke, and warm firelight on rough ground
5. A brand item on a dusty trail with worn leather, canvas, and metal hardware in harsh outdoor light
6. A product photographed in rain or splashing water with gritty industrial concrete and steel textures
7. A tough product on a workshop bench with raw wood, metal tools, and heavy-duty materials
8. A brand item on a person in athletic gear at a gritty outdoor training ground with dirt and sweat
9. A product in a dense forest setting with moss-covered logs, rough bark, and dappled wilderness light

Negative prompts:

1. A product arranged with silk ribbons and rose petals on a pastel pink soft fabric surface

2. A brand item in a gentle flat lay with macarons, dried flowers, and soft pastel watercolor tones
3. A product on fine lace doily with lavender sprigs and white porcelain in delicate soft lighting
4. A brand item in a serene spa setting with white robes, orchids, and warm diffused candlelight
5. A product on a fluffy cloud-like blanket with baby pink accents and soft plush textures
6. A delicate product arrangement with cherry blossom branches and fine porcelain in gentle light
7. A brand item with tulle fabric and ballet-slipper-pink tones on a polished wooden surface
8. A product surrounded by floating flower petals in soft pastel hues with dreamy bokeh background
9. A product in a pristine manicured setting with fine crystal and gossamer fabrics in quiet warm light