SEMANTICS AS DISTRIBUTIONS: STRUCTURAL MEASUREMENT FOR VISION-LANGUAGE REPRESENTATION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantics in vision-language representation spaces is distributional: a semantic concept is expressed through a distribution of cross-modal correspondence patterns rather than a single pairwise response. How such distributional structure should enter alignment measurement, however, remains largely unexplored. Empirically, correspondence vectors collected from image-text pairs sharing the same semantic concept form reproducible and concept-dependent distributions, whose organization we term *semantic structural fingerprints*. Motivated by this observation, we introduce a decision-theoretic formulation that recasts cross-modal similarity measurement as discrimination between latent matched and mismatched correspondence distributions. By locally approximating the Bayes-optimal log-likelihood ratio to second order, we derive a principled measurement of correspondence strength and structural organization. Based on this formulation, we develop *Distributional Structural Measurement* (DSM), a computational realization that learns match- and mismatch-associated structures through signed low-rank components at $\mathcal{O}(rd)$ complexity. Experiments across architectures and retrieval benchmarks consistently demonstrate the effectiveness of DSM, while statistical analyses validate the reproducibility and distinctiveness of semantic structural fingerprints. Together, these results support a distributional view of cross-modal semantic measurement. Code will be released.

1 INTRODUCTION

Vision-language alignment learns a shared representation space in which semantically matched visual and textual content is brought together and mismatched content is separated. This capability underpins cross-modal retrieval, zero-shot recognition, and large-scale vision-language pretraining. Progress has largely been driven by stronger encoders, larger datasets, and improved learning objectives. By contrast, the measurement rule at the end of the pipeline has remained strikingly stable, i.e., most methods reduce visual-textual correspondence to a scalar score, typically cosine similarity (Radford et al., 2021; Jia et al., 2021). Although often treated as a routine final operation, this measurement not only determines how cross-modal correspondence is scored, but also shapes the learning signal used to organize the representation space.

Consider semantic content c expressed by a caption word and its attended visual evidence. Let $\mathbf{x}^c, \mathbf{y}^c \in \mathbb{R}^d$ denote the unit-normalized visual and textual features. Their coordinate-wise products $z_i^c = x_i^c y_i^c$ form the correspondence vector $\mathbf{z}^c = [z_1^c, \ldots, z_d^c]^\top$. Conventional *point-wise measurement*, such as cosine similarity, collapses this vector to the scalar response $\mathbf{1}^\top \mathbf{z}^c$. We instead consider correspondence vectors observed across instances expressing the same semantics:

$$\mathcal{Z}_c = \{\mathbf{z}^c \mid \mathbf{z}^c \text{ arises from semantic content } c\}, \qquad \mathcal{Z}_c \sim P_c.$$

Here, P_c denotes the latent distribution underlying these correspondence observations. As illustrated in Figure 1, repeated observations of the same semantic concept exhibit characteristic dimension-wise correspondence patterns, while different concepts exhibit distinct profiles. We refer to this reproducible organization as a *semantic structural fingerprint*. Figure 1 shows three representative examples; an expanded analysis of fingerprint reproducibility and distinctiveness is in Appendix A.7.

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

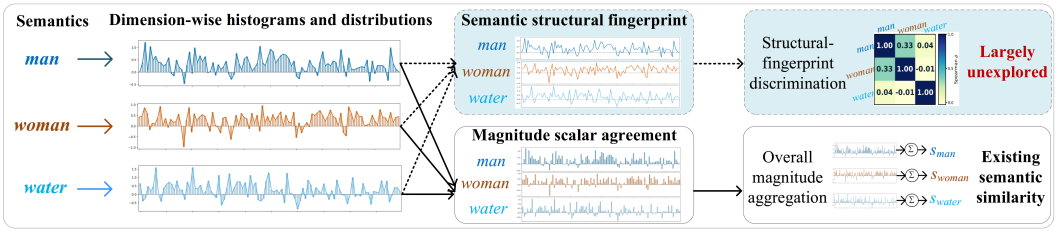

Figure 1: **Motivation for semantic structural fingerprints.** Dimension-wise correspondence observations for three representative semantics, *man*, *woman*, and *water*, exhibit characteristic profiles. Conventional scalar aggregation compresses these patterns into an overall magnitude, whereas the structural-fingerprint view retains their dimension-wise organization for semantic comparison.

This observation suggests that semantic correspondence contains information not only in its overall magnitude, but also in how evidence is organized across representation dimensions. Two semantic conditions may therefore have similar aggregate responses while exhibiting different correspondence structures. For example, $\mathbf{z}^{(1)} = [0.5, 0.5, 0, 0]^\top$ and $\mathbf{z}^{(2)} = [1, 0, 0, 0]^\top$ both yield an aggregate response of 1, although their evidence is distributed differently across dimensions. Point-wise scalar aggregation cannot preserve this distinction. More generally, correspondence distributions may exhibit similar first-order evidence but different structural organization, suggesting that structural information can provide complementary discriminative evidence beyond aggregate similarity.

This leads to a central question: *how should cross-modal similarity be measured when semantic evidence exhibits distributional structure?* As illustrated in Figure 2, conventional *point-wise measurement* aggregates correspondence evidence into a scalar score, whereas our *distributional structural measurement* additionally captures how such evidence is structurally organized. We formulate alignment measurement as discrimination between matched and mismatched correspondence distributions. A local second-order approximation of the Bayes-optimal log-likelihood ratio yields a measurement containing both first-order correspondence evidence and second-order structural evidence. We further prove that when two correspondence distributions have equal means but different second-order moments, any linear measurement has the same expected response, whereas even a rank-one structural response can distinguish them.

Based on this analysis, we propose **Distributional Structural Measurement (DSM)**. DSM realizes the second-order term through complementary positive and negative low-rank structural responses: the positive branch captures structures associated with matched correspondence, while the negative branch captures structures associated with mismatched correspondence. Together with first-order evidence, these responses form an efficient structural measurement with $\mathcal{O}(rd)$ complexity, where $r \ll d$. Importantly, the latent distributions provide the theoretical basis for deriving the measurement; DSM does not explicitly estimate these distributions during training, but learns directly from matched and mismatched correspondence observations through the alignment objective.

Our contributions are fourfold: (1) We identify *semantic structural fingerprints*, revealing reproducible and concept-dependent structure in cross-modal correspondence beyond conventional point-wise similarity. (2) We formulate cross-modal measurement as matched-mismatched distribution discrimination and derive a principled first- and second-order measurement with a distinguishability guarantee. (3) We propose DSM, an efficient signed low-rank structural measurement that combines first-order evidence with complementary positive and negative structural responses at $\mathcal{O}(rd)$ cost. (4) Extensive experiments show consistent gains across backbones, datasets, and zero-shot tasks, while controlled ablations, fingerprint analyses, and near-tie evaluations provide converging evidence for the role of structural information beyond point-wise evidence.

2 Related Work

Vision-language alignment. Its goal is to learn representations in which matched visual and textual content receives higher similarity than mismatched content. Early visual-semantic embedding methods optimize global image-sentence representations with ranking objectives (Faghri et al., 2018), while subsequent work improves global aggregation (Chen et al., 2021). Large-scale contrastive pretraining, exemplified by CLIP and ALIGN, further establishes the effectiveness of dual-encoder alignment (Radford et al., 2021; Jia et al., 2021). Fine-grained methods model region-word or patch-

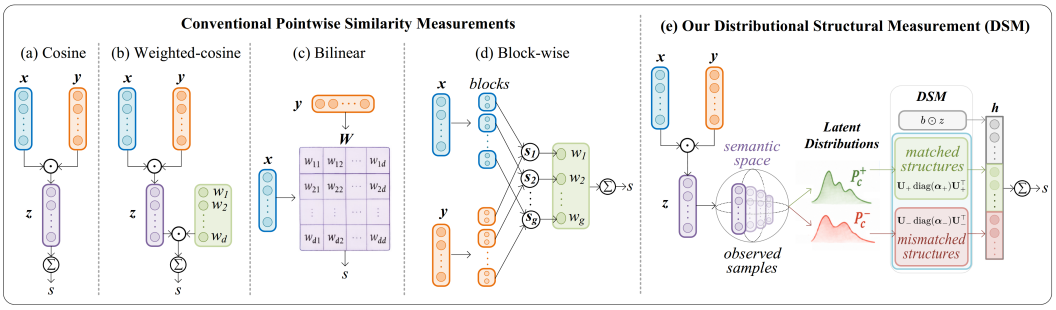

Figure 2: Comparison of conventional point-wise methods and our structural measurement DSM.

word correspondence through cross-attention, recurrent matching, graph reasoning, and negative-aware interaction (Lee et al., 2018; Chen et al., 2020; Liu et al., 2020; Diao et al., 2021; Zhang et al., 2022). Recent approaches further improve alignment through dense semantic supervision, selective patch interaction, set-based matching, graded relevance modeling, and semantic calibration (Liu et al., 2025a; Mao et al., 2026; Alomari et al., 2025; Jang et al., 2025; Lu & Feng, 2026; Kang et al., 2025). Overall, prior work primarily advances representations, interactions, or supervision, while cross-modal correspondence is still typically reduced to a scalar similarity. We instead study how this correspondence itself should be measured.

Cross-modal semantic measurement. Most vision-language methods adopt *point-wise* measurement, aggregating coordinate-wise correspondence between visual and textual features into a scalar similarity, typically through inner product, cosine similarity, Euclidean distance, or weighted variants (Radford et al., 2021; Jia et al., 2021; Lee et al., 2018; Zhang et al., 2022). Several studies relax this formulation by introducing richer measurement units or aggregation rules. GSMN partitions representations into feature blocks and learns to combine block-wise similarities (Liu et al., 2020; Yang et al., 2022). ESL further generalizes fixed single-dimensional measure units into learnable multi-dimensional units, enabling similarity to be examined at richer granularities (Zhang et al., 2024). These methods enrich the measurement of an individual visual-textual correspondence, but do not explicitly characterize the structural fingerprint implied by semantic correspondence distributions. DSM addresses this gap by deriving a structural measurement from that distributional view and realizing it with complementary positive and negative structural responses.

3 Distributional Structural Measurement

Our goal is to derive a semantic similarity measure that captures the structural organization of cross-modal correspondence under a distributional view of semantics. We first formalize semantic correspondence distributions and derive the measurement from optimal discrimination between matched and mismatched observations. We then establish its discriminative capacity, develop the signed low-rank DSM parameterization, and describe its use in image-text alignment. Algorithm 1 summarizes the computation, with detailed proofs in Appendices B-D.

3.1 Correspondence distributions and structural fingerprints

Definition 1 (Fine-grained semantic input). To characterize structural differences associated with identifiable semantic content, we use caption-word features rather than globally pooled sentence representations, which may obscure concept-specific correspondence patterns. Let $\{\mathbf{t}_k\}_{k=1}^{L_T} \subset \mathbb{R}^d$ denote unit-normalized word features of caption T, with $\|\mathbf{t}_k\|_2 = 1$, and $\{\mathbf{f}_j(I)\}_{j=1}^{N_I} \subset \mathbb{R}^d$ the region or patch features of image I. For word k, we obtain its conditioned visual representation as $\mathbf{v}_k(I) = \sum_{j=1}^{N_I} \beta_{kj}\mathbf{f}_j(I)$, where $\beta_{kj} = \mathrm{softmax}_j\big(\mathbf{t}_k^\top \mathbf{f}_j(I)\big)$ measures the relevance of visual element j to word k. Its unit-normalized form is denoted by $\widehat{\mathbf{v}}_k(I)$.

Definition 2 (Correspondence distributions). For an image-caption pair (I, T) and word position k, define the *correspondence vector* as:

$$\mathbf{z}_k = \widehat{\mathbf{v}}_k(I) \odot \mathbf{t}_k \in \mathbb{R}^d, \qquad z_{k,j} = \widehat{v}_{k,j}(I)t_{k,j}. \tag{1}$$

Each coordinate measures agreement along one representation dimension, and cosine similarity is recovered by $\mathbf{1}^\top \mathbf{z}_k$. Let c denote the semantic meaning expressed by a word occurrence, and let $\mathbf{Z}_c$ collect correspondence vectors from occurrences of c across image-text pairs. We define the matched and mismatched correspondence distributions as:

$$P_c^+ = \mathcal{L}(\mathbf{Z}_c \mid M = 1, c), \qquad P_c^- = \mathcal{L}(\mathbf{Z}_c \mid M = 0, c), \tag{2}$$

where $M = 1$ and $M = 0$ denote matched and mismatched observations, respectively.

Definition 3 (Semantic structural fingerprint). For theoretical analysis, let $\boldsymbol{\mu}_c^\pm = \mathbb{E}_{P_c^\pm}[\mathbf{Z}_c]$ and $\mathbf{G}_c^\pm = \mathbb{E}_{P_c^\pm}[\mathbf{Z}_c \mathbf{Z}_c^\top]$ denote the mean and second-order matrix of $P_c^\pm$, respectively. We characterize the *semantic structural fingerprint* of concept c by the contrasts:

$$\Delta\boldsymbol{\mu}_c = \boldsymbol{\mu}_c^+ - \boldsymbol{\mu}_c^-, \qquad \Delta\mathbf{G}_c = \mathbf{G}_c^+ - \mathbf{G}_c^-. \tag{3}$$

Here, $\Delta\boldsymbol{\mu}_c$ captures changes in average coordinate responses, while $\Delta\mathbf{G}_c$ captures second-order structural information, i.e., its diagonal entries reflect changes in individual response strength, and its off-diagonal entries reflect how different correspondence dimensions vary jointly. Together, they characterize both the average response and the structural organization of the latent distributions without assuming a specific distributional form.

3.2 A MEASUREMENT DERIVED FROM DISTRIBUTION DISCRIMINATION

We next ask what score best separates matched and mismatched correspondence patterns for a given semantic condition c. Let $p_c^+(\mathbf{z})$ and $p_c^-(\mathbf{z})$ denote the densities of P_c^+ and P_c^-, respectively, and let $\pi_c^+ = \Pr(M = 1 \mid c)$ and $\pi_c^- = \Pr(M = 0 \mid c)$ denote their prior probabilities.

Proposition 1 (Bayes-optimal correspondence score). For the binary task of determining whether $\mathbf{z}$ comes from the matched or mismatched distribution, Bayes decision theory gives the optimal discriminative score:

$$\ell_c^*(\mathbf{z}) = \log \frac{p_c^+(\mathbf{z})}{p_c^-(\mathbf{z})} + \log \frac{\pi_c^+}{\pi_c^-}. \tag{4}$$

The first term compares the likelihood of $\mathbf{z}$ under the two distributions, while the second accounts for their prior probabilities. For fixed c, the prior term is constant and therefore does not affect the ordering of correspondence patterns. Appendix B gives the derivation. Since the densities are unknown, we derive a tractable local approximation of ℓ_c^* around a reference pattern $\mathbf{z}_0$, assuming positive densities and continuous third derivatives nearby.

Proposition 2 (Local second-order measurement). A second-order Taylor expansion around $\mathbf{z}_0$, followed by collecting terms in $\mathbf{z}$, yields:

$$\ell_c^*(\mathbf{z}) = a + \mathbf{b}^\top \mathbf{z} + \mathbf{z}^\top \mathbf{A} \mathbf{z} + \mathcal{O}(\|\mathbf{z} - \mathbf{z}_0\|_2^3), \qquad \mathbf{A} = \mathbf{A}^\top. \tag{5}$$

Here, $\mathbf{A} = \frac{1}{2}\nabla^2 \ell_c^*(\mathbf{z}_0)$. The linear term captures evidence from individual dimensions, while the quadratic term captures their interactions, which may either support or oppose a match. This motivates the signed structural measurement used in DSM. Details are provided in Appendix B.

3.3 A DISTINGUISHABILITY GUARANTEE

We next show when second-order structure distinguishes correspondence patterns that linear scores cannot. Let m_1 and m_2 denote two semantic correspondence conditions with distributions P_{m_1} and P_{m_2}. For each condition m, let $\boldsymbol{\mu}_m = \mathbb{E}_{P_m}[\mathbf{Z}]$ and $\mathbf{G}_m = \mathbb{E}_{P_m}[\mathbf{Z}\mathbf{Z}^\top]$ denote its mean and second-order matrix.

Theorem 1 (Second-order distinguishability). If $\boldsymbol{\mu}_{m_1} = \boldsymbol{\mu}_{m_2}$ but $\mathbf{G}_{m_1} \neq \mathbf{G}_{m_2}$, then any linear score $\mathbf{b}^\top \mathbf{Z}$ has the same expectation. In contrast, there exists a unit vector $\mathbf{u}$ such that:

$$\mathbb{E}_{P_{m_1}}[(\mathbf{u}^\top \mathbf{Z})^2] - \mathbb{E}_{P_{m_2}}[(\mathbf{u}^\top \mathbf{Z})^2] = \mathbf{u}^\top (\mathbf{G}_{m_1} - \mathbf{G}_{m_2}) \mathbf{u} \neq 0. \tag{6}$$

The reason is simple, that is, the expectation of a linear score depends only on the mean, so equal means yield equal expected linear responses. However, different second-order matrices imply that some direction $\mathbf{u}$ has different response energy under the two conditions. Thus, even a rank-one structural response can distinguish patterns that linear scores cannot separate in expectation. Appendix C provides the full proof.

Input: matched minibatch $\{(I_i, T_i)\}_{i=1}^{B}$; encoders; DSM parameters $\mathbf{b}, \mathbf{U}_\pm, \boldsymbol{\alpha}_\pm$; margin γ.
Output: similarities $S(I_i, T_j)$ and loss $\mathcal{L}_{\text{trip}}$.

1. **Encode:** compute image region/patch features $\{\mathbf{f}_{i,p}\}_p$ and caption word features $\{\mathbf{t}_{j,k}\}_k$.
2. **Form correspondence:** for each word, attend to $\{\mathbf{f}_{i,p}\}_p$ with $\mathbf{t}_{j,k}$, obtain $\mathbf{z}_{i,j,k} = \widehat{\mathbf{v}}_{i,j,k} \odot \mathbf{t}_{j,k}$.
3. **DSM structural responses:** $\mathbf{q}_\pm \leftarrow \boldsymbol{\alpha}_\pm \odot (\mathbf{U}_\pm^\top \mathbf{z}_{i,j,k})^{\odot 2}$.
4. **DSM score:** form $\mathbf{h} \leftarrow [\mathbf{b} \odot \mathbf{z}_{i,j,k}; \mathbf{q}_+; -\mathbf{q}_-]$ and compute $s_{i,j,k}$ by Eq. (10).
5. **Aggregate and learn:** average $s_{i,j,k}$ over caption words to obtain $S(I_i, T_j)$; select valid in-batch negatives and optimize Eq. (11).

Algorithm 1: Distributional Structural Measurement Training Framework.

3.4 Signed low-rank structural measurement

Proposition 2 identifies $\mathbf{z}^\top \mathbf{A}\mathbf{z}$ as the structural component of the locally derived measurement, while Theorem 1 establishes its ability to distinguish second-order differences. Although derived under a semantic condition c, DSM shares this functional form across semantics, with semantic context carried by the word-conditioned correspondence vector $\mathbf{z}$. For semantic condition c, let $\Delta\mathbf{G}_c = \mathbf{G}_c^+ - \mathbf{G}_c^-$ denote the second-order contrast between matched and mismatched observations. Using $\mathbb{E}[\mathbf{Z}^\top \mathbf{A}\mathbf{Z}] = \operatorname{tr}(\mathbf{A}\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top])$, the average separation induced by the quadratic component is:

$$\mathcal{J}_c(\mathbf{A}) = \mathbb{E}_{P_c^+}\left[\mathbf{Z}_c^\top \mathbf{A}\mathbf{Z}_c\right] - \mathbb{E}_{P_c^-}\left[\mathbf{Z}_c^\top \mathbf{A}\mathbf{Z}_c\right] = \operatorname{tr}(\mathbf{A}\Delta\mathbf{G}_c). \tag{7}$$

Positive spectral directions of $\Delta\mathbf{G}_c$ contribute more strongly to matched observations, whereas negative directions respond more strongly to mismatched observations. Both therefore provide complementary evidence for alignment measurement.

Motivated by this signed contrast, DSM represents $\mathbf{A}$ as the difference of two positive-semidefinite low-rank components:

$$\mathbf{A} = \mathbf{A}_+ - \mathbf{A}_-, \qquad \mathbf{A}_\pm = \mathbf{U}_\pm \operatorname{diag}(\boldsymbol{\alpha}_\pm)\mathbf{U}_\pm^\top, \qquad \boldsymbol{\alpha}_\pm \geq \mathbf{0}, \qquad r = r_+ + r_- \ll d. \tag{8}$$

Here, $\mathbf{U}_\pm = [\mathbf{u}_{\pm,1}, \ldots, \mathbf{u}_{\pm,r_\pm}] \in \mathbb{R}^{d \times r_\pm}$ contains the learned structural directions, and $\boldsymbol{\alpha}_\pm$ contains their nonnegative weights. Combining this structural response with the linear term derived in Eq. (5) gives:

$$g(\mathbf{z}) = \mathbf{b}^\top \mathbf{z} + \sum_{\ell=1}^{r_+} \alpha_{+,\ell}(\mathbf{u}_{+,\ell}^\top \mathbf{z})^2 - \sum_{\ell=1}^{r_-} \alpha_{-,\ell}(\mathbf{u}_{-,\ell}^\top \mathbf{z})^2. \tag{9}$$

The positive branch increases the score for match-associated structures, while the negative branch suppresses mismatch-associated structures.

To control the magnitude of the combined responses during training, we collect its component-wise contributions into $\mathbf{h}(\mathbf{z}) = [\mathbf{b} \odot \mathbf{z};\ \boldsymbol{\alpha}_+ \odot (\mathbf{U}_+^\top \mathbf{z})^{\odot 2};\ -\boldsymbol{\alpha}_- \odot (\mathbf{U}_-^\top \mathbf{z})^{\odot 2}] \in \mathbb{R}^{d+r}$, where $(\cdot)^{\odot 2}$ denotes element-wise squaring. Since $g(\mathbf{z}) = \mathbf{1}^\top \mathbf{h}(\mathbf{z})$, we define the normalized DSM score as:

$$s_{\text{DSM}}(\mathbf{z}) = \sqrt{\frac{d}{d+r}}\, \frac{g(\mathbf{z})}{\|\mathbf{h}(\mathbf{z})\|_2 + \varepsilon}, \qquad \varepsilon > 0. \tag{10}$$

The numerator $g(\mathbf{z})$ follows the derived first- and second-order form, whereas the denominator is an implementation-level scale stabilization and is not part of the local Bayes approximation. The factor $\sqrt{d/(d+r)}$ keeps the score bound independent of r. Indeed, by Cauchy-Schwarz, $|g(\mathbf{z})| \leq \sqrt{d+r}\,\|\mathbf{h}(\mathbf{z})\|_2$, and therefore $|s_{\text{DSM}}(\mathbf{z})| < \sqrt{d}$.

DSM costs $\mathcal{O}(rd)$ per correspondence, lying between the $\mathcal{O}(d)$ cost of cosine-style measurement and the $\mathcal{O}(d^2)$ cost of a full quadratic form. With $r \ll d$, it retains linear scaling in the representation dimension. In implementation, $\mathbf{U}_\pm$, $\boldsymbol{\alpha}_\pm$, and $\mathbf{b}$ are optimized end-to-end through the alignment objective (see Algorithm 1); they induce $\mathbf{A}_\pm$ and the effective matrix $\mathbf{A} = \mathbf{A}_+ - \mathbf{A}_-$. Proposition 3 in Appendix D provides further theoretical support for this signed low-rank form.

Table 1: Controlled comparison of similarity measurements on Flickr30K and MS-COCO. Within each backbone, all methods follow the same pipeline in Algorithm 1 and differ only in the measurement function; bold indicates column maxima and shaded rows denote DSM.

Backbone	Method	Flickr30K							MS-COCO 1K (5-fold mean)						
		Image-to-Text			Text-to-Image			Rsum	Image-to-Text			Text-to-Image			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Region+BERT	Cosine	76.2	94.3	97.7	58.8	82.8	89.0	498.8	77.2	97.2	99.3	64.2	91.4	96.4	525.7
	Weighted cosine	79.4	95.7	98.2	62.1	85.2	91.1	511.7	80.8	97.6	99.5	65.0	91.7	96.6	531.2
	Bilinear (full)	82.9	95.7	97.7	64.2	86.7	92.3	519.5	84.1	98.3	**99.8**	68.6	93.4	**97.7**	541.9
	Bilinear ($r = 32$)	81.2	95.3	97.6	58.4	82.8	88.7	504.0	82.1	97.7	**99.8**	65.9	92.6	97.3	535.4
	Block-wise	82.9	95.2	98.1	61.2	84.4	90.8	512.6	83.9	97.9	99.7	68.3	93.3	97.4	540.5
	DSM ($r = 32$)	**85.7**	**96.8**	98.3	**65.4**	87.7	**92.7**	**526.7**	**86.3**	98.1	99.7	70.3	93.4	**97.7**	545.5
	DSM (full)	85.0	96.4	**98.4**	65.2	**87.8**	**92.7**	525.4	86.2	**98.5**	99.7	**70.7**	**93.7**	97.5	**546.3**
Swin-224+BERT	Cosine	82.6	96.9	99.1	70.1	90.8	94.6	534.1	81.9	97.3	99.4	69.6	92.8	96.9	537.8
	Weighted cosine	83.3	97.5	98.8	70.8	91.6	95.1	537.2	84.5	97.6	99.4	71.6	93.4	97.1	543.7
	Bilinear (full)	87.8	98.4	**99.4**	72.7	92.9	96.2	547.4	85.9	97.8	99.3	72.7	93.9	97.5	546.9
	Bilinear ($r = 32$)	81.4	96.7	98.5	68.5	91.5	95.1	531.7	78.5	95.9	98.6	65.2	91.5	96.4	526.1
	Block-wise	78.8	96.2	98.5	67.2	90.4	94.8	525.9	82.1	97.5	99.3	69.6	93.4	97.4	539.3
	DSM ($r = 32$)	**89.8**	**98.6**	99.3	**75.4**	**93.3**	**96.3**	**552.6**	**86.9**	98.1	**99.5**	**73.6**	**94.2**	97.6	**549.8**
	DSM (full)	88.6	97.9	98.9	74.6	93.2	96.2	549.4	86.6	**98.2**	**99.5**	73.3	94.1	**97.7**	549.5
Swin-384+BERT	Cosine	84.0	97.4	99.3	72.2	92.5	96.0	541.3	81.9	97.4	99.3	71.0	93.5	97.2	540.3
	Weighted cosine	84.8	97.8	99.3	72.3	92.9	96.3	543.4	83.1	97.9	99.4	71.5	94.0	97.6	543.6
	Bilinear (full)	87.4	98.3	99.2	75.8	94.1	97.0	551.8	85.8	98.1	99.5	73.6	94.5	97.9	549.4
	Bilinear ($r = 32$)	77.1	95.9	98.5	66.1	90.6	94.7	522.9	83.4	97.7	99.3	69.9	93.7	97.6	541.6
	Block-wise	81.7	96.3	98.9	69.8	92.4	95.6	534.7	84.9	97.7	99.4	72.5	94.3	97.8	546.6
	DSM ($r = 32$)	**92.0**	**99.3**	**99.9**	**79.5**	**95.9**	**97.9**	**564.5**	**88.2**	**98.5**	**99.6**	**75.6**	**95.0**	**98.1**	**555.0**
	DSM (full)	91.0	99.2	99.8	79.4	95.4	97.8	562.7	87.9	**98.5**	99.5	75.3	94.8	98.0	554.1
CLIP ViT-B/16 fine-tuning	Cosine	89.9	**99.1**	99.8	77.7	94.0	96.9	557.3	83.4	97.0	99.0	71.1	93.0	97.2	540.7
	Weighted cosine	89.4	98.7	**99.9**	78.3	94.3	96.8	557.4	83.3	96.9	99.0	71.2	93.2	97.2	540.8
	Bilinear (full)	91.4	98.9	99.8	78.8	94.6	96.9	560.4	84.7	96.9	99.0	71.6	93.1	**97.4**	542.7
	Bilinear ($r = 32$)	88.4	98.5	99.6	77.3	93.5	96.3	553.7	83.5	**97.3**	99.0	70.8	93.0	97.3	541.0
	Block-wise	81.7	97.8	99.1	69.9	90.5	94.2	533.2	82.1	96.6	98.9	69.5	92.0	96.3	535.4
	DSM ($r = 32$)	**92.1**	99.0	99.8	79.1	94.8	97.2	562.1	**85.7**	97.2	99.1	**72.0**	**93.3**	**97.4**	**544.7**
	DSM (full)	**92.1**	99.0	99.8	**79.3**	**95.1**	**97.3**	**562.6**	85.5	**97.3**	**99.2**	71.9	93.2	**97.4**	544.5
CLIP ViT-L/14 fine-tuning	Cosine	94.0	99.3	**100.0**	82.4	96.7	98.4	570.8	86.2	97.6	99.2	74.7	94.4	97.8	549.9
	Weighted cosine	93.3	99.4	99.9	81.7	96.5	98.3	569.1	86.7	97.8	**99.4**	75.2	94.6	**98.0**	551.7
	Bilinear (full)	93.4	99.5	99.8	82.6	96.7	98.3	570.3	87.7	97.9	99.3	75.8	**94.8**	**98.0**	553.5
	Bilinear ($r = 32$)	92.2	99.5	99.9	81.7	96.5	98.2	568.1	87.6	**98.1**	99.2	75.2	94.7	97.9	552.8
	Block-wise	90.8	99.0	99.7	80.0	96.0	97.9	563.3	86.5	**98.1**	**99.4**	74.5	94.5	97.8	550.7
	DSM ($r = 32$)	**95.1**	**99.7**	**100.0**	84.4	**97.0**	**98.7**	574.9	**88.7**	98.0	99.2	**76.1**	**94.8**	**98.0**	**554.8**
	DSM (full)	94.9	**99.7**	**100.0**	**85.0**	**97.0**	98.5	**575.1**	88.1	97.9	99.3	**76.1**	**94.8**	**98.0**	554.2

3.5 Image-text similarity and learning objective

For a candidate pair (I, T), DSM yields a score for each valid caption word, and the image-text similarity is defined as their arithmetic mean, i.e., $S(I,T) = \frac{1}{L_T}\sum_{k=1}^{L_T} s_{\mathrm{DSM}}(\mathbf{z}_k)$. For a minibatch of matched pairs $\{(I_i, T_i)\}_{i=1}^{B}$, let T_i^- and I_i^- denote a caption and an image that do not match I_i and T_i, respectively. With margin $\gamma > 0$ and $[u]_+ = \max(u, 0)$, we optimize:

$$\mathcal{L}_{\mathrm{trip}} = \frac{1}{B}\sum_{i=1}^{B}\Big([\gamma - S(I_i,T_i) + S(I_i,T_i^-)]_+ + [\gamma - S(I_i,T_i) + S(I_i^-,T_i)]_+\Big). \tag{11}$$

Note that P_c^+ and P_c^- are latent constructs that motivate the measurement and need not be explicitly sampled or estimated. Training observes matched and mismatched correspondence vectors $\mathbf{z}$ within each minibatch and learns the shared DSM parameters directly through the ranking objective.

4 Experiments

To evaluate the effectiveness of DSM, we organize the experiments around three questions. **Q1** *asks whether DSM improves vision-language alignment*. **Q2** *asks whether the gains arise from structural measurement*. **Q3** *asks whether DSM exploits structural information as predicted*. Sections 4.2-4.4 address these questions in turn, followed by a qualitative view of the learned correspondence space in Section 4.5. Overall, DSM improves alignment consistently across backbones and transfer settings; the gains are driven primarily by low-rank structural responses that remain informative where point-wise evidence is insufficient.

4.1 Experimental Setup

We evaluate image-text retrieval on Flickr30K and MS-COCO using Recall@K ($K \in \{1, 5, 10\}$) and Rsum, with the standard five-fold 1K protocol on MS-COCO. Experiments cover Re-

Table 2: Comparison with specialized image-text retrieval systems on Flickr30K and MS-COCO. Results are grouped by backbone; bold indicates column maxima and shaded rows DSM.

Method	Flickr30K							MS-COCO 1K (5-fold mean)						
	Image-to-Text			Text-to-Image			Rsum	Image-to-Text			Text-to-Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Region + BERT														
ESL (TCSVT 2024) (Zhang et al., 2024)	83.4	95.8	97.9	64.5	88.1	92.7	522.4	82.2	96.9	98.9	66.5	92.1	96.7	533.4
MaxMatch (ACL 2025) (Alomari et al., 2025)	84.2	96.1	97.9	63.2	87.3	92.2	520.8	83.0	96.9	98.9	66.4	91.9	96.6	533.8
AdaCL (ICLR 2025) (Pan et al., 2025)	82.0	96.5	98.2	**65.6**	**88.3**	**93.3**	523.9	82.1	97.2	98.9	67.9	92.5	97.5	536.1
DSM ($r = 32$)	**85.7**	**96.8**	98.3	65.4	87.7	92.7	**526.7**	**86.3**	98.1	**99.7**	70.3	93.4	**97.7**	545.5
DSM (full)	85.0	96.4	**98.4**	65.2	87.8	92.7	525.4	86.2	**98.5**	**99.7**	**70.7**	**93.7**	97.5	**546.3**
Swin-224 + BERT														
AVSE (AAAI 2025) (Liu et al., 2025b)	83.9	97.4	99.4	70.0	92.4	95.6	538.7	84.9	98.0	99.3	72.1	94.0	97.4	545.7
D2S-VSE (ICCV 2025) (Liu et al., 2025a)	87.2	98.4	**99.9**	73.0	**93.5**	**96.7**	548.7	82.4	97.6	99.3	70.3	93.7	97.4	540.7
CDDS (AAAI 2026) (Ma et al., 2026)	83.0	98.1	99.7	70.3	92.1	95.8	539.0	84.6	97.9	99.4	72.1	93.5	97.6	545.1
MG-Net (AAAI 2026) (Ji & Liu, 2026)	87.8	**98.6**	99.7	72.8	92.6	95.7	547.2	84.4	97.9	**99.5**	73.1	93.3	97.2	545.4
DSM ($r = 32$)	**89.8**	**98.6**	99.3	**75.4**	93.3	96.3	**552.6**	**86.9**	98.1	**99.5**	**73.6**	**94.2**	97.6	**549.8**
DSM (full)	88.6	97.9	98.9	74.6	93.2	96.2	549.4	86.6	**98.2**	**99.5**	73.3	94.1	**97.7**	549.5
Swin-384 + BERT														
AVSE (AAAI 2025) (Liu et al., 2025b)	87.1	98.3	99.2	73.6	93.5	96.5	548.2	85.1	98.2	99.5	71.6	94.0	97.5	545.9
D2S-VSE (ICCV 2025) (Liu et al., 2025a)	87.8	99.0	99.7	75.7	94.1	96.9	553.2	83.8	97.9	99.4	71.9	94.2	97.9	544.7
CDDS (AAAI 2026) (Ma et al., 2026)	86.8	98.3	99.6	76.3	94.3	97.2	552.5	84.9	98.0	**99.6**	73.5	94.6	98.0	548.6
MG-Net (AAAI 2026) (Ji & Liu, 2026)	89.1	**99.5**	**100.0**	76.4	94.4	97.0	556.4	84.6	**98.6**	99.5	73.3	94.0	97.5	546.8
DSM ($r = 32$)	**92.0**	99.3	99.9	**79.5**	**95.9**	**97.9**	**564.5**	**88.2**	98.5	**99.6**	**75.6**	**95.0**	**98.1**	**555.0**
DSM (full)	91.0	99.2	99.8	79.4	95.4	97.8	562.7	87.9	98.5	99.5	75.3	94.8	98.0	554.1

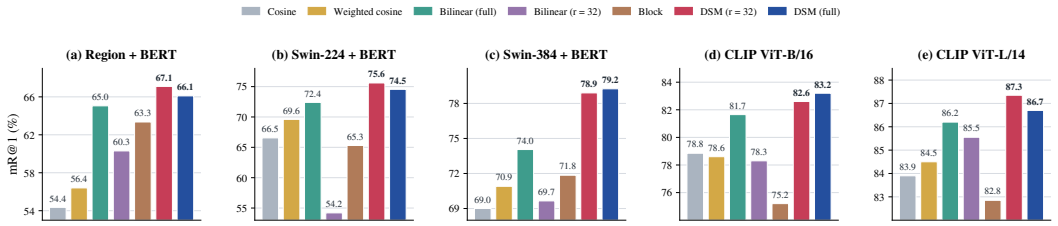

Figure 3: MS-COCO→Flickr30K transfer measured by bidirectional mean R@1 (mR@1).

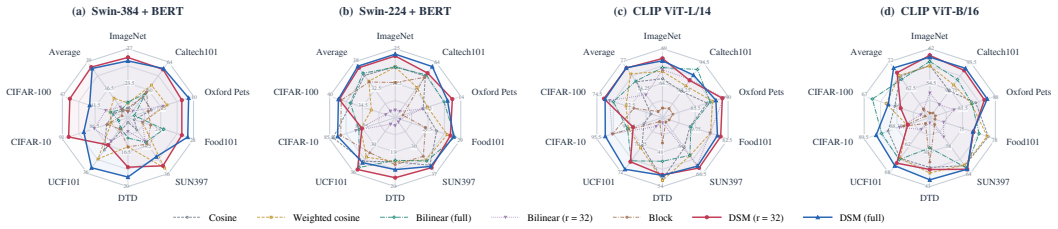

Figure 4: Zero-shot classification on nine benchmarks across four backbones.

gion+BERT, Swin-224/384+BERT, and CLIP ViT-B/16 and ViT-L/14, comparing Cosine, Weighted cosine, full and rank-32 Bilinear, Block-wise interaction, and DSM under the same representation pipeline. Unless otherwise stated, DSM uses $r = 32$, while DSM (full) uses $r = d$. Generalization is evaluated by MS-COCO→Flickr30K transfer and zero-shot classification on nine benchmarks. Full experimental details are provided in Appendix A.1.

4.2 Q1: Does DSM Improve Vision-Language Alignment?

Controlled comparison with point-wise measurements. Table 1 isolates the contribution of the measurement function by keeping the representation pipeline fixed within each backbone. DSM consistently improves vision-language alignment, with average **R@1** gains of **4.64** percentage points over Cosine and **1.69** points over the strongest conventional alternative across architectures, datasets, and retrieval directions. The relatively smaller gains on CLIP are consistent with its representations having already been shaped by cosine-based contrastive pretraining (Radford et al., 2021), leaving less room for further improvement through measurement-level fine-tuning.

Comparison with specialized retrieval systems. Table 2 provides a complementary system-level comparison. Unlike the controlled study in Table 1, the compared methods may introduce dedicated alignment modules, auxiliary supervision, or specialized training strategies. This comparison

Table 3: Component ablations and normalization control for Region+BERT DSM ($r = 32$). Bold denotes the best result.

Method	Flickr30K							MS-COCO 1K (5-fold mean)						
	Image-to-Text			Text-to-Image			Rsum	Image-to-Text			Text-to-Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
DSM w/o second-order	79.4	95.7	98.2	62.1	85.2	91.1	511.7	80.8	97.6	99.5	65.0	91.7	96.6	531.2
DSM w/o first-order	84.2	96.7	**98.4**	**65.8**	**88.2**	**93.2**	526.6	85.0	**98.5**	99.5	**70.6**	**93.8**	**97.9**	545.3
DSM (frozen U)	85.2	96.7	98.2	65.0	87.3	92.4	524.8	84.8	**98.5**	**99.8**	69.9	93.1	97.5	543.6
DSM w/o normalization	85.1	95.9	98.0	64.5	88.0	92.1	523.6	85.6	97.7	99.4	69.8	92.8	97.7	543.0
DSM ($r = 32$)	**85.7**	**96.8**	98.3	65.4	87.7	92.7	**526.7**	**86.3**	98.1	99.7	70.3	93.4	97.7	**545.5**

therefore asks whether a measurement-level modification alone can remain competitive with more elaborate retrieval systems. As in Table 2, DSM attains the highest Rsum among various backbones.

Generalization across datasets and tasks. The advantage persists beyond in-domain retrieval. In MS-COCO→Flickr30K transfer (Figure 3), DSM improves bidirectional mean R@1 over full bilinear by 1.0-4.9 points across the five backbones and by 3.5-12.8 points over Cosine. Beyond retrieval, the best DSM variant improves the nine-dataset zero-shot average over the strongest non-DSM measurement for all four Swin/CLIP backbones by 0.60-2.79 percentage points (Figure 4). Thus, the benefit of structural measurement transfers across both datasets and downstream recognition tasks; complete results are reported in Appendices A.2 and the zero-shot classification subsection.

4.3 Q2: Are the Gains Attributable to Structural Measurement?

Structural contribution. Table 3 isolates the roles of the first-order term, structural term, learned structural basis, and score normalization. Removing the second-order component reduces Rsum from 526.7 to 511.7 on Flickr30K and from 545.5 to 531.2 on MS-COCO, whereas the structure-only variant remains within 0.1/0.2 Rsum of full DSM. Freezing the structural basis also degrades performance, indicating that adapting the learned structural directions is beneficial. Removing the normalization lowers Rsum to 523.6 and 543.0, respectively, supporting its role as a practical scale-stabilization component. Overall, the results show the second-order response as the primary source of the gain, with the first-order term and normalization providing complementary contributions.

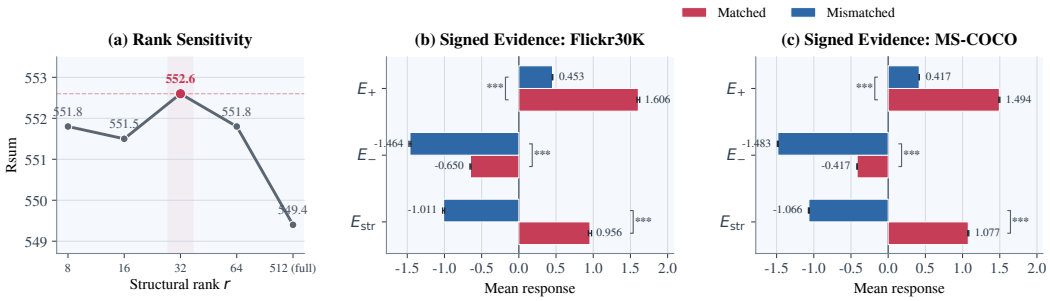

Figure 5: Structural-rank sensitivity and signed structural evidence (Swin-224+BERT). (a) Rank ablation; (b,c) positive and negative responses for matched and mismatched pairs. ***: $p < 0.001$.

Capacity and efficiency. The improvement is not explained by a larger measurement head. As a higher-capacity unstructured control, we replace the structured low-rank form with an unconstrained full matrix $W \in \mathbb{R}^{d\times d}$ (denoted Full-W). Compared with Full-W, DSM uses 15.5×-31.0× fewer parameters while improving Rsum in all six evaluated settings by 1.5-5.2 points (Figure 6). Exact parameter and FLOP counts are reported in Appendix A.4.

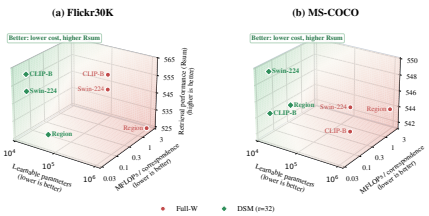

Figure 6: Accuracy-cost trade-off between Full-W and DSM ($r = 32$).

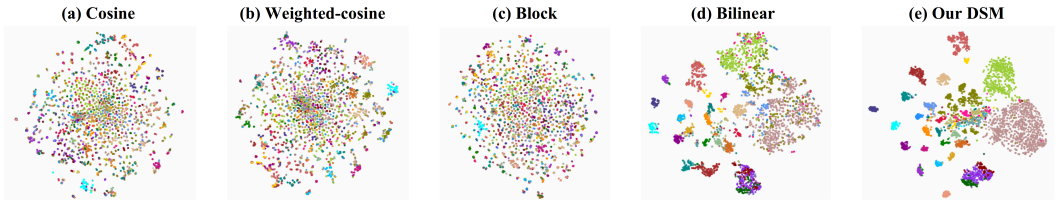

Figure 8: Impact of different measurements on cross-modal representation learning.

Low-rank sufficiency. Ranks 8-64 vary by only 1.1 Rsum, while full rank is 3.2 points below $r = 32$ (Figure 5(a)). This is consistent with the discriminative structural signal being captured effectively by a low-rank parameterization; complete directional recalls are provided in Appendix A.5.

4.4 Q3: Does DSM Functionally Exploit Semantic Structural Information?

Reproducible structure and signed evidence. The six prespecified semantic fingerprints in Appendix A.7 show high split-half reproducibility ($\rho > 0.97$) and remain distinguishable under permutation testing, supporting concept-dependent correspondence structure. DSM's signed responses also behave as intended: matched pairs show stronger positive evidence, while mismatched pairs show stronger negative evidence (Figure 5(b,c)); both separations are significant under paired sign-flip tests. Full statistics are provided in Appendix A.8.

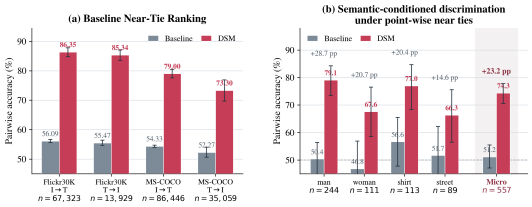

Figure 7: Structural discrimination under point-wise ambiguity.

Theorem 1 motivates examining whether structural responses remain informative when conventional point-wise evidence is ambiguous. We therefore examine *near-tie* pairs selected to have nearly equal scores under an independently trained cosine baseline. Because this subset is conditioned on baseline ambiguity, the analysis is a diagnostic of difficult baseline cases rather than an unbiased estimate of overall ranking performance. On these selected pairs, DSM improves discrimination across both datasets and retrieval directions (Figure 7(a)). The same diagnostic within the semantic concepts used in the fingerprint study increases pooled micro accuracy from 51.2% for cosine to 74.3% for DSM (Figure 7(b)). Together with the component and signed-evidence analyses, these results are consistent with structural responses providing complementary information when point-wise evidence is weak. Detailed protocols and statistics are provided in Appendices A.9 and A.10.

4.5 Qualitative View of the Learned Correspondence Space

Finally, we examine how the measurement affects representation learning. Using 30 concepts in Flickr30K with Swin-384+BERT, Figure 8 shows that conventional measurements lead to greater semantic mixing, whereas DSM induces more coherent within-semantic organization. This pattern suggests that DSM shapes cross-modal representation learning through structurally informed signals. The complete sampling and projection protocol is provided in Appendix A.11.

5 Conclusion

This work studies how cross-modal semantic correspondence should be measured in vision-language learning. We show that aggregate similarity alone overlooks reproducible structural organization in correspondence distributions, and derive a principled structural measurement from matched-mismatched distribution discrimination. DSM provides an efficient realization of this idea through signed low-rank structural responses and consistently improves alignment across diverse settings. Beyond the specific method, our results highlight measurement as an active component of representation learning rather than a passive readout of learned features. This distributional perspective supports a broader view of how semantic correspondence can be represented, measured, and learned in multimodal systems.

AI USE STATEMENT

We used generative AI tools solely for language editing and polishing of author-written text, including improvements to grammar, clarity, and presentation. We did not use generative AI to generate research ideas, formulate the proposed method, derive theoretical results, conduct experiments, analyze data, or produce empirical claims. All AI-assisted edits were carefully reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and conclusions.

ETHICS STATEMENT

This work does not involve newly collected human-subject data, private information, or personally identifiable information. All experiments are conducted on publicly available benchmark datasets under their established research-use protocols. The proposed method concerns general vision-language representation alignment and does not introduce additional data collection or intervention involving human participants. We therefore do not identify any specific ethical concerns associated with this study.

REPRODUCIBILITY STATEMENT

We provide the methodological, theoretical, and experimental details required to reproduce the main results throughout the paper and appendix. The main text specifies the formulation of fine-grained semantic correspondence, the DSM measurement and learning objective, as well as the evaluation protocols. The appendix further documents implementation details, dataset and preprocessing settings, experimental configurations, statistical analyses of semantic structural fingerprints, transfer and zero-shot evaluations, and complete proofs of the theoretical results. Because our study focuses on fine-grained semantic correspondence, the construction and analysis of semantic-level observations are described explicitly in the corresponding sections and appendices.

REFERENCES

Hani Alomari, Anushka Sivakumar, Andrew Zhang, and Chris Thomas. Maximal matching matters: Preventing representation collapse for robust cross-modal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31769–31785, 2025. URL https://aclanthology.org/2025.acl-long.1533/.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12655–12663, 2020.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15789–15798, 2021.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1218–1226, 2021.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018.

Jinhyun Jang, Jiyeong Lee, and Kwanghoon Sohn. Descriptive image-text matching with graded contextual similarity, 2025. URL https://arxiv.org/abs/2505.09997.

Linlin Ji and Li Liu. Multi-view differential mixing and graph-guided structural region selection for cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22200–22208, 2026. doi: 10.1609/aaai.v40i26.39376.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916, 2021.

Bin Kang, Bin Chen, Junjie Wang, Yulin Li, Junzhi Zhao, Junle Wang, and Zhuotao Tian. CalibCLIP: Contextual calibration of dominant semantics for text-driven image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3755765. URL https://arxiv.org/abs/2510.05586.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pp. 201–216, 2018.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10921–10930, 2020.

Yang Liu, Wentao Feng, Zhuoyao Liu, Shudong Huang, and Jiancheng Lv. Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Aligning_Information_Capacity_Between_Vision_and_Language_via_Dense-to-Sparse_Feature_ICCV_2025_paper.html.

Yang Liu, Mengyuan Liu, Shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5676–5684, 2025b. doi: 10.1609/aaai.v39i6.32605.

Can Lu and Muye Feng. Enhancing image–text retrieval via region–grid interaction and semantic calibration. *Sensors*, 26(14):4483, 2026. doi: 10.3390/s26144483.

Xiang Ma, Lexin Fang, Litian Xu, and Caiming Zhang. Aligning the true semantics: Constrained decoupling and distribution sampling for cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24290–24298, 2026. doi: 10.1609/aaai.v40i29.39609.

Xinyu Mao, Junsi Li, Haoji Zhang, Yu Liang, and Ming Sun. SEPS: Semantic-enhanced patch slimming framework for fine-grained cross-modal alignment. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zfrb8tikzS.

Renjie Pan, Jihao Dong, and Hua Yang. Discovering clone negatives via adaptive contrastive learning for image-text matching. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b010241b9f1cdfc7d4c392db899cef86-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Song Yang, Qiang Li, Wenhui Li, Xuanya Li, and An-An Liu. Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):8037–8050, 2022.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670, 2022.

Kun Zhang, Bo Hu, Huatian Zhang, Zhe Li, and Zhendong Mao. Enhanced semantic similarity learning framework for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2973–2988, 2024. doi: 10.1109/TCSVT.2023.3307554.

A Additional Experimental Results

A.1 Implementation and Result Sources

Implementation. Unless otherwise specified, DSM uses $r = 32$ and AdamW. Region+BERT is trained on two GPUs with a batch size of 128 per GPU and a learning rate of 5×10^{-4} for 25 epochs on Flickr30K and 20 epochs on MS-COCO. Swin-224 and Swin-384 are trained on four GPUs for 20 epochs, using per-GPU batch sizes of 64 and 32 and learning rates of 4×10^{-4} and 2×10^{-4}, respectively. CLIP ViT-B/16 and ViT-L/14 are trained for 20 epochs with batch sizes of 128 and 32; the learning rates for the CLIP encoder, measurement head, and adapter are 1×10^{-6}, 3×10^{-5}, and 5×10^{-5}, respectively.

For the controlled measurement baselines, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ denote the normalized visual and textual features and $\mathbf{z} = \mathbf{x} \odot \mathbf{y}$. Cosine uses $s = \mathbf{1}^\top \mathbf{z}$, while weighted cosine uses $s = \mathbf{b}^\top \mathbf{z}$. Full bilinear measurement uses $s = \mathbf{x}^\top \mathbf{W} \mathbf{y}$, and the rank-$r$ bilinear variant factorizes the interaction as $s = (\mathbf{U}\mathbf{x})^\top (\mathbf{V}\mathbf{y})$. For block-wise measurement, $\mathbf{x}$ and $\mathbf{y}$ are evenly partitioned into g feature blocks, cosine similarity is computed within each corresponding block, and the resulting g block scores are linearly mapped to the final similarity. The block number is selected on the validation set, with $g = 16$ for the 1024-dimensional Region+BERT features and $g = 8$ for the 512-dimensional Swin and CLIP features.

A.2 Cross-Dataset Transfer

Table 4: MS-COCO→Flickr30K retrieval without Flickr30K fine-tuning.

Method	Image-to-Text			Text-to-Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
Region + BERT							
Cosine	60.6	86.9	91.6	48.1	73.6	82.5	443.2
Weighted cosine	63.0	88.1	93.7	49.8	75.4	83.1	453.1
Bilinear (full)	74.6	**92.6**	96.0	55.5	80.4	87.1	486.2
Bilinear ($r = 32$)	69.2	90.1	94.7	51.4	77.2	85.0	467.5
Block	72.1	91.7	95.0	54.6	79.2	86.1	478.8
DSM ($r = 32$)	**76.5**	92.4	95.4	**57.7**	**81.8**	88.0	**491.9**
DSM (full)	74.8	92.0	**96.5**	57.4	81.7	**88.1**	490.4
Swin-224 + BERT							
Cosine	72.2	92.4	96.9	60.9	85.1	90.9	498.4
Weighted cosine	75.5	94.2	97.4	63.7	86.7	91.8	509.2
Bilinear (full)	78.8	95.0	97.6	66.0	87.5	92.9	517.8
Bilinear ($r = 32$)	59.9	86.0	91.9	48.5	75.4	84.4	446.1
Block-wise	69.4	91.0	96.1	61.2	85.3	91.2	494.1
DSM ($r = 32$)	**83.3**	96.7	**98.7**	67.9	89.7	94.1	**530.4**
DSM (full)	81.1	**97.5**	98.5	**68.0**	**90.0**	**94.4**	529.5
Swin-384 + BERT							
Cosine	74.8	94.5	97.5	63.2	87.1	92.7	509.8
Weighted cosine	76.8	95.0	98.2	65.0	88.3	93.8	517.1
Bilinear (full)	79.8	96.2	98.2	68.3	90.0	94.5	526.9
Bilinear ($r = 32$)	77.0	94.7	98.1	62.3	86.9	92.8	511.7
Block-wise	77.4	94.5	97.6	66.3	89.6	93.8	519.2
DSM ($r = 32$)	84.7	97.7	**99.2**	**73.1**	**92.3**	**96.0**	**542.9**
DSM (full)	**86.0**	**97.8**	99.1	72.5	91.8	95.7	542.8
CLIP ViT-B/16							
Cosine	83.3	97.3	99.7	74.4	92.6	96.2	543.5
Weighted cosine	83.1	97.1	99.2	74.1	93.2	96.0	542.7
Bilinear (full)	87.2	**98.4**	**99.9**	76.1	93.5	**96.8**	551.8
Bilinear ($r = 32$)	83.4	97.8	99.4	73.2	93.2	96.4	543.3
Block-wise	79.3	96.0	98.3	71.1	91.3	95.0	531.0
DSM ($r = 32$)	89.1	97.7	99.0	76.1	**93.7**	**96.8**	552.4
DSM (full)	**89.4**	97.8	99.5	**77.0**	93.6	96.5	**553.8**
CLIP ViT-L/14							
Cosine	88.5	98.7	99.4	79.3	95.5	97.5	559.0
Weighted cosine	88.7	98.5	99.3	80.3	95.8	**98.1**	560.7
Bilinear (full)	91.0	99.1	99.6	81.4	95.8	97.9	564.7
Bilinear ($r = 32$)	90.0	99.2	99.6	81.1	**96.0**	**98.1**	564.1
Block-wise	86.0	97.9	99.3	79.7	95.4	97.7	555.9
DSM ($r = 32$)	**92.4**	**99.3**	**99.7**	**82.3**	**96.0**	98.0	**567.7**
DSM (full)	91.3	99.0	99.6	82.1	95.9	98.0	565.9

A.3 Zero-Shot Classification

All zero-shot results use MS-COCO-trained checkpoints without target-dataset training. For ImageNet, Caltech101, Oxford Pets, Food101, SUN397, DTD, and UCF101, we use the fixed CoOp evaluation splits and class-name prompt templates; CIFAR-10/100 use their official test sets. Each test image is scored against all class-name prompts using the checkpoint's trained alignment measurement, and the highest-scoring class is selected. The same splits and prompts are used for all measurement functions within each backbone, and Avg9 is the unweighted mean over the nine datasets. Table 5 reports top-1 accuracy.

Table 5: Zero-shot top-1 accuracy (%) with MS-COCO-trained checkpoints.

Method	IN	Cal	Pets	Food	SUN	DTD	UCF	C10	C100	Avg9
Swin-224 + BERT										
Cosine	21.64	59.39	8.83	23.47	34.97	14.95	33.39	**85.00**	36.41	35.34
Weighted cosine	21.55	55.70	9.43	25.90	33.61	14.89	32.51	81.00	33.00	34.18
Bilinear (full)	21.50	59.80	11.69	**27.66**	33.89	14.66	34.73	81.25	36.55	35.75
Bilinear ($r = 32$)	13.40	48.03	8.75	15.12	24.10	7.62	25.40	81.72	30.59	28.30
Block-wise	18.62	58.99	6.05	20.92	32.70	15.54	31.54	83.90	38.05	34.03
DSM ($r = 32$)	23.63	60.37	**12.84**	26.62	**35.56**	**18.20**	**35.16**	80.82	38.38	36.84
DSM (full)	**23.97**	**62.47**	12.18	27.34	35.24	16.55	33.68	84.34	**38.53**	**37.14**
Swin-384 + BERT										
Cosine	20.99	58.95	6.46	21.67	33.65	13.48	33.07	84.93	38.28	34.61
Weighted cosine	21.49	60.65	7.77	22.55	**35.71**	15.96	34.18	84.31	39.03	35.74
Bilinear (full)	21.57	59.43	5.04	24.83	34.32	15.07	30.29	83.26	38.87	34.74
Bilinear ($r = 32$)	20.58	56.96	7.44	22.54	34.26	14.42	31.93	86.62	39.73	34.94
Block-wise	20.69	59.68	6.21	23.42	34.43	15.96	30.43	84.81	37.01	34.74
DSM ($r = 32$)	**26.11**	62.92	9.02	26.75	35.59	18.03	32.43	**90.14**	**45.75**	**38.53**
DSM (full)	25.76	**63.04**	**9.59**	**27.38**	35.10	**19.03**	**35.29**	88.06	42.43	38.41
CLIP ViT-B/16 Fine-tuning										
Cosine	58.40	90.63	84.11	**77.36**	62.33	36.05	64.76	88.66	64.20	69.61
Weighted cosine	58.27	91.24	83.29	77.25	62.22	38.06	64.76	88.00	64.98	69.79
Bilinear (full)	59.23	91.76	83.40	76.51	62.78	28.96	65.03	88.89	**66.25**	69.20
Bilinear ($r = 32$)	52.38	89.29	79.80	73.27	56.13	19.86	58.52	88.26	61.82	64.37
Block-wise	48.04	88.44	76.12	72.48	53.46	34.16	57.26	88.12	61.41	64.39
DSM ($r = 32$)	**60.54**	92.66	85.83	76.00	**63.11**	36.82	65.79	87.97	63.66	70.26
DSM (full)	60.16	**92.86**	**86.40**	75.95	63.09	**40.54**	**66.35**	**89.08**	64.79	**71.02**
CLIP ViT-L/14 Fine-tuning										
Cosine	65.08	93.18	87.14	79.30	64.19	51.30	70.26	93.71	71.20	75.04
Weighted cosine	66.04	93.43	87.90	81.01	64.90	**51.71**	70.39	94.29	72.73	75.82
Bilinear (full)	66.52	**94.16**	87.54	79.38	64.88	44.44	71.21	94.33	73.50	75.11
Bilinear ($r = 32$)	66.26	93.71	**89.40**	80.99	65.77	29.91	69.05	94.92	73.72	73.75
Block-wise	61.24	92.41	82.42	77.57	61.90	37.53	68.75	94.90	72.25	72.11
DSM ($r = 32$)	**67.72**	93.67	89.21	**81.88**	**66.01**	49.59	71.27	94.12	**74.07**	76.39
DSM (full)	67.42	93.91	88.33	81.61	65.84	49.76	**71.74**	**95.18**	74.00	**76.42**

IN: ImageNet; Cal: Caltech101; Pets: Oxford Pets; Food: Food101; SUN: SUN397; DTD: Describable Textures; UCF: UCF101; C10/C100: CIFAR-10/100. Avg9 is the unweighted mean over all nine displayed datasets. Bold indicates the maximum over the seven displayed runs for each backbone.

A.4 Capacity and Efficiency Comparison

We compare the analytic cost of representative cross-modal measurements under a common setting with d-dimensional visual and textual features. The comparison concerns only the measurement head and excludes shared encoders, attention, correspondence construction, and parameter-only transformations that can be cached. A multiply-accumulate is counted as two FLOPs.

Table 6: **Analytic complexity of representative cross-modal measurements.** Here, d is the feature dimension, $r \ll d$ the low-rank dimension, and g the number of feature blocks. Complexity refers to the measurement head for one visual-textual correspondence.

Measurement	Score form	Complexity
Cosine	$\mathbf{v}^\top \mathbf{t}$	$\mathcal{O}(d)$
Weighted cosine	$\mathbf{b}^\top (\mathbf{v} \odot \mathbf{t})$	$\mathcal{O}(d)$
Block-wise	$\phi([\cos(\mathbf{v}^{(1)}, \mathbf{t}^{(1)}), \ldots, \cos(\mathbf{v}^{(g)}, \mathbf{t}^{(g)})])$	$\mathcal{O}(d+g)$
Bilinear (full)	$\mathbf{v}^\top \mathbf{W} \mathbf{t}$	$\mathcal{O}(d^2)$
Bilinear (rank r)	$(\mathbf{U}\mathbf{v})^\top (\mathbf{V}\mathbf{t})$	$\mathcal{O}(rd)$
Full-W structural	$\mathbf{z}^\top \mathbf{W} \mathbf{z}$	$\mathcal{O}(d^2)$
DSM (rank r)	linear + signed low-rank structural responses	$\mathcal{O}(rd)$

Cosine and weighted cosine scale linearly with the representation dimension. For block-wise measurement, the d dimensions are partitioned into g blocks, cosine similarity is computed within each block, and the resulting g similarities are mapped to the final score; its overall cost remains linear in d. A full bilinear or unrestricted quadratic measurement requires $\mathcal{O}(d^2)$ computation, whereas low-rank bilinear measurement reduces this to $\mathcal{O}(rd)$.

Under the same counting convention, Full-W requires $2d^2 + 3d$ FLOPs, while DSM requires $2dr + 5d + 5r$. Thus, DSM retains the linear-in-d scaling of low-rank measurements when $r \ll d$, while introducing signed second-order structural responses without the quadratic cost of an unrestricted structural matrix.

Table 7: Parameter, FLOP, and Rsum comparison between Full-W and DSM. FLOPs are counted per correspondence vector; bold marks the favorable value.

Backbone	Method	Parameters	MFLOPs/corr.	Flickr30K Rsum	MS-COCO 1K Rsum
Region + BERT	Full-W	1,048,576	2.100224	525.2	543.9
	DSM ($r = 32$)	**33,824**	**0.070816**	**526.7**	**545.5**
Swin-224 + BERT	Full-W	262,144	0.525824	549.9	544.6
	DSM ($r = 32$)	**16,928**	**0.035488**	**552.6**	**549.8**
CLIP ViT-B/16	Full-W	262,144	0.525824	558.3	541.5
	DSM ($r = 32$)	**16,928**	**0.035488**	**562.1**	**544.7**

A.5 Structural-Rank Ablation

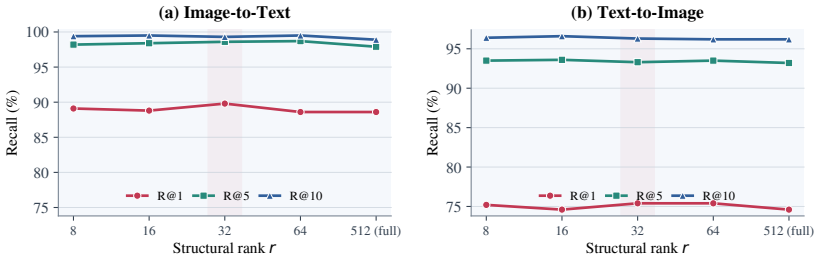

Figure 9: Directional recall sensitivity to structural rank on Flickr30K with Swin-224+BERT. The shaded band marks $r = 32$.

Table 8: Structural-rank ablation on Flickr30K with Swin-224+BERT.

Rank	Image-to-Text			Text-to-Image			Rsum	$\Delta_{r=32}$
	R@1	R@5	R@10	R@1	R@5	R@10		
8	89.1	98.2	99.4	75.2	93.5	96.4	551.8	-0.8
16	88.8	98.4	**99.5**	74.6	**93.6**	**96.6**	551.5	-1.1
32	**89.8**	98.6	99.3	**75.4**	93.3	96.3	**552.6**	0.0
64	88.6	**98.7**	**99.5**	**75.4**	93.5	96.2	551.8	-0.8
512 (full)	88.6	97.9	98.9	74.6	93.2	96.2	549.4	-3.2

A.6 Spectral Structure of DSM Across Backbones

We further examine how the signed structural component of DSM is realized across different vision-language backbones. DSM parameterizes

$$\mathbf{A} = \mathbf{A}_+ - \mathbf{A}_-, \qquad \mathbf{A}_\pm = \mathbf{U}_\pm \operatorname{diag}(\boldsymbol{\alpha}_\pm)\mathbf{U}_\pm^\top, \qquad \boldsymbol{\alpha}_\pm \geq \mathbf{0},$$

where $\boldsymbol{\alpha}_+$ and $\boldsymbol{\alpha}_-$ are the signed structural coefficients associated with match-enhancing and mismatch-suppressing directions, respectively.

Figure 10 shows these learned structural coefficients for five Flickr30K DSM models with total rank $r = 32$. For visualization, positive and negative branches are displayed as $+\alpha_+$ and $-\alpha_-$, although $\alpha_\pm$ themselves are nonnegative. All backbones learn nonzero coefficients in both branches, showing that both components are retained across architectures.

Because the absolute coefficient scale is architecture-dependent, we further analyze the normalized singular spectrum of the induced operator $\mathbf{A} = \mathbf{A}_+ - \mathbf{A}_-$. Figure 11 plots $\sigma_i(\mathbf{A})/\sigma_{\max}(\mathbf{A})$ for the same checkpoints.

Swin-B/224, Swin-B/384, and Region-BERT exhibit spectra dominated by a small number of leading structural modes, whereas the CLIP backbones show more slowly decaying and less concentrated normalized spectra. Together with the smaller performance gains on CLIP, this pattern is consistent with representations already shaped by cosine-based contrastive pretraining leaving less room for downstream measurement-level adaptation. This interpretation is suggestive rather than causal; the figures provide parameter-level evidence of different structural organization across backbones.

A.7 Semantic Structural Fingerprint Analysis

We extend the semantic structural fingerprint analysis in Figure 1 to six concepts selected before the statistical analysis from frequent nouns with sufficient image-level support: *man*, *woman*, *shirt*, *dog*, *street*, and *water*. They span human, object, animal, scene, and material/environmental semantics.

Fingerprint construction. For each occurrence of a target semantic concept, we extract the cross-modal correspondence vector defined in Eq. (1). Repeated correspondence observations of the same concept are aggregated dimension by dimension. For visualization, we apply a signed logarithmic transform independently to each aggregated dimension, $\phi(x) = \operatorname{sign}(x)\log(1 + |x|)$, yielding a dimension-wise signed-log correspondence profile. This profile provides an observable summary of the semantic structural fingerprint by retaining its dimension-wise organization, whereas Definition 3 characterizes the full fingerprint through $(\Delta\boldsymbol{\mu}_c, \Delta\mathbf{G}_c)$, including joint second-order structure.

All semantics are displayed using the same ordering of the 512 representation dimensions, without smoothing or semantic-specific dimension reordering.

Because Flickr30K provides five captions for each image, repeated captions describing the same visual scene are not treated as statistically independent observations. Instead, the analyses below are conducted at the image-semantic level, such that repeated occurrences associated with the same image are first aggregated within the same group.

Figure 12 extends the three representative examples in Figure 1 to all six analyzed semantics. For visualization clarity, Figure 1 displays only the first 100 dimensions, whereas Figure 12 presents the

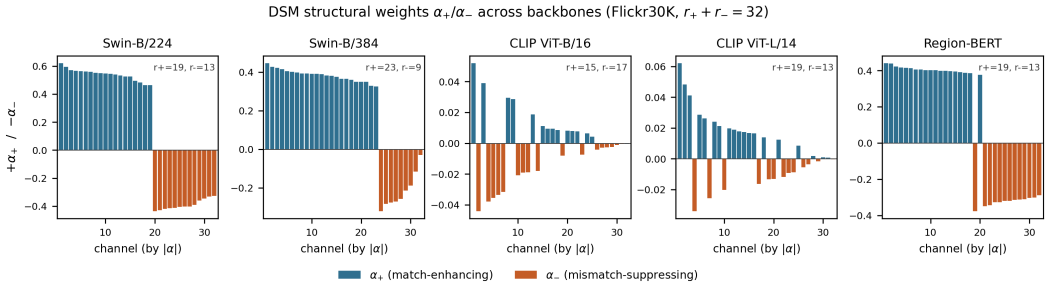

Figure 10: **Signed structural coefficients of DSM across backbones.** Learned α_+/α_- coefficients on Flickr30K with total rank $r = 32$. Blue and orange bars denote $+\alpha_+$ and $-\alpha_-$, respectively, and channels are ordered by $|\alpha|$.

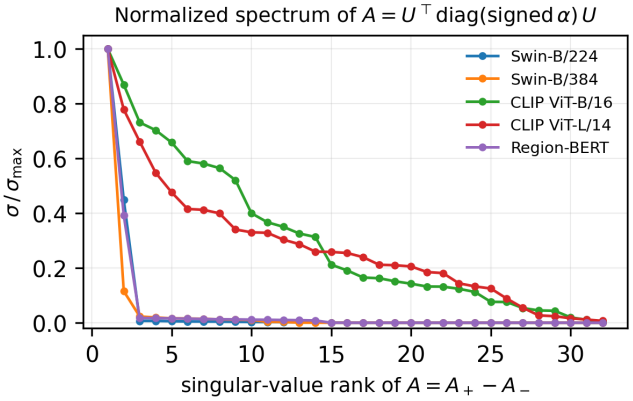

Figure 11: **Normalized singular spectrum of the induced DSM operator.** Singular values of $\mathbf{A} = \mathbf{A}_+ - \mathbf{A}_-$ are normalized by $\sigma_{\max}$ for each backbone.

complete 512-dimensional fingerprints. Across concepts, the full profiles exhibit distinct patterns of dimension-wise agreement.

Within-semantic reproducibility. We next examine whether the observed fingerprints are reproducible when estimated from different subsets of images. For each semantic concept, we randomly divide its image-semantic observations into two non-overlapping halves and independently reconstruct the fingerprint from each half. Reproducibility is measured by the Spearman correlation between the two resulting 512-dimensional profiles. We repeat this split-half procedure 1,000 times for each semantic concept.

As shown in Figure 13, the median split-half Spearman correlations are 0.989, 0.983, 0.987, 0.982, 0.980, and 0.973 for *man*, *woman*, *shirt*, *dog*, *street*, and *water*, respectively. Thus, all six semantic fingerprints achieve split-half correlations above 0.97. The narrow variation across repeated splits further indicates that the observed dimension-wise profiles are not determined by a particular partition of the image population.

Between-semantic distinctiveness. High within-semantic reproducibility does not by itself imply that different semantic concepts exhibit different fingerprints. We therefore compare the full 512-dimensional fingerprints across semantic concepts using pairwise Spearman correlation.

As shown in Figure 14, between-semantic correlations are substantially lower than the within-semantic split-half correlations, indicating that the reproducible fingerprints are concept dependent.

To complement the descriptive fingerprint correlations, we test whether the dimension-wise fingerprint profiles remain distinguishable across semantic concepts under random reassignment of semantic labels. For each semantic pair, the test statistic is the observed dissimilarity between their 512-dimensional signed-log fingerprint profiles, computed as (1-ρ), where (ρ) is the Spearman correlation between the two profiles. Permutations are performed at the image–semantic level, preserving

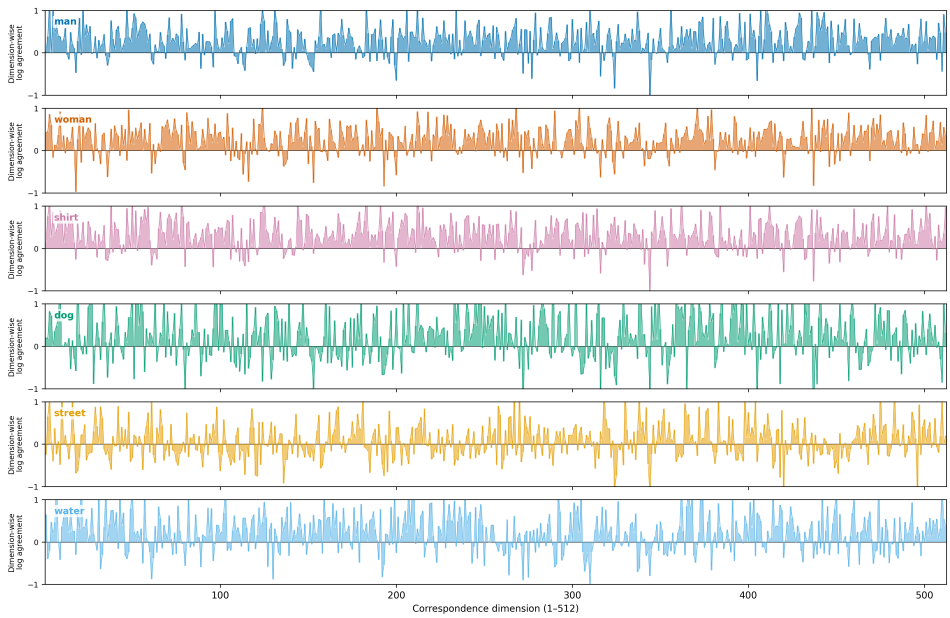

Figure 12: **Semantic structural fingerprints across six concepts.** Each row shows the dimension-wise log agreement profile over the same 512 representation dimensions.

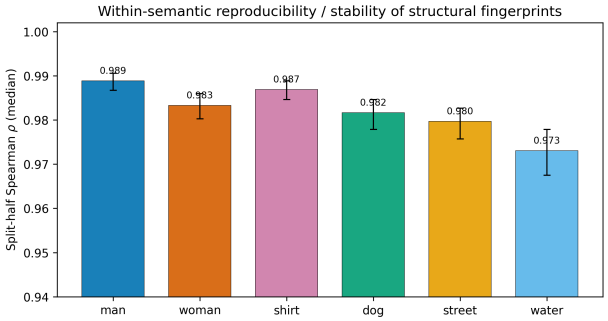

Figure 13: **Within-semantic fingerprint reproducibility.** Bars show median split-half Spearman correlation over 1,000 repetitions; error bars denote the 2.5th-97.5th percentiles.

the within-image aggregation described above. Pairwise tests use 10,000 permutations, and Holm correction is applied jointly over all 15 semantic pairs.

Together, these analyses provide the detailed empirical support for the two properties summarized in the main text: semantic structural fingerprints are reproducible across independent subsets of observations and remain distinguishable across semantic concepts.

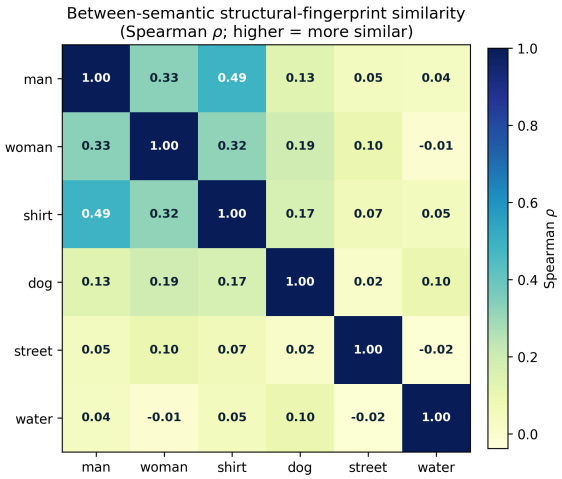

Figure 14: **Between-semantic fingerprint similarity.** Each entry reports the Spearman correlation between two 512-dimensional semantic fingerprints.

A.8 Full-Pool Signed Structural Evidence

For each image query, signed responses are averaged over its annotated captions and compared with mismatched captions; caption queries are evaluated analogously against mismatched images. All within-fold pairs are used, with image-cluster bootstrap resampling and paired sign-flip tests for uncertainty assessment.

Table 9: Query-averaged signed DSM evidence on Flickr30K and MS-COCO. $E_{\mathrm{str}} = E_{+} + E_{-}$.

Dataset	Pair type	E_{+}	E_{-}	E_{str}
Flickr30K	Matched	1.6060	-0.6500	0.9560
Flickr30K	Mismatched	0.4527	-1.4637	-1.0111
MS-COCO	Matched	1.4942	-0.4175	1.0766
MS-COCO	Mismatched	0.4174	-1.4832	-1.0658

A.9 Global Baseline Near-Tie Analysis

Protocol. Positive and negative candidates are selected when $|s_{\mathrm{base}}(p) - s_{\mathrm{base}}(n)| \leq 0.02$ using an independently trained cosine baseline, and the same selected pairs are then evaluated by DSM. This conditioning deliberately targets cases ambiguous to cosine and is therefore used as a diagnostic rather than an unbiased estimate of overall ranking performance. COCO folds and retrieval directions are pooled by selected-comparison counts. Uncertainty is estimated with 3,000 image-cluster bootstrap resamples; score-gap intervals are reported in Table 11.

Table 10: Pairwise ranking accuracy (%) on baseline-selected near ties.

Dataset	Direction	Comparisons	Baseline	DSM
Flickr30K	I→T	67,323	56.09	**86.35**
Flickr30K	T→I	13,929	55.47	**85.34**
MS-COCO	I→T	86,446	54.33	**79.00**
MS-COCO	T→I	35,059	52.27	**73.30**

Table 11: Total-score separation on baseline-selected near ties. The last column reports the 95% image-cluster bootstrap interval.

Dataset	Method	Matched	Mismatched	Difference	95% CI
Flickr30K	Baseline	0.4030	0.4014	0.0016	$[0.0015, 0.0017]$
	DSM	0.1593	-0.3199	0.4792	$[0.4456, 0.5082]$
MS-COCO	Baseline	-0.2820	-0.2830	0.0010	$[0.0009, 0.0012]$
	DSM	0.5045	0.2312	0.2734	$[0.2495, 0.3002]$

A.10 Semantic-Conditioned Near-Tie Analysis

This analysis connects the semantic fingerprint study with the ambiguity analysis in Section 4.4. For each semantic group, one matched word-level correspondence is paired with $K = 20$ mismatched correspondences drawn from images in which that semantic is absent. Near ties are selected using the independently trained cosine baseline,

$$|s_{\cos}(p) - s_{\cos}(n)| \leq \varepsilon,$$

and the same selected pairs are then evaluated by DSM. Because selection is conditioned on cosine ambiguity, this experiment is a semantic-conditioned diagnostic of baseline-hard cases rather than an unbiased estimate of overall ranking performance. To ensure reliable estimates, we report only semantic groups with at least 80 selected pairs, i.e., $n_{\text{near}} \geq 80$. Under this criterion, four concepts are retained: *man*, *woman*, *shirt*, and *street*. Image-cluster bootstrap intervals use 2,000 resamples.

Table 12: Semantic-conditioned near-tie discrimination at $\varepsilon = 0.02$ for concepts with $n_{\text{near}} \geq 80$.

Semantic	n_{near}	Cosine	95% CI	DSM	95% CI	Δ (pp)
man	244	50.4	[44.1, 56.4]	79.1	[73.5, 84.3]	+28.7
woman	111	46.8	[37.8, 56.9]	67.6	[58.6, 76.5]	+20.7
shirt	113	56.6	[47.7, 65.5]	77.0	[68.4, 84.7]	+20.4
street	89	51.7	[41.2, 62.2]	66.3	[56.5, 75.6]	+14.6
Micro	557	51.2	-	74.3	-	+23.2

Across the four sufficiently supported semantic groups, DSM increases pooled micro accuracy from 51.2% for cosine to 74.3% on the same baseline-selected pairs. The gain is also stable across $\varepsilon \in \{0.01, 0.02, 0.05, 0.10\}$. We interpret this as complementary evidence that DSM can separate cases that are ambiguous to the cosine baseline, not as an unconditional estimate of its overall advantage.

A.11 Cross-modal Correspondence Visualization Details

This section details the correspondence-space visualization in Figure 8. The analysis qualitatively compares the semantic organization of representations learned under different measurements and is used only as a complementary visualization, not as an additional performance metric.

Experimental setting. The visualization uses the Flickr30K test split and the Swin-Base-384 visual backbone with BERT-base-uncased as the text encoder. The cross-modal correspondence dimension is 512. We compare five models trained under the same retrieval setting but using different measurement functions: Cosine, Weighted cosine, rank-32 Bilinear, Block-wise interaction, and DSM. For each measurement, the checkpoint selected by retrieval Rsum is used for visualization.

Semantic sampling. The Flickr30K test split contains 1,000 images and 5,000 captions. We use 30 frequent noun semantics: *man*, *woman*, *shirt*, *dog*, *street*, *water*, *hat*, *background*, *building*, *jacket*, *soccer*, *beach*, *sidewalk*, *camera*, *grass*, *table*, *guitar*, *city*, *bike*, *wall*, *snow*, *hair*, *bench*, *jeans*, *mouth*, *park*, *rock*, *horse*, *food*, and *car*. Each selected noun corresponds to a single BERT WordPiece.

For each caption and selected noun, if the corresponding token occurs in the sequence, one occurrence is retained using its first matching token position. Each retained occurrence is paired with the ground-truth image associated with that caption, yielding 5,001 semantic occurrences in total. The same captions, semantic labels, token positions, and matched images are used for every compared measurement model. No class-balanced downsampling is applied, so the visualization preserves the original frequency distribution of the selected concepts.

Visualized representation. For each retained semantic occurrence, we extract the cross-modal correspondence representation immediately before the final measurement operation. Specifically, the corresponding textual representation and word-conditioned visual context produce the same element-wise correspondence vector defined in Eq. (1). The final scalar similarity score is not visualized, and no subsequent measurement-specific scoring transformation is applied to the saved feature.

Consequently, Figure 8 compares correspondence representations learned under different measurement functions rather than directly visualizing the output scores of those functions.

Projection protocol. For each trained model, the 5,001 correspondence vectors are stacked into a 5001×512 feature matrix. We first reduce the representation from 512 to 50 dimensions using PCA and then independently apply t-SNE to obtain the two-dimensional visualization. We use perplexity 40, PCA initialization, an automatic learning rate, 1,000 optimization iterations, and random seed 20260923.

Interpretation. Because t-SNE is fitted independently for each model, comparisons are restricted to within-panel semantic organization. Under this common protocol, DSM exhibits more coherent semantic neighborhoods, consistent with the qualitative observation in Figure 8.

A.12 Global-feature measurement control

The main formulation operates on fine-grained word-visual correspondences so that semantic content remains individually identifiable. We additionally test whether the advantage of structural measurement persists when correspondence is instead formed from globally pooled image and text representations.

Using Swin-B/224 and BERT on Flickr30K, we remove the word-patch correspondence path and obtain global image and text features by mean pooling their encoder outputs, followed by ℓ_2 normalization. Their correspondence is defined as $\mathbf{z}_g = \mathbf{v}_g \odot \mathbf{t}_g$. We then apply the same family of measurement functions used in the controlled comparison of the main paper. No cross-attention, sparse selection, token aggregation, or auxiliary sparsity loss is used.

Table 13: Global-feature measurement on Flickr30K with mean-pooled Swin-B/224+BERT representations.

Method	Image-to-Text			Text-to-Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
Cosine	84.0	97.3	99.2	69.9	91.4	94.9	536.8
Weighted cosine	84.1	97.3	98.9	69.8	91.5	95.0	536.7
DSM ($r = 32$)	84.4	97.2	98.7	68.6	91.4	95.5	535.8
Block-wise	80.2	95.9	98.6	66.7	90.2	94.4	526.1
Bilinear ($r = 32$)	67.0	90.6	95.0	55.3	84.0	90.0	481.9

Under global pooling, Cosine, Weighted cosine, and DSM achieve comparable Rsum, whereas the fine-grained word-visual DSM counterpart reaches 552.6 with the same Swin-B/224+BERT backbone. This controlled comparison supports the use of fine-grained correspondence as the input to structural measurement under the evaluated protocol.

B Derivation of the Second-Order Measurement

From match probabilities to a score. Proposition 1 concerns the simplest decision faced by a correspondence measurement: under a fixed semantic condition, does an observed vector come from matched or mismatched visual evidence? Let $M \in \{0, 1\}$ denote the correspondence label, and suppose both classes admit densities at the observed $\mathbf{z}$ and have positive priors. Bayes' rule gives

$$\frac{\Pr(M = 1 \mid \mathbf{Z} = \mathbf{z})}{\Pr(M = 0 \mid \mathbf{Z} = \mathbf{z})} = \frac{p_+(\mathbf{z})\pi_+}{p_-(\mathbf{z})\pi_-}. \tag{12}$$

Under equal costs of the two errors, choosing the larger posterior is Bayes optimal. Taking logarithms preserves the ordering and gives Eq. (4). This derivation specifies the *target* discriminative

score; it does not assume that the class densities can be estimated exactly. If the densities are zero on parts of the support, the corresponding log ratio is understood where it is defined, or through the usual limiting convention.

Why the first quadratic correction appears. To obtain a tractable measurement, select a reference correspondence $\mathbf{z}_0$ within a region of interest and write $\boldsymbol{\delta} = \mathbf{z} - \mathbf{z}_0$. Proposition 2 assumes that the log posterior odds are three times continuously differentiable in a neighborhood of this point. With $\mathbf{g}_0 = \nabla \ell_c^*(\mathbf{z}_0)$ and $\mathbf{H}_0 = \nabla^2 \ell_c^*(\mathbf{z}_0)$, Taylor's theorem yields

$$\ell_c^*(\mathbf{z}) = \ell_c^*(\mathbf{z}_0) + \mathbf{g}_0^\top \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{H}_0 \boldsymbol{\delta} + \mathcal{R}_3(\mathbf{z}; \mathbf{z}_0). \tag{13}$$

This expansion shows that, locally, the optimal score is governed by first-order variation and local curvature. Expanding $\boldsymbol{\delta} = \mathbf{z} - \mathbf{z}_0$ and collecting terms in $\mathbf{z}$ yields the quadratic and linear coefficients

$$\mathbf{A} = \frac{1}{2}\mathbf{H}_0, \qquad \mathbf{b} = \mathbf{g}_0 - \mathbf{H}_0 \mathbf{z}_0,$$

while all terms independent of $\mathbf{z}$ are absorbed into a. Since the Hessian $\mathbf{H}_0$ is symmetric, so is $\mathbf{A}$. This gives Eq. (5).

For completeness, if the operator norm of the third derivative is bounded by L along the segment from $\mathbf{z}_0$ to $\mathbf{z}$, then $|\mathcal{R}_3(\mathbf{z}; \mathbf{z}_0)| \leq (L/6)\|\mathbf{z} - \mathbf{z}_0\|_2^3$. The error is thus of higher order near the reference point. Away from that neighborhood, a fixed quadratic need not approximate the optimal density ratio well; the measurement is learned from data rather than assumed to be exact everywhere.

An exact special case. If P_+ and P_- are Gaussian with nonsingular covariances $\boldsymbol{\Sigma}_+$ and $\boldsymbol{\Sigma}_-$, expanding their log densities produces $\mathbf{A} = \frac{1}{2}(\boldsymbol{\Sigma}_-^{-1} - \boldsymbol{\Sigma}_+^{-1})$. The quadratic form is then exact, and the covariance difference can make $\mathbf{A}$ indefinite. This example helps interpret the signed structural terms; Gaussianity is not part of the general derivation or the DSM model.

C Proof and Scope of Structural Distinguishability

Theorem 1 separates two claims: a linear score cannot produce a *difference in expected response* when the means are identical, while a quadratic response can detect a difference in second moments. Its premise requires only finite second moments.

For any $\mathbf{b} \in \mathbb{R}^d$, linearity of expectation gives $\mathbb{E}_{P_{m_1}}[\mathbf{b}^\top \mathbf{Z}] - \mathbb{E}_{P_{m_2}}[\mathbf{b}^\top \mathbf{Z}] = \mathbf{b}^\top(\boldsymbol{\mu}_{m_1} - \boldsymbol{\mu}_{m_2}) = 0$. The mean assumption is exactly what makes the result hold for *every* choice of first-order weights; retraining $\mathbf{b}$ cannot change this expectation.

Now set $\mathbf{D} = \mathbf{G}_{m_1} - \mathbf{G}_{m_2}$. It is real, symmetric, and nonzero, so the spectral theorem guarantees a unit eigenvector $\mathbf{u}$ with nonzero eigenvalue ν. For either mode m,

$$\mathbb{E}_{P_m}[(\mathbf{u}^\top \mathbf{Z})^2] = \mathbf{u}^\top \mathbb{E}_{P_m}[\mathbf{Z}\mathbf{Z}^\top]\mathbf{u} = \mathbf{u}^\top \mathbf{G}_m \mathbf{u}. \tag{14}$$

Subtracting the two instances of Eq. (14) yields $\mathbf{u}^\top \mathbf{D}\mathbf{u} = \nu \neq 0$, proving Eq. (6). In concrete terms, two modes can activate each coordinate equally often on average while differing in the energy of a particular combination of coordinates; squaring the projected response reveals that difference.

Approximately equal means. The exact-equality premise makes the theorem clean, but its mechanism persists when mean vectors are close. Cauchy-Schwarz bounds the expected gap of a linear score by $\|\mathbf{b}\|_2 \|\boldsymbol{\mu}_{m_1} - \boldsymbol{\mu}_{m_2}\|_2$. A unit eigenvector $\mathbf{u}$ of $\mathbf{D}$, meanwhile, yields a quadratic expected gap $|\nu|$. These quantities are directly comparable only after choosing a scale or norm constraint for the scores; the statement does not promise better classification for every dataset or every trained parameter choice. Also, different distributions can give different *distributions* of linear scores despite identical expected values. The theorem is a precise moment-level justification for including structural responses.

D Why the Discriminative Structural Form Can Be Low Rank

Eq. (7) expresses the benefit of a quadratic measurement as its inner product with the matched-mismatched second-moment contrast. A full $d \times d$ matrix could respond to every direction, but

a constrained measurement should spend its capacity on directions with the largest contrast. The following idealized result makes that choice precise.

Proposition 3 (Optimal rank-constrained quadratic contrast). Let $\Delta\mathbf{G} = \sum_{i=1}^{d} \nu_i \mathbf{v}_i \mathbf{v}_i^\top$ be symmetric, with $|\nu_1| \geq \cdots \geq |\nu_d|$, and assume $\Delta\mathbf{G} \neq \mathbf{0}$ and $r \geq 1$. Among symmetric matrices $\mathbf{A}$ of rank at most r and Frobenius norm at most one, an optimizer of $\mathrm{tr}(\mathbf{A}\Delta\mathbf{G})$ is

$$\mathbf{A}^* = \frac{\sum_{i=1}^{r} \nu_i \mathbf{v}_i \mathbf{v}_i^\top}{\left(\sum_{i=1}^{r} \nu_i^2\right)^{1/2}}. \tag{15}$$

The statement identifies the best *expected quadratic contrast* under its stated constraints. It does not guarantee that empirical training recovers the population eigendirections.

Proof. The singular-value form of the von Neumann trace inequality, followed by Cauchy-Schwarz, bounds the objective of any feasible rank-r matrix by

$$\mathrm{tr}(\mathbf{A}\Delta\mathbf{G}) \leq \sum_{i=1}^{r} \sigma_i(\mathbf{A})|\nu_i| \leq \|\mathbf{A}\|_{\mathrm{F}} \Big(\sum_{i=1}^{r} \nu_i^2\Big)^{1/2} \leq \Big(\sum_{i=1}^{r} \nu_i^2\Big)^{1/2}. \tag{16}$$

The matrix in Eq. (15) has rank at most r and Frobenius norm one. It aligns each retained eigenvector with its corresponding signed eigenvalue, so its trace objective attains the last bound in Eq. (16). This proves the proposition, including the need for both positive and negative coefficients when the dominant eigenvalues have different signs.

The proposition explains the two-branch parameterization in Eq. (9): positive directions contribute match-enriched responses, while negative directions subtract mismatch-enriched responses. In DSM the branch matrices are trained through the chosen vision-language loss, and the normalized score in Eq. (10) is not the linear trace objective optimized by the proposition. The spectral solution is therefore a principled design guide for a compact signed measurement, not an assertion of exact parameter recovery.