A Unified Spectral Theory of Multimodal Losses

Yu-Ang Cheng[1] **Sixuan Chen**[1] **Zhouyang Lu**[2]
Xizheng Yu[1] **Grégoire Dhimoïla**[1] **Thomas Serre**[1]

[1]Department of Cognitive and Psychological Sciences, Brown University
[2]School of Engineering, Brown University

Abstract

Vision-language models are trained with very different objectives: contrastive alignment of paired embeddings (CLIP), match/no-match prediction on pairs (SigLIP), token-space generation of one modality from the other (Flamingo, LLaVA), and latent-space prediction of its embedding (VL-JEPA). Which objective should one use, and how does the answer depend on the data? Existing theory has focused on the contrastive case. Using closed-form solutions of the linearized losses, we show that contrastive alignment, match/no-match prediction, and latent-space prediction all reduce to canonical correlation analysis, whereas token-space generation reduces to reduced-rank regression. This difference matters because nuisance variation, which is specific to one modality and should ideally be discarded, is often entangled with the content an image and its caption share, so a model relying on that content also picks up the nuisance. We show that strong entanglement in either modality can prevent those three objectives from recovering shared content free of nuisance, whereas token-space generation can tolerate it in its input as long as its target is weakly entangled. When entanglement is mild, the direction of latent-space prediction matters for retrieval: scoring candidate captions by how well they predict an image's embedding gives the highest signal-to-noise ratio of the six losses. Results on numerical and synthetic data, and on retrieval in Flickr30k, COCO, and CC3M, are consistent with these predictions.

1 Introduction

Modern vision-language models are trained with very different objectives. CLIP [1] learns by contrastive alignment of paired image and text embeddings; SigLIP [2] replaces the cross-batch softmax with a per-pair binary classifier; text-generative models such as Flamingo [3] and LLaVA [4] learn by predicting text tokens from images; and VL-JEPA [5] predicts a latent representation of text from its paired image. Often treated as interchangeable, these objectives define different learning problems. Which should one train with, and how does the answer depend on the data? Answering this requires separating how a model is trained from what it is later used for. We ask two things of each training objective: what shared information it recovers from the data, and how well it then retrieves the caption that matches an image from a set of candidates, or the reverse.

These models instantiate four recurring families of cross-modal objectives: contrastive alignment, match/no-match prediction, token-space generation (predicting raw tokens of one modality from the other), and latent-space prediction (predicting a latent representation instead). Because the latter two predict one modality from the other, each can run in either direction, image to text or text to image. The four families therefore yield six losses, which are the objects of our analysis; Table 1 shows which of them representative models use.[1]

[1]Appendix A.1 provides a broader version of this table covering additional models.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

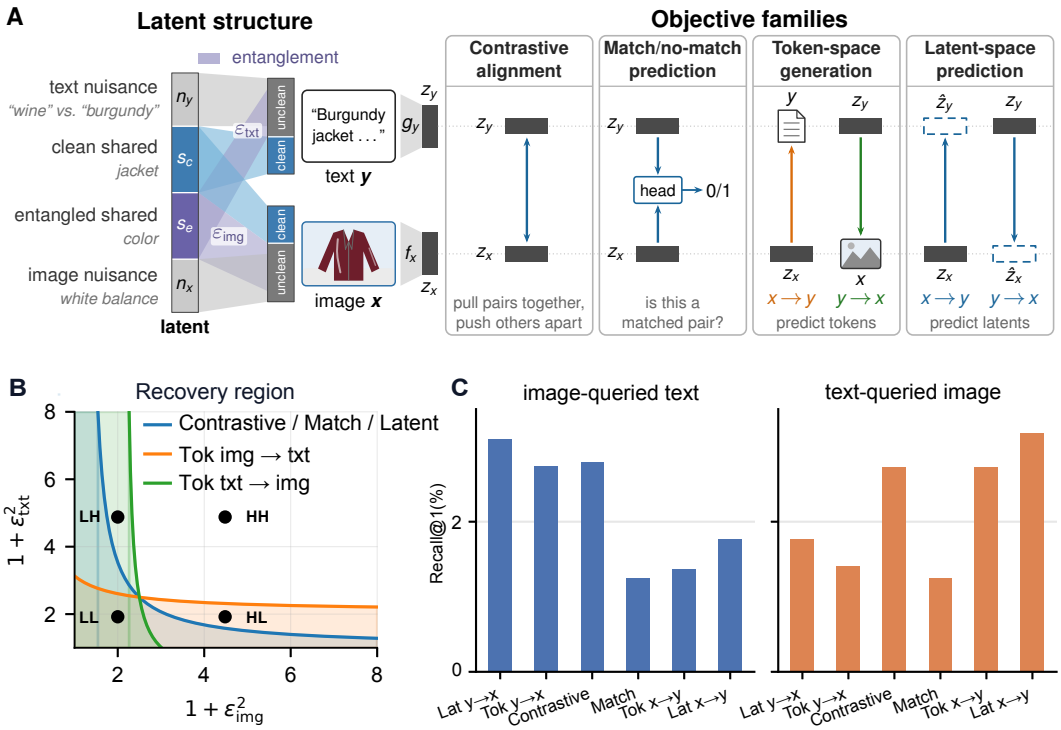


Figure 1: **(A)** Entangled-shared latent structure and the four objective families: contrastive alignment, match/no-match prediction, token-space generation (two directional variants), and latent-space prediction (two directional variants). Clean coordinates carry only clean shared factors, and unclean coordinates carry the rest; x denotes the image and y the text. **(B)** Clean-coordinate recovery boundaries in entanglement-strength space (ε_{img}, ε_{txt}), with four settings: both low (LL), image high only (HL), text high only (LH), and both high (HH), with dot positions schematic. Shaded areas are the recovery regions, where recovery is guaranteed. Contrastive, match/no-match, and latent-space prediction share a single boundary that depends on entanglement on both sides, while Tok $x \to y$ and Tok $y \to x$ are limited mainly along their target axes. **(C)** Recall@1 at LL in the numerical experiment for **image-queried text** (left) and **text-queried image** (right) retrieval. Here, query-matched latent-space prediction has the highest Recall@1 in both directions (Lat $y \to x$ for image queries, Lat $x \to y$ for text queries).

To the best of our knowledge, existing theory of multimodal training has focused almost exclusively on one of these families: contrastive alignment. Spectral analyses with linear encoders characterize the cross-modal subspace that contrastive learning recovers [6], and identifiability results characterize what semantic content it provably recovers [7, 8]; neither extends to the other three families.[2] Here we extend both lines of analysis to all four objective families, and characterize when they are equivalent, when they diverge, and how each determines retrieval.

Our main contributions are as follows:

- A linear spectral analysis showing that contrastive alignment, match/no-match prediction, and latent-space prediction recover the same whitened cross-modal eigenspaces, the solution of canonical correlation analysis, while token-space generation recovers target-weighted eigenspaces, the solution of reduced-rank regression (Section 2).
- A formalization of the latent structure in image–text data, introducing entangled-shared factors: shared semantic content observed through coordinates that also carry modality-specific nuisance (Section 3).
- Objective-specific recovery conditions: contrastive alignment, match/no-match prediction, and latent-space prediction can fail when *either* of the two modalities has high entanglement. Directed token-space objectives can tolerate high entanglement on the *source* side when the target side is

[2] Appendix A.2 provides a more detailed discussion about related work.

Table 1: Four objective families (six losses) used by representative vision-language models.

Loss	LXMERT [12]	UNITER [13]	CLIP [1]	ALBEF [14]	FLAVA [15]	CoCa [16]	Flamingo [3]	SigLIP [2]	BLIP-2 [17]	LLaVA [4]	VL-JEPA [5]	SigLIP 2 [18]
Contrastive align.	–	–	✓	✓	✓	✓	–	–	✓	–	–	–
Match/no-match	✓	✓	–	✓	✓	–	–	✓	✓	–	–	✓
Token img→txt	✓	✓	–	✓	✓	✓	✓	–	✓	✓	–	✓
Token txt→img	–	–	–	–	✓	–	–	–	–	–	–	–
Latent pred. img→txt	–	–	–	–	–	–	–	–	–	–	✓	–
Latent pred. txt→img	✓	✓	–	–	–	–	–	–	–	–	–	–

weakly entangled, so they can remain identifiable in asymmetric regimes where the other three families fail (Section 3).

- A retrieval ranking on the recovered subspace: even when different objectives all recover shared clean coordinates, the highest pairwise signal-to-noise ratio is given by the latent-space prediction objective whose target is the modality used as the query, for example predicting the image when searching with an image (Section 3).
- Empirical validation on numerical data, synthetic image–text data with vision and language encoders trained from scratch, and Flickr30k [9], COCO [10], and CC3M [11] with off-the-shelf vision and language backbones (Section 4).

2 Four Objective Families Collapse to Two Under Linearization

We now compare the four objective families by the statistical information that each encourages a representation to preserve, treating directional variants separately. Real models differ in architecture, scale, tokenization, and optimization, so no single system can be analyzed literally. We therefore study a local linear setting. We do not claim that vision–language models are linear; this setting simply makes the objectives analytically comparable.

Let $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ denote centered features from two modalities (e.g. image and text). Each modality is mapped to an m-dimensional embedding space by encoders $z_x = f_\theta(x)$ and $z_y = g_\phi(y)$, subject to whitening constraints $\mathbb{E}[z_x z_x^\top] = \mathbb{E}[z_y z_y^\top] = I_m$ to rule out collapse and scale degeneracies. [3]

Proposition 1 (Optimal solutions of linearized loss functions)**.** *For the linearized objectives in Table 2, with $f_\theta(x) = W_x x$ and $g_\phi(y) = W_y y$, after the auxiliary parameters are optimized out, the optimal encoders satisfy* $\mathrm{row}(W_x) = \mathrm{span}\{a_1^\top, \ldots, a_m^\top\}$ *and* $\mathrm{row}(W_y) = \mathrm{span}\{b_1^\top, \ldots, b_m^\top\}$*, where a_i, b_i are the top-m generalized eigenvectors of the corresponding eigenproblems. Token-space generation constrains only the encoder of its input modality. Detailed proofs are given in Appendix A.3.*

Here Σ_{xx}, Σ_{yy} are the covariances of x, y and Σ_{xy} their cross-covariance; $\rho_1 \geq \rho_2 \geq \cdots$ are the singular values of $A = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$, and λ_i are the squared singular values of $\Sigma_{xx}^{-1/2}\Sigma_{xy}$ (Tok $x \to y$) or $\Sigma_{yy}^{-1/2}\Sigma_{yx}$ (Tok $y \to x$).[4] The original objectives and their notation are described in Appendix A.1.

The next theorem (proved in Appendix A.3.2) shows when the restriction to linear encoders is without loss of optimality.

Theorem 1 (Beyond linear encoders)**.** *Let (x, y) be jointly Gaussian, and replace $W_x x$ and $W_y y$ in the surrogates of Table 2 by* ***arbitrary*** *centered whitened encoders $f(x)$ and $g(y)$, keeping B, V, and M as matrices acting on the encoder outputs. For token-space generation, and for contrastive and latent-space prediction whenever $\rho_m > \rho_1^2$, the optimum is attained by linear encoders, so*

[3] Whitening only fixes the effective rank and the coordinates of the representation: any centered full-rank representation is whitened by an invertible linear map, which leaves the subspace it represents unchanged. Our recovery results concern this subspace, not its coordinates.

[4] Only the match row involves fourth moments, which we evaluate for jointly Gaussian (x, y); the other rows depend only on second moments.

Table 2: Linearized solutions for the four objective families (six losses). Each nonlinear auxiliary map is replaced by its leading-order form, the loss by a quadratic surrogate, and the optimal encoders solve the listed eigenproblem.

Loss	Original objective	Linearized surrogate	Encoder eigenproblem
Contrastive	$\mathrm{InfoNCE}\big(f(x), g(y)\big)$	$\mathbb{E}\|W_x x - W_y y\|^2$	$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\, a_i = \rho_i^2\, \Sigma_{xx} a_i$ $\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\, b_i = \rho_i^2\, \Sigma_{yy} b_i$
Match	$\mathrm{BCE}\big(s_\psi(f(x), g(y)), 1\big)$	$\mathbb{E}_+\big(1 - (W_x x)^\top M W_y y\big)^2$	
$+$	$\mathrm{BCE}\big(s_\psi(f(x), g(\tilde{y})), 0\big)$	$+\,\mathbb{E}_-\big((W_x x)^\top M W_y \tilde{y}\big)^2$	
Lat $x \to y$	$\mathbb{E}\|g(y) - h_{y\leftarrow x}(f(x))\|^2$	$\mathbb{E}\|W_y y - B_{y\leftarrow x} W_x x\|^2$	
Lat $y \to x$	$\mathbb{E}\|f(x) - h_{x\leftarrow y}(g(y))\|^2$	$\mathbb{E}\|W_x x - B_{x\leftarrow y} W_y y\|^2$	
Tok $x \to y$	$\mathrm{CE}\big(y, \tilde{h}_{y\leftarrow x}(f(x))\big)$	$\mathbb{E}\|y - V_{y\leftarrow x} W_x x\|^2$	$\Sigma_{xy}\Sigma_{yx}\, a_i = \lambda_i\, \Sigma_{xx} a_i$
Tok $y \to x$	$\mathrm{CE}\big(x, \tilde{h}_{x\leftarrow y}(g(y))\big)$	$\mathbb{E}\|x - V_{x\leftarrow y} W_y y\|^2$	$\Sigma_{yx}\Sigma_{xy}\, b_i = \lambda_i\, \Sigma_{yy} b_i$

Match: $M \in \mathbb{R}^{m\times m}$ is a bilinear head; a multihead variant leads to the same eigenproblem (Appendix A.3.8).

Contrastive: following Cai et al. [8], a regularizer on the marginals, here whitening, replaces the negatives of InfoNCE.

Proposition 1 gives the global optimum over nonlinear encoders. For match, the same holds, with no condition on ρ, when f and g are restricted to the block-structured class of Appendix A.3.2.

Takeaway Our analysis shows that contrastive, match, and latent-prediction losses, which we refer to as joint objectives, all extract the same statistical structure from the data, despite their very different implementations. In the linear setting, each performs canonical correlation analysis, which ranks directions by how strongly the two modalities correlate along them. Token-space generation is the exception. It performs reduced-rank regression, which also accounts for how much variance each direction carries in the target. Latent-space prediction is therefore a notable case: it resembles token-space generation in architecture, since both predict one modality from the other, but its solution matches that of contrastive alignment.

This suggests that if a contrastive model underperforms on a given dataset, switching to a match or latent-prediction objective is unlikely to recover a different subspace, because all three look for the same signal. A failure to recover the shared structure probably lies in the data (e.g. weak cross-modal correlation). A meaningful change in what is recovered requires moving to the other family—token-space generation—which is sensitive to a different aspect of the data.

3 Shared-Private Entanglement Governs Clean Recovery and Retrieval Performance

Section 2 leaves open two questions: (i) what data properties distinguish the two classes of objective families, and (ii) whether the four losses sharing a linearized solution perform equally well at test time. Both questions turn on a property of real image–text data that is often overlooked: modality-specific nuisance is rarely semantically neutral. Even how a photo is lit, or how its caption is phrased, can hint at what the pair is about.

3.1 Shared Semantics Can Entangle Modality-Specific Nuisance

We think of an image and its caption as generated by two kinds of factors: shared semantic factors (s), which both of them describe, and each modality's own nuisance (n_x and n_y), which the other modality does not carry. Ideally, the two occupy separate coordinates ($x = (s, n_x); y = (s, n_y)$). In that case, discarding the nuisance coordinates would leave the shared semantics intact.

In real image–text data, shared semantic content is seldom cleanly separated from modality-specific nuisance. In an online fashion catalog, for example, object categories such as *shirt*, *jacket*, or *sneakers* are clean shared factors: the image and the caption refer unambiguously to the same concept. Color is different. It appears as pixel values in the image and as words in the caption, but the pixels also depend on camera exposure, white balance, and post-processing, and the words on the retailer's color taxonomy ("wine" versus "burgundy") and caption template. Color is therefore still

shared, but it is observed in coordinates that also carry modality-specific nuisance. Discarding these coordinates, which was harmless in the ideal case, would now also remove part of the shared signal.

To model this, we split the shared factors into $s = (s_c, s_e)$: *clean shared* factors s_c, which, as in the ideal case, occupy only coordinates of their own, and *entangled-shared* factors s_e, which occupy coordinates of their own but are, in addition, observed in the coordinates that carry n_x or n_y (Figure 1A). That is, the ideal $x = (s, n_x)$ becomes $x = (s_c, s_e, C_x s_e + n_x)$, and likewise for y, where for $a \in \{x, y\}$, the matrix C_a encodes the entanglement of s_e with n_a. We measure the entanglement strength by $\varepsilon_a := \|C_a\|_2$, that is, how far modality a departs from the ideal case, which is recovered at $\varepsilon_a = 0$. The central object of interest is s_e, and in particular how strong it is compared with s_c.

Formally, we take s_c, s_e, n_x, and n_y to be four zero-mean, mutually independent latent blocks, with

$$\begin{aligned} &s_c \in \mathbb{R}^k, && \mathrm{Cov}(s_c) = \mathrm{diag}(\kappa_i)_{i=1}^k, && \kappa_1 \geq \cdots \geq \kappa_k > 0, \\ &s_e \in \mathbb{R}^p, && \mathrm{Cov}(s_e) = \mathrm{diag}(\tau_i)_{i=1}^p, && \tau_1 \geq \cdots \geq \tau_p > 0, \\ &n_x, n_y \in \mathbb{R}^r, && \mathrm{Cov}(n_x), \mathrm{Cov}(n_y) \text{ arbitrary.} \end{aligned} \tag{1}$$

Allowing arbitrary orthogonal coordinate mixing and coordinate-dependent observation noise, the observations take the form

$$x = Q_x\big((s_c, s_e, C_x s_e + n_x) + \eta_x\big), \qquad y = Q_y\big((s_c, s_e, C_y s_e + n_y) + \eta_y\big), \tag{2}$$

where $Q_x, Q_y \in \mathbb{R}^{d\times d}$ are orthogonal matrices, $d = k + p + r$, and $C_x, C_y \in \mathbb{R}^{r\times p}$. We call the first k latent coordinates, which carry s_c, the *clean coordinates*. The observation noises η_x, η_y are mutually independent and independent of the latent blocks, with

$$\mathrm{Cov}(\eta_a) = \mathrm{diag}(\sigma_{a,1}^2, \ldots, \sigma_{a,d}^2), \qquad \nu_a := \min_{k<j\leq d} \sigma_{a,j}^2 > 0. \tag{3}$$

Here ν_a is the smallest observation-noise variance on the non-clean coordinates of modality a.

3.2 Directed Token Objectives Escape the Two-Sided Recovery Bottleneck

Our first question is what shared structure a rank-m multimodal objective can recover when the learned representation is required to avoid coordinates that also carry modality-specific variation. We call a solution *clean-coordinate identifiable* if its learned row space has no support on any coordinate that carries n_x or n_y. This is the property that matters when the association between nuisance and semantics changes at test time: a representation that uses those coordinates also encodes, for instance, white balance or word choice, and its cross-modal matches degrade once that association shifts.

Since both s_c and s_e are shared across modalities, one might hope that a clean-coordinate solution could recover both. However, s_e is observed not only in its own coordinates but also in coordinates that carry n_a. Recovering s_e through those coordinates would therefore also admit modality-specific nuisance. The next theorem shows that, under full-rank entanglement, imposing clean-coordinate identifiability forces the admissible shared representation to lie entirely in the clean shared block s_c.

Theorem 2 (Clean-coordinate identifiability under full-rank entanglement)**.** *Consider the entangled-shared latent model in Equation* (2)*. Assume that the entanglement matrices have full column rank:*

$$\mathrm{rank}(C_x) = p, \qquad \mathrm{rank}(C_y) = p.$$

Then any optimal rank-m solution of an objective in Table 2 that is clean-coordinate identifiable must lie entirely in the clean shared block s_c. In particular, $m \leq k$, and, for each encoder the objective trains,

$$\mathrm{row}(W_x^\star Q_x) \subseteq \mathrm{span}\{e_1^\top, \ldots, e_k^\top\}, \qquad \mathrm{row}(W_y^\star Q_y) \subseteq \mathrm{span}\{e_1^\top, \ldots, e_k^\top\}.$$

Theorem 2 (proved in Appendix A.3) shows that, under full-rank entanglement, any optimal clean-coordinate identifiable solution must lie in the s_c block. We now ask when the optima of our objectives actually select this block. The answer is governed by a spectral competition between the clean shared factors and the entangled-shared factors.

Theorem 3 (Objective-specific clean-coordinate identifiability condition)**.** *Consider the entangled-shared latent model in Equation* (2)*, and fix $m \le k$. For $i \le k$ and $a \in \{x, y\}$, let*

$$\beta_i^a := \frac{\kappa_i}{\sqrt{\kappa_i + \sigma_{a,i}^2}}, \qquad \gamma_a := \frac{(1+\varepsilon_a^2)\tau_1}{(1+\varepsilon_a^2)\tau_1 + \nu_a}.$$

Let $(\beta^x\beta^y/\kappa)_{[m]}$ and $((\beta^a)^2)_{[m]}$ denote the m-th largest values of $\{\beta_i^x\beta_i^y/\kappa_i\}_{i=1}^k$ and $\{(\beta_i^a)^2\}_{i=1}^k$, respectively, and let $W_x^\star, W_y^\star$ denote any optimal rank-m encoders of the corresponding objective in Table 2: (i) *contrastive alignment, match/no-match prediction, and latent-space prediction;* (ii) *token-space generation $x \to y$; and* (iii) *token-space generation $y \to x$. Then:*

$$\text{(i)} \quad \gamma_x\gamma_y < \left((\beta^x\beta^y/\kappa)_{[m]}\right)^2 \;\Rightarrow\; \mathrm{row}(W_x^\star Q_x), \mathrm{row}(W_y^\star Q_y) \subseteq \mathrm{span}\{e_1^\top, \dots, e_k^\top\};$$

$$\text{(ii)} \quad \frac{\nu_y\gamma_x\gamma_y}{1-\gamma_y} < ((\beta^x)^2)_{[m]} \;\Rightarrow\; \mathrm{row}(W_x^\star Q_x) \subseteq \mathrm{span}\{e_1^\top, \dots, e_k^\top\};$$

$$\text{(iii)} \quad \frac{\nu_x\gamma_x\gamma_y}{1-\gamma_x} < ((\beta^y)^2)_{[m]} \;\Rightarrow\; \mathrm{row}(W_y^\star Q_y) \subseteq \mathrm{span}\{e_1^\top, \dots, e_k^\top\}.$$

Moreover, each inequality is also necessary in the worst case: if it fails, some model with the same κ_i, $\sigma_{a,i}^2$, τ_1, ε_a, and ν_a has an optimal encoder that violates the corresponding inclusion.[5]

Theorem 3 (proved in Appendix A.3) gives a worst-case necessary and sufficient condition for clean-block recovery: joint objectives are controlled by the two-sided bottleneck $(\beta^x\beta^y/\kappa)_{[m]}$, while token-space generation is controlled by the one-sided bottleneck $((\beta^a)^2)_{[m]}$ of its source modality a. The corresponding recovery regions in entanglement-strength space let us compare objectives directly: if one recovery region contains another, the objective with the smaller region is dominated for clean-coordinate identifiability; if they overlap only partially, neither uniformly dominates the other.

Let the recovery regions be

$$\begin{aligned} \mathcal{F}_{\text{joint}} &:= \{(\gamma_x, \gamma_y) : \gamma_x\gamma_y < ((\beta^x\beta^y/\kappa)_{[m]})^2\}, \\ \mathcal{F}_{\text{tok},\, x\to y} &:= \{(\gamma_x, \gamma_y) : \nu_y\gamma_x\gamma_y/(1-\gamma_y) < ((\beta^x)^2)_{[m]}\}, \\ \mathcal{F}_{\text{tok},\, y\to x} &:= \{(\gamma_x, \gamma_y) : \nu_x\gamma_x\gamma_y/(1-\gamma_x) < ((\beta^y)^2)_{[m]}\}. \end{aligned} \tag{4}$$

Theorem 4 (Isotropic noise implies token-space coverage)**.** *Assume that*

$$\sigma_{x,i}^2 = \sigma_x^2 > 0, \qquad \sigma_{y,i}^2 = \sigma_y^2 > 0, \qquad \forall i \le d.$$

Then, for every rank $m \le k$,

$$\mathcal{F}_{\text{joint}} \subseteq \mathcal{F}_{\text{tok},\, x\to y} \cup \mathcal{F}_{\text{tok},\, y\to x}.$$

Theorem 4 (proved in Appendix A.3) shows that, in the isotropic baseline, every entanglement configuration certified by the joint condition is already certified by at least one token-space condition, so joint objectives create no additional guaranteed clean recovery regimes.

Takeaway To guarantee clean recovery, contrastive alignment, match/no-match prediction, and latent-space prediction need clean shared factors to spectrally dominate entangled-shared factors *jointly* across both modalities, so strong entanglement on one side must be offset by weak entanglement on the other; token-space generation is mainly limited by entanglement on its *target* side. When entanglement is strong in one modality but weak in the other, token-space generation training to predict the low-entanglement side can still recover clean coordinates, while every joint objective can fail.

3.3 Query-Matched Latent Prediction Has the Highest SNR After Clean Recovery

Section 3.2 suggests that if a joint objective fails to recover clean shared structure, switching to token-space generation toward the low-entanglement modality can restore clean recovery without changing the data or architecture. However, when entanglement is mild enough that all objectives succeed, the

[5] E.g., $C_a = \varepsilon_a e_1 e_1^\top$, $\mathrm{Cov}(n_a) = 0$, and $\sigma_{a,j}^2 = \nu_a$ for all $j > k$ (Appendix A.3).

identifiability analysis above no longer distinguishes them: they all recover clean coordinates, but not necessarily the same ones. This brings us to our second question. Once several objectives have recovered clean coordinates, do they perform equally well at test time? We address this question under a classic image-text retrieval setting.

In retrieval, a model searches with an input from one modality, the query modality, for its match in the other. Given a query image x with positive text y, a model is presented with y and N negative texts $\tilde{y}_1, \ldots, \tilde{y}_N$, drawn i.i.d. from the marginal distribution of y and independently of (x, y). A retrieval score $q_\ell(x, \cdot)$ ranks the $N + 1$ candidates, with smaller values preferred. Each $q_\ell(x, y)$ is the per-pair loss of objective ℓ in Table 2, evaluated on the candidate pair as if it were a positive pair (Appendix A.3 lists all six scores). The query is correctly retrieved when the true positive text achieves the smallest score among all candidates. The fraction of queries for which this happens is the standard Recall@1 metric, denoted as $\mathrm{R@1}^{\ell}_{x\text{-queried } y,N}$ where $\ell \in \{\mathrm{contrastive}, \mathrm{match}, \mathrm{tok}_{x\to y}, \mathrm{tok}_{y\to x}, \mathrm{lat}_{x\to y}, \mathrm{lat}_{y\to x}\}$.

Recall@1 is a rank statistic that is difficult to compare across objectives in closed form. We therefore introduce a pairwise signal-to-noise ratio (SNR) that is analytically tractable and, as Lemma 1 will show, bounds Recall@1.

For x-queried y retrieval with a single negative text $\tilde{y}$, define

$$\Delta^{\ell}_{x\text{-queried } y} := q_\ell(x, \tilde{y}) - q_\ell(x, y), \qquad \mathrm{SNR}^{\ell}_{x\text{-queried } y} := \frac{\mathbb{E}[\Delta^{\ell}_{x\text{-queried } y}]}{\sqrt{\mathrm{Var}(\Delta^{\ell}_{x\text{-queried } y})}}. \tag{5}$$

Lemma 1 (Recall@1–SNR correspondence)**.** *Suppose that, conditional on (x, y), the negatives $\tilde{y}_1, \ldots, \tilde{y}_N$ are i.i.d., and write $\Delta^{\ell}_{x\text{-queried } y} = \sum_{i=1}^{m} Z^{\ell}_i$, where the independent Z^{ℓ}_i's satisfy the Lindeberg condition. If $\mathrm{SNR}^{\ell}_{x\text{-queried } y}$ converges to a finite limit, then, for fixed N,*

$$\Phi(\mathrm{SNR}^{\ell}_{x\text{-queried } y})^N + o(1) \le \mathrm{R@1}^{\ell}_{x\text{-queried } y,N} \le \Phi(\mathrm{SNR}^{\ell}_{x\text{-queried } y}) + o(1).$$

Moreover, if $\mathrm{Var}\Big(\mathbb{P}(\Delta^{\ell}_{x\text{-queried } y} > 0 \mid x, y)\Big) \to 0$, then

$$\mathrm{R@1}^{\ell}_{x\text{-queried } y,N} = \Phi(\mathrm{SNR}^{\ell}_{x\text{-queried } y})^N + o(1).$$

By Lemma 1 (proved in Appendix A.3), once clean coordinates have been recovered, we compare retrieval performance through the pairwise SNRs induced by the different objective-specific scores, under jointly Gaussian assumptions. We obtain two structural results: (i) each directional objective (token-space generation or latent-space prediction) favors the retrieval direction whose query modality matches its prediction target, and (ii) for a fixed retrieval direction, query-matched latent-space prediction has the largest SNR among all six objectives.

Theorem 5 (Query-matched direction advantage)**.** *Under clean-coordinate identifiability of the optimal encoders,*

$$\mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried } y} \ge \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{y\text{-queried } x}, \qquad \mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y} \ge \mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{y\text{-queried } x}.$$

The same inequalities hold with x and y swapped (proof in Appendix A.3).

Theorem 6 (Latent prediction has the highest retrieval SNR)**.** *Under clean-coordinate identifiability of the optimal encoders, for x-queried y retrieval,*

$$\mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y} \ge \max\Big\{\mathrm{SNR}^{\mathrm{contrastive}}_{x\text{-queried } y}, \mathrm{SNR}^{\mathrm{match}}_{x\text{-queried } y}, \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried } y}, \mathrm{SNR}^{\mathrm{tok}_{x\to y}}_{x\text{-queried } y}, \mathrm{SNR}^{\mathrm{lat}_{x\to y}}_{x\text{-queried } y}\Big\}.$$

The same inequality holds with x and y swapped (proof in Appendix A.3).

Both theorems follow from how a score uses the query (proofs in Appendix A.3). A $y \to x$ objective scores a pair by how well y predicts x. With x as the query, all candidates are judged against the same target; with y as the query, the target changes from candidate to candidate, and the candidates' own variation enters the comparison. This gives Theorem 5, and it is also why the $x \to y$ objectives lose in Theorem 6. Among the remaining scores, $\mathrm{lat}_{y\to x}$ compares the whitened query with its best linear prediction from the candidate, which shrinks each coordinate by its correlation ρ_i; contrastive alignment omits this shrinkage, token-space generation skips the whitening, and match/no-match keeps only the bilinear cross term.

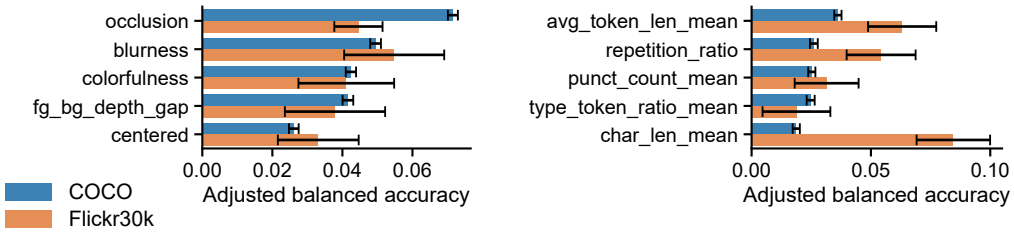


Figure 2: Adjusted balanced accuracy (zero = chance) of logistic-regression probes predicting semantic super-categories from a single **visual** (left) or **text** (right) nuisance attribute, on COCO and Flickr30k. Nuisances predict shared semantics above chance in both modalities, providing correlational evidence consistent with the entangled-shared structure assumed in Section 3. See Appendix A.4 for additional results.

Takeaway Once clean coordinates are recovered, choose the latent-space prediction objective whose *prediction target matches the query modality*: train $\mathrm{lat}_{y\to x}$ for image-queried text retrieval, and $\mathrm{lat}_{x\to y}$ for text-queried image retrieval. Symmetric scores (contrastive, match/no-match) cannot exploit the choice of query modality.

Together, these results yield a prescription: when one modality carries substantially stronger entanglement, move from the joint objectives to token-space generation toward the low-entanglement modality to restore clean recovery; when entanglement is mild enough that all objectives succeed, select the query-matched latent prediction objective, which shares the clean-recovery guarantee of the other joint objectives yet achieves the highest retrieval SNR.

4 Empirical Validation

4.1 Numerical validation

First, we validate the closed-form spectral calculations. Starting from random initialization, we optimize each objective by gradient descent and compare the learned spectra and row spaces with the analytic solutions in Section 2: All cases converge to their closed-form targets (see Appendix A.4). Second, we sweep the entanglement strengths $(\varepsilon_x, \varepsilon_y)$ and compare the theoretical recovery boundaries with empirical projector distances. The low-entanglement regions match Theorem 3: directed token-space objectives remain accurate in asymmetric regimes where joint objectives fail (see Appendix A.4). Finally, in low-entanglement regimes where all objectives recover the clean block, we evaluate Recall@1 at the ground-truth rank k. The ordering is consistent with Theorem 6 and Theorem 5: the query-matched latent-space prediction score gives the highest Recall@1, while the opposite latent direction performs worse, as predicted (see Figure 1C).

4.2 Real Image–Text Data Contains Semantically Predictive Modality-Specific Nuisance

The theory assumes that modality-specific nuisances can spuriously carry shared semantic information. We test this premise on Flickr30k and COCO, chosen for their rich object- and caption-level annotations, by asking whether modality-specific nuisances predict semantic super-categories. For each image–caption pair, we define the semantic super-category as that of an object both annotated in the image and mentioned in the caption, ensuring that the label reflects cross-modal shared content.

For each dataset and modality, we train a logistic-regression probe from a single nuisance to the semantic super-category label. We report adjusted balanced accuracy, where zero corresponds to chance and positive values indicate above-chance prediction. Figure 2 shows a subset of results; additional results are reported in Appendix A.4. The above-chance prediction is correlational evidence consistent with the modeling assumption that shared semantic content is not cleanly separated from modality-specific variation in real image–text datasets, whatever its source (acquisition bias, annotation conventions, curation, or intrinsic interactions).

Table 3: R@1 (%) under four entanglement settings (LL/HL/LH/HH, corresponding to dots in Figure 1B), averaged over five seeds. Percentage changes are relative to the LL baseline; absolute R@1 in parentheses. I→T denotes image-query text retrieval, and T→I denotes text-query image retrieval. Contrastive, match, and latent-space prediction degrade at both HL and LH. Token-space objectives degrade only when their target side is high. See standard deviations in Table A2.

	img L, txt L (baseline)		**img H, txt L**		**img L, txt H**		**img H, txt H**	
Loss	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
Lat $x \to y$	0.47	**1.17**	-9.3% (0.43)	-31.0% (0.81)	-6.3% (0.44)	-25.9% (0.87)	-6.8% (0.44)	-28.3% (0.84)
Tok $x \to y$	0.52	1.14	**+0.0% (0.52)**	**-5.1% (1.08)**	-54.2% (0.24)	-47.4% (0.60)	-48.5% (0.27)	-45.8% (0.62)
Match	0.80	0.85	-26.3% (0.59)	-29.0% (0.61)	-25.3% (0.60)	-32.3% (0.58)	-27.5% (0.58)	-32.6% (0.58)
Contrastive	0.95	0.94	-27.4% (0.69)	-37.9% (0.58)	-31.0% (0.65)	-24.3% (0.71)	-29.1% (0.67)	-27.2% (0.68)
Tok $y \to x$	1.16	0.40	-51.0% (0.57)	-32.2% (0.27)	**-1.5% (1.15)**	**+13.6% (0.45)**	-54.0% (0.54)	-30.2% (0.28)
Lat $y \to x$	**1.19**	0.57	-32.7% (0.80)	-30.3% (0.40)	-28.5% (0.85)	-23.9% (0.43)	-32.3% (0.81)	-35.9% (0.36)

4.3 Controlled Entanglement Degrades the Objectives as Predicted

We next test the identifiability predictions in a controlled synthetic image–text dataset. Each example (see Figure 3) is generated from shared semantic factors and modality-specific private factors. The clean shared factors determine object attributes, including shape, color, and size. Additional shared semantic factors, including coarse position and style, are allowed to entangle with private image factors, including background pattern, badge corner, rotation, and jitter, and with private text factors, including template family, wrapper style, pseudo-word, and detailedness. We consider four entanglement regimes by setting the image-side and text-side entanglement strengths to low or high values, yielding LL/HL/LH/HH settings for ε_x and ε_y, respectively (corresponding to the dots in Figure 1B).

For each entanglement setting, we generate paired images and captions and train image and text encoders from scratch. The image encoder is a four-block strided CNN; the text encoder is a three-layer Transformer. We train each model with one of the four objective families. Evaluation is performed on a clean test distribution with no entanglement. This directly tests identifiability: if a loss recovers the clean coordinates, its retrieval performance should remain stable on clean data.

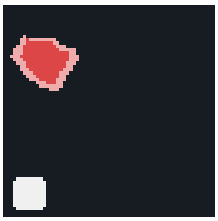


Figure 3: Caption: "Shown is a red egg of large size in the left region. It sits in the upper band and has a solid look." See Table 3 for corresponding results.

Table 3 confirms the predicted identifiability pattern at all four settings: contrastive, match, and latent-space prediction degrade at both HL and LH while token-space generation degrades only when its *target* side is high, consistent with Theorem 3.

At the LL baseline, where all objectives recover clean coordinates, query-matched latent-space prediction achieves the highest R@1 in both retrieval directions ($\mathcal{L}^{\text{lat}}_{y\to x}$: R@1 $= 1.19\%$ for I→T; $\mathcal{L}^{\text{lat}}_{x\to y}$: R@1 $= 1.17\%$ for T→I), consistent with Theorem 6.

4.4 Query-Matched Latent Prediction Gives the Best Retrieval Performance

We next test the retrieval prediction on real image–text data, comparing all six losses on Flickr30k, COCO, and CC3M with a DINOv2 [19] backbone for images and a ModernBERT [20] backbone for text. On top of these frozen features, we fit the retrieval scores induced by each loss mode, using nonlinear heads for Flickr30k and COCO and linear heads for CC3M due to resource constraints. Because the true number of shared latent factors is unknown for real data, we sweep over bottleneck ranks and report R@1 at the best rank (peak), a cross-validated rank, and full rank $k = 768$. Since the true data-generating process is unknown, we cannot directly measure clean-coordinate identification as in the controlled experiment, and instead test the retrieval-level prediction: the latent prediction score whose target modality matches the query modality should give the highest R@1. Table 4 shows that the directional prediction of Theorem 6 holds across all three datasets. For

Table 4: R@1 (%) on CC3M, COCO, and Flickr30k across all six losses, with the bottleneck rank k chosen in three ways: the best R@1 over the rank sweep (peak), k selected on four of five query folds and evaluated on the held-out fold (cross-validated), and the full rank $k = 768$ (fixed). I→T denotes image-query text retrieval, and T→I denotes text-query image retrieval. For each retrieval direction and each choice of k, the query-matched latent-space prediction objective achieves the strongest R@1 (bolded).

	Peak						**Cross-validated**						**Fixed**					
	CC3M		COCO		Flickr30k		CC3M		COCO		Flickr30k		CC3M		COCO		Flickr30k	
Loss	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
Lat $x \to y$	1.3	**9.9**	9.2	**23.4**	24.7	**44.7**	1.3	**9.9**	9.2	**23.4**	24.7	**44.3**	0.0	**9.9**	1.5	**23.2**	5.2	**44.3**
Tok $x \to y$	2.2	8.2	10.3	19.0	28.9	40.6	2.2	8.2	10.0	19.0	28.2	40.6	2.2	8.2	10.2	18.9	28.1	40.1
Match	6.3	6.8	24.0	20.3	53.5	43.0	6.3	6.8	23.7	20.1	52.2	43.0	6.3	6.8	23.8	20.3	53.1	43.0
Contrastive	5.0	7.3	24.2	18.3	46.0	38.4	5.0	7.3	24.2	18.3	44.4	38.4	1.9	1.0	17.2	6.8	31.6	19.4
Tok $y \to x$	8.1	1.5	22.0	3.9	53.0	17.6	8.1	1.5	21.9	3.8	52.4	17.6	8.1	1.5	21.9	3.9	53.0	17.3
Lat $y \to x$	**9.6**	2.4	**28.4**	6.9	**55.0**	18.2	**9.6**	2.4	**28.4**	6.9	**54.4**	18.2	**9.6**	0.0	**27.7**	0.2	**54.9**	0.8

image-queried text retrieval, $\mathcal{L}^{\text{lat}}_{y\to x}$ achieves the highest R@1 on all datasets while for text-queried image retrieval, $\mathcal{L}^{\text{lat}}_{x\to y}$ achieves the highest R@1 on all datasets.

5 Conclusion & Limitations

Vision-language objectives are not interchangeable. In the linear setting, contrastive alignment, match/no-match prediction, and latent-space prediction perform canonical correlation analysis, whereas token-space generation performs reduced-rank regression. When shared semantics are entangled with modality-specific nuisance, recovering them may also bring in that nuisance. In this case, the first three objectives can fail in asymmetric regimes where a directed token-space objective can still succeed. Finally, once all objectives recover clean coordinates, query-matched latent-space prediction gives the strongest retrieval performance.

A limitation of our analysis is the reliance on a linear approximation. While this may not fully capture the behavior of modern deep neural networks, similar linearized analyses have proven useful for understanding multimodal contrastive learning [6] and the inductive biases of unimodal self-supervised objectives [21–24].

Our formal results concern population optima of the surrogates in Table 2 (or, under Theorem 1, of nonlinear encoders), not complete CLIP-, SigLIP-, or JEPA-style systems, and do not model finite batches, learned temperatures, optimization dynamics, or autoregressive decoding. The entangled-shared model is a covariance-level abstraction, and $\varepsilon_x, \varepsilon_y$ are not directly observable, so our advice to switch objective families is meant for use after a failure is observed. Our experiments support the predictions but do not prove them. Our retrieval results compare each objective's own score; we do not rank the objectives on generation, reasoning, or instruction following [3, 4], or compare them through large-scale pretraining from scratch.

Our results may appear to reverse findings from unimodal self-supervised learning, where latent-space prediction is preferred under high noise and token-space generation under low noise [22–24]. The difference is that unimodal noise is generic within-modality randomness that corrupts the reconstruction target uniformly, whereas multimodal entanglement couples shared semantics with modality-specific nuisance in specific coordinates. A directed token-space objective can avoid this coupling by predicting the cleaner modality, an option with no unimodal analogue.

Acknowledgments and Disclosure of Funding

This work was supported by ONR (N00014-24-1-2026), the NSF AI Research Institute on Interaction for AI Assistants (ARIA; NSF Cooperative Agreement 2433429), and ANITI (France 2030, ANR-23-IACL-0002). Computing hardware was supported by NIH grant S10OD036341 (Carney Condo, RRID:SCR_029056).

References

[1] 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 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[2] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[5] Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, 2026.

[6] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4348–4380. PMLR, 2023.

[7] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

[8] Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[9] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics.

[12] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, 2019.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705, 2021.

[15] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[16] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

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

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[20] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria, 2025. Association for Computational Linguistics.

[21] Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 19511–19522, 2022.

[22] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

[23] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[24] Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint-embedding vs reconstruction: Provable benefits of latent space prediction for self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020.

[27] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. In *International Conference on Learning Representations*, 2023.

[28] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[29] Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650(8101):327–333, 2026.

[30] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

[31] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011, 2021.

[32] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR, 2021.

[33] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467, 2021.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Appendix

A.1 Background

This section reviews the six dominant losses (four objective families) used in modern vision–language pretraining. We organize them by whether the loss scores image–text pairs (contrastively or by binary classification) or predicts one modality from the other (in either direction), and whether that prediction is made in token space or latent space.

Contrastive alignment. This family of losses is built on InfoNCE [25, 26], which pulls the embeddings $f(x)$ and $g(y)$ of a paired image x and text y toward each other while pushing negatives apart. InfoNCE is an $(N+1)$-way softmax cross-entropy in which, for each image, its paired text is the single positive among $N+1$ candidate texts and the other N are negatives. Negative pairs are not assigned separate classification targets; instead, they enter the softmax normalization, and their scores are penalized relative to the positive pair. CLIP [1] fixes the implementation choices that InfoNCE leaves open: the candidates are the other pairs in the mini-batch, the score is cosine similarity with a learnable temperature, the image and text embeddings come from two separate encoders that do not interact before the score is computed, and the loss is applied symmetrically in both directions. The same contrastive loss is also retained as one component of multi-objective models [14–17, 27].

Match/no-match prediction. A second family replaces batch-relative multiclass discrimination with binary match/no-match supervision, assigning positive and negative labels to image–text pairs without normalizing their scores against a shared candidate set. Instead of asking which text in the batch belongs to an image, it asks each pair in isolation, so every score becomes its own binary classification, with loss $\mathrm{BCE}(s_\psi(f(x), g(y)), 1)$ for a matched pair and $\mathrm{BCE}(s_\psi(f(x), g(\tilde{y})), 0)$ for an unmatched one, where s_ψ is the score, BCE the binary cross-entropy, and $\tilde{y}$ a caption drawn from the marginal distribution of y independently of x. This image–text matching (ITM) formulation was introduced on *fused* cross-modal representations [12–15, 17, 27]. SigLIP [2] and SigLIP 2 [18] later apply the same objective directly to pooled unimodal embeddings via a sigmoid loss.

Image-to-text token-space generation. A third family treats the caption itself as the target: given the image, the model is trained to assign probability to the actual text tokens. Unlike the two families above, which score a pair, this one scores every token of one modality conditioned on the other, so the loss is cross-entropy $\mathrm{CE}(y, \tilde{h}_{y\leftarrow x}(f(x)))$ of the caption under a decoder $\tilde{h}_{y\leftarrow x}$, rather than a comparison of embeddings. Masked language modeling (MLM) on image–text pairs [12–15, 27, 28] predicts masked caption tokens from the image and the unmasked tokens. More recently, autoregressive language modeling (LM) has largely replaced masked variants [3, 4, 16–18, 29, 30], predicting each token from the image and the preceding tokens. Both variants learn image-conditioned predictions of text tokens, but differ in their probabilistic factorization: MLM reconstructs masked tokens conditioned on visible context, whereas autoregressive LM explicitly factorizes the conditional likelihood of complete text sequences and supports sequential generation.

Text-to-image token-space generation. Symmetrically, the same family predicts image contents conditioned on text, with loss $\mathrm{CE}(x, \tilde{h}_{x\leftarrow y}(g(y)))$. Masked image modeling (MIM) conditioned on text predicts held-out patches using the caption as context, either on discrete visual tokens, as in FLAVAs masked multimodal modeling [15] and BEiT-3 [28], or directly in pixel space, as in MaskVLM [27]. The fully autoregressive analogue generates image tokens left-to-right conditioned on text, as in Emu3 [29]; Transfusion [30] instead applies a diffusion loss to the image's VAE latents, which we count as token-space because they are autoencoder codes rather than features of a semantic encoder.

Image-to-text latent-space prediction. Another family keeps the predictive structure and changes only the target: instead of the text tokens, the model predicts a feature encoder's embedding of the text, with loss $\|g(y) - h_{y\leftarrow x}(f(x))\|^2$ for a predictor $h_{y\leftarrow x}$, so the comparison happens in a continuous embedding space rather than as a likelihood over the vocabulary. VL-JEPA [5] makes this framing explicit: a query-conditioned predictor maps visual embeddings directly to the text encoder's output embedding, optimized in embedding space with an InfoNCE loss, whose uniformity

Table A1: Extended taxonomy of loss families used by representative vision–language models. Rows list individual models and columns indicate whether each model uses contrastive alignment, match/no-match prediction, token-space generation in either direction, or latent-space prediction in either direction.

Model	**1.** Contrastive	**2.** Match	**3.** Token img→txt	**4.** Token txt→img	**5.** Latent img→txt	**6.** Latent txt→img
LXMERT [12]	–	✓	✓	–	–	✓
UNITER [13]	–	✓	✓	–	–	✓
CLIP [1]	✓	–	–	–	–	–
ALBEF [14]	✓	✓	✓	–	–	–
FLAVA [15]	✓	✓	✓	✓	–	–
CoCa [16]	✓	–	✓	–	–	–
Flamingo [3]	–	–	✓	–	–	–
BEiT-3 [28]	–	–	✓	✓	–	–
SigLIP [2]	–	✓	–	–	–	–
BLIP-2 [17]	✓	✓	✓	–	–	–
LLaVA [4]	–	–	✓	–	–	–
MaskVLM [27]	✓	✓	✓	✓	–	–
Emu3 [29]	–	–	✓	✓	–	–
Transfusion [30]	–	–	✓	✓	–	–
VL-JEPA [5]	–	–	–	–	✓	–
SigLIP 2 [18]	–	✓	✓	–	–	–

term provides the anti-collapse regularization. The key distinction from token-space objectives is in the prediction target. Token-space objectives supervise predictions of discrete text tokens, so semantically equivalent captions with different wordings correspond to different target sequences. In contrast, latent-space objectives predict continuous text representations in which such captions may be mapped to nearby embeddings.

Text-to-image latent-space prediction. Symmetrically, the same family predicts continuous image latents from text, with loss $\|f(x) - h_{x \leftarrow y}(g(y))\|^2$, comparing the prediction to a feature encoder's representation of the image (or of an image region) rather than reconstructing its pixels or an autoencoder code of it. An important early instantiation is masked region modeling: a subset of detector-proposed image regions is masked, and the transformer output at each masked region is projected by a fully-connected layer and regressed with an L_2 loss onto the region's frozen detector region-of-interest (RoI) feature, conditioned on the text and the visible regions [12, 13].

No principled ordering of objectives. Even before considering combinations, the field has no principled basis for choosing which loss to use. As Table A1 shows,[6] leading models pick very different objectives: CLIP [1] relies on contrastive alignment, VL-JEPA [5] on latent-space prediction, Flamingo [3] on autoregressive text generation, and SigLIP [2] on a sigmoid matching loss—yet each was reported as state-of-the-art for its target benchmark. Models that combine multiple losses (UNITER [13], FLAVA [15], MaskVLM [27]) likewise justify their specific combination empirically, through ablations and benchmarks, with little agreement across models on which combination to use. What is missing is a principled ordering: an axis along which the families can be compared on equal footing, independent of architectural and dataset confounds.

A.2 Related Work

Theory of single-modality multi-view learning. In self-supervised learning, paired views of the same input are generated within a single modality by augmentation, masking, or corruption; a growing body of work analyzes such objectives through the lens of provable guarantees. HaoChen et al. [31] introduce an augmentation graph and a spectral contrastive loss whose population optimum corresponds to spectral embeddings of that graph, yielding representations with provable linear-probe

[6] We count only image–text losses in each model's own pretraining, not objectives used only to pretrain a frozen component (e.g., Flamingo's vision encoder). For directional losses, targets that are raw content or its tokenizer/autoencoder codes count as token-space, and outputs of any other encoder or detector as latent-space.

guarantees under suitable assumptions. Tian [21] shows that, with linear encoders, contrastive learning performs PCA on a contrastive covariance matrix. Identifiability analyses instead ask which latent factors the learned representations provably recover: Zimmermann et al. [32] show that contrastive learning can invert the data-generating process when views share a common latent cause, and von Kügelgen et al. [33] model data augmentations as changing only a style block while keeping a content block fixed, proving that contrastive learning block-identifies the content variables. Beyond contrastive learning, a few linear analyses touch on reconstruction and latent-space prediction. Balestriero and LeCun [22] show that different denoising tasks align the features learned by reconstruction with perception tasks to different degrees. Littwin et al. [23] find that JEPA-style self-distillation is implicitly biased toward features with high regression coefficients. Van Assel et al. [24] prove that joint-embedding prediction requires a weaker condition on the augmentation than reconstruction when irrelevant features are strong, while the reverse holds when they are weak.

Theory of multimodal learning. Multimodal learning can be viewed as multi-view learning in which the two views come from distinct modalities, but it is more than a change of data source. Unlike views produced from the same input, which share one generative mechanism, two modalities can be asymmetric: they are produced by distinct generative mechanisms, carry their own modality-specific variation, and need not be equally informative about the shared content. Theory in this setting has focused even more heavily on contrastive learning. Nakada et al. [6] show that, with linear encoders, CLIP-type contrastive learning reduces to a singular value decomposition of the cross-covariance between the two modalities. Building on the content–style formulation of von Kügelgen et al. [33], Daunhawer et al. [7] introduce distinct, modality-specific generative mechanisms alongside a shared content block, proving that a symmetric contrastive objective block-identifies the invariant content factors even under modality heterogeneity. Cai et al. [8] further relax the strict content-invariance assumption underlying prior work by formally modeling cross-modal misalignment: they introduce *selection bias*, whereby some semantic variables are absent from text, and *perturbation bias*, whereby present variables are stochastically corrupted, and they prove that minimizing the symmetric objective block-identifies precisely the semantic variables unaffected by either form of bias.

Limitations of prior theory and our departures. Existing theoretical results share three notable limitations. First, prior multimodal theory focuses almost exclusively on contrastive objectives, whereas modern multimodal learning employs four distinct families of objectives—contrastive alignment, match/no-match prediction, token-space generation, and latent-space prediction—and it remains unknown whether these guarantees extend beyond the contrastive case or how objectives should be chosen in practice. Second, prior multimodal theory treats modality-specific variation as separate from the shared content—as private latent blocks or as independent noise. In practice, however, shared semantics can be entangled with modality-specific nuisance (e.g., occlusion, blur, token length, punctuation count), for which we find correlational evidence on COCO and Flickr30k (Section 4.2)—a setting that prior theory does not address. Third, existing identifiability results characterize what is recovered up to equivalence but are silent on whether two objectives that recover the same latent factors use them equally well at test time, leaving open the question of which objective should be preferred for downstream retrieval even after identifiability is established. We address all three gaps.

A.3 Proofs

A.3.1 Proof of Proposition 1

We prove each part of Proposition 1 in turn. All four proofs share a common set of notation, which we collect here to avoid repetition.

Shared notation. Throughout this subsection, $\Sigma_{xx}, \Sigma_{yy} \succ 0$, $m \leq \min(d_x, d_y)$, and the whitening constraints are

$$W_x \Sigma_{xx} W_x^\top = I_m, \qquad W_y \Sigma_{yy} W_y^\top = I_m, \tag{A1}$$

and we write

$$z_x = W_x x, \qquad z_y = W_y y, \qquad \tilde{z}_y = W_y \tilde{y},$$
$$A = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}, \qquad U = W_x\Sigma_{xx}^{1/2}, \qquad V = W_y\Sigma_{yy}^{1/2}, \tag{A2}$$
$$UU^\top = I_m, \qquad VV^\top = I_m, \qquad C = \mathbb{E}[z_x z_y^\top] = W_x\Sigma_{xy}W_y^\top = UAV^\top.$$

Let

$$A = P\,\mathrm{diag}(\rho_1, \rho_2, \ldots)Q^\top, \qquad \rho_1 > \cdots > \rho_m > \rho_{m+1} > \cdots > 0, \tag{A3}$$

be a singular value decomposition, and let P_m, Q_m denote the matrices of the top-m left and right singular vectors.

Contrastive alignment. For contrastive alignment, the local linear surrogate in Table 2 is

$$\mathcal{L}_{\text{align}} = \mathbb{E}\|W_x x - W_y y\|_2^2. \tag{A4}$$

The optimal $W_x^\star$ and $W_y^\star$ are the matrices whose rows form the Σ_{xx}- and Σ_{yy}-orthonormal top-m generalized eigenspaces of

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a_i = \rho_i^2\Sigma_{xx}a_i, \qquad \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b_i = \rho_i^2\Sigma_{yy}b_i. \tag{A5}$$

Proof. Using the shared notation,

$$\begin{aligned}\mathcal{L}_{\text{align}} &= \mathrm{tr}(W_x\Sigma_{xx}W_x^\top) + \mathrm{tr}(W_y\Sigma_{yy}W_y^\top) - 2\,\mathrm{tr}(W_x\Sigma_{xy}W_y^\top)\\ &= m + m - 2\,\mathrm{tr}(C)\\ &= 2m - 2\,\mathrm{tr}(UAV^\top).\end{aligned} \tag{A6}$$

Hence

$$\begin{aligned}\arg\min_{W_x, W_y}\mathcal{L}_{\text{align}} &= \arg\max_{UU^\top = I_m,\, VV^\top = I_m}\mathrm{tr}(UAV^\top)\\ &= (P_m^\top, Q_m^\top) \quad \text{(von Neumann's trace inequality)},\end{aligned} \tag{A7}$$

up to a common rotation of the rows of U and V. Therefore

$$W_x^\star = P_m^\top\Sigma_{xx}^{-1/2}, \qquad W_y^\star = Q_m^\top\Sigma_{yy}^{-1/2}, \tag{A8}$$

and its rows $a_i^\top = p_i^\top\Sigma_{xx}^{-1/2}$ solve (A5) since $AA^\top p_i = \rho_i^2 p_i$; likewise for b_i. □

Match/no-match prediction. For match/no-match prediction, the nonlinear objective uses a sigmoid score head s_ψ. In the local linear regime, the leading cross-modal interaction is bilinear. The most direct linearization replaces s_ψ by a single bilinear score $z_x^\top M z_y$ and the Binary Cross-Entropy objective by the squared label loss with targets 1 and 0:

$$\mathcal{L}_{\text{match}} = \mathbb{E}_+\big[(1 - z_x^\top M z_y)^2\big] + \mathbb{E}_-\big[(z_x^\top M\tilde{z}_y)^2\big]. \tag{A9}$$

After optimizing the bilinear head M, $W_x^\star$ is the matrix whose rows form the Σ_{xx}-orthonormal top-m generalized eigenspace of

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a_i = \rho_i^2\Sigma_{xx}a_i. \tag{A10}$$

Similarly, $W_y^\star$ is obtained from

$$\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b_i = \rho_i^2\Sigma_{yy}b_i. \tag{A11}$$

The optimized bilinear head is

$$\mathrm{vec}(M^\star) = \Big(2I_{m^2} + \mathrm{vec}(C^\star)\,\mathrm{vec}(C^\star)^\top + ((C^\star)^\top \otimes C^\star)K_{m,m}\Big)^{-1}\mathrm{vec}(C^\star), \tag{A12}$$

where $K_{m,m}$ is the commutation matrix and

$$C^\star = W_x^\star\Sigma_{xy}(W_y^\star)^\top. \tag{A13}$$

Proof. Let

$$\begin{aligned} &\phi = \mathrm{vec}(z_x z_y^\top), \qquad \tilde{\phi} = \mathrm{vec}(z_x \tilde{z}_y^\top), \qquad m_M = \mathrm{vec}(M), \\ &\mu = \mathbb{E}_+[\phi] = \mathrm{vec}(C), \qquad Q_+ = \mathbb{E}_+[\phi\phi^\top], \qquad Q_- = \mathbb{E}_-[\tilde{\phi}\tilde{\phi}^\top]. \end{aligned} \tag{A14}$$

Then

$$\begin{aligned} \mathcal{L}_{\mathrm{match}} &= \mathbb{E}_+\left[(1 - m_M^\top \phi)^2\right] + \mathbb{E}_-\left[(m_M^\top \tilde{\phi})^2\right] \\ &= 1 - 2m_M^\top \mu + m_M^\top (Q_+ + Q_-) m_M, \end{aligned} \tag{A15}$$

so

$$m_M^\star = (Q_+ + Q_-)^{-1}\mu. \tag{A16}$$

For negative pairs, and for the vectorized index (i,j) corresponding to the entry $(z_x)_i(\tilde{z}_y)_j$,

$$\begin{aligned} [Q_-]_{(i,j),(k,\ell)} &= \mathbb{E}_-\left[(z_x)_i(\tilde{z}_y)_j(z_x)_k(\tilde{z}_y)_\ell\right] \\ &= \mathbb{E}[(z_x)_i(z_x)_k]\,\mathbb{E}[(\tilde{z}_y)_j(\tilde{z}_y)_\ell] \\ &= \delta_{ik}\delta_{j\ell}. \end{aligned} \tag{A17}$$

$$Q_- = I_{m^2}. \tag{A18}$$

For positive pairs, the joint Gaussian closure gives, by Isserlis' formula,

$$\begin{aligned} [Q_+]_{(i,j),(k,\ell)} &= \mathbb{E}_+\left[(z_x)_i(z_y)_j(z_x)_k(z_y)_\ell\right] \\ &= \mathbb{E}[(z_x)_i(z_x)_k]\,\mathbb{E}[(z_y)_j(z_y)_\ell] + \mathbb{E}[(z_x)_i(z_y)_j]\,\mathbb{E}[(z_x)_k(z_y)_\ell] \\ &\quad + \mathbb{E}[(z_x)_i(z_y)_\ell]\,\mathbb{E}[(z_x)_k(z_y)_j] \\ &= \delta_{ik}\delta_{j\ell} + C_{ij}C_{k\ell} + C_{i\ell}C_{kj}, \end{aligned} \tag{A19}$$

where

$$\begin{aligned} \delta_{ik}\delta_{j\ell} &= [I_{m^2}]_{(i,j),(k,\ell)}, \\ C_{ij}C_{k\ell} &= [\mathrm{vec}(C)\,\mathrm{vec}(C)^\top]_{(i,j),(k,\ell)}, \\ C_{i\ell}C_{kj} &= [(C^\top \otimes C)K_{m,m}]_{(i,j),(k,\ell)}. \end{aligned} \tag{A20}$$

Thus

$$Q_+ = I_{m^2} + \mathrm{vec}(C)\,\mathrm{vec}(C)^\top + (C^\top \otimes C)K_{m,m}. \tag{A21}$$

$$\begin{aligned} \mathrm{vec}(M^\star) &= (Q_+ + Q_-)^{-1}\mu \\ &= \left(2I_{m^2} + \mathrm{vec}(C)\,\mathrm{vec}(C)^\top + (C^\top \otimes C)K_{m,m}\right)^{-1}\mathrm{vec}(C). \end{aligned} \tag{A22}$$

$$\mathcal{L}^\star_{\mathrm{match}}(C) = 1 - \mathrm{vec}(C)^\top \left(2I_{m^2} + \mathrm{vec}(C)\,\mathrm{vec}(C)^\top + (C^\top \otimes C)K_{m,m}\right)^{-1}\mathrm{vec}(C). \tag{A23}$$

Let

$$C = RDS^\top, \qquad D = \mathrm{diag}(s_1, \ldots, s_m). \tag{A24}$$

Then

$$\begin{aligned} \mathrm{vec}(C) &= (S \otimes R)\,\mathrm{vec}(D), \\ (C^\top \otimes C)K_{m,m} &= (SDR^\top \otimes RDS^\top)K_{m,m} \\ &= (S \otimes R)(D \otimes D)(R^\top \otimes S^\top)K_{m,m} \\ &= (S \otimes R)(D \otimes D)K_{m,m}(S^\top \otimes R^\top) \\ 2I_{m^2} + \mathrm{vec}(C)\,\mathrm{vec}(C)^\top &+ (C^\top \otimes C)K_{m,m} \\ &= (S \otimes R)\left[2I_{m^2} + \mathrm{vec}(D)\,\mathrm{vec}(D)^\top + (D \otimes D)K_{m,m}\right](S^\top \otimes R^\top), \\ \mathrm{vec}(C)^\top \left(2I_{m^2} + \mathrm{vec}(C)\,\mathrm{vec}(C)^\top + (C^\top \otimes C)K_{m,m}\right)^{-1}&\mathrm{vec}(C) \\ &= \mathrm{vec}(D)^\top \left(2I_{m^2} + \mathrm{vec}(D)\,\mathrm{vec}(D)^\top + (D \otimes D)K_{m,m}\right)^{-1}\mathrm{vec}(D). \end{aligned} \tag{A25}$$

By the Sherman–Morrison formula,

$$\begin{aligned}\operatorname{vec}(D)^\top \Big(2I_{m^2} + \operatorname{vec}(D)\operatorname{vec}(D)^\top + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D)\\ = \frac{\operatorname{vec}(D)^\top \Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D)}{1+\operatorname{vec}(D)^\top \Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D)}.\end{aligned} \tag{A26}$$

It remains to compute $\operatorname{vec}(D)^\top \Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D)$.

Note

$$\Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)\operatorname{vec}\Big(\operatorname{diag}\Big(\frac{s_1}{2+s_1^2},\dots,\frac{s_m}{2+s_m^2}\Big)\Big) = \operatorname{vec}(D). \tag{A27}$$

Thus

$$\Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D) = \operatorname{vec}\Big(\operatorname{diag}\Big(\frac{s_1}{2+s_1^2},\dots,\frac{s_m}{2+s_m^2}\Big)\Big). \tag{A28}$$

Consequently,

$$\operatorname{vec}(D)^\top \Big(2I_{m^2} + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D) = \sum_{i=1}^m s_i\frac{s_i}{2+s_i^2} = \sum_{i=1}^m \frac{s_i^2}{2+s_i^2},$$

$$\operatorname{vec}(D)^\top \Big(2I_{m^2} + \operatorname{vec}(D)\operatorname{vec}(D)^\top + (D\otimes D)K_{m,m}\Big)^{-1}\operatorname{vec}(D) = \frac{\displaystyle\sum_{i=1}^m \frac{s_i^2}{2+s_i^2}}{1+\displaystyle\sum_{i=1}^m \frac{s_i^2}{2+s_i^2}},$$

$$\begin{aligned}\mathcal{L}^\star_{\text{match}}(C) &= 1 - \operatorname{vec}(C)^\top \Big(2I_{m^2} + \operatorname{vec}(C)\operatorname{vec}(C)^\top + (C^\top\otimes C)K_{m,m}\Big)^{-1}\operatorname{vec}(C)\\ &= 1 - \frac{\sum_{i=1}^m \frac{s_i(C)^2}{2+s_i(C)^2}}{1+\sum_{i=1}^m \frac{s_i(C)^2}{2+s_i(C)^2}}\\ &= \frac{1}{1+\sum_{i=1}^m \frac{s_i(C)^2}{2+s_i(C)^2}}\\ &= \frac{1}{1+\sum_{i=1}^m \frac{s_i(UAV^\top)^2}{2+s_i(UAV^\top)^2}}.\end{aligned} \tag{A29}$$

Since $t^2/(2+t^2)$ is increasing on $t\ge 0$ and $s_i(UAV^\top)\le \rho_i$ by interlacing, the loss is minimized by the compression $UAV^\top$ with maximal leading singular values. Thus, up to rotations of the rows of U and V,

$$(U^\star, V^\star) = (P_m^\top, Q_m^\top), \qquad W_x^\star = P_m^\top \Sigma_{xx}^{-1/2}, \quad W_y^\star = Q_m^\top \Sigma_{yy}^{-1/2}, \tag{A30}$$

or equivalently

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a_i = \rho_i^2\Sigma_{xx}a_i, \qquad \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b_i = \rho_i^2\Sigma_{yy}b_i. \tag{A31}$$

At the optimum, $C = C^\star$, which gives the stated formula for $M^\star$. □

Latent-space prediction. For latent-space prediction, the target is the other modality's latent representation. For the $x\to y$ direction, the local linear surrogate in Table 2 is

$$\mathcal{L}^{\text{lat}}_{x\to y} = \mathbb{E}\|W_y y - B_{y\leftarrow x}W_x x\|_2^2. \tag{A32}$$

After optimizing $B_{y\leftarrow x}$, $W_x^\star$ is the matrix whose rows form the Σ_{xx}-orthonormal top-m generalized eigenspace of

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a_i = \rho_i^2\Sigma_{xx}a_i. \tag{A33}$$

The corresponding $W_y^\star$ is obtained from

$$\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b_i = \rho_i^2\Sigma_{yy}b_i. \tag{A34}$$

The optimized latent predictor is

$$B_{y\leftarrow x}^\star = W_y^\star\Sigma_{yx}(W_x^\star)^\top. \tag{A35}$$

Proof. For fixed W_x and W_y,

$$\begin{aligned} B_{y\leftarrow x}^\star &= \mathbb{E}[z_y z_x^\top]\mathbb{E}[z_x z_x^\top]^{-1} \\ &= W_y\Sigma_{yx}W_x^\top\left(W_x\Sigma_{xx}W_x^\top\right)^{-1} \\ &= W_y\Sigma_{yx}W_x^\top. \end{aligned} \tag{A36}$$

Substitution gives

$$\begin{aligned} (\mathcal{L}_{x\to y}^{\mathrm{lat}})^\star &= \mathbb{E}\|z_y\|_2^2 - \mathrm{tr}\left(\mathbb{E}[z_y z_x^\top]\mathbb{E}[z_x z_x^\top]^{-1}\mathbb{E}[z_x z_y^\top]\right) \\ &= m - \mathrm{tr}\left(W_y\Sigma_{yx}W_x^\top W_x\Sigma_{xy}W_y^\top\right) \\ &= m - \|W_x\Sigma_{xy}W_y^\top\|_F^2 \\ &= m - \|UAV^\top\|_F^2. \end{aligned} \tag{A37}$$

Hence

$$\begin{aligned} \underset{W_x,W_y}{\arg\min}(\mathcal{L}_{x\to y}^{\mathrm{lat}})^\star &= \underset{UU^\top=I_m,\, VV^\top=I_m}{\arg\max} \|UAV^\top\|_F^2 \\ &= \underset{UU^\top=I_m,\, VV^\top=I_m}{\arg\max} \sum_{i=1}^{m} s_i(UAV^\top)^2 \\ &= (P_m^\top, Q_m^\top) \quad (\text{interlacing, } s_i(UAV^\top) \le \rho_i), \end{aligned} \tag{A38}$$

up to rotations of the rows of U and V. Therefore,

$$W_x^\star = P_m^\top\Sigma_{xx}^{-1/2}, \qquad W_y^\star = Q_m^\top\Sigma_{yy}^{-1/2}, \tag{A39}$$

or equivalently

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}a_i = \rho_i^2\Sigma_{xx}a_i, \qquad \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}b_i = \rho_i^2\Sigma_{yy}b_i. \tag{A40}$$

Finally,

$$B_{y\leftarrow x}^\star = W_y^\star\Sigma_{yx}(W_x^\star)^\top. \tag{A41}$$

The $y \to x$ direction follows by exchanging x and y, with $B_{x\leftarrow y}^\star = W_x^\star\Sigma_{xy}(W_y^\star)^\top$. □

Token-space generation. For token-space generation in the $x \to y$ direction, the local linear surrogate is

$$\mathcal{L}_{x\to y}^{\mathrm{tok}} = \mathbb{E}\|y - V_{y\leftarrow x}W_x x\|_2^2. \tag{A42}$$

After optimizing $V_{y\leftarrow x}$, $W_x^\star$ is the matrix whose rows form the Σ_{xx}-orthonormal top-m generalized eigenspace of

$$\Sigma_{xy}\Sigma_{yx}a_i = \lambda_i\Sigma_{xx}a_i. \tag{A43}$$

The optimized token-space decoder is

$$V_{y\leftarrow x}^\star = \Sigma_{yx}(W_x^\star)^\top. \tag{A44}$$

Proof. For fixed W_x,

$$\begin{aligned} V_{y\leftarrow x}^\star &= \mathbb{E}[yx^\top W_x^\top]\left(W_x\Sigma_{xx}W_x^\top\right)^{-1} \\ &= \Sigma_{yx}W_x^\top. \end{aligned} \tag{A45}$$

Substitution gives

$$\begin{aligned} (\mathcal{L}_{x\to y}^{\mathrm{tok}})^\star &= \mathbb{E}\|y\|_2^2 - \mathrm{tr}\left(\mathbb{E}[yx^\top W_x^\top]\left(W_x\Sigma_{xx}W_x^\top\right)^{-1}\mathbb{E}[W_x xy^\top]\right) \\ &= \mathrm{tr}(\Sigma_{yy}) - \mathrm{tr}\left(\Sigma_{yx}W_x^\top W_x\Sigma_{xy}\right) \\ &= \mathrm{tr}(\Sigma_{yy}) - \mathrm{tr}\left(W_x\Sigma_{xy}\Sigma_{yx}W_x^\top\right). \end{aligned} \tag{A46}$$

Hence

$$\arg\min_{W_x}(\mathcal{L}^{\mathrm{tok}}_{x\to y})^\star = \arg\max_{W_x\Sigma_{xx}W_x^\top=I_m} \operatorname{tr}\left(W_x\Sigma_{xy}\Sigma_{yx}W_x^\top\right). \tag{A47}$$

By the Rayleigh–Ritz theorem for generalized eigenvalue problems, the rows of $W_x^\star$ are, up to a rotation, the Σ_{xx}-orthonormal top-m generalized eigenvectors of

$$\Sigma_{xy}\Sigma_{yx}a_i = \lambda_i\Sigma_{xx}a_i. \tag{A48}$$

Therefore

$$V^\star_{y\leftarrow x} = \Sigma_{yx}(W_x^\star)^\top. \tag{A49}$$

The loss does not involve W_y, so only W_x is constrained. Exchanging x and y gives the $y \to x$ direction, with $\Sigma_{yx}\Sigma_{xy}b_i = \lambda_i\Sigma_{yy}b_i$ and $V^\star_{x\leftarrow y} = \Sigma_{xy}(W_y^\star)^\top$. □

A.3.2 Proof of Theorem 1

Shared notation. Throughout this subsection, (x, y) is centered jointly Gaussian, the encoders f and g have finite second moments, and whitening means

$$\mathbb{E}[f(x)] = \mathbb{E}[g(y)] = 0, \qquad \mathbb{E}[f(x)f(x)^\top] = \mathbb{E}[g(y)g(y)^\top] = I_m. \tag{A50}$$

With P and Q as in Appendix A.3.1, the canonical coordinates

$$\xi = P^\top\Sigma_{xx}^{-1/2}x, \qquad \zeta = Q^\top\Sigma_{yy}^{-1/2}y \tag{A51}$$

are standard Gaussian with $\mathbb{E}[\xi\zeta^\top] = P^\top AQ = \operatorname{diag}(\rho_1, \rho_2, \ldots)$. We append independent standard Gaussian coordinates to the shorter of ξ and ζ and set $\rho_i = 0$ for $i > \min(d_x, d_y)$, so that both have length $d = \max(d_x, d_y)$ and the pairs (ξ_i, ζ_i) are independent across i.

Hermite expansion. Let He_j be the Hermite polynomials, $e^{tu-t^2/2} = \sum_{j\ge0}\mathrm{He}_j(u)t^j/j!$, and let $h_j = \mathrm{He}_j/\sqrt{j!}$. For standard Gaussian (u, v) with correlation ρ,

$$\begin{aligned}\sum_{j,k\ge0}\mathbb{E}[\mathrm{He}_j(u)\mathrm{He}_k(v)]\frac{t^j s^k}{j!\,k!} &= \mathbb{E}\big[e^{tu-t^2/2}e^{sv-s^2/2}\big]\\ &= e^{\rho ts}\\ &= \sum_{j\ge0}\rho^j\frac{t^j s^j}{j!},\end{aligned} \tag{A52}$$

so $\mathbb{E}[h_j(u)h_k(v)] = \delta_{jk}\rho^j$. For a multi-index $\mathbf{j} \in \mathbb{N}_0^d$, write $H_{\mathbf{j}}(\xi) = \prod_{i=1}^d h_{j_i}(\xi_i)$, $|\mathbf{j}| = \sum_i j_i$, and $\rho^{\mathbf{j}} = \prod_i \rho_i^{j_i}$ with $0^0 = 1$. Independence across i gives

$$\mathbb{E}[H_{\mathbf{j}}(\xi)H_{\mathbf{k}}(\zeta)] = \prod_{i=1}^d \mathbb{E}[h_{j_i}(\xi_i)h_{k_i}(\zeta_i)] = \mathbf{1}\{\mathbf{j}=\mathbf{k}\}\rho^{\mathbf{j}}, \tag{A53}$$

and, with $\rho = 1$, $\mathbb{E}[H_{\mathbf{j}}(\xi)H_{\mathbf{k}}(\xi)] = \mathbf{1}\{\mathbf{j}=\mathbf{k}\}$. Since the $H_{\mathbf{j}}$ form an orthonormal basis of square-integrable functions, whitened encoders expand in mean square as

$$f(x) = \sum_{|\mathbf{j}|\ge1}\hat{f}_{\mathbf{j}}H_{\mathbf{j}}(\xi), \qquad g(y) = \sum_{|\mathbf{j}|\ge1}\hat{g}_{\mathbf{j}}H_{\mathbf{j}}(\zeta), \tag{A54}$$

with $\hat{f}_{\mathbf{j}} = \mathbb{E}[f(x)H_{\mathbf{j}}(\xi)]$ and $\hat{g}_{\mathbf{j}} = \mathbb{E}[g(y)H_{\mathbf{j}}(\zeta)]$; there is no constant term since f and g are centered. Hence

$$\sum_{|\mathbf{j}|\ge1}\hat{f}_{\mathbf{j}}\hat{f}_{\mathbf{j}}^\top = \sum_{|\mathbf{j}|\ge1}\hat{g}_{\mathbf{j}}\hat{g}_{\mathbf{j}}^\top = I_m, \qquad \mathbb{E}[f(x)g(y)^\top] = \sum_{|\mathbf{j}|\ge1}\rho^{\mathbf{j}}\hat{f}_{\mathbf{j}}\hat{g}_{\mathbf{j}}^\top. \tag{A55}$$

The first-order terms are linear, since $H_{e_i}(\xi) = \xi_i$:

$$\sum_{i=1}^{d_x}\hat{f}_{e_i}\xi_i = F_1P^\top\Sigma_{xx}^{-1/2}x, \qquad F_1 = [\hat{f}_{e_1}, \ldots, \hat{f}_{e_{d_x}}], \tag{A56}$$

and a linear encoder $f(x) = W_x x$ has $F_1 = UP$ and no other terms; likewise for g, with $G_1 = [\hat{g}_{e_1}, \ldots, \hat{g}_{e_{d_y}}] = VQ$. Finally, the weights satisfy

$$\rho^{e_i} = \rho_i, \qquad \rho^{\mathbf{j}} \le \rho_1^{|\mathbf{j}|} \le \rho_1^2 \quad (|\mathbf{j}| \ge 2). \tag{A57}$$

Reduced objectives. Write $C = \mathbb{E}[f(x)g(y)^\top]$ as in Appendix A.3.1, so that

$$C = \sum_{|\mathbf{j}|\geq 1} \rho^{\mathbf{j}} \hat{f}_{\mathbf{j}} \hat{g}_{\mathbf{j}}^\top. \tag{A58}$$

Optimizing the linear heads as in Appendix A.3.1, with z_x and z_y replaced by $f(x)$ and $g(y)$, uses only whitening and gives

$$\begin{aligned} \mathcal{L}_{\mathrm{align}} &= 2m - 2\,\mathrm{tr}(C), \\ (\mathcal{L}^{\mathrm{lat}}_{x\to y})^\star &= m - \|C\|_F^2, \\ (\mathcal{L}^{\mathrm{tok}}_{x\to y})^\star &= \mathrm{tr}(\Sigma_{yy}) - \left\|\mathbb{E}[f(x)y^\top]\right\|_F^2. \end{aligned} \tag{A59}$$

Since $y = \Sigma_{yy}^{1/2} Q\zeta$ has only first-order terms, $\mathbb{E}[f(x)\zeta_i] = \rho_i \hat{f}_{e_i}$, and

$$\begin{aligned} \mathbb{E}[f(x)y^\top] &= \mathbb{E}[f(x)\zeta^\top] Q^\top \Sigma_{yy}^{1/2} \\ &= F_1 P^\top A Q Q^\top \Sigma_{yy}^{1/2} \\ &= F_1 P^\top \Sigma_{xx}^{-1/2} \Sigma_{xy}. \end{aligned} \tag{A60}$$

Thus the contrastive and latent objectives depend on the encoders only through C, and the token objective only through the first-order coefficients F_1. The $y \to x$ directions follow by exchanging x and y.

Contrastive alignment and latent-space prediction. We bound $\mathrm{tr}(C)$ and $\|C\|_F^2$ over all whitened encoders and show that the linear optima of Appendix A.3.1 attain the bounds. Let $\omega_{\mathbf{j}}^f = \|\hat{f}_{\mathbf{j}}\|_2^2$ and $\omega_{\mathbf{j}}^g = \|\hat{g}_{\mathbf{j}}\|_2^2$. Since $\sum_{\mathbf{j}} \hat{f}_{\mathbf{j}} \hat{f}_{\mathbf{j}}^\top = I_m$, the sequences $((\hat{f}_{\mathbf{j}})_r)_{\mathbf{j}}$, $r = 1, \ldots, m$, are orthonormal, so Bessel's inequality and the trace give

$$0 \leq \omega_{\mathbf{j}}^f \leq 1, \qquad \sum_{|\mathbf{j}|\geq 1} \omega_{\mathbf{j}}^f = m, \tag{A61}$$

and likewise for $\omega_{\mathbf{j}}^g$. Under $\rho_m > \rho_1^2$, every weight other than $\rho_1, \ldots, \rho_m$ is at most ρ_m, so for any $c_{\mathbf{j}} \in [0,1]$ with $\sum_{\mathbf{j}} c_{\mathbf{j}} = m$,

$$\begin{aligned} \sum_{|\mathbf{j}|\geq 1} \rho^{\mathbf{j}} c_{\mathbf{j}} - \sum_{i=1}^m \rho_i &= \sum_{i=1}^m \rho_i (c_{e_i} - 1) + \sum_{\mathbf{j} \notin \{e_1, \ldots, e_m\}} \rho^{\mathbf{j}} c_{\mathbf{j}} \\ &\leq \rho_m \Big(\sum_{i=1}^m (c_{e_i} - 1) + \sum_{\mathbf{j} \notin \{e_1, \ldots, e_m\}} c_{\mathbf{j}} \Big) \\ &= 0, \end{aligned} \tag{A62}$$

and the same holds with $\rho^{\mathbf{j}}$ and ρ_i replaced by their squares. Hence, by Cauchy–Schwarz and $ab \leq (a^2 + b^2)/2$,

$$\begin{aligned} \mathrm{tr}(C) &= \sum_{|\mathbf{j}|\geq 1} \rho^{\mathbf{j}} \hat{f}_{\mathbf{j}}^\top \hat{g}_{\mathbf{j}} \\ &\leq \sum_{|\mathbf{j}|\geq 1} \rho^{\mathbf{j}} \, \frac{\omega_{\mathbf{j}}^f + \omega_{\mathbf{j}}^g}{2} \\ &\leq \sum_{i=1}^m \rho_i, \end{aligned} \tag{A63}$$

and, by Bessel's inequality for the orthonormal sequences $((\hat{g}_{\mathbf{j}})_s)_{\mathbf{j}}$,

$$\begin{aligned}\|C\|_F^2 &= \sum_{r,s=1}^{m} \Big(\sum_{|\mathbf{j}|\geq 1} \rho^{\mathbf{j}} (\hat{f}_{\mathbf{j}})_r (\hat{g}_{\mathbf{j}})_s \Big)^2 \\ &\leq \sum_{r=1}^{m} \sum_{|\mathbf{j}|\geq 1} (\rho^{\mathbf{j}})^2 (\hat{f}_{\mathbf{j}})_r^2 \\ &= \sum_{|\mathbf{j}|\geq 1} (\rho^{\mathbf{j}})^2 \omega_{\mathbf{j}}^f \\ &\leq \sum_{i=1}^{m} \rho_i^2 .\end{aligned} \tag{A64}$$

The linear encoders $W_x^\star = P_m^\top \Sigma_{xx}^{-1/2}$ and $W_y^\star = Q_m^\top \Sigma_{yy}^{-1/2}$ of Appendix A.3.1 give $C = P_m^\top A Q_m = \operatorname{diag}(\rho_1, \dots, \rho_m)$, so they attain both bounds. By the reduced objectives, they therefore minimize $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{x\to y}^{\text{lat}}$ over all whitened encoders.

Token-space generation. By the reduced objectives, it suffices to show that every whitened encoder is matched by a linear one in $\|F_1 M\|_F^2$, where $M = P^\top \Sigma_{xx}^{-1/2} \Sigma_{xy}$. Since

$$F_1 F_1^\top = \sum_{i=1}^{d_x} \hat{f}_{e_i} \hat{f}_{e_i}^\top \preceq \sum_{|\mathbf{j}|\geq 1} \hat{f}_{\mathbf{j}} \hat{f}_{\mathbf{j}}^\top = I_m, \tag{A65}$$

F_1 has a singular value decomposition $F_1 = R \operatorname{diag}(\sigma_1, \dots, \sigma_m) S^\top$ with R orthogonal, $S^\top S = I_m$, and $\sigma_i \in [0, 1]$. With s_i the columns of S,

$$\begin{aligned}\|F_1 M\|_F^2 &= \sum_{i=1}^{m} \sigma_i^2 \|M^\top s_i\|_2^2 \\ &\leq \sum_{i=1}^{m} \|M^\top s_i\|_2^2 \\ &= \|S^\top M\|_F^2 .\end{aligned} \tag{A66}$$

The right-hand side is the value of the linear encoder $W_x = S^\top P^\top \Sigma_{xx}^{-1/2}$, which is whitened since $W_x \Sigma_{xx} W_x^\top = S^\top S = I_m$ and has $F_1 = UP = S^\top$. Hence the linear optimum of Appendix A.3.1 minimizes $\mathcal{L}_{x\to y}^{\text{tok}}$ over all whitened encoders, with no condition on ρ.

Match/no-match prediction. The optimal head depends on fourth moments, which Gaussianity of (x, y) does not determine for nonlinear encoders. We therefore restrict f and g to the block-structured class below, bound the loss over this class, and show that the linear optimum of Appendix A.3.1 attains the bound. The block-structured class consists of

$$f(x) = \Phi\, h_{\boldsymbol{\ell}}(\xi), \qquad g(y) = \Psi\, h_{\boldsymbol{\ell}'}(\zeta), \qquad \Phi\Phi^\top = \Psi\Psi^\top = I_m, \tag{A67}$$

where $h_{\boldsymbol{\ell}}(\xi) = (h_{\ell_1}(\xi_1), \dots, h_{\ell_{d_x}}(\xi_{d_x}))^\top$ and $h_{\boldsymbol{\ell}'}(\zeta) = (h_{\ell'_1}(\zeta_1), \dots, h_{\ell'_{d_y}}(\zeta_{d_y}))^\top$ for some degrees $\ell_i, \ell'_i \geq 1$, so that each canonical coordinate enters through a single Hermite degree. Here $\Phi \in \mathbb{R}^{m\times d_x}$ and $\Psi \in \mathbb{R}^{m\times d_y}$, and in sums over $i \leq d$ we set $\Phi_{ri} = 0$ for $i > d_x$ and $\Psi_{ri} = 0$ for $i > d_y$. These encoders are whitened, since $\mathbb{E}[h_{\boldsymbol{\ell}}(\xi)] = 0$ and $\mathbb{E}[h_{\boldsymbol{\ell}}(\xi) h_{\boldsymbol{\ell}}(\xi)^\top] = I_{d_x}$, and whitened linear encoders are exactly the case $\ell_i = \ell'_i = 1$, with $\Phi = UP$ and $\Psi = VQ$, since $W_x x = UP\xi$ and $\Phi\Phi^\top = UU^\top$.

We first show that, compared with jointly Gaussian (z_x, z_y) with the same C, the loss only gains a fourth-cumulant term. Write $a = h_{\boldsymbol{\ell}}(\xi)$, $b = h_{\boldsymbol{\ell}'}(\zeta)$ and $N = \Phi^\top M \Psi$, so that $f(x)^\top M g(y) = a^\top N b$. The pairs (a_i, b_i) are centered with unit variances and independent across i, and by the Hermite identity $\mathbb{E}[a_i b_j] = \delta_{ij} \bar{\rho}_i$ with $\bar{\rho}_i = \mathbf{1}\{\ell_i = \ell'_i\} \rho_i^{\ell_i} \leq \rho_i$. Hence

$$C = \mathbb{E}[f(x) g(y)^\top] = \Phi \bar{D} \Psi^\top, \tag{A68}$$

where $\bar{D} \in \mathbb{R}^{d_x \times d_y}$ is diagonal with entries $\bar{\rho}_i$. For centered scalar random variables X, Y let

$$\mathcal{K}(X,Y) = \mathbb{E}[X^2Y^2] - \mathbb{E}[X^2]\,\mathbb{E}[Y^2] - 2\big(\mathbb{E}[XY]\big)^2, \tag{A69}$$

and $\mathcal{K}_i = \mathcal{K}(a_i, b_i)$. Since the pairs are centered and independent across i, a fourth moment vanishes unless its indices pair up, and

$$\mathbb{E}[a_i b_j a_k b_l] = \delta_{ik}\delta_{jl} + \delta_{ij}\delta_{kl}\bar{\rho}_i\bar{\rho}_k + \delta_{il}\delta_{jk}\bar{\rho}_i\bar{\rho}_j + \mathbf{1}\{i=j=k=l\}\mathcal{K}_i. \tag{A70}$$

Hence

$$\begin{aligned}
\mathbb{E}_+\big[(f(x)^\top M g(y))^2\big] &= \sum_{i,j,k,l} N_{ij}N_{kl}\,\mathbb{E}[a_i b_j a_k b_l] \\
&= \|N\|_F^2 + \big(\mathrm{tr}(N\bar{D}^\top)\big)^2 + \mathrm{tr}(N\bar{D}^\top N\bar{D}^\top) + \sum_i N_{ii}^2\mathcal{K}_i \\
&= \|M\|_F^2 + \big(\mathrm{tr}(MC^\top)\big)^2 + \mathrm{tr}(MC^\top MC^\top) + \sum_i N_{ii}^2\mathcal{K}_i \\
&= \mathrm{vec}(M)^\top Q_+\,\mathrm{vec}(M) + \sum_i N_{ii}^2\mathcal{K}_i,
\end{aligned} \tag{A71}$$

with Q_+ as in Appendix A.3.1. The mean $\mathbb{E}_+[f(x)^\top M g(y)] = \mathrm{vec}(M)^\top\,\mathrm{vec}(C)$ and, by whitening and independence, the negative term $\mathbb{E}_-[(f(x)^\top M g(\tilde{y}))^2] = \|M\|_F^2$ are as in Appendix A.3.1. Hence, if every $\mathcal{K}_i \geq 0$, the loss is at least its Gaussian counterpart for every M, and minimizing over M as in Appendix A.3.1 gives

$$\mathcal{L}^\star_{\mathrm{match}} \geq \frac{1}{1 + \sum_{r=1}^m \frac{s_r(C)^2}{2+s_r(C)^2}}. \tag{A72}$$

We next show $\mathcal{K}_i \geq 0$. Comparing the coefficients of $s^p t^p$ in $e^{sw-s^2/2}e^{tw-t^2/2} = e^{st}e^{(s+t)w-(s+t)^2/2}$ gives $\mathrm{He}_p^2 = \sum_{n=0}^p (p-n)!\binom{p}{n}^2\mathrm{He}_{2n}$. Together with $\mathbb{E}[\mathrm{He}_a(\xi_i)\mathrm{He}_b(\zeta_i)] = \delta_{ab}\,a!\,\rho_i^a$, this gives for $p, q \geq 1$

$$\begin{aligned}
\mathbb{E}[h_p(\xi_i)^2 h_q(\zeta_i)^2] &= \frac{1}{p!\,q!}\sum_{n=0}^p\sum_{n'=0}^q (p-n)!\,(q-n')!\binom{p}{n}^2\binom{q}{n'}^2\mathbb{E}[\mathrm{He}_{2n}(\xi_i)\mathrm{He}_{2n'}(\zeta_i)] \\
&= \sum_{n=0}^{p\wedge q}\frac{(p-n)!\,(q-n)!\,(2n)!}{p!\,q!}\binom{p}{n}^2\binom{q}{n}^2\rho_i^{2n} \\
&= \sum_{n=0}^{p\wedge q}\binom{p}{n}\binom{q}{n}\binom{2n}{n}\rho_i^{2n},
\end{aligned} \tag{A73}$$

so, since $\mathbb{E}[h_p(\xi_i)h_q(\zeta_i)] = \mathbf{1}\{p=q\}\rho_i^p$,

$$\mathcal{K}\big(h_p(\xi_i), h_q(\zeta_i)\big) = \sum_{n=1}^{p\wedge q}\binom{p}{n}\binom{q}{n}\binom{2n}{n}\rho_i^{2n} - 2\cdot\mathbf{1}\{p=q\}\rho_i^{2p} \geq 0, \tag{A74}$$

because for $p=q$ the term $n=p$ equals $\binom{2p}{p}\rho_i^{2p} \geq 2\rho_i^{2p}$.

Finally, $\Phi\Phi^\top = \Psi\Psi^\top = I_m$ gives $s_r(C) = s_r(\Phi\bar{D}\Psi^\top) \leq s_r(\bar{D}) \leq \rho_r$ by interlacing, since $\bar{\rho}_i \leq \rho_i$ and ρ is decreasing. Hence

$$\mathcal{L}^\star_{\mathrm{match}} \geq \frac{1}{1 + \sum_{r=1}^m \frac{s_r(C)^2}{2+s_r(C)^2}} \geq \frac{1}{1 + \sum_{r=1}^m \frac{\rho_r^2}{2+\rho_r^2}}. \tag{A75}$$

The linear minimizer $U = P_m^\top$, $V = Q_m^\top$ of Appendix A.3.1 lies in the class, with $\Phi = UP$, $\Psi = VQ$ and $\ell_i = \ell_i' = 1$. It has $\mathcal{K}_i = 0$ by Isserlis' formula and $s_r(C) = \rho_r$, so it attains the right-hand side. It therefore minimizes $\mathcal{L}_{\mathrm{match}}$ over the block-structured class, with no condition on ρ.

A.3.3 Proof of Theorem 2

Setup For $a \in \{x, y\}$, write $K_c = \mathrm{Cov}(s_c)$, $K_e = \mathrm{Cov}(s_e)$, and $H_a = \mathrm{Cov}(n_a)$ in Equation (2), and split the noise covariance into

$$\Sigma_{\eta,a}^{(c)} = \mathrm{diag}(\sigma_{a,i}^2)_{i \le k}, \qquad \Sigma_{\eta,a}^{(e,n)} = \mathrm{diag}(\sigma_{a,i}^2)_{i > k}, \qquad \Sigma_{\eta,a}^{(e)} = \mathrm{diag}(\sigma_{a,i}^2)_{k < i \le k+p}. \tag{A76}$$

Then

$$\Sigma_{aa} = Q_a \begin{bmatrix} K_c + \Sigma_{\eta,a}^{(c)} & 0 \\ 0 & \Omega_a + \Sigma_{\eta,a}^{(e,n)} \end{bmatrix} Q_a^\top, \qquad \Sigma_{xy} = Q_x \begin{bmatrix} K_c & 0 \\ 0 & \Omega_{xy} \end{bmatrix} Q_y^\top, \tag{A77}$$

where, since $(s_e,\, C_a s_e + n_a) = P_a s_e + N_a n_a$,

$$\Omega_a = P_a K_e P_a^\top + N_a H_a N_a^\top, \qquad \Omega_{xy} = P_x K_e P_y^\top, \qquad P_a = \begin{bmatrix} I_p \\ C_a \end{bmatrix}, \qquad N_a = \begin{bmatrix} 0 \\ I_r \end{bmatrix}. \tag{A78}$$

Contrastive, Match, Latent Prediction By Appendix A.3.1, the rows of $U^\star = W_x^\star \Sigma_{xx}^{1/2}$ span the top-m left singular subspace of

$$A = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} = Q_x \begin{bmatrix} D_c & 0 \\ 0 & B \end{bmatrix} Q_y^\top, \tag{A79}$$

where, with $R_{a,c} = K_c + \Sigma_{\eta,a}^{(c)}$ and $S_a = \Omega_a + \Sigma_{\eta,a}^{(e,n)}$,

$$D_c = R_{x,c}^{-1/2} K_c R_{y,c}^{-1/2} = \mathrm{diag}\left(\frac{\kappa_i}{\sqrt{(\kappa_i + \sigma_{x,i}^2)(\kappa_i + \sigma_{y,i}^2)}} \right)_{i=1}^{k}, \tag{A80}$$

$$B = S_x^{-1/2} P_x K_e P_y^\top S_y^{-1/2}.$$

Hence, for some orthogonal O,

$$\bar{W}_x := W_x^\star Q_x = O \mathcal{U}^\top \begin{bmatrix} R_{x,c}^{-1/2} & 0 \\ 0 & S_x^{-1/2} \end{bmatrix} = \begin{bmatrix} \bar{W}_{x,c} & \bar{W}_{x,e} & \bar{W}_{x,n} \end{bmatrix}, \tag{A81}$$

where the columns of $\mathcal{U} = \begin{bmatrix} \mathcal{U}_c \\ \mathcal{U}_b \end{bmatrix}$ are left singular vectors of $\mathrm{diag}(D_c, B)$ for its m largest singular values, and the blocks of $\bar{W}_x$ have k, p, and r columns. Clean-coordinate identifiability means $\bar{W}_{x,n} = 0$, and the conclusion of Theorem 2 means $\bar{W}_{x,e} = \bar{W}_{x,n} = 0$. Since $[\bar{W}_{x,e}\ \bar{W}_{x,n}] = O \mathcal{U}_b^\top S_x^{-1/2}$, it suffices to show that a column $w = S_x^{-1/2} u$ of $S_x^{-1/2} \mathcal{U}_b$ with zero bottom block, $w = [v^\top\ 0]^\top$, vanishes.

If the singular value ρ of u is positive, then $u = B\tilde{v}/\rho$ for the matching block $\tilde{v}$ of the right singular vector, so $S_x w = P_x q$ with $q = K_e P_y^\top S_y^{-1/2} \tilde{v}/\rho$, and

$$\begin{bmatrix} q \\ C_x q \end{bmatrix} = P_x q = S_x \begin{bmatrix} v \\ 0 \end{bmatrix} = \begin{bmatrix} (K_e + \Sigma_{\eta,x}^{(e)}) v \\ C_x K_e v \end{bmatrix}. \tag{A82}$$

Hence

$$C_x \Sigma_{\eta,x}^{(e)} v = C_x (q - K_e v) = 0, \tag{A83}$$

and $v = 0$ since $\mathrm{rank}(C_x) = p$ and $\Sigma_{\eta,x}^{(e)} \succeq \nu_x I_p \succ 0$. If $\rho = 0$, then $B^\top u = S_y^{-1/2} P_y K_e P_x^\top w = 0$, and since $S_y^{-1/2} P_y K_e$ has full column rank, $v = P_x^\top w = 0$. In both cases $w = 0$, so $\bar{W}_{x,e} = \bar{W}_{x,n} = 0$. As $\bar{W}_x$ has rank m and is supported on the first k columns, $m \le k$. The y-side follows by symmetry.

Token generation For token generation $x \to y$, by Appendix A.3.1 the rows of $U^\star$ span the top-m eigenspace of

$$\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2} = Q_x \begin{bmatrix} R_{x,c}^{-1/2} K_c^2 R_{x,c}^{-1/2} & 0 \\ 0 & B_{x,\mathrm{tok}} \end{bmatrix} Q_x^\top, \tag{A84}$$

where $B_{x,\mathrm{tok}} = S_x^{-1/2} P_x K_e P_y^\top P_y K_e P_x^\top S_x^{-1/2}$. The argument above applies with eigenvectors in place of singular vectors and $B_{x,\mathrm{tok}}$ in place of B: if $B_{x,\mathrm{tok}} u = \lambda u$ with $\lambda > 0$, then $S_x w = P_x q$ with $q = K_e P_y^\top P_y K_e P_x^\top w/\lambda$; if $\lambda = 0$, then $\|P_y K_e P_x^\top w\|_2^2 = u^\top B_{x,\mathrm{tok}} u = 0$, so $P_x^\top w = 0$ since $P_y K_e$ has full column rank. Token generation $y \to x$ follows by exchanging x and y.

A.3.4 Proof of Theorem 3

Contrastive, Match, Latent Prediction We first bound the non-clean block B of A from the proof of Theorem 2, used in (i). With $S_a = \Omega_a + \Sigma_{\eta,a}^{(e,n)}$ as there,

$$B = U_x U_y^\top, \qquad U_a = S_a^{-1/2} P_a K_e^{1/2}. \tag{A85}$$

Since

$$\begin{aligned} \Omega_x = P_x K_e P_x^\top + N_x H_x N_x^\top \succeq P_x K_e P_x^\top, \\ \Sigma_{\eta,x}^{(e,n)} \succeq \nu_x I, \end{aligned} \tag{A86}$$

Thus, by order reversal for positive definite inverses,

$$S_x^{-1} \preceq (P_x K_e P_x^\top + \nu_x I)^{-1}. \tag{A87}$$

$$\begin{aligned} \|U_x\|_2^2 &= \lambda_{\max}\Big(K_e^{1/2} P_x^\top S_x^{-1} P_x K_e^{1/2}\Big) \\ &\le \lambda_{\max}\Big(K_e^{1/2} P_x^\top (P_x K_e P_x^\top + \nu_x I)^{-1} P_x K_e^{1/2}\Big) \\ &= \frac{\|P_x K_e^{1/2}\|_2^2}{\|P_x K_e^{1/2}\|_2^2 + \nu_x}. \end{aligned} \tag{A88}$$

Moreover, since $P_x^\top P_x = I + C_x^\top C_x$,

$$\|P_x K_e^{1/2}\|_2^2 \le \|P_x\|_2^2 \|K_e\|_2 = (1 + \|C_x\|_2^2)\tau_1 \le (1 + \varepsilon_x^2)\tau_1. \tag{A89}$$

Since $t \mapsto t/(t + \nu_x)$ is increasing on $t \ge 0$,

$$\|U_x\|_2^2 \le \frac{(1 + \varepsilon_x^2)\tau_1}{(1 + \varepsilon_x^2)\tau_1 + \nu_x} = \gamma_x. \tag{A90}$$

The same argument gives

$$\|U_y\|_2^2 \le \frac{(1 + \varepsilon_y^2)\tau_1}{(1 + \varepsilon_y^2)\tau_1 + \nu_y} = \gamma_y. \tag{A91}$$

Hence

$$\|B\|_2^2 \le \|U_x\|_2^2 \|U_y\|_2^2 \le \gamma_x \gamma_y. \tag{A92}$$

Recall that the clean block of A is

$$D_c = \operatorname{diag}\left(\frac{\kappa_i}{\sqrt{(\kappa_i + \sigma_{x,i}^2)(\kappa_i + \sigma_{y,i}^2)}}\right)_{i=1}^{k} = \operatorname{diag}\left(\frac{\beta_i^x \beta_i^y}{\kappa_i}\right)_{i=1}^{k}. \tag{A93}$$

If $\gamma_x \gamma_y < \big((\beta^x \beta^y / \kappa)_{[m]}\big)^2$, then the first m singular directions of A lie in the clean block. This proves (i).

Token generation For token generation $x \to y$, the encoder $W_x^\star$ is determined by the top-m eigenvectors of

$$A_{\mathrm{tok},x} = Q_x \begin{bmatrix} D_{x,c} & 0 \\ 0 & B_{x,\mathrm{tok}} \end{bmatrix} Q_x^\top, \tag{A94}$$

where

$$D_{x,c} = R_{x,c}^{-1/2} K_c^2 R_{x,c}^{-1/2} = \operatorname{diag}\big((\beta_i^x)^2\big)_{i=1}^{k} \tag{A95}$$

and

$$B_{x,\mathrm{tok}} = S_x^{-1/2} \Omega_{xy} \Omega_{xy}^\top S_x^{-1/2} = S_x^{-1/2} P_x K_e (P_y^\top P_y) K_e P_x^\top S_x^{-1/2}. \tag{A96}$$

Using

$$U_x = S_x^{-1/2} P_x K_e^{1/2}, \tag{A97}$$

$$B_{x,\mathrm{tok}} = U_x \left(K_e^{1/2} P_y^\top P_y K_e^{1/2} \right) U_x^\top. \tag{A98}$$

Therefore

$$\begin{aligned} \|B_{x,\mathrm{tok}}\|_2 &\le \|U_x\|_2^2 \left\| K_e^{1/2} P_y^\top P_y K_e^{1/2} \right\|_2 \\ &= \|U_x\|_2^2 \|P_y K_e^{1/2}\|_2^2 \\ &\le \gamma_x (1+\varepsilon_y^2)\tau_1 \\ &= \frac{\nu_y \gamma_x \gamma_y}{1-\gamma_y}, \end{aligned} \tag{A99}$$

where the last step uses $\gamma_y/(1-\gamma_y) = (1+\varepsilon_y^2)\tau_1/\nu_y$.

If $\frac{\nu_y \gamma_x \gamma_y}{1-\gamma_y} < ((\beta^x)^2)_{[m]}$, the first m eigenvectors of $A_{\mathrm{tok},x}$ lie in the clean block. This proves (ii).

Token-space generation $y \to x$ just needs to swap x and y.

Worst-case necessity It remains to show that each inequality is attained by some model. Keep κ_i and $\sigma_{a,i}^2$ for $i \le k$, τ_1, ε_a, and ν_a, take any $\tau_2, \dots, \tau_p \le \tau_1$, and set

$$C_a = \varepsilon_a E_{11}, \qquad H_a = 0, \qquad \Sigma_{\eta,a}^{(e,n)} = \nu_a I_{p+r}, \tag{A100}$$

where $E_{11} \in \mathbb{R}^{r \times p}$ has a single unit entry at $(1,1)$. Then $S_a = P_a K_e P_a^\top + \nu_a I$, and with

$$T_a = K_e^{1/2} P_a^\top P_a K_e^{1/2} = \mathrm{diag}\big((1+\varepsilon_a^2)\tau_1, \tau_2, \dots, \tau_p\big), \tag{A101}$$

the push-through identity gives

$$U_a^\top U_a = K_e^{1/2} P_a^\top (P_a K_e P_a^\top + \nu_a I)^{-1} P_a K_e^{1/2} = T_a (T_a + \nu_a I)^{-1}, \tag{A102}$$

a diagonal matrix whose largest entry, γ_a, is in position 1. Hence

$$\begin{aligned} \|B\|_2^2 &= \lambda_{\max}\big(U_x^\top U_x \, U_y^\top U_y\big) = \gamma_x \gamma_y, \\ \|B_{x,\mathrm{tok}}\|_2 &= \lambda_{\max}\big(T_y \, U_x^\top U_x\big) = \gamma_x (1+\varepsilon_y^2)\tau_1 = \frac{\nu_y \gamma_x \gamma_y}{1-\gamma_y}, \end{aligned} \tag{A103}$$

so the bounds above are attained. If the inequality in (i) fails, then $\|B\|_2 \ge (\beta^x \beta^y/\kappa)_{[m]}$, so a top singular pair of B can be taken among the m leading singular pairs of $\mathrm{diag}(D_c, B)$. The corresponding optimal encoder has $\mathcal{U}_b \neq 0$ in the proof of Theorem 2, so $\bar{W}_x$ does not vanish on the non-clean coordinates, and likewise for $\bar{W}_y$. The same argument with $B_{x,\mathrm{tok}}$ in place of B gives (ii), and exchanging x and y gives (iii).

A.3.5 Proof of Theorem 4

Proof. Under isotropic noise, $\nu_a = \sigma_a^2$, and both $\beta_i^x \beta_i^y/\kappa_i$ and $(\beta_i^a)^2$ increase with κ_i, so their m-th largest values are attained at $i = m$. With $c_a = \kappa_m/(\kappa_m + \sigma_a^2) \in (0,1)$,

$$\begin{aligned} \big((\beta^x \beta^y/\kappa)_{[m]}\big)^2 &= \frac{\kappa_m^2}{(\kappa_m + \sigma_x^2)(\kappa_m + \sigma_y^2)} = c_x c_y, \\ \frac{((\beta^x)^2)_{[m]}}{\nu_y} &= \frac{\kappa_m^2}{(\kappa_m + \sigma_x^2)\,\sigma_y^2} = \frac{c_x c_y}{1-c_y}. \end{aligned} \tag{A104}$$

Hence

$$(\gamma_x, \gamma_y) \in \mathcal{F}_{\mathrm{tok},\, x \to y} \iff \gamma_x \gamma_y < c_x c_y \, \frac{1-\gamma_y}{1-c_y}, \tag{A105}$$

and likewise $(\gamma_x, \gamma_y) \in \mathcal{F}_{\mathrm{tok},\, y \to x}$ if and only if $\gamma_x \gamma_y < c_x c_y (1-\gamma_x)/(1-c_x)$. Now let $(\gamma_x, \gamma_y) \in \mathcal{F}_{\mathrm{joint}}$, that is, $\gamma_x \gamma_y < c_x c_y$. Then $\gamma_y < c_y$ or $\gamma_x < c_x$. In the first case,

$$\gamma_x \gamma_y < c_x c_y < c_x c_y \, \frac{1-\gamma_y}{1-c_y}, \tag{A106}$$

so $(\gamma_x, \gamma_y) \in \mathcal{F}_{\mathrm{tok},\, x \to y}$. In the second case, $(\gamma_x, \gamma_y) \in \mathcal{F}_{\mathrm{tok},\, y \to x}$ in the same way. □

A.3.6 Proof of Lemma 1

Proof. Define $\rho^{\ell}_{x,y} := \mathbb{P}(\Delta^{\ell}_{x\text{-queried } y} > 0 \mid x, y)$ and $p^{\ell}_{x\text{-queried } y} := \mathbb{E}[\rho^{\ell}_{x,y}] = \mathbb{P}(\Delta^{\ell}_{x\text{-queried } y} > 0)$. Since the negatives $\tilde{y}_1, \ldots, \tilde{y}_N$ are conditionally i.i.d. given (x, y),

$$\mathrm{R@1}^{\ell}_{x\text{-queried } y,N} = \mathbb{E}\big[(\rho^{\ell}_{x,y})^N\big]\,. \tag{A107}$$

Using Jensen's inequality (convexity of $t \mapsto t^N$) on the left and $t^N \le t$ on the right,

$$\big(p^{\ell}_{x\text{-queried } y}\big)^N \;\le\; \mathrm{R@1}^{\ell}_{x\text{-queried } y,N} \;\le\; p^{\ell}_{x\text{-queried } y}. \tag{A108}$$

By assumption $\Delta^{\ell}_{x\text{-queried } y} = \sum_{i=1}^{m} Z^{\ell}_i$ with independent summands satisfying the Lindeberg condition, so the Lindeberg CLT and $\mathrm{SNR}^{\ell}_{x\text{-queried } y} := \mathbb{E}[\Delta]/\sqrt{\mathrm{Var}(\Delta)} \to$ finite limit give

$$p^{\ell}_{x\text{-queried } y} = \mathbb{P}\left(\frac{\Delta - \mathbb{E}[\Delta]}{\sqrt{\mathrm{Var}(\Delta)}} > -\mathrm{SNR}^{\ell}_{x\text{-queried } y}\right) = \Phi\big(\mathrm{SNR}^{\ell}_{x\text{-queried } y}\big) + o(1). \tag{A109}$$

$$\Phi\big(\mathrm{SNR}^{\ell}_{x\text{-queried } y}\big)^N + o(1) \;\le\; \mathrm{R@1}^{\ell}_{x\text{-queried } y,N} \;\le\; \Phi\big(\mathrm{SNR}^{\ell}_{x\text{-queried } y}\big) + o(1). \tag{A110}$$

Assume $\mathrm{Var}(\rho^{\ell}_{x,y}) \to 0$. Since $t \mapsto t^N$ is N-Lipschitz on $[0, 1]$,

$$\Big|\mathbb{E}\big[(\rho^{\ell}_{x,y})^N\big] - \big(\mathbb{E}[\rho^{\ell}_{x,y}]\big)^N\Big| \;\le\; N\,\mathbb{E}\big|\rho^{\ell}_{x,y} - \mathbb{E}[\rho^{\ell}_{x,y}]\big| \;\le\; N\sqrt{\mathrm{Var}(\rho^{\ell}_{x,y})} \;=\; o(1). \tag{A111}$$

$$\mathrm{R@1}^{\ell}_{x\text{-queried } y,N} = \big(p^{\ell}_{x\text{-queried } y}\big)^N + o(1) = \Phi\big(\mathrm{SNR}^{\ell}_{x\text{-queried } y}\big)^N + o(1). \tag{A112}$$

□

A.3.7 Proof of Theorem 5 and Theorem 6

Proof of Theorem 5 and Theorem 6. Throughout, (x, y) is jointly Gaussian, so distinct coordinates are independent and the fourth moments below follow from Isserlis' formula. Under clean-coordinate identifiability, objective ℓ retains a set S_ℓ of m clean coordinates: the joint objectives the m largest ρ_i (call it S_{J}), $\mathrm{tok}_{y\to x}$ the m largest $(\beta^y_i)^2$, and $\mathrm{tok}_{x\to y}$ the m largest $(\beta^x_i)^2$. Write $x_i = (Q_x^\top x)_i$ and $y_i = (Q_y^\top y)_i$; each $\sum_i$ below runs over the S_ℓ of the objective at hand. For each retained coordinate,

$$\begin{aligned} x_i &= s_i + \eta_{x,i}, & y_i &= s_i + \eta_{y,i}, \\ \mathrm{Var}(s_i) &= \kappa_i, & \mathrm{Var}(\eta_{x,i}) &= \sigma^2_{x,i}, & \mathrm{Var}(\eta_{y,i}) &= \sigma^2_{y,i}, \\ a_i &:= \kappa_i + \sigma^2_{x,i}, & b_i &:= \kappa_i + \sigma^2_{y,i}. \end{aligned} \tag{A113}$$

After whitening,

$$\begin{aligned} \rho_i &= \frac{\mathrm{Cov}(x_i, y_i)}{\sqrt{\mathrm{Var}(x_i)\,\mathrm{Var}(y_i)}} = \frac{\kappa_i}{\sqrt{a_i b_i}}, \\ \xi_i &:= \frac{x_i}{\sqrt{a_i}}, \quad \zeta_i := \frac{y_i}{\sqrt{b_i}}, \qquad \mathrm{Var}(\xi_i) = \mathrm{Var}(\zeta_i) = 1, \mathrm{Cov}(\xi_i, \zeta_i) = \rho_i. \end{aligned} \tag{A114}$$

Use the clean token eigenvalue notation

$$\lambda^{(x)}_i := \frac{\kappa^2_i}{a_i} = b_i \rho^2_i = (\beta^x_i)^2, \quad \lambda^{(y)}_i := \frac{\kappa^2_i}{b_i} = a_i \rho^2_i = (\beta^y_i)^2. \tag{A115}$$

Margins use the convention

$$\begin{aligned} \Delta_{x\text{-queried } y} &= q(x, \tilde{y}) - q(x, y), \quad \Delta_{y\text{-queried } x} = q(\tilde{x}, y) - q(x, y), \\ \mathrm{SNR} &= \frac{\mathbb{E}[\Delta]}{\sqrt{\mathrm{Var}(\Delta)}}. \end{aligned} \tag{A116}$$

The six scores of Section 3.3 are the per-pair terms of the linearized surrogates at their optima. In the original coordinates and in the coordinates above, they are

$$\begin{aligned}
q_{\text{contrastive}}(x,y) &= \|W_x^\star x - W_y^\star y\|_2^2 = \sum_i (\xi_i - \zeta_i)^2,\\
q_{\text{match}}(x,y) &= -(W_x^\star x)^\top M^\star W_y^\star y,\\
q_{\text{lat}_{y\to x}}(x,y) &= \|W_x^\star x - B^\star_{x\leftarrow y} W_y^\star y\|_2^2 = \sum_i (\xi_i - \rho_i\zeta_i)^2,\\
q_{\text{lat}_{x\to y}}(x,y) &= \|W_y^\star y - B^\star_{y\leftarrow x} W_x^\star x\|_2^2 = \sum_i (\zeta_i - \rho_i\xi_i)^2,\\
q_{\text{tok}_{y\to x}}(x,y) &= \|x - V^\star_{x\leftarrow y} W_y^\star y\|_2^2 = \sum_i \Big(x_i - \frac{\kappa_i}{b_i} y_i\Big)^2 + R_x(x),\\
q_{\text{tok}_{x\to y}}(x,y) &= \|y - V^\star_{y\leftarrow x} W_x^\star x\|_2^2 = \sum_i \Big(y_i - \frac{\kappa_i}{a_i} x_i\Big)^2 + R_y(y),
\end{aligned} \tag{A117}$$

where $R_x(x) = \sum_{j\notin S_{\text{tok}_{y\to x}}} x_j^2$ and $R_y(y) = \sum_{j\notin S_{\text{tok}_{x\to y}}} y_j^2$ collect the coordinates outside the retained set, and q_{match} is written in canonical coordinates below.

Token space generation $\text{tok}_{y\to x}$**.** The closed-form token objective and decoder are

$$\mathcal{L}^{\text{tok}}_{y\to x} = \mathbb{E}\|x - V_{x\leftarrow y} W_y y\|_2^2, \quad V^\star_{x\leftarrow y} = \Sigma_{xy}(W_y^\star)^\top. \tag{A118}$$

In canonical coordinates, up to a rotation of $W_y^\star y$,

$$W_y^\star y = (\zeta_i)_{i\in S_{\text{tok}_{y\to x}}}, \qquad (Q_x^\top V^\star_{x\leftarrow y} W_y^\star y)_i = \frac{\kappa_i}{\sqrt{b_i}}\zeta_i = \frac{\kappa_i}{b_i} y_i \tag{A119}$$

for retained i and 0 otherwise, which gives $q_{\text{tok}_{y\to x}}$ in (A117). For x-queried y, the residual cancels:

$$\begin{aligned}
\Delta^{\text{tok}_{y\to x}}_{x\text{-queried } y} &= \sum_i \left[\left(x_i - \frac{\kappa_i}{b_i}\tilde{y}_i\right)^2 - \left(x_i - \frac{\kappa_i}{b_i} y_i\right)^2\right],\\
U_i &:= x_i - \frac{\kappa_i}{b_i}\tilde{y}_i, \quad V_i := x_i - \frac{\kappa_i}{b_i} y_i.
\end{aligned} \tag{A120}$$

The one-coordinate moments are

$$\begin{aligned}
\text{Var}(U_i) &= a_i + \frac{\kappa_i^2}{b_i} = a_i + \lambda_i^{(y)},\\
\text{Var}(V_i) &= a_i + \frac{\kappa_i^2}{b_i} - 2\frac{\kappa_i}{b_i}\,\text{Cov}(x_i, y_i) = a_i - \lambda_i^{(y)},\\
\text{Cov}(U_i, V_i) &= a_i - \frac{\kappa_i}{b_i}\,\text{Cov}(x_i, y_i) = a_i - \lambda_i^{(y)}.
\end{aligned} \tag{A121}$$

Thus

$$\begin{aligned}
\mathbb{E}[U_i^2 - V_i^2] &= 2\lambda_i^{(y)},\\
\text{Var}(U_i^2 - V_i^2) &= 2(a_i + \lambda_i^{(y)})^2 + 2(a_i - \lambda_i^{(y)})^2 - 4(a_i - \lambda_i^{(y)})^2\\
&= 8 a_i \lambda_i^{(y)}.
\end{aligned} \tag{A122}$$

Independence across retained coordinates gives

$$\text{SNR}^{\text{tok}_{y\to x}}_{x\text{-queried } y} = \frac{\sum_i \lambda_i^{(y)}}{\sqrt{2\sum_i a_i \lambda_i^{(y)}}}. \tag{A123}$$

For y-queried x, the residual $R_x(\tilde{x}) - R_x(x)$ has zero mean and is independent of the retained part, so it only adds variance and the retained part upper-bounds the full SNR:

$$\begin{aligned}
\Delta^{\text{tok}_{y\to x}}_{y\text{-queried } x} &= \sum_i \left[\left(\tilde{x}_i - \frac{\kappa_i}{b_i} y_i\right)^2 - \left(x_i - \frac{\kappa_i}{b_i} y_i\right)^2\right] + R_x(\tilde{x}) - R_x(x),\\
\widetilde{U}_i &:= \tilde{x}_i - \frac{\kappa_i}{b_i} y_i, \qquad V_i := x_i - \frac{\kappa_i}{b_i} y_i.
\end{aligned} \tag{A124}$$

The covariance cancels:

$$\begin{aligned}\mathrm{Var}(\widetilde{U}_i) = a_i + \lambda_i^{(y)}, \qquad \mathrm{Var}(V_i) = a_i - \lambda_i^{(y)},\\ \mathrm{Cov}(\widetilde{U}_i, V_i) = -\frac{\kappa_i}{b_i}\,\mathrm{Cov}(y_i, x_i) + \frac{\kappa_i^2}{b_i^2}\,\mathrm{Var}(y_i) = -\lambda_i^{(y)} + \lambda_i^{(y)} = 0.\end{aligned} \tag{A125}$$

Hence

$$\begin{aligned}\mathbb{E}[\widetilde{U}_i^2 - V_i^2] &= 2\lambda_i^{(y)},\\ \mathrm{Var}(\widetilde{U}_i^2 - V_i^2) &= 2(a_i + \lambda_i^{(y)})^2 + 2(a_i - \lambda_i^{(y)})^2 = 4\left(a_i^2 + (\lambda_i^{(y)})^2\right),\\ \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{y\text{-queried } x} &\le \frac{\sum_i \lambda_i^{(y)}}{\sqrt{\sum_i \left(a_i^2 + (\lambda_i^{(y)})^2\right)}}.\end{aligned} \tag{A126}$$

Latent space prediction $\mathrm{lat}_{y\to x}$**.** The latent objective and predictor are

$$\mathcal{L}^{\mathrm{lat}}_{y\to x} = \mathbb{E}\|W_x x - B_{x\leftarrow y} W_y y\|_2^2, \quad B^\star_{x\leftarrow y} = W_x^\star \Sigma_{xy} (W_y^\star)^\top. \tag{A127}$$

Clean-coordinate identifiability gives, up to rotations of $W_x^\star x$ and $W_y^\star y$ that $B^\star_{x\leftarrow y}$ absorbs,

$$W_x^\star x = (\xi_i)_{i\in S_{\mathrm{J}}}, \qquad W_y^\star y = (\zeta_i)_{i\in S_{\mathrm{J}}}, \qquad B^\star_{x\leftarrow y} = \mathrm{diag}(\rho_i)_{i\in S_{\mathrm{J}}}, \tag{A128}$$

which gives $q_{\mathrm{lat}_{y\to x}}$ in (A117). For x-queried y,

$$\begin{aligned}\Delta^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y} &= \sum_i \left[(\xi_i - \rho_i\tilde{\zeta}_i)^2 - (\xi_i - \rho_i\zeta_i)^2\right],\\ U_i &:= \xi_i - \rho_i\tilde{\zeta}_i, \quad V_i := \xi_i - \rho_i\zeta_i.\end{aligned} \tag{A129}$$

The moments are

$$\begin{aligned}\mathrm{Var}(U_i) = 1 + \rho_i^2, \quad \mathrm{Var}(V_i) = 1 - \rho_i^2, \quad \mathrm{Cov}(U_i, V_i) = 1 - \rho_i^2,\\ \mathbb{E}[U_i^2 - V_i^2] = 2\rho_i^2,\\ \mathrm{Var}(U_i^2 - V_i^2) = 2(1+\rho_i^2)^2 + 2(1-\rho_i^2)^2 - 4(1-\rho_i^2)^2 = 8\rho_i^2.\end{aligned} \tag{A130}$$

Therefore

$$\mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y} = \sqrt{\frac{1}{2}\sum_i \rho_i^2}. \tag{A131}$$

For y-queried x,

$$\begin{aligned}\Delta^{\mathrm{lat}_{y\to x}}_{y\text{-queried } x} &= \sum_i \left[(\tilde{\xi}_i - \rho_i\zeta_i)^2 - (\xi_i - \rho_i\zeta_i)^2\right],\\ \widetilde{U}_i &:= \tilde{\xi}_i - \rho_i\zeta_i, \quad V_i := \xi_i - \rho_i\zeta_i.\end{aligned} \tag{A132}$$

Here

$$\begin{aligned}\mathrm{Var}(\widetilde{U}_i) = 1 + \rho_i^2, \quad \mathrm{Var}(V_i) = 1 - \rho_i^2,\\ \mathrm{Cov}(\widetilde{U}_i, V_i) = -\rho_i\,\mathrm{Cov}(\zeta_i, \xi_i) + \rho_i^2\,\mathrm{Var}(\zeta_i) = -\rho_i^2 + \rho_i^2 = 0,\\ \mathbb{E}[\widetilde{U}_i^2 - V_i^2] = 2\rho_i^2,\\ \mathrm{Var}(\widetilde{U}_i^2 - V_i^2) = 2(1+\rho_i^2)^2 + 2(1-\rho_i^2)^2 = 4(1+\rho_i^4).\end{aligned} \tag{A133}$$

Thus

$$\mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{y\text{-queried } x} = \frac{\sum_i \rho_i^2}{\sqrt{\sum_i (1+\rho_i^4)}}. \tag{A134}$$

Directional mismatch penalty. Using $\lambda_i^{(y)} = a_i\rho_i^2$,

$$\begin{aligned}a_i^2 + (\lambda_i^{(y)})^2 = a_i^2(1+\rho_i^4) \ge 2a_i^2\rho_i^2 = 2a_i\lambda_i^{(y)},\\ \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{y\text{-queried } x} \le \frac{\sum_i \lambda_i^{(y)}}{\sqrt{\sum_i \left(a_i^2 + (\lambda_i^{(y)})^2\right)}} \le \frac{\sum_i \lambda_i^{(y)}}{\sqrt{2\sum_i a_i\lambda_i^{(y)}}} = \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried } y}.\end{aligned} \tag{A135}$$

For the latent score,

$$\sum_i (1+\rho_i^4) \geq 2\sum_i \rho_i^2, \quad \mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{y\text{-queried } x} \leq \mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y}. \tag{A136}$$

The swapped inequalities follow by exchanging x and y.

Latent score dominance. First,

$$\left(\mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried } y}\right)^2 = \frac{1}{2}\sum_i \rho_i^2. \tag{A137}$$

For $\mathrm{tok}_{y\to x}$,

$$\begin{aligned}
\left(\mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried } y}\right)^2 &= \frac{\left(\sum_i \lambda_i^{(y)}\right)^2}{2\sum_i a_i \lambda_i^{(y)}} = \frac{\left(\sum_i a_i\rho_i^2\right)^2}{2\sum_i a_i^2\rho_i^2}, \\
\left(\sum_i a_i\rho_i^2\right)^2 &\leq \left(\sum_i a_i^2\rho_i^2\right)\left(\sum_i \rho_i^2\right), \\
\left(\mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried } y}\right)^2 &\leq \frac{1}{2}\sum_i \rho_i^2.
\end{aligned} \tag{A138}$$

For $\mathrm{tok}_{x\to y}$, exchanging x and y in the y-queried computation for $\mathrm{tok}_{y\to x}$ gives

$$\mathrm{SNR}^{\mathrm{tok}_{x\to y}}_{x\text{-queried } y} \leq \frac{\sum_i \lambda_i^{(x)}}{\sqrt{\sum_i \left(b_i^2 + (\lambda_i^{(x)})^2\right)}}. \tag{A139}$$

Since $\lambda_i^{(x)} = b_i\rho_i^2$,

$$\begin{aligned}
\left(\mathrm{SNR}^{\mathrm{tok}_{x\to y}}_{x\text{-queried } y}\right)^2 &\leq \frac{\left(\sum_i b_i\rho_i^2\right)^2}{\sum_i b_i^2(1+\rho_i^4)} \\
&\leq \frac{\left(\sum_i b_i\rho_i^2\right)^2}{2\sum_i b_i^2\rho_i^2} \leq \frac{1}{2}\sum_i \rho_i^2.
\end{aligned} \tag{A140}$$

For $\mathrm{lat}_{x\to y}$, exchanging x and y in the y-queried computation for $\mathrm{lat}_{y\to x}$ gives

$$\begin{aligned}
\mathrm{SNR}^{\mathrm{lat}_{x\to y}}_{x\text{-queried } y} &= \frac{\sum_i \rho_i^2}{\sqrt{\sum_i (1+\rho_i^4)}}, \\
\left(\mathrm{SNR}^{\mathrm{lat}_{x\to y}}_{x\text{-queried } y}\right)^2 &= \frac{\left(\sum_i \rho_i^2\right)^2}{\sum_i (1+\rho_i^4)} \leq \frac{1}{2}\sum_i \rho_i^2.
\end{aligned} \tag{A141}$$

For contrastive retrieval,

$$\begin{aligned}
\Delta^{\mathrm{contrastive}}_{x\text{-queried } y} &= \sum_i \left[(\xi_i - \tilde{\zeta}_i)^2 - (\xi_i - \zeta_i)^2\right], \\
U_i &:= \xi_i - \tilde{\zeta}_i, \quad V_i := \xi_i - \zeta_i.
\end{aligned} \tag{A142}$$

The moments are

$$\begin{aligned}
\mathrm{Var}(U_i) = 2, \quad \mathrm{Var}(V_i) &= 2(1-\rho_i), \quad \mathrm{Cov}(U_i, V_i) = 1-\rho_i, \\
\mathbb{E}[U_i^2 - V_i^2] &= 2\rho_i, \\
\mathrm{Var}(U_i^2 - V_i^2) &= 4(3 - 2\rho_i + \rho_i^2), \\
\mathrm{SNR}^{\mathrm{contrastive}}_{x\text{-queried } y} &= \frac{\sum_i \rho_i}{\sqrt{\sum_i (3 - 2\rho_i + \rho_i^2)}}.
\end{aligned} \tag{A143}$$

Then

$$\left(\sum_i \rho_i\right)^2 \le m\sum_i \rho_i^2, \quad 3-2\rho_i+\rho_i^2 = 2+(1-\rho_i)^2 \ge 2,$$
$$\left(\mathrm{SNR}^{\mathrm{contrastive}}_{x\text{-queried }y}\right)^2 \le \frac{1}{2}\sum_i \rho_i^2. \tag{A144}$$

For match/no-match, in canonical coordinates $(W_x^\star x)^\top M^\star W_y^\star y = \sum_{i,j} M_{ij}\xi_i\zeta_j$, where M is $M^\star$ written in these coordinates, so

$$\Delta^{\mathrm{match}}_{x\text{-queried }y} = \sum_i M_{ii}\,\xi_i(\zeta_i-\tilde\zeta_i) + \sum_{i\ne j} M_{ij}\,\xi_i(\zeta_j-\tilde\zeta_j). \tag{A145}$$

By Isserlis' formula,

$$\begin{aligned}\mathbb{E}[\xi_i(\zeta_i-\tilde\zeta_i)] &= \rho_i,\\ \mathrm{Var}(\xi_i\zeta_i) = 1+\rho_i^2, \quad \mathrm{Var}(\xi_i\tilde\zeta_i) &= 1, \quad \mathrm{Cov}(\xi_i\zeta_i, \xi_i\tilde\zeta_i) = 0,\\ \mathrm{Var}\big(\xi_i(\zeta_i-\tilde\zeta_i)\big) &= 2+\rho_i^2.\end{aligned} \tag{A146}$$

The diagonal summands are independent across i, and the off-diagonal part has zero mean and is uncorrelated with each of them, so it only adds variance. If $\sum_i M_{ii}\rho_i \le 0$, the SNR is nonpositive; otherwise, by the Cauchy–Schwarz inequality,

$$\mathrm{SNR}^{\mathrm{match}}_{x\text{-queried }y} \le \frac{\sum_i M_{ii}\rho_i}{\sqrt{\sum_i M_{ii}^2(2+\rho_i^2)}} \le \sqrt{\sum_i \frac{\rho_i^2}{2+\rho_i^2}} \le \sqrt{\frac{1}{2}\sum_i \rho_i^2}. \tag{A147}$$

Each bound above on the squared SNR is $\frac{1}{2}\sum_{i\in S_\ell}\rho_i^2$, and $\sum_{i\in S_\ell}\rho_i^2 \le \sum_{i\in S_{\mathrm{J}}}\rho_i^2$ because S_{J} holds the m largest ρ_i. Combining the comparisons,

$$\mathrm{SNR}^{\mathrm{lat}_{y\to x}}_{x\text{-queried }y} \ge \max\Big\{\mathrm{SNR}^{\mathrm{contrastive}}_{x\text{-queried }y}, \mathrm{SNR}^{\mathrm{match}}_{x\text{-queried }y}, \mathrm{SNR}^{\mathrm{tok}_{y\to x}}_{x\text{-queried }y}, \mathrm{SNR}^{\mathrm{tok}_{x\to y}}_{x\text{-queried }y}, \mathrm{SNR}^{\mathrm{lat}_{x\to y}}_{x\text{-queried }y}\Big\}. \tag{A148}$$

The symmetric statement follows by exchanging x and y. □

A.3.8 Multihead match/no-match

The single bilinear head of Table 2 shares one scalar label per pair across all m coordinate products. A multihead variant instead keeps m rank-one heads $t_r(x,y) = \alpha_r(u_r^\top z_x)(v_r^\top z_y)$, with $\{u_r\}$ and $\{v_r\}$ orthonormal, and applies the squared label loss to each head. We show that Proposition 1 and Theorem 1 extend to this variant, so Theorems 2 and 3, which use only the row spaces of the optimal encoders, extend as well, and that Theorem 6 covers its score. We use the notation of Appendices A.3.1 and A.3.2.

Linear encoders. The multihead loss is

$$\mathcal{L}^{\mathrm{MH}}_{\mathrm{match}} = \sum_{r=1}^m \Big(\mathbb{E}_+\big[(1-t_r(x,y))^2\big] + \mathbb{E}_-\big[t_r(x,\tilde y)^2\big]\Big). \tag{A149}$$

After optimizing the heads, the rows of $W_x^\star$ and $W_y^\star$ span the top-m generalized eigenspaces of (A5). The optimized gains are

$$\alpha_r^\star = \frac{c_r}{2(1+c_r^2)}, \tag{A150}$$

where

$$c_r = \mathbb{E}_+\big[(u_r^\top z_x)(v_r^\top z_y)\big]. \tag{A151}$$

Proof. For any orthogonal R,

$$(Ru_r)^\top(RW_x)x = u_r^\top R^\top R W_x x = u_r^\top W_x x, \qquad (RW_x)\Sigma_{xx}(RW_x)^\top = RR^\top = I_m, \tag{A152}$$

so replacing (W_x, u_r) by (RW_x, Ru_r) leaves every t_r, hence the loss, and the whitening constraint unchanged; likewise for (W_y, v_r). Hence only the row spaces of $W_x^\star$ and $W_y^\star$ can be identified. In particular, with the orthogonal matrices

$$H_u = [u_1, \dots, u_m], \qquad H_v = [v_1, \dots, v_m], \tag{A153}$$

replacing (W_x, u_r) by $(H_u^\top W_x, e_r)$ and (W_y, v_r) by $(H_v^\top W_y, e_r)$ keeps the loss and the row spaces. We may therefore take $u_r = v_r = e_r$, so that

$$c_r = \mathbb{E}_+\big[(z_x)_r (z_y)_r\big] = C_{rr} = (UAV^\top)_{rr}. \tag{A154}$$

For negative pairs, the independence of $\tilde{y}$ and whitening give

$$\begin{aligned} \mathbb{E}_-\big[(z_x)_r^2 (\tilde{z}_y)_r^2\big] &= \mathbb{E}[(z_x)_r^2]\,\mathbb{E}[(\tilde{z}_y)_r^2] \\ &= 1. \end{aligned} \tag{A155}$$

For positive pairs, the joint Gaussian closure gives, by Isserlis' formula,

$$\begin{aligned} \mathbb{E}_+\big[(z_x)_r^2 (z_y)_r^2\big] &= \mathbb{E}[(z_x)_r^2]\,\mathbb{E}[(z_y)_r^2] + 2\big(\mathbb{E}_+[(z_x)_r (z_y)_r]\big)^2 \\ &= 1 + 2c_r^2. \end{aligned} \tag{A156}$$

Substitution gives

$$\begin{aligned} \mathcal{L}_{\text{match}}^{\text{MH}} &= \sum_{r=1}^m \Big(\mathbb{E}_+\big[(1 - \alpha_r (z_x)_r (z_y)_r)^2\big] + \alpha_r^2\, \mathbb{E}_-\big[(z_x)_r^2 (\tilde{z}_y)_r^2\big]\Big) \\ &= \sum_{r=1}^m \big(1 - 2\alpha_r c_r + \alpha_r^2 (1 + 2c_r^2) + \alpha_r^2\big) \\ &= \sum_{r=1}^m \big(1 - 2\alpha_r c_r + 2\alpha_r^2 (1 + c_r^2)\big) \\ &= \sum_{r=1}^m \Big(1 - J(c_r) + 2(1 + c_r^2)\Big(\alpha_r - \frac{c_r}{2(1 + c_r^2)}\Big)^2\Big), \end{aligned} \tag{A157}$$

where

$$J(s) = \frac{s^2}{2(1 + s^2)}, \tag{A158}$$

so

$$\alpha_r^\star = \frac{c_r}{2(1 + c_r^2)}, \qquad (\mathcal{L}_{\text{match}}^{\text{MH}})^\star = m - \sum_{r=1}^m J(c_r). \tag{A159}$$

Hence

$$\arg\min_{W_x, W_y} (\mathcal{L}_{\text{match}}^{\text{MH}})^\star = \arg\max_{UU^\top = I_m,\, VV^\top = I_m} \sum_{r=1}^m J\big((UAV^\top)_{rr}\big). \tag{A160}$$

We next find one minimizer by bounding this objective for every (U, V) and attaining the bound. For each k,

$$\begin{aligned} \sum_{r=1}^k |c|_{[r]} &\le \sum_{i=1}^k s_i(C) \quad \text{(Ky Fan's inequality)} \\ &\le \sum_{i=1}^k \rho_i \quad \text{(interlacing, since } C = UAV^\top\text{)}, \end{aligned} \tag{A161}$$

where $|c|_{[1]} \ge \cdots \ge |c|_{[m]}$ are $|c_1|, \dots, |c_m|$ in decreasing order. Hence

$$(|c_1|, \dots, |c_m|) \prec_w (\rho_1, \dots, \rho_m). \tag{A162}$$

By Marshall–Olkin–Arnold, there exists $\bar{w} \in \mathbb{R}^m$ such that $\bar{w}_r \ge |c_r|$ for all r and $\bar{w} \prec (\rho_1, \dots, \rho_m)$. By the real Schur–Horn theorem, there exists an orthogonal O such that

$$\operatorname{diag}(O^\top D_m O) = \bar{w}, \qquad D_m = \operatorname{diag}(\rho_1, \dots, \rho_m). \tag{A163}$$

Since J is even and increasing on $s \geq 0$,

$$\begin{aligned}\sum_{r=1}^{m} J(c_r) &= \sum_{r=1}^{m} J(|c_r|) \\ &\leq \sum_{r=1}^{m} J(\bar{w}_r) \\ &= \sum_{r=1}^{m} J\big((O^\top D_m O)_{rr}\big) \\ &\leq \max_{\tilde{O}^\top \tilde{O} = I_m} \sum_{r=1}^{m} J\big((\tilde{O}^\top D_m \tilde{O})_{rr}\big).\end{aligned} \tag{A164}$$

The left-hand side attains the right-hand side at $U = \tilde{O}_\star^\top P_m^\top$ and $V = \tilde{O}_\star^\top Q_m^\top$, where $\tilde{O}_\star$ attains the maximum, since there

$$C = \tilde{O}_\star^\top P_m^\top A Q_m \tilde{O}_\star = \tilde{O}_\star^\top D_m \tilde{O}_\star. \tag{A165}$$

It remains to show that every other minimizer has the same row spaces. At any minimizer, the left-hand side equals the right-hand side, so both inequalities are equalities. Since J is strictly increasing on $s \geq 0$, the first one gives $|c_r| = \bar{w}_r$ for all r, so, as $\bar{w} \prec (\rho_1, \dots, \rho_m)$,

$$\sum_{r=1}^{m} |c_r| = \sum_{r=1}^{m} \bar{w}_r = \sum_{i=1}^{m} \rho_i. \tag{A166}$$

Flipping the signs of the r-th row of V and of α_r whenever $c_r < 0$ keeps the loss and gives $c_r \geq 0$, so

$$\begin{aligned}\operatorname{tr}(UAV^\top) &= \sum_{r=1}^{m} c_r \\ &= \sum_{r=1}^{m} |c_r| \\ &= \sum_{i=1}^{m} \rho_i.\end{aligned} \tag{A167}$$

Thus (U, V) attains equality in von Neumann's trace inequality, and, as in contrastive alignment, up to a common rotation of the rows of U and V,

$$(U^\star, V^\star) = (P_m^\top, Q_m^\top), \qquad W_x^\star = P_m^\top \Sigma_{xx}^{-1/2}, \quad W_y^\star = Q_m^\top \Sigma_{yy}^{-1/2}. \tag{A168}$$

Undoing the sign flips and the replacement by H_u and H_v keeps the row spaces, so the rows of $W_x^\star$ and $W_y^\star$ span the top-m generalized eigenspaces of (A5). □

Nonlinear encoders. We extend Theorem 1 to the multihead loss on the block-structured class of Appendix A.3.2, again with no condition on ρ. The class is closed under left multiplication by orthogonal matrices, so the heads can be absorbed as in the linear case, and we take $u_r = v_r = e_r$. With $\bar{\rho}_i$ as in Appendix A.3.2,

$$c_r = \mathbb{E}_+[f_r(x) g_r(y)] = \sum_i \bar{\rho}_i \Phi_{ri} \Psi_{ri}. \tag{A169}$$

Let $\mathcal{K}_r = \mathcal{K}(f_r(x), g_r(y))$ under positive pairs, with $\mathcal{K}$ as in Appendix A.3.2. Since $\mathbb{E}_-[f_r(x)^2 g_r(\tilde{y})^2] = 1$ and $\mathbb{E}_+[f_r(x)^2 g_r(y)^2] = 1 + 2c_r^2 + \mathcal{K}_r$, the r-th term is

$$\begin{aligned}&\mathbb{E}_+\big[(1 - \alpha_r f_r(x) g_r(y))^2\big] + \alpha_r^2\, \mathbb{E}_-\big[f_r(x)^2 g_r(\tilde{y})^2\big] \\ &\quad = 1 - 2\alpha_r c_r + \alpha_r^2 (2 + 2c_r^2 + \mathcal{K}_r) \\ &\quad = 1 - \frac{c_r^2}{2 + 2c_r^2 + \mathcal{K}_r} + (2 + 2c_r^2 + \mathcal{K}_r)\Big(\alpha_r - \frac{c_r}{2 + 2c_r^2 + \mathcal{K}_r}\Big)^2,\end{aligned} \tag{A170}$$

where $2 + 2c_r^2 + \mathcal{K}_r = 1 + \mathbb{E}_+[f_r(x)^2 g_r(y)^2] \geq 1$. Hence

$$(\mathcal{L}_{\mathrm{match}}^{\mathrm{MH}})^\star = m - \sum_{r=1}^{m} \frac{c_r^2}{2 + 2c_r^2 + \mathcal{K}_r}, \tag{A171}$$

and each summand is at most $J(c_r)$ whenever $\mathcal{K}_r \geq 0$.

We next show $\mathcal{K}_r \geq 0$. Write $f_r(x) = \sum_i X_i$ and $g_r(y) = \sum_i Y_i$ with $X_i = \Phi_{ri} h_{\ell_i}(\xi_i)$ and $Y_i = \Psi_{ri} h_{\ell_i'}(\zeta_i)$. The pairs (X_i, Y_i) are centered and independent across i, so

$$\mathbb{E}\Big[\Big(\sum_i X_i\Big)^2 \Big(\sum_i Y_i\Big)^2\Big] = \sum_i \mathbb{E}[X_i^2 Y_i^2] + \sum_{i \neq i'} \big(\mathbb{E}[X_i^2]\,\mathbb{E}[Y_{i'}^2] + 2\,\mathbb{E}[X_i Y_i]\,\mathbb{E}[X_{i'} Y_{i'}]\big), \tag{A172}$$

and the cross terms cancel in $\mathcal{K}$:

$$\mathcal{K}_r = \sum_i \mathcal{K}(X_i, Y_i) = \sum_i \Phi_{ri}^2 \Psi_{ri}^2\, \mathcal{K}\big(h_{\ell_i}(\xi_i), h_{\ell_i'}(\zeta_i)\big) \geq 0, \tag{A173}$$

since each $\mathcal{K}\big(h_{\ell_i}(\xi_i), h_{\ell_i'}(\zeta_i)\big) \geq 0$ by Appendix A.3.2.

Finally, for any set $\mathcal{S}$ of $k \leq m$ heads, let $\omega_i^{\mathcal{S}} = \frac{1}{2} \sum_{r \in \mathcal{S}} (\Phi_{ri}^2 + \Psi_{ri}^2)$. The rows of Φ and Ψ are orthonormal, so $0 \leq \omega_i^{\mathcal{S}} \leq 1$ and $\sum_i \omega_i^{\mathcal{S}} = k$. Since $\bar{\rho}_i \leq \rho_i$ and $ab \leq (a^2 + b^2)/2$,

$$\begin{aligned}
\sum_{r \in \mathcal{S}} |c_r| &\leq \sum_i \rho_i \omega_i^{\mathcal{S}} \\
&= \sum_{i=1}^{k} \rho_i + \sum_{i \leq k} \rho_i (\omega_i^{\mathcal{S}} - 1) + \sum_{i > k} \rho_i \omega_i^{\mathcal{S}} \\
&\leq \sum_{i=1}^{k} \rho_i + \rho_k \Big(\sum_i \omega_i^{\mathcal{S}} - k\Big) \\
&= \sum_{i=1}^{k} \rho_i.
\end{aligned} \tag{A174}$$

Taking $\mathcal{S}$ to be the k heads with the largest $|c_r|$ gives $(|c_1|, \ldots, |c_m|) \prec_w (\rho_1, \ldots, \rho_m)$, as in the linear case. The Marshall–Olkin–Arnold and Schur–Horn steps there use only this weak majorization, and give

$$\begin{aligned}
(\mathcal{L}_{\mathrm{match}}^{\mathrm{MH}})^\star &= m - \sum_{r=1}^{m} \frac{c_r^2}{2 + 2c_r^2 + \mathcal{K}_r} \\
&\geq m - \sum_{r=1}^{m} J(c_r) \\
&\geq m - \max_{\tilde{O}^\top \tilde{O} = I_m} \sum_{r=1}^{m} J\big((\tilde{O}^\top D_m \tilde{O})_{rr}\big).
\end{aligned} \tag{A175}$$

The linear minimizer $U = \tilde{O}_\star^\top P_m^\top$, $V = \tilde{O}_\star^\top Q_m^\top$ of the linear case lies in the class, with $\Phi = UP$, $\Psi = VQ$ and $\ell_i = \ell_i' = 1$. It has $\mathcal{K}_r = 0$ by Isserlis' formula and $c_r = (\tilde{O}_\star^\top D_m \tilde{O}_\star)_{rr}$, so it attains the right-hand side. It therefore minimizes $\mathcal{L}_{\mathrm{match}}^{\mathrm{MH}}$ over the block-structured class, with no condition on ρ.

Retrieval. The multihead score is again bilinear,

$$\sum_{r=1}^{m} t_r(x, y) = z_x^\top H_u \,\mathrm{diag}(\alpha_1^\star, \ldots, \alpha_m^\star) H_v^\top z_y, \tag{A176}$$

with H_u and H_v as in the linear case, and the match part of the proof of Theorem 6 uses only that the score is bilinear in (z_x, z_y). Hence Theorem 6 also holds with $q_{\mathrm{match}} = -\sum_r t_r$.

A.3.9 Unequal weighting and hard negatives.

The balanced loss above weights positive and negative pairs equally and draws $\tilde{y}$ independently of x. We now relax both. Pooling n_+ positive and n_- negative pairs into one average weights the negative term by $\lambda = n_-/n_+$. A hard-negative sampler draws $\tilde{y}$ for the query x with density proportional to $p(\tilde{y}) \exp(\beta z_x^\top M \tilde{z}_y)$, $\beta \geq 0$, so candidates that the current head scores higher are drawn more often; $\beta = 0$ recovers independent negatives. The loss becomes

$$\mathcal{L}^{\lambda,\beta}_{\text{match}} = \mathbb{E}_+\big[(1 - z_x^\top M z_y)^2\big] + \lambda\, \mathbb{E}_-^\beta\big[(z_x^\top M \tilde{z}_y)^2\big]. \tag{A177}$$

For every $\lambda > 0$ and $\beta \geq 0$, the rows of $W_x^\star$ and $W_y^\star$ still span the top-m generalized eigenspaces of (A5); only the optimal head changes. For $\beta = 0$ it is

$$\operatorname{vec}(M^\star) = \Big((1+\lambda) I_{m^2} + \operatorname{vec}(C^\star)\operatorname{vec}(C^\star)^\top + ((C^\star)^\top \otimes C^\star) K_{m,m}\Big)^{-1} \operatorname{vec}(C^\star). \tag{A178}$$

This claim concerns only the sampler above, not adaptive samplers in general.

Proof. Tilting the Gaussian $\tilde{y}$ by $\exp(\beta z_x^\top M W_y \tilde{y})$ shifts its mean by $\beta \Sigma_{yy} W_y^\top M^\top z_x$, so by whitening $\tilde{z}_y \mid x \sim \mathcal{N}(\beta M^\top z_x, I_m)$. With $z_x \sim \mathcal{N}(0, I_m)$,

$$\begin{aligned} \mathbb{E}_-^\beta\big[(z_x^\top M \tilde{z}_y)^2\big] &= \mathbb{E}\big[z_x^\top M M^\top z_x + \beta^2 (z_x^\top M M^\top z_x)^2\big] \\ &= \|M\|_F^2 + \beta^2\big(\|M\|_F^4 + 2\|M M^\top\|_F^2\big), \end{aligned} \tag{A179}$$

and, with Q_+ as in the proof above,

$$\begin{aligned} \mathcal{L}^{\lambda,\beta}_{\text{match}} = 1 - 2\operatorname{vec}(M)^\top \operatorname{vec}(C) + \operatorname{vec}(M)^\top (Q_+ + \lambda I_{m^2}) \operatorname{vec}(M) \\ + \lambda\beta^2\big(\|M\|_F^4 + 2\|M M^\top\|_F^2\big). \end{aligned} \tag{A180}$$

For $\beta = 0$ the loss is quadratic in M, which gives the stated $M^\star$, and the computation above with 2 replaced by $1 + \lambda$ gives

$$\min_M \mathcal{L}^{\lambda,0}_{\text{match}} = \frac{1}{1 + \sum_{i=1}^m \frac{s_i(C)^2}{1+\lambda+s_i(C)^2}}, \tag{A181}$$

which, as above, is minimized by $s_i(UAV^\top) = \rho_i$ for $i \leq m$.

For $\beta > 0$ there is no closed form, so we show that the minimized loss again depends on C only through its singular values and decreases in each of them. The loss is strictly convex in M: $Q_+ + \lambda I_{m^2} \succ 0$, $\|M\|_F^4$ is the square of a nonnegative convex function, and $\|MM^\top\|_F^2 = \sum_i s_i(M)^4$ is a convex spectral function. It has a unique minimizer, and it is unchanged under $(M, C) \mapsto (RMS^\top, RCS^\top)$ for orthogonal R, S. Hence, with $C = RDS^\top$ and $D = \operatorname{diag}(s_1, \ldots, s_m)$, the minimizer is $RM_DS^\top$, where M_D minimizes the loss at $C = D$. Flipping the sign of the i-th row and column of M leaves the loss at $C = D$ unchanged, so uniqueness forces $M_D = \operatorname{diag}(w_1, \ldots, w_m)$, and

$$\begin{aligned} \mathcal{L}^{\lambda,\beta}_{\text{match}} = 1 - 2\sum_i w_i s_i + (1+\lambda)\sum_i w_i^2 + \Big(\sum_i w_i s_i\Big)^2 + \sum_i w_i^2 s_i^2 \\ + \lambda\beta^2\Big(\Big(\sum_i w_i^2\Big)^2 + 2\sum_i w_i^4\Big). \end{aligned} \tag{A182}$$

Write $\tau = 1 - \sum_j w_j s_j$. At the minimizer, stationarity in w_i gives

$$s_i \tau = w_i g_i, \qquad g_i = 1 + \lambda + s_i^2 + 2\lambda\beta^2 \sum_j w_j^2 + 4\lambda\beta^2 w_i^2 > s_i^2, \tag{A183}$$

so $w_i s_i = s_i^2 \tau / g_i$ has the sign of τ. If $\tau \leq 0$, then $\sum_j w_j s_j \leq 0$ and $\tau \geq 1$, a contradiction. Hence $\tau > 0$, and for $s_i > 0$ we get $0 < w_i s_i < \tau$, so $w_i > 0$. By the envelope theorem,

$$\frac{\partial}{\partial s_i} \min_M \mathcal{L}^{\lambda,\beta}_{\text{match}} = -2w_i + 2w_i \sum_j w_j s_j + 2w_i^2 s_i = -2w_i(\tau - w_i s_i) < 0 \qquad (s_i > 0). \tag{A184}$$

The minimized loss therefore decreases in each $s_i(C)$, and since $s_i(UAV^\top) \leq \rho_i$ by interlacing, it is minimized by $s_i(UAV^\top) = \rho_i$ for $i \leq m$, which gives the same row spaces as above. □

A.4 Additional experiment results

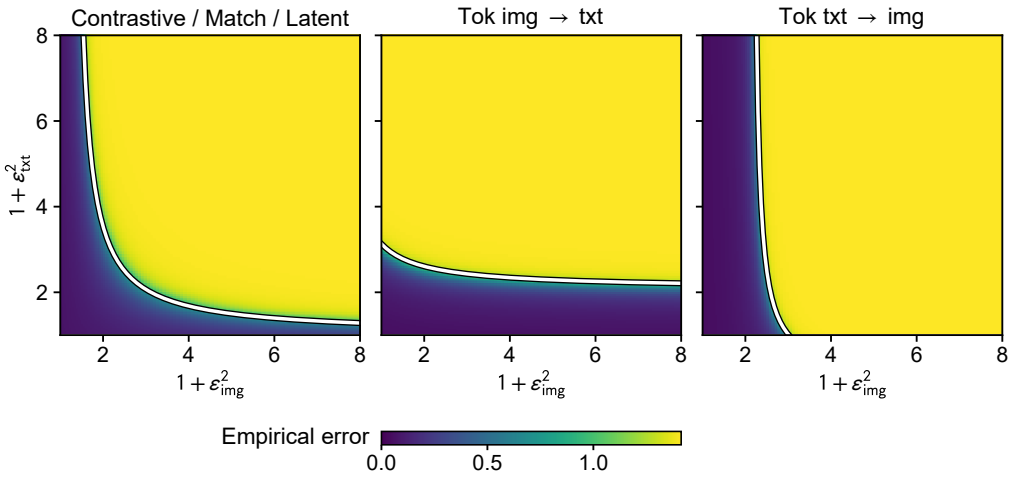


Figure A1: Empirical recovery error as a function of the image and text entanglement strengths, plotted as $1+\varepsilon^2_{\text{img}}$ and $1+\varepsilon^2_{\text{txt}}$, for the joint objectives (contrastive, match/no-match, and latent-space prediction, which share one solution) and for token-space generation in each direction (img→txt and txt→img). Each cell reports the projected Frobenius distance between the estimated and ground-truth clean subspaces under synthetic Gaussian data ($n = 10{,}000$): for the joint objectives, the larger of the image-side and text-side distances; for token-space generation, the distance for its source encoder. Overlaid curves show the recovery boundaries of Theorem 3; the recovery region is the side of each curve that contains the lower-left corner. Yellow indicates high error (recovery fails); dark purple indicates near-perfect recovery.

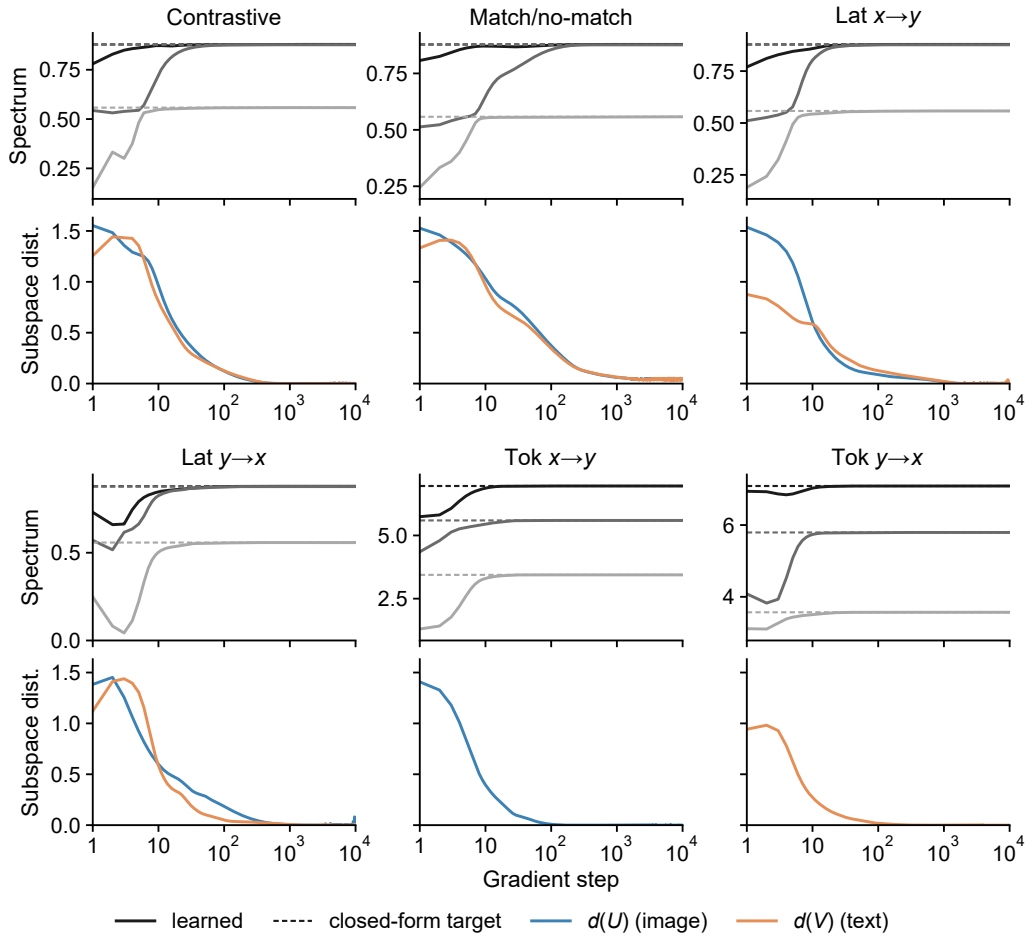


Figure A2: Gradient-descent training of the linearized objectives on the numerical model of Section 4.1, with steps on a log scale. For each loss, the upper panel shows the learned spectrum (solid) and its closed-form target (dashed): singular values for the joint objectives and eigenvalues for token-space generation. The final spectral errors are 1×10^{-12}, 2×10^{-4}, 9×10^{-10}, 1×10^{-8}, 1×10^{-7}, and 1×10^{-7} in reading order. The lower panel shows the subspace distances of the image encoder, $d(U)$, and the text encoder, $d(V)$, to their closed-form row spaces. Token-space generation trains only its source encoder, so a single distance is shown.

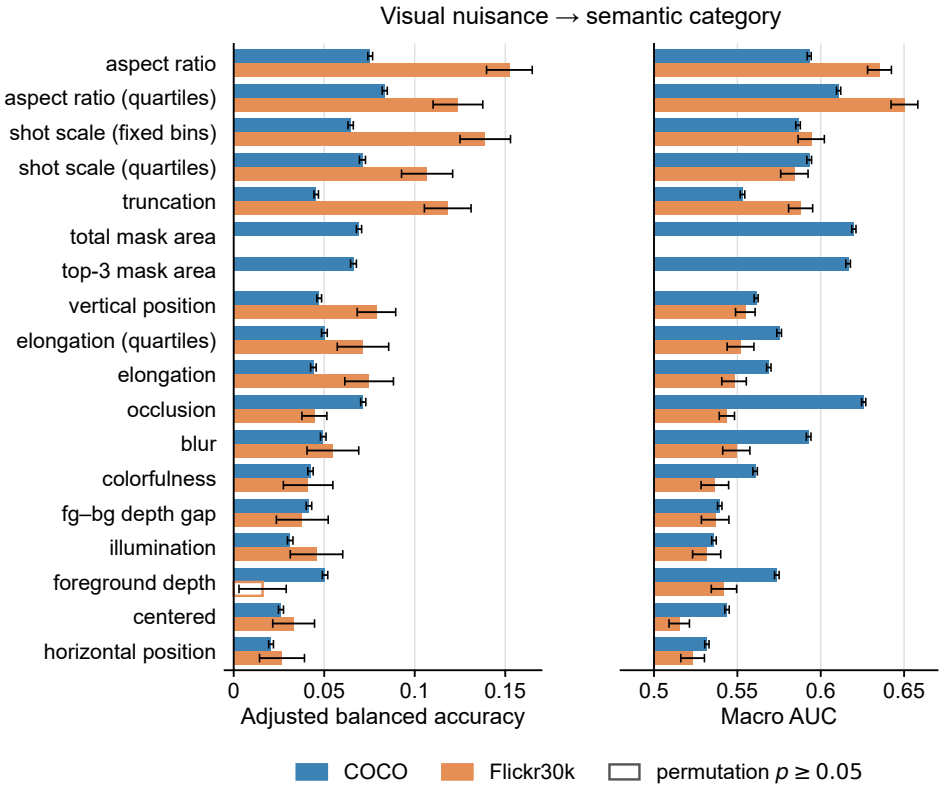


Figure A3: Probes predicting the semantic category of the primary object shared by an image and its captions (12 COCO super-categories; 8 Flickr30k Entities phrase types) from a single **visual** nuisance attribute, for every visual attribute we probed. Each probe is a class-balanced logistic regression evaluated with five-fold cross-validation. *Left:* adjusted balanced accuracy (zero = chance). *Right:* macro-averaged one-vs-rest AUC (bars start at chance, 0.5). Error bars show the standard deviation of a 200-permutation null; hollow bars mark probes with permutation $p \geq 0.05$. Quartile variants bin an attribute at its quartiles; the two mask-area attributes are available only for COCO. Every probe is above chance on both metrics except foreground depth on Flickr30k (balanced accuracy only), providing correlational evidence consistent with the entangled-shared structure of Section 3; Figure 2 shows a subset.

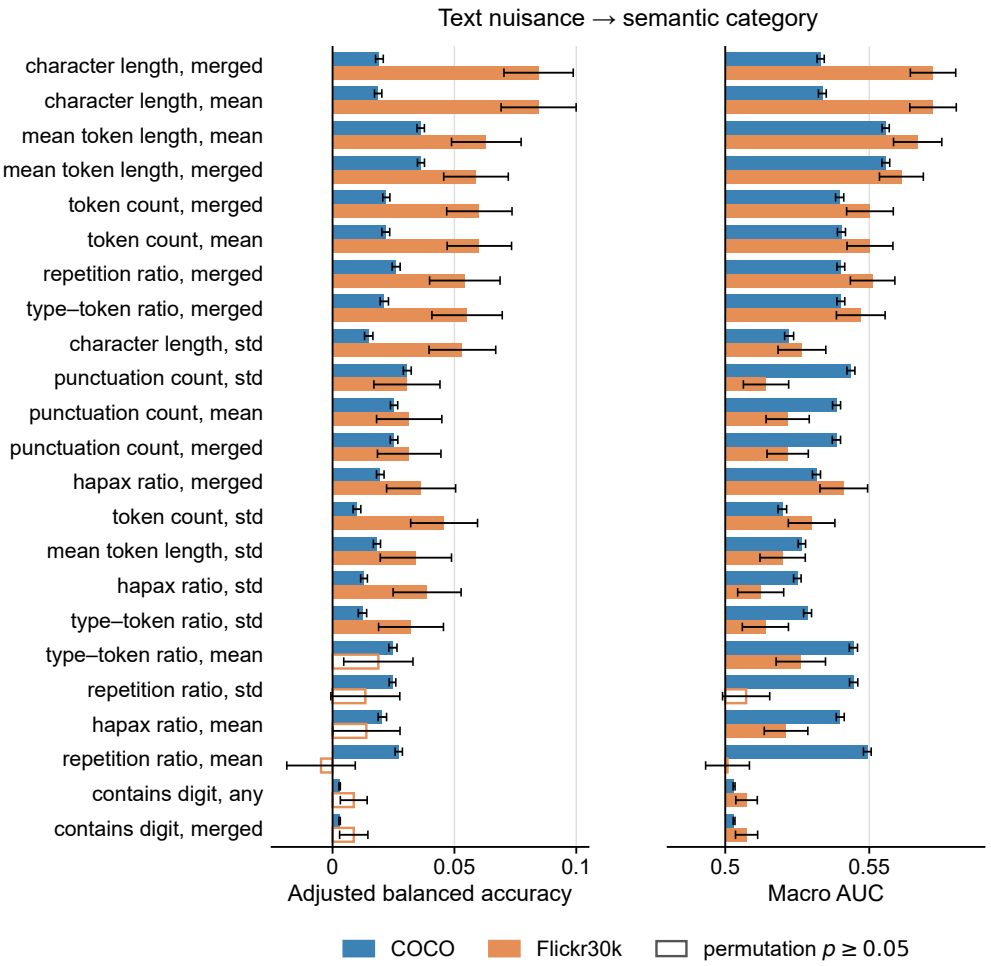


Figure A4: Same probes as Figure A3, from a single **text** nuisance attribute. For each attribute, *mean* and *std* summarize it over the captions of an image, and *merged* computes it on their concatenation. All probes are above chance on COCO. On Flickr30k, six probes are not significant for balanced accuracy, two of which (repetition ratio, mean and std) are also not significant for AUC; the remaining probes predict the shared semantics above chance.

Table A2: R@1 (%) under four entanglement settings (LL/HL/LH/HH, corresponding to dots in Figure 1B), averaged over five seeds. Percentage changes are relative to the LL baseline; absolute R@1 in parentheses. I→T denotes image-query text retrieval, and T→I denotes text-query image retrieval. Contrastive, match, and latent-space prediction degrade at both HL and LH. Token-space objectives degrade only when their *target* side is high.

	img low, txt low (baseline)		**img high, txt low**		**img low, txt high**		**img high, txt high**	
Loss	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
Lat $x \to y$	0.47±0.07	**1.17±0.05**	-9.3% (0.43±0.05)	-31.0% (0.81±0.12)	-6.3% (0.44±0.08)	-25.9% (0.87±0.17)	-6.8% (0.44±0.06)	-28.3% (0.84±0.11)
Tok $x \to y$	0.52±0.13	1.14±0.08	**+0.0% (0.52±0.10)**	**-5.1% (1.08±0.04)**	-54.2% (0.24±0.07)	-47.4% (0.60±0.08)	-48.5% (0.27±0.05)	-45.8% (0.62±0.07)
Match	0.80±0.09	0.85±0.05	-26.3% (0.59±0.06)	-29.0% (0.61±0.11)	-25.3% (0.60±0.07)	-32.3% (0.58±0.10)	-27.5% (0.58±0.07)	-32.6% (0.58±0.06)
Contrastive	0.95±0.13	0.94±0.08	-27.4% (0.69±0.08)	-37.9% (0.58±0.05)	-31.0% (0.65±0.18)	-24.3% (0.71±0.08)	-29.1% (0.67±0.14)	-27.2% (0.68±0.07)
Tok $y \to x$	1.16±0.14	0.40±0.05	-51.0% (0.57±0.04)	-32.2% (0.27±0.07)	**-1.5% (1.15±0.19)**	**+13.6% (0.45±0.03)**	-54.0% (0.54±0.07)	-30.2% (0.28±0.05)
Lat $y \to x$	**1.19±0.18**	0.57±0.09	-32.7% (0.80±0.11)	-30.3% (0.40±0.07)	-28.5% (0.85±0.15)	-23.9% (0.43±0.08)	-32.3% (0.81±0.04)	-35.9% (0.36±0.04)

Table A3: R@1 (%) on Flickr30k after fine-tuning both backbones end-to-end under each loss ($k = 256$; best epoch per direction; mean ± SD over five seeds).

Loss	I→T	T→I
Lat $x \to y$	18.26±0.09	**47.27±0.04**
Tok $x \to y$	47.42±0.13	46.68±0.07
Match	53.24±0.21	43.94±0.03
Contrastive	53.44±0.18	41.14±0.05
Tok $y \to x$	52.90±0.12	27.40±0.09
Lat $y \to x$	**60.44±0.13**	10.50±0.05

A.5 Hyperparameters and runtimes

All experiments were run on NVIDIA L40S GPUs; the time estimates below assume a single L40S with pre-staged data and exclude scheduler queue time.

A.5.1 Hyperparameters for Section 4.1

For the gradient-based numerical validation, we optimized each objective with Adam [34] using learning rate 0.1 for 10,000 steps. Unless otherwise noted, experiments used random seed 42. For the retrieval evaluation, we used 1,000,000 training samples, 100,000 evaluation samples, and retrieval group size 1000. For the entanglement-boundary plots, we evaluated a 100×100 grid over the two entanglement strengths.

The full numerical-experiments pipeline in this subsection takes approximately 2–4 hours end-to-end, with the 100×100 entanglement grid sweep accounting for most of the time.

A.5.2 Hyperparameters for Section 4.2

For each dataset and modality, we trained a logistic-regression probe from a single nuisance attribute to the semantic super-category label. We used L_2-regularized logistic regression with regularization strength $C = 1.0$ and evaluated performance with 5-fold stratified cross-validation. All experiments used random seed 42.

The full pipeline over COCO and Flickr30k takes approximately 2–3 hours end-to-end, with COCO image-side depth feature extraction dominating runtime.

A.5.3 Hyperparameters for Section 4.3

For each entanglement setting, we generated 50,000 training examples and 10,000 clean test examples at image resolution 64×64. The low- and high-entanglement settings used entanglement strengths 1.0 and 6.0, respectively, on the image and text sides, yielding the LL, HL, LH, and HH conditions. Results were averaged over 5 random seeds.

The image encoder was a 4-block strided CNN with hidden width 256. The text encoder was a 3-layer Transformer with model dimension 192, 4 attention heads, and maximum sequence length 64. Both encoders were trained with AdamW [35] using weight decay 10^{-4} and batch size 256. The image encoder used learning rate 3×10^{-4} and the text encoder used learning rate 5×10^{-4}. Both were trained for 875 steps per run.

The full sweep averaged over 5 seeds takes approximately 5 hours end-to-end, or about 15 minutes per setting-seed run, with encoder training dominating runtime.

A.5.4 Hyperparameters for Section 4.4

We extracted frozen image and text features using `facebook/dinov2-base` [19] and `Alibaba-NLP/gte-modernbert-base`, respectively. On top of these features, we fit each retrieval objective while sweeping the bottleneck rank $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 768\}$ and report the best R@1 over the rank sweep. Following the setup in Section 4.4, we used nonlinear heads for Flickr30k and COCO, and linear heads for CC3M. For the nonlinear setting, the encoder and decoder heads were one-hidden-layer residual MLPs with dropout 0.1.

The full retrieval-performance pipeline takes approximately 1–1.5 days end-to-end when CC3M feature extraction is included, whereas the downstream rank sweeps on already cached features typically finish in under 1 hour.

A.6 Asset Licenses and Terms of Use

In accordance with the NeurIPS guidelines, we detail the licenses and terms of use for all existing datasets and pre-trained models utilized in this work. All assets were used strictly for non-commercial, academic research purposes, respecting their original terms.

Table A4: Licenses and terms of use for datasets and models used in our paper.

Asset	**Type**	**License / Terms of Use**
COCO[10]	Dataset	CC BY 4.0 (Annotations); Images bound by Flickr terms
Flickr30k[9]	Dataset	Custom (Available upon request for non-commercial research)
CC3M[11]	Dataset	Custom (Freely available for academic research)
DINOv2[19]	Model	Apache License 2.0
ModernBERT[20]	Model	Apache License 2.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both abstract and introduction accurately covers the theoretical framework for multimodal losses and empirical results on synthetic and real datasets which matches the paper's contributions and scope in the main text.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The authors explicitly acknowledge the limitation of analyzing a local linear setting at Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper lists all assumptions clearly in the main text Section 3 and provided complete mathematical proof in Appendix A.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper lists specific model architecture and datasets and experimental settings for both numerical and real data evaluations at Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provide open access to both data and code as a Supplementary Zip File on OpenReview.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all hyperparameter details in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports run-to-run variability for experiments by averaging results over random seeds and providing corresponding standard deviations in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides all compute time estimation details and resources in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves theoretical mathematical analysis and experiments on standard, publicly available vision-language datasets, which conforms to standard ethical guidelines without introducing novel ethical violations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is theoretical research focusing on mathematical analysis. Because it analyzes abstract objectives rather than deploying real-world applications, direct societal impacts are inapplicable.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper investigates the theoretical properties and retrieval performance of training objectives and does not release a new pre-trained generative model or high-risk dataset.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets (datasets and models) are properly cited in the main text, and their respective licenses and terms of use are explicitly detailed in Appendix A.6

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper relies on existing open-source datasets and theoretical mathematical contributions; it does not introduce new standalone datasets or large models for public release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study is purely theoretical and computational, involving no crowdsourcing or human subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects, making IRB approval inapplicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this paper is originally from analyzing multimodal objectives theoretically.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.