UNIMODAL AUGMENTATION-BASED SSL OBJECTIVES INDUCE CROSS-MODAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent works have observed implicit cross-modal alignment between independently trained unimodal models. However, the mechanisms underlying the emergence of this phenomenon remain unresolved. In this paper, we demonstrate that unimodal augmentation-based self-supervised objectives implicitly induce cross-modal alignment. We show that low unimodal losses force high cross-modal alignment, whereas a high unimodal loss prevents such alignment. We theoretically connect the widely used InfoNCE loss to local CKA, the popular measure of cross-modal alignment. We further provide extensive empirical evidence supporting this claim, both in controlled settings and across 53 unconstrained frozen text, vision, and audio unimodal encoders. In both cases, unimodal augmentation-based objectives control cross-modal alignment. Overall, our results provide a theoretical and empirical foundation for understanding the emergence of implicit cross-modal alignment by establishing a direct connection between unimodal training objectives and cross-modal structure.

1 INTRODUCTION

Recent discoveries in representation learning have provided empirical evidence that independently trained unimodal vision and language models converge toward a unified representation, particularly in terms of their local neighborhood structure (Huh et al., 2024; Gröger et al., 2026a). This phenomenon is known as the Platonic Representation Hypothesis (PRH), and has been further supported by extensive subsequent work (Li et al., 2024; Acevedo et al., 2025; He et al., 2025), including extensions to additional modalities (Duraphe et al., 2025; Li et al., 2025; Edamadaka et al., 2025; Ngo & Kim, 2024) and practical applications (Yacobi et al., 2026; Yoon et al., 2025; Zhang et al., 2025; Gröger et al., 2026b; Jha et al., 2026; Gupta et al., 2026; Schnaus et al., 2025; Li & Walsh, 2026).

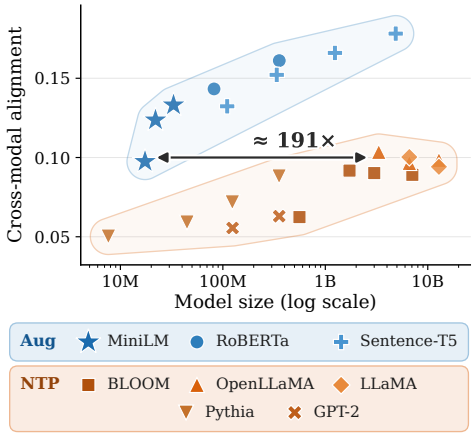


Figure 1: **Augmentation-based objectives induce cross-modal alignment.** Cross-modal text-vision alignment (mutual k-nearest neighbors) as a function of model size (number of parameters) across different text models to DINOv2. Notably, augmentation-based models achieve higher cross-modal alignment for a given model size, and yield superior alignment even with significantly fewer parameters.

However, the understanding of the causes underlying this emergent alignment remains limited. Most works either investigate the effect of scaling-related attributes of model size and capacity or the amount of training data on the strength of this phenomenon (Umar & Laio, 2026; Lobashev, 2025), or empirically address this question with explicit cross-modal factors. For example, shared input statistics and architectural biases (Kapoor et al., 2026; Tjandrasuwita et al., 2025), or overlap in datasets and downstream tasks (Li et al., 2025).

Meanwhile, Augmentation-based Self-Supervised Learning (Aug-SSL) has emerged as a leading training paradigm for training large models (Gui et al., 2024). Popular models such as DINO for

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

vision (Oquab et al., 2023), Sentence-BERT for text (Reimers & Gurevych, 2019), and BYOL-A for audio (Niizumi et al., 2021), have all been trained via Aug-SSL. While its power in unimodal settings is well-studied (Wang & Isola, 2020; Balestriero & LeCun, 2022), to the best of our knowledge, its effect on implicit cross-modal alignment has remained unexplored.

In this paper, we show that emergent cross-modal alignment (CMA) is a direct consequence of unimodal Aug-SSL objectives. That is, the unimodal objectives implicitly induce cross-modal alignment, even in the absence of explicit cross-modal supervision. The augmentations act as view-specific information removers, filtering out view-specific variation. The encoders then steer each embedding space toward the shared semantic structure underlying the modalities, thereby promoting cross-modal similarity. To the best of our knowledge, *this is the first work to explain implicit cross-modal alignment via unimodal properties, connecting them both theoretically and in practice.*

We begin with a theoretical investigation, directly connecting the widely used InfoNCE objective (Oord et al., 2018) to local CKA, a popular measure of CMA (Cristianini et al., 2001). We prove that low unimodal Aug-SSL losses force high CMA (Thm. 1), whereas even a single high unimodal loss restricts the alignment (Thm. 2). These results hold independently of data size, model architecture, and other factors previously proposed to explain the emergence of cross-modal alignment.

We further provide extensive empirical evidence for these results in real-world settings, both in controlled experiments and across large-scale pretrained models and modalities. For example, Fig. 1 shows that pre-trained text augmentation-based models consistently achieve higher CMA with a vision model than models of comparable size trained with next-token prediction (NTP). Moreover, augmentation-based models reach high CMA with substantially fewer parameters.

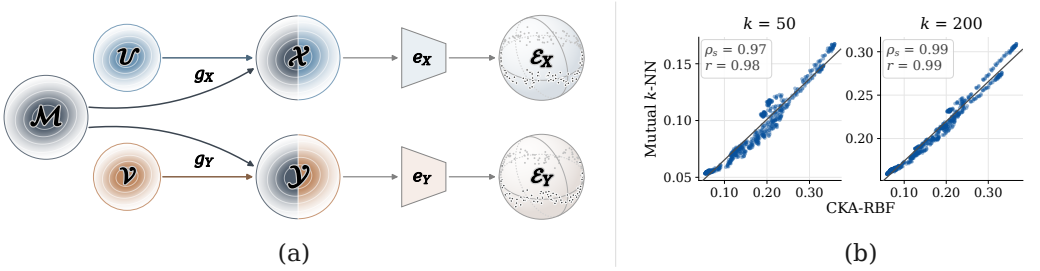
Contributions. Our contributions are summarized as follows: (1) We show that unimodal self-supervised objectives implicitly drive cross-modal alignment; (2) We theoretically establish a connection between the unimodal InfoNCE loss and cross-modal alignment as measured by local CKA; and (3) We empirically demonstrate that the unimodal Aug-SSL loss predicts cross-modal alignment across controlled experiments and large-scale pretrained models.

2 Related Work

Implicit Cross-Modal Alignment. Numerous studies have documented the emergence of implicit CMA in vision-language models (Huh et al., 2024; Li et al., 2024; Acevedo et al., 2025; Miller, 2026; Fan et al., 2025; He et al., 2025). Subsequent work has shown that similar alignment phenomena arise across a broader range of modalities, including video-text (Ovsjanikov et al., 2026), astronomical data (Duraphe et al., 2025), and scientific data (Li & Walsh, 2026; Edamadaka et al., 2025), among others. Beyond documenting its existence, several studies have examined how CMA depends on properties of the data and training setup, attributing its emergence to factors such as dataset and task overlap (Li et al., 2025) or shared data characteristics (Tjandrasuwita et al., 2025). However, these explanations remain empirical and do not establish a theoretical account of why CMA emerges.

Theory of Implicit CMA. Several recent works have developed theoretical perspectives on implicit cross-modal alignment. Lu et al. (2026) formalize the alignment hypothesis using category theory. Insulla et al. (2025) establish a connection between representation alignment and downstream task performance. However, these do not characterize the mechanisms underlying its emergence. Other studies investigate potential mechanisms that may give rise to this phenomenon (Bangachev et al., 2026; Ziyin et al., 2026; Ziyin & Chuang, 2025); however, their analyses rely on strong linearity assumptions about the data or learned representations, limiting their applicability to the highly nonlinear models used in practice. Lobashev (2025) further assumes that the two modalities share input statistics, imposing an additional restriction on the data-generating process. In contrast, this paper provides a theoretical account of how implicit cross-modal alignment emerges directly from the unimodal training objective in general settings.

Augmentation-based SSL. Self-supervised learning (SSL) has emerged as a dominant training paradigm across a wide range of domains, including vision (Chen et al., 2020; Oquab et al., 2023; He et al., 2022), language (Devlin et al., 2019; Gao et al., 2021; Liu et al., 2019), audio (Baevski



(a) (b)

Figure 2: **(a) Theoretical setup visualization.** Three manifolds $\mathcal{M}, \mathcal{U}$, and $\mathcal{V}$ generate two observation spaces $\mathcal{X}$ and $\mathcal{Y}$ with shared semantics $\mathcal{M}$. Two trainable encoders then map the observations to the corresponding embedding spaces $\mathcal{E}_X$ and $\mathcal{E}_Y$. **(b) Correlation between CKA-RBF and mkNN.** mkNN is highly correlated with CKA-RBF across different choices of k and the corresponding σ. Both Spearman (ρ_s) and Pearson (r) correlations are close to 1.

et al., 2020; Hsu et al., 2021; Saeed et al., 2021; Niizumi et al., 2021), video (Tong et al., 2022; Bardes et al., 2024), and scientific data (Wang et al., 2022). In particular, augmentation-based SSL provides a simple and effective mechanism for learning from large numbers of self-generated training examples through different transformations of the same underlying samples. Prior works have investigated the properties induced by SSL training within individual modalities (Wang & Isola, 2020; Balestriero & LeCun, 2022). Conversely, other works have demonstrated the effectiveness of explicit cross-modal SSL objectives for learning aligned representations (Radford et al., 2021; Wu et al., 2022). In contrast, this paper considers the implications of SSL on *implicit cross-modal* alignment, without any cross-modal objective.

3 Theoretical Analysis: From Unimodal InfoNCE to CMA

This section presents our theoretical results connecting unimodal InfoNCE and cross-modal alignment (CMA). We first formally define the setup, CMA, and unimodal augmentation-based self-supervised learning (Aug-SSL) (Sec. 3.1). We then establish their connection through two complementary results (Sec. 3.2). In Thm. 1, we show that low unimodal losses force high CMA. In Thm. 2, we show that a high unimodal loss for even one model restricts CMA, independently of other factors such as model architecture and training data size. Together, these results establish that unimodal augmentation-based training is a fundamental driver of implicit CMA. For brevity, we defer the complete derivations and proofs to App. A.

3.1 Setup

Notations (visualized in Fig. 2a)**.** Let $\mathcal{M}, \mathcal{U}, \mathcal{V}$ be manifolds representing shared semantics and view-specific information, respectively. Observations are generated by $g_X : \mathcal{M} \times \mathcal{U} \to \mathcal{X}$ and $g_Y : \mathcal{M} \times \mathcal{V} \to \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}, \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Two trainable unimodal encoders, $e_X : \mathcal{X} \to \mathcal{E}_X, e_Y : \mathcal{Y} \to \mathcal{E}_Y$, map the observations to normalized embedding spaces $\mathcal{E}_X \subseteq \mathbb{S}^{d_X-1}, \mathcal{E}_Y \subseteq \mathbb{S}^{d_Y-1}$, respectively. Throughout, $0 < \tau \leq 1$ denotes the fixed temperature of the training objective.

The phenomenon of implicit CMA dictates that independently trained unimodal encoders naturally develop highly aligned representations. To measure the alignment between the embedding spaces $\mathcal{E}_X$ and $\mathcal{E}_Y$, we use the well-established Centered Kernel Alignment (CKA) metric (Kornblith et al., 2019), following recent works on implicit cross-modal alignment (Huh et al., 2024; Gröger et al., 2026a). We denote the observations $x = g_X(m, u), x' = g_X(m', u'), y = g_Y(m, v), y' = g_Y(m', v')$. That is, x and y share the same underlying semantics m, and likewise x' and y' share m'. In practice, this shared semantic structure is obtained through paired data. The CKA is then often defined using the Hilbert-Schmidt Independence Criterion (HSIC), which is

$$\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_Y) = \mathbb{E}\left[k_X^c\left(e_X(x), e_X(x')\right)\ k_Y^c\left(e_Y(y), e_Y(y')\right)\right]$$

where k^c is the doubly centered kernel $k^c = k - \mathbb{E}[k|x] - \mathbb{E}[k|x'] + \mathbb{E}[k]$. CKA then measures the correlation between the centered kernels, giving:

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) = \frac{\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_Y)}{\sqrt{\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_X)\ \mathrm{HSIC}(\mathcal{E}_Y, \mathcal{E}_Y)}} \in [0, 1], \tag{1}$$

Throughout the paper, we use the CKA-RBF version of CKA (Kornblith et al., 2019), with the Gaussian kernel $k(z, z') = e^{-\|z-z'\|^2/2\sigma^2}$, and denote the kernel variances $\sigma_X^2 = \mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_X), \sigma_Y^2 = \mathrm{HSIC}(\mathcal{E}_Y, \mathcal{E}_Y)$. In particular, we assume $\sigma_X, \sigma_Y > 0$ so the CKA is defined.

The Gaussian kernel emphasizes local structure, making this measure particularly suitable for studying the local nature of implicit cross-modal alignment observed in recent work (Gröger et al., 2026a). Moreover, as shown in Fig. 2b, CKA-RBF is strongly correlated with the mutual k-nearest-neighbors (mkNN) metric used in recent works, with the kernel bandwidth σ playing a role analogous to the neighborhood parameter k (see App. J).

Augmentation-based SSL. We define an *augmentation* as a resampling of both the semantics and the view, while preserving the underlying semantic structure. Specifically:

$$x^+ := \mathrm{aug}(x) = g_X(m^+, u^+), \qquad m^+ \sim P(m, \cdot), \quad u^+ \sim \nu_{\mathcal{U}} \text{ independent} \tag{2}$$

where $P(m, \cdot) := \mathsf{J}(m, \cdot)\,\mu$ is a Markov kernel on $\mathcal{M}$ with density J. This definition aligns with prior theoretical studies of augmentation-based SSL (Von Kügelgen et al., 2021), while not restricting the augmentation to exactly preserve the semantics. We exclude the degenerate case $\mathsf{J} \equiv 1$, in which the augmentation resamples the semantics independently and carries no signal.

An Aug-SSL model is trained to maximize the similarity between augmented samples, while preventing collapse to a constant representation, for instance by contrasting against randomly sampled observations. A popular objective for this purpose is the InfoNCE loss (Oord et al., 2018), which has been widely adopted in vision (Chen et al., 2020; He et al., 2020), text (Reimers & Gurevych, 2019; Gao et al., 2021), audio (Niizumi et al., 2021), and other domains (Qian et al., 2021; Wang et al., 2022). Let $x, x' \in \mathcal{X}$ denote observations and let $x^+ = \mathrm{aug}(x)$ be an augmentation of x. The K-negative InfoNCE loss is

$$\ell_X^{(K)} = \mathbb{E}\left[-\log \frac{\exp\big(\langle e_X(x), e_X(x^+)\rangle/\tau\big)}{\exp\big(\langle e_X(x), e_X(x^+)\rangle/\tau\big) + \sum_{k=1}^{K} \exp\big(\langle e_X(x), e_X(x_k)\rangle/\tau\big)}\right], \tag{3}$$

where K is a hyperparameter controlling the number of negative samples. Following Wang & Isola (2020) (Thm. 1), we use its population form:

$$N_X = \lim_{K\to\infty} \big(\ell_X^{(K)} - \log K\big) = -\frac{1}{\tau}\,\mathbb{E}\big[\langle e_X(x), e_X(x^+)\rangle\big] + \mathbb{E}_x\Big[\log \mathbb{E}_{x'}\big[e^{\langle e_X(x), e_X(x')\rangle/\tau}\big]\Big].$$

For convenience, we normalize the objective as follows, with both ends attainable (Lemma 8):

$$L_X = \frac{N_X + 1/\tau}{R_\tau} \in [0, 1],$$

where $R_\tau = \frac{2}{\tau} + \log\cosh\frac{1}{\tau}$. L_Y is analogously defined for $\mathcal{Y}$.

3.2 Connecting InfoNCE and local CKA

We now present the theoretical results that formalize the connection between the unimodal Aug-SSL losses (InfoNCE) and CMA under the setup above. The two results establish complementary directions of this relationship. Specifically, Thm. 1 shows that low losses in both modalities guarantee high CMA, while Thm. 2 shows that high loss in either modality prevents alignment.

Implicit alignment. The bound in Thm. 1 utilizes the concentration strength of the augmentations, $\|J - 1\|_{L^2} =: \chi$, where $J(m, m^+)$ is the likelihood ratio of m^+ being an augmentation of m versus an independent sample. Thus, χ quantifies the semantic information retained by augmentations: $\chi = 0$ when $J \equiv 1$ and augmentations ignore the original semantics, and $\chi = \infty$ when they preserve them exactly. We consider the practical regime $0 < \chi < \infty$, where augmentations alter but retain semantics. Together with the InfoNCE temperature τ and CKA-RBF scale σ, this allows us to lower-bound CKA in terms of the unimodal losses.

Theorem 1 (Implicit alignment)**.**

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\geq\; f(L_X, L_Y \mid \tau, \sigma, \chi),$$

with f non-increasing in both L_X and L_Y, and converging to 1 *as both losses approach* 0*.*

This theorem establishes that when both ***unimodal*** Aug-SSL losses (L_X and L_Y) are effectively minimized, a high degree of cross-modal alignment is guaranteed. Importantly, the bound depends only on the attained loss values, not on the training procedure. The monotonic convergence of f to 1 strictly forces the CKA toward its maximum value when the losses are sufficiently small. Specifically,

$$f(L_X, L_Y) = \cos(\theta_X + \theta_Y), \quad \theta(L) = \text{clamp}\left(\arccos\left(\frac{a_\sigma - b_{\sigma,\tau} L}{\chi\sqrt{D_{\sigma,\tau} L}}\right),\ 0,\ \frac{\pi}{2}\right),$$

where clamp restricts $\theta(L)$ to $[0, \frac{\pi}{2}]$, and $\theta_X = \theta(L_X), \theta_Y = \theta(L_Y)$. In particular, f depends on the augmentations concentration strength χ, and on the InfoNCE temperature τ and the CKA-RBF scale σ through the hyperparameter-dependent constants: $a_\sigma, b_{\sigma,\tau}$ and $D_{\sigma,\tau}$. For brevity, their exact definitions are deferred to App. A.8.

Intuitively, augmentations remove view-specific information, filtering out variations that are not shared across modalities. Low loss then encourages augmented samples to cluster together, guiding each embedding space toward the shared semantic structure underlying the modalities. Thereby, promoting cross-modal similarity, i.e., high CKA.

Implicit misalignment. While the above highlights the geometric convergence driven by successful unimodal optimization, evaluating the implications of high unimodal loss requires the complementary perspective addressed in Thm. 2. For this, we assume the unimodal deep SSL encoders e_X, e_Y exhibit Lipschitz continuity. That is, $\|e_X(x) - e_X(x')\| \leq \ell_X\, d_{\mathcal{X}}(x, x')$ for all $x, x' \in \mathcal{X}$, with an analogous bound ℓ_Y for e_Y. This is consistent with recent literature, proving both empirically and theoretically that this happens in practice (Virmaux & Scaman, 2018; Lee et al., 2021). This theorem utilizes an augmentation-dependent constant, the semantic displacement:

$$\varepsilon_X = \mathbb{E}_{\pi,u}\Big[d_{\mathcal{X}}\big(g_X(m,u), g_X(m^+,u)\big)^2\Big],$$

where $\pi(m, m^+) = J(m, m^+)\mu(m)\mu(m^+)$ is the joint law of the semantic pair. Intuitively, ε_X measures the extent to which the semantic variations caused by the augmentation process are manifested in the observation space.

Theorem 2 (Implicit misalignment)**.** *If the encoder e_X is Lipschitz with a constant ℓ_X, and $\tau \leq \sigma^2$, then*

$$\text{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \ \leq\ g(L_X \mid \tau, \sigma, \ell_X, \varepsilon_X),$$

where g is strictly decreasing in L_X. This holds symmetrically for L_Y with ℓ_Y, ε_Y.

Thm. 2 shows that a sufficiently high ***unimodal*** loss strictly upper-bounds CMA, preventing convergence. In particular, for practical hyperparameters, the bound remains bounded away from 1 as L_X approaches its maximum (App. K). Consequently, a poorly optimized encoder strictly precludes the emergence of shared geometric structure, regardless of the optimization quality in the counterpart modality. The condition $\tau \leq \sigma^2$ is consistent with standard InfoNCE practice, where τ is a small fraction and σ^2 is typically larger for normalized embeddings.

The bound is given by

$$g(L_X) = \frac{d_{\sigma,\tau} + \ell_X^2 \varepsilon_X}{2\sigma^2 \sigma_X} + \frac{h_{\sigma,\tau}}{\sigma_X}(1 - L_X).$$

In particular, g depends on the Lipschitz constant ℓ_X, the semantic displacement ε_X, and on the InfoNCE temperature τ and the CKA-RBF scale σ via the hyperparameter-dependent constants: $d_{\sigma,\tau}$ and $h_{\sigma,\tau}$ (App. A.8).

Intuitively, only semantic variation can align across modalities, so CKA is bounded by the fraction of embedding variation that is semantic. As the unimodal loss increases, less variation reflects semantics, limiting the achievable CKA. Both bounds are visualized in App. K.

Together, these theoretical results provide a rigorous mathematical foundation, showing that implicit cross-modal alignment is a geometric consequence of the unimodal InfoNCE objective. In the next section, we empirically demonstrate these predictions, showing that Aug-SSL models exhibit higher CMA than their counterparts.

Message of Section 3: Low unimodal augmentation-based SSL losses induce cross-modal convergence, while a high loss in either modality forces misalignment.

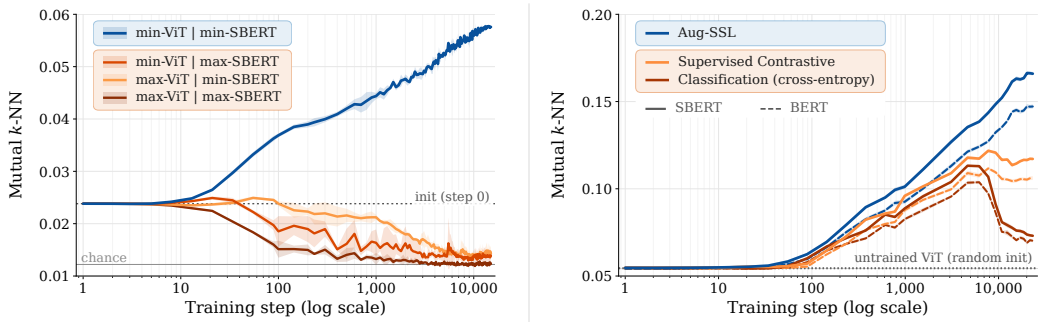


Figure 3: **(a) Empirical validation of Thm. 1 and 2.** Vision (ViT) and text (SBERT) models trained to minimize/maximize the Aug-SSL loss (Eq. 3). Minimizing both unimodal losses improves CMA (Thm. 1), whereas maximizing either loss prevents alignment, with CMA converging to the chance level (Thm. 2). **(b) Augmentations vs. Supervision.** CMA throughout training of a vision model (ViT) against frozen text models: an augmentation-based model (SBERT) and a non-augmentation-based model (BERT). Training ViT with augmentations achieves substantially higher CMA, including compared to the BERT model.

4 Aug-SSL is Key and Predictor of CMA in Practice

In this section, we empirically validate the core premise of this paper. We consider both controlled experiments (Sec. 4.1) and frozen pretrained models (Sec. 4.2), providing evidence that unimodal augmentation-based objectives play a central role in the emergence of implicit CMA. Further details concerning the experimental setup and implementation are provided in App. L. Code for reproducing all results will be publicly available upon acceptance.

Datasets. We utilize four benchmark datasets. Following Huh et al. (2024) and Gröger et al. (2026a), we use the Wikipedia image-text (WIT) dataset (Srinivasan et al., 2021). We additionally use the MS-COCO image-text benchmark (Lin et al., 2014), which provides semantic labels that allow us to more closely examine the effect of augmentations on cross-modal alignment. To further validate the phenomenon across modalities, we use the audio-text CLOTHO (Drossos et al., 2020) and AudioCaps (Kim et al., 2019) benchmarks.

Evaluating Cross-Modal Alignment. Following Gröger et al. (2026a), we evaluate CMA using the mutual k-nearest neighbors (mkNN) metric. Given a set of paired observations $\mathcal{S} = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{E}_X \times \mathcal{E}_Y$, let $N_X^k(x)$ and $N_Y^k(y)$ denote the sets of the k nearest neighbors of x and y within their respective embedding spaces. We define

$$\mathrm{m}k\mathrm{NN}(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|N_X^k(x_i) \cap N_Y^k(y_i)\right|}{k}. \tag{4}$$

As shown in Fig. 2b, mkNN can be viewed as a discrete approximation of the CKA-RBF considered in Sec. 3, making it a natural choice for evaluating CMA in our experiments. Unless stated otherwise, we use $k = 50$. Additional results using different values of k, and CKNNA are provided in App. G, and exhibit the same qualitative trends.

4.1 Direct Validation of the Theory

In this section, we isolate the effect of unimodal Aug-SSL training by controlling the model architectures and training setup. We train vision and text encoders from scratch using a ViT (Dosovitskiy et al., 2020) for vision and an SBERT architecture (Reimers & Gurevych, 2019) for text, initialized from a pretrained BERT model (Devlin et al., 2019).

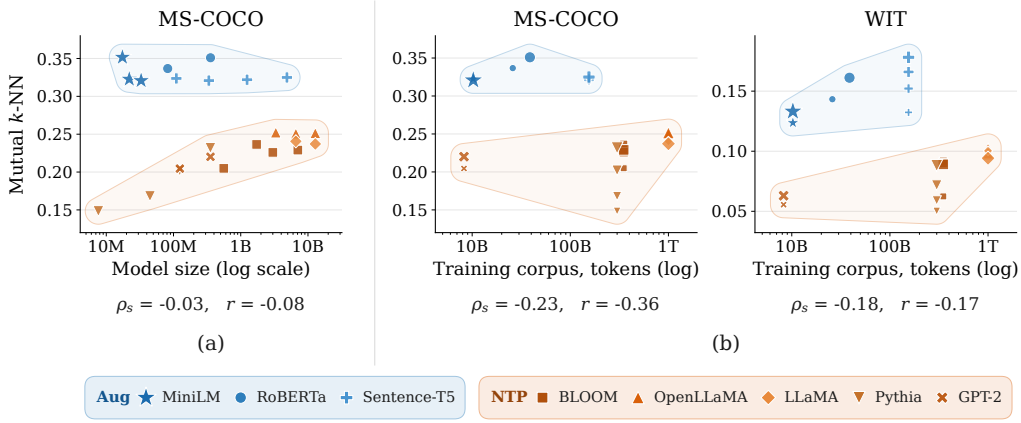


Figure 4: **Alignment by model/corpus size.** CMA between text models and the DINOv2 vision encoder, using augmentation-based (Aug) and next-token prediction-based (NTP) models. Shown as a function of **(a) model size** (number of parameters), and **(b) corpus size** (amount of tokens the model was trained on). Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

Isolating the InfoNCE effect. To empirically validate Thm. 1 and 2, we train two vision and two text encoders on COCO as described above, with one model per modality minimizing and the other maximizing the unimodal Aug-SSL loss. We use the InfoNCE objective from Sec. 3. For vision, we employ standard image augmentations (e.g., random crops, horizontal flips). For text, we use an LLM (Qwen2.5 (Qwen et al., 2025)) to generate five rephrased versions of each text.

Fig. 3a presents the resulting CMA for all pairings of the four encoders. The pairing of the two loss-minimized encoders corresponds to Thm. 1: minimizing the unimodal Aug-SSL objectives leads to increasing CMA. In contrast, every pairing involving a loss-maximized encoder corresponds to Thm. 2, demonstrating that a high unimodal loss in either modality prevents cross-modal alignment, with CMA approaching the chance level. Training on WIT shows the same trends (App. D).

Aug-SSL Outperforms Explicit Supervision. To isolate the contribution of unimodal augmentations to CMA, we compare augmentation-based SSL with supervised learning on the COCO dataset. Specifically, we consider two supervised objectives: Supervised Contrastive Learning (SupCon) (Khosla et al., 2020) and cross-entropy classification, trained using the labels in COCO. In these experiments, we freeze the pretrained text encoder (BERT/SBERT) and train only the vision encoder.

Fig. 3b presents the resulting CMA. Most notably, the Aug-SSL-trained vision encoder achieves the highest CMA with both text encoders. Its alignment with SBERT, which is an augmentation-based model, is consistently stronger than with BERT, further supporting that augmentation-based training promotes alignment. Importantly, however, the Aug-SSL vision encoder also achieves stronger CMA with BERT than the supervised baselines. Thus, the effect cannot be attributed simply to matching the training paradigm across modalities: augmentation-based training promotes CMA even with a text encoder trained using a different objective. These results further support the role of augmentations as a driver of implicit CMA.

4.2 Universal Emergence in Foundation Models

We extend our analysis to unconstrained, frozen pre-trained models from diverse model kinds and modalities. We show that augmentations implicitly induce CMA across architectures, training datasets, and model scales.

Models. We consider 53 pretrained unimodal encoders across text, vision, and audio. Specifically, we use eight families of 24 text encoders from two training paradigms: augmentation-based models, including MiniLM (Wang et al., 2020), RoBERTa (Reimers & Gurevych, 2019), and Sentence-T5

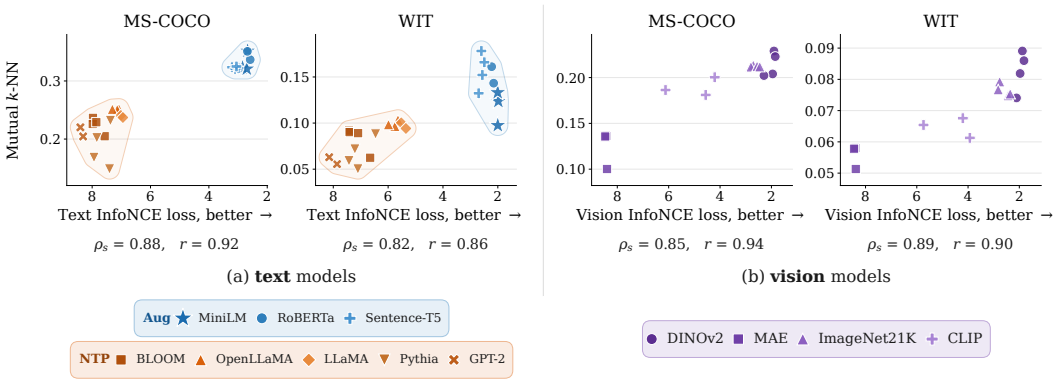


Figure 5: **Alignment by unimodal loss.** Image-text CMA between text models and DINOv2, and vision models and BLOOM as a function of unimodal **(a) text InfoNCE loss**, and **(b) image InfoNCE loss**. In contrast to model or corpus size, the unimodal Aug-SSL loss serves as a robust predictor of CMA. It cleanly separates low- and high-alignment models, exhibiting strong Spearman (ρ_s) and Pearson (r) correlations.

Table 1: **Image-text CMA is highly correlated with unimodal InfoNCE loss.** Spearman (ρ_s) and Pearson (r) correlations between mkNN and unimodal properties over 24/14 text/vision models. Results are mean $\pm$ std over all second modality models. Notably, CMA is consistently highly correlated with the negative text and vision unimodal InfoNCE loss.

	MS-COCO		WIT	
ρ_s / r	Text models	Vision models	Text models	Vision models
Model size	$-0.05_{\pm 0.02}$ / $-0.14_{\pm 0.06}$	$-0.23_{\pm 0.06}$ / $-0.24_{\pm 0.01}$	$0.09_{\pm 0.02}$ / $0.08_{\pm 0.04}$	$-0.25_{\pm 0.12}$ / $-0.28_{\pm 0.10}$
Training-corpus size	$-0.32_{\pm 0.04}$ / $-0.43_{\pm 0.04}$	$0.20_{\pm 0.05}$ / $0.58_{\pm 0.03}$	$-0.18_{\pm 0.01}$ / $-0.21_{\pm 0.03}$	$0.34_{\pm 0.06}$ / $0.45_{\pm 0.16}$
InfoNCE loss ($-\mathcal{L}$)	$\mathbf{0.85}_{\pm 0.03}$ / $\mathbf{0.94}_{\pm 0.03}$	$\mathbf{0.70}_{\pm 0.14}$ / $\mathbf{0.92}_{\pm 0.02}$	$\mathbf{0.84}_{\pm 0.02}$ / $\mathbf{0.88}_{\pm 0.02}$	$\mathbf{0.86}_{\pm 0.09}$ / $\mathbf{0.90}_{\pm 0.03}$

(Ni et al., 2022); and next-token prediction (NTP)-based models, including BLOOM (Le Scao et al., 2022), OpenLLaMA (Geng & Liu, 2023), LLaMA (Touvron et al., 2023), Pythia (Biderman et al., 2023), and GPT-2 (Radford et al., 2019). Also, we consider four families of 14 vision encoders: DINOv2 (Oquab et al., 2023), MAE (He et al., 2022), ImageNet21K (Russakovsky et al., 2015), and CLIP (Radford et al., 2021). For each family, we evaluate multiple models of different sizes. We compare the text models with the DINOv2 vision encoder and the BYOL-A v2 audio encoder (Niizumi et al., 2022), with the other vision encoders and a total of 15 audio encoders shown in App. E and App. F. Further details regarding the models are provided in App. L.1.

Alignment vs. model/corpus size. Fig. 4a (and 1) shows the CMA of text pre-trained models to DINOv2 as a function of model size. Similarly, Fig. 4b shows it as a function of the corpus size the model was trained on. Interestingly, augmentation-based models consistently achieve higher CMA than NTP-based models of comparable size, and often outperform even substantially larger NTP models, trained with significantly more data. This provides evidence that unimodal augmentations play a key role in inducing CMA. Indicated by both Pearson and Spearman correlations, when considering different model families jointly, model or corpus sizes are uncorrelated with CMA.

Unimodal InfoNCE predicts CMA. Fig. 5 and Tab. 1 show the CMA as a function of the unimodal InfoNCE loss. We apply the same standard image and LLM-based text augmentations described in Sec. 4.1 post-hoc to the frozen pre-trained models and evaluate their unimodal InfoNCE loss. Interestingly, lower *unimodal InfoNCE* loss is highly correlated with higher CMA, indicated by both high

Table 2: **Audio-text CMA is highly correlated with unimodal InfoNCE loss.** Spearman (ρ_s) and Pearson (r) correlations between mkNN and text unimodal properties over 24 text models, averaged $\pm$ std over 15 audio models. Notably, CMA is consistently highly correlated with the negative unimodal InfoNCE loss.

ρ_s / r	Clotho	AudioCaps
Model size	$0.05_{\pm 0.02}$ / $0.18_{\pm 0.06}$	$-0.08_{\pm 0.05}$ / $-0.21_{\pm 0.03}$
Training-corpus size	$-0.19_{\pm 0.01}$ / $-0.10_{\pm 0.04}$	$-0.22_{\pm 0.04}$ / $-0.30_{\pm 0.04}$
InfoNCE loss ($-\mathcal{L}$)	$\mathbf{0.92}_{\pm 0.01}$ / $\mathbf{0.85}_{\pm 0.03}$	$\mathbf{0.80}_{\pm 0.03}$ / $\mathbf{0.97}_{\pm 0.01}$

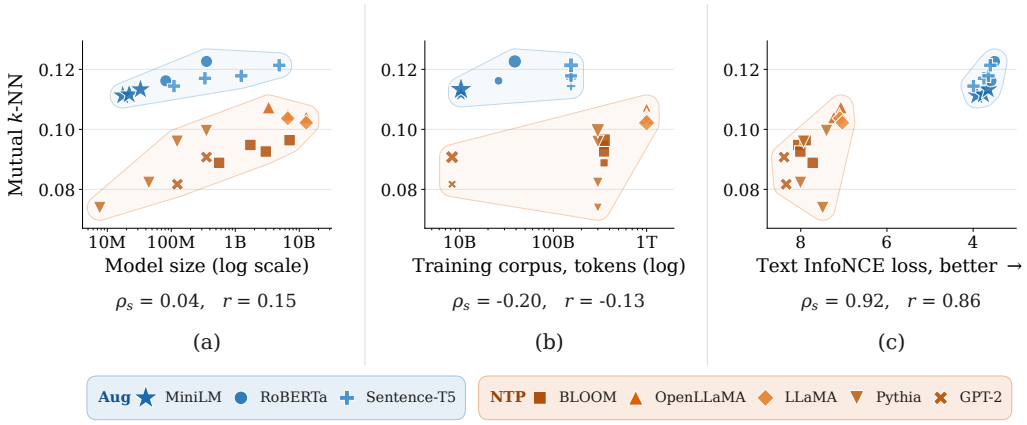


Figure 6: **Audio-text results.** CMA between text models and the BYOL-A audio encoder as a function of **(a) model size** (number of parameters), **(b) corpus size** (amount of tokens the model was trained on), and **(c) unimodal text InfoNCE loss**. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, while the unimodal loss is highly correlated, indicated by both Spearman (ρ_s) and Pearson (r) correlations.

Spearman and Pearson correlations. This further validates the theory presented in Sec. 3, showing that InfoNCE can serve as a unimodal predictor of cross-modal alignment.

Beyond image-text. Fig. 6 and Tab. 2 show that the same trends observed in vision-text hold robustly to audio-text. Specifically, model/corpus sizes are uncorrelated with CMA, showing augmentation-based models experience higher alignment nonetheless. Also, the unimodal InfoNCE loss is a strong predictor of the cross-modal alignment. This provides further evidence for the universality of these findings across modalities.

Ablations. Due to lack of space, we defer ablations to the appendix. Specifically, we show the results are consistent across different kinds of text augmentations (App. D), other vision encoders (App. E), other audio encoders (App. F), other CMA metrics (App. G), different test set sizes (App. H), and different choices of temperature τ (App. I).

Message of Section 4: Across controlled training, frozen pre-trained encoders, and modalities, unimodal augmentation-based SSL loss predicts cross-modal alignment.

5 Conclusions and Summary

In this paper, we presented an explanation for the emergence of implicit cross-modal alignment through unimodal augmentation-based objectives. We established theoretical connections between the widely used InfoNCE loss and cross-modal alignment, as measured by local CKA. In particular, we showed that minimizing Aug-SSL unimodal objectives promotes high cross-modal alignment (Thm. 1), while sufficiently high unimodal loss restricts cross-modal alignment (Thm. 2). We then empirically validated these results in both controlled experiments (Sec. 4.1) and unconstrained pre-trained models (Sec. 4.2), demonstrating that unimodal augmentations are a strong driving force behind cross-modal alignment.

We believe this work provides a foundation for understanding the emergence of CMA from both theoretical and empirical perspectives. While Aug-SSL is commonly motivated by its ability to leverage large-scale data, our results reveal an additional, unexplored benefit: augmentation-based objectives can induce cross-modal alignment implicitly, and much more efficiently, without explicitly incorporating any cross-modal information into the training objective.

AI USE STATEMENT

Generative AI tools assisted in the development of the mathematical claims, proofs, and code presented in this work. All AI-generated outputs were subjected to extensive human oversight: mathematical content and proofs were manually verified line-by-line and rewritten, while all code was comprehensively reviewed and tested for correctness. The authors take full and final responsibility for the integrity of all text, claims, derivations, and code contained in this manuscript.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our theoretical and empirical results. All theoretical claims, including Theorems 1 and 2, Propositions 1 and 2, and Lemmas 1-11, are accompanied by complete proofs and derivations in App. A, with the underlying assumptions stated explicitly. The experimental setup is described at a high level in Sec. 4 and in complete, hyperparameter-level detail in App. L. All model implementations, including the source repositories, pre-trained checkpoints used, supervised baselines, and alignment metrics, are documented in App. L.1. Dataset sources, preprocessing pipelines, and text augmentation prompts are detailed in App. L.2. Hardware and software infrastructure used for all experiments are described in App. L.4. The code for all experiments will be made publicly available upon acceptance.

REFERENCES

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images. *arXiv preprint arXiv:2505.17101*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Kiril Bangachev, Guy Bresler, and Yury Polyanskiy. Representation alignment rests on linear structure. *arXiv preprint arXiv:2605.28870*, 2026.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Kshitij Duraphe, Michael J Smith, Shashwat Sourav, John F Wu, et al. The platonic universe: Do foundation models see the same sky? *arXiv preprint arXiv:2509.19453*, 2025.

Sathya Edamadaka, Soojung Yang, Ju Li, and Rafael Gómez-Bombarelli. Universally converging representations of matter across scientific foundation models. *arXiv preprint arXiv:2512.03750*, 2025.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–13. IEEE, 2025.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL `https://github.com/openlm-research/open_llama`.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026a.

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

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *International Conference on Learning Representations*, volume 2026, pp. 73970–74025, 2026.

Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, Atabak Ashfaq, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, et al. Phi-3 safety post-training: Aligning language models with a" break-fix" cycle. *arXiv preprint arXiv:2407.13833*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35645–35660, 2025.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in neural information processing systems*, 35:28708–28720, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Francesco Maria Gabriele Insulla, Shuo Huang, and Lorenzo Rosasco. Towards a learning theory of representation alignment. In *International Conference on Learning Representations*, volume 2025, pp. 37594–37608, 2025.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John Morris. Harnessing the universal geometry of embeddings. *Advances in Neural Information Processing Systems*, 38:45963–45987, 2026.

Chaitanya Kapoor, Sudhanshu Srivastava, and Meenakshi Khosla. Bridging critical gaps in convergent learning: How representational alignment evolves across layers, training, and distribution shifts. *Advances in Neural Information Processing Systems*, 38:140352–140366, 2026.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2022.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024.

Zeyu Michael Li, Hung Anh Vu, Damilola Awofisayo, and Emily Wenger. Causes and consequences of representational similarity in machine learning models. *arXiv preprint arXiv:2505.13899*, 2025.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, pp. 1–11, 2026.

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*, pp. 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Alexander Lobashev. An information-geometric view of the platonic hypothesis. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. The indra representation hypothesis for multimodal alignment. *Advances in Neural Information Processing Systems*, 38:156980–157000, 2026.

JW Miller. Testing the platonic representation hypothesis via representation-controlled tokenization. 2026.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

Jerry Ngo and Yoon Kim. What do language models hear? probing for auditory representations in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5435–5448, 2024.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the association for computational linguistics: ACL 2022*, pp. 1864–1874, 2022.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Exploring pre-trained general-purpose audio representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:137–151, 2022.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maks Ovsjanikov, Viorica Patraucean, Leonidas Guibas, Tyler Zhu, and Tengda Han. Dynamic reflections: Probing video representations with text alignment. In *International Conference on Learning Representations*, volume 2026, pp. 71224–71250, 2026.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6960–6970. IEEE, 2021.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pp. 3982–3992, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3875–3879. IEEE, 2021.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It'sa (blind) match! towards vision-language correspondence without parallel data. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24983–24992. IEEE, 2025.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36:48382–48402, 2023.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. *arXiv preprint arXiv:2502.16282*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ali Hussaini Umar and Alessandro Laio. Signal-to-noise ratio and sample size govern representational alignment in neural networks. *arXiv preprint arXiv:2605.26973*, 2026.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in neural information processing systems*, 31, 2018.

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. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. *Advances in Neural Information Processing Systems*, 38:46634–46666, 2026.

Lauren Hyoseo Yoon, Yisong Yue, and Been Kim. Escaping plato's cave: Jam for aligning independently trained vision and language models. *arXiv preprint arXiv:2507.01201*, 2025.

Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14604–14614. IEEE, 2025.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.

Liu Ziyin, Yizhou Xu, and Isaac Chuang. Neural thermodynamics: Entropic forces in deep and universal representation learning. *Advances in Neural Information Processing Systems*, 38:56057–56089, 2026.

A Proofs

A.1 Rationale

This section gives an informal overview of the argument: the objects it involves, the role played by each lemma, and how they chain into the two theorems. The formal settings are given in Sec. A.2, and the formal statements and proofs follow in Sec. A.3 and Sec. A.4.

Overview. The centered kernel k_X^c (Eq. 6) lies in L^2 with norm σ_X, and Eq. 7 expresses CKA as the cosine of the angle between k_X^c and k_Y^c. Since the encoders never interact directly, alignment arises through their shared semantic structure: each encoder is trained against its own augmentations, which act as view-specific information removers by filtering out view-specific variation. Minimizing the unimodal losses therefore steers each embedding space toward the shared semantics underlying the modalities, promoting cross-modal similarity. To formalize this connection, we introduce the augmentation gap Δ_X, which measures the difference between a sample's similarity to its augmentations and its similarity to random samples. The proofs of both theorems follow chains of inequalities, bounding the loss with the gap, and the gap with the CKA.

Theorem 1: low losses force alignment. To prove Thm. 1, we bound the angle between each kernel and the shared semantics separately, quantifying the fraction of the encoder structure explained by the shared semantics, and then combine the two bounds via the triangle inequality. In details, Lemma 1 involves the basic components underlying both loss and the CKA: the inner products between a sample and its augmentations, and between a sample and independently sampled points, both of which are bounded in terms of the loss. Lemma 2 then lower-bounds the augmentation gap by the loss, while Lemma 3 relates the augmentation gap to the fraction of the encoder structure explained by the semantics (induced by the augmentations). Lemma 4 shows that the kernel variance, which appears in the denominator of CKA, is upper-bounded in terms of the loss. Finally, Lemma 9 and Prop. 1 establish that the bound in Thm. 1 is non-increasing in both losses and attains 1 when both reach their minima.

Theorem 2: a high loss prevents alignment. To prove Thm. 2, we bound the fraction of each encoder's structure that is able to carry cross-modal alignment at all, and then bound that fraction by the loss. In details, Lemma 5 shows that conditioned on the shared semantics the two embedding spaces are independent, so only the semantic part of each kernel contributes to the CKA. It therefore upper-bounds the CKA by the semantic variance β_X, measuring how much the representation varies with the semantics, normalized by the kernel variance. Lemma 6 then relates the semantic variance to the augmentation gap, up to a term measuring how much the augmentations displace the observations, which is where the Lipschitz assumption (Assumption 1) is used. Lemma 7 completes

the chain by upper-bounding the augmentation gap in terms of the loss, complementing Lemma 2. Finally, Prop. 2 establishes that the bound in Thm. 2 is strictly decreasing in the loss, and since it involves a single modality, a high loss in either one alone caps the CKA.

A.2 SETTING

Every statement made for X holds verbatim for Y.

Data. We fix a probability measure μ on the semantic manifold $\mathcal{M}$, and nuisance laws $\nu_{\mathcal{U}}$ on $\mathcal{U}$ and $\nu_{\mathcal{V}}$ on $\mathcal{V}$. Observations are $g_X(m,u) \in \mathcal{X}$ and $g_Y(m,v) \in \mathcal{Y}$, and the trained encoders send them to the unit spheres of the embedding spaces $\mathcal{E}_X, \mathcal{E}_Y$,

$$x = e_X\big(g_X(m,u)\big) \in \mathcal{E}_X \subseteq \mathbb{S}^{d_X-1}, \qquad y = e_Y\big(g_Y(m,v)\big) \in \mathcal{E}_Y \subseteq \mathbb{S}^{d_Y-1}.$$

Note that for readability, the appendix refers to $x \in \mathcal{E}_X$ and $y \in \mathcal{E}_Y$ as embedding points rather than observations. We write ρ_X for the law of x and $\bar{x} = \mathbb{E}[x]$. Throughout, (m,u,v) and (m',u',v') are two independent draws, and all four of x, x', y, y' are built from them, so that x, y share the semantics m and x', y' share m'.

Every object below is a square-integrable function of the pair $\big((m,u,v),(m',u',v')\big)$. We write L^2 for the space of such functions, with $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ its inner product and norm.

Kernel. Fix $\omega > 0$. On the sphere $\|x - x'\|^2 = 2 - 2\langle x, x'\rangle$, so the Gaussian kernel of bandwidth $\omega^{-1/2}$ is exactly

$$k(x,x') = e^{\omega(\langle x,x'\rangle - 1)}. \tag{5}$$

We abbreviate $k_X = k(x,x')$ and $k_Y = k(y,y')$, and use ω rather than a bandwidth symbol so that nothing collides with the kernel scales σ_X, σ_Y. The centred kernel and the alignment are

$$k_X^c = k_X - \mathbb{E}[k_X \mid x] - \mathbb{E}[k_X \mid x'] + \mathbb{E}[k_X], \qquad \sigma_X^2 = \mathbb{E}\big[(k_X^c)^2\big], \tag{6}$$

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) = \frac{\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_Y)}{\sqrt{\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_X)\mathrm{HSIC}(\mathcal{E}_Y, \mathcal{E}_Y)}} = \frac{\mathbb{E}[k_X^c k_Y^c]}{\sigma_X \sigma_Y} = \langle \tfrac{k_X^c}{\sigma_X}, \tfrac{k_Y^c}{\sigma_Y} \rangle, \tag{7}$$

the latter defined whenever $\sigma_X \sigma_Y > 0$. Equation 7 exhibits the CKA as the cosine of an angle in L^2 between two unit vectors; both directions exploit this.

Three scalars.

$$\hat{A}_X := \mathbb{E}\big[k(x,x^+)\big], \qquad \hat{S}_X := \mathbb{E}\big[k(x,x')\big], \qquad \Delta_X := \hat{A}_X - \hat{S}_X. \tag{8}$$

$\hat{A}_X$ is how similar a sample looks to its own augmentation, $\hat{S}_X$ how similar two unrelated samples look, and the *augmentation gap* Δ_X the excess of the former over the latter.

Definition 1 (Augmentation)**.** An *augmentation* redraws both the semantics and the nuisance, while keeping the semantics close. That is,

$$x^+ = e_X\big(g_X(m^+, u^+)\big), \qquad m^+ \sim P(m,\cdot), \quad u^+ \sim \nu_{\mathcal{U}} \text{ independent}, \tag{9}$$

where $P(m,\cdot) = \mathsf{J}(m,\cdot)\,\mu$ is a Markov kernel on $\mathcal{M}$ with density J.

Write $\pi(m,m^+) = \mathsf{J}(m,m^+)\mu(m)\mu(m^+)$ for the joint law of the semantic pair.

The *concentration strength* of the augmentation is

$$\chi := \|\mathsf{J} - 1\|_{L^2(\mu\otimes\mu)}. \tag{10}$$

We assume $\mathsf{J} \in L^2(\mu\otimes\mu)$ and $\chi > 0$; the degenerate case $\chi = 0$ means $\mathsf{J} \equiv 1$, i.e. the augmentation resamples the semantics independently and carries no signal. Both marginals of π are μ, so x and x^+ each have marginal law ρ_X, a fact used repeatedly below.

The objective. With $0 < \tau \leq 1$ the temperature, the population InfoNCE loss and its normalisation are

$$N_X = -\frac{A_X}{\tau} + U_X, \qquad A_X := \mathbb{E}\big[\langle x, x^+\rangle\big], \qquad U_X := \mathbb{E}_x\Big[\log \mathbb{E}_{x'}\big[e^{\langle x,x'\rangle/\tau}\big]\Big], \tag{11}$$

$$L_X = \frac{N_X + 1/\tau}{R}, \qquad R = \tfrac{2}{\tau} + \log\cosh \tfrac{1}{\tau}. \tag{12}$$

The derivations often use

$$\tau R = 2 + \tau \log\cosh \tfrac{1}{\tau} \;>\; 2. \tag{13}$$

The constants of Section A.3 are

$$a := e^{-\omega}\sinh\omega \in \big(0, \tfrac{1}{2}\big), \quad b := \big(\omega + e^{-\omega}\sinh\omega\big)\tau R, \quad D := \big(1 - e^{-\omega}\big)^2\Big(1 + \tfrac{4}{\log 2}\Big)\tau R. \tag{14}$$

A.3 Implicit Alignment

This section proves the implicit alignment theorem. That is, small unimodal losses restrict the CKA to be large (Thm. 1). The key is the augmentations, which are accessed via $J - 1$, which depends only on the shared semantics and is therefore common to both modalities. Since the CKA is the cosine of the angle between the centered kernels k_X^c and k_Y^c (Eq. 7), it suffices to show that each kernel is close in angle to $J - 1$. This angle is determined by two quantities: the inner product $\langle k_X^c, J - 1\rangle$, measuring how much of k_X^c points in the direction of $J - 1$, and the length of k_X^c, namely the kernel scale σ_X. Lemma 3 shows that the inner product equals the augmentation gap Δ_X, i.e., how much more similar a sample is to its augmentation than to a random sample. Lemma 2 shows that a low loss forces this gap to be large, while Lemma 4 shows that it forces σ_X to be small, both using the moment bounds of Lemma 1. Together, a low loss forces k_X^c to align with $J-1$, and likewise for k_Y^c. Thm. 1 then combines the two angles via the triangle inequality, forcing the two kernels to align with each other.

Lemma 1 (Moments)**.**

$$\|\bar{x}\|^2 \leq \tau R\, L_X, \qquad \mathbb{E}\big[\langle x, x'\rangle^2\big] \leq \Big(1 + \tfrac{4}{\log 2}\Big)\tau R\, L_X, \qquad A_X \;\geq\; 1 - \tau R\, L_X.$$

Proof. Write $r(x) = \mathbb{E}[\langle x, x'\rangle \mid x] = \langle x, \bar{x}\rangle$ and $V(x) = \mathrm{Var}(\langle x, x'\rangle \mid x)$. Conditionally on x, the variable $\delta = \langle x, x'\rangle - r(x)$ has mean zero and lies in an interval of length 2, so Lemma 11 with $\lambda = 1/\tau \geq 1$ gives $\log \mathbb{E}[e^{\delta/\tau} \mid x] \geq \frac{\log 2}{4\tau} V(x)$. Averaging and using $\mathbb{E}[r(x)] = \|\bar{x}\|^2$,

$$U_X = \frac{\mathbb{E}[r(x)]}{\tau} + \mathbb{E}\left[\log \mathbb{E}[e^{\delta/\tau} \mid x]\right] \geq \;\frac{\|\bar{x}\|^2}{\tau} + \frac{\log 2}{4\tau}\,\mathbb{E}[V] \;\geq\; 0. \tag{15}$$

Since $A_X \leq 1$, Eq. 11 gives $U_X = N_X + A_X/\tau \leq N_X + 1/\tau = R\, L_X$. Both terms on the last inequality of Eq. 15 are nonnegative, so each is at most RL_X: this is the first bound, and $\mathbb{E}[V] \leq \frac{4}{\log 2}\tau R L_X$. The second bound follows from $\mathbb{E}[\langle x, x'\rangle^2] = \mathbb{E}[V] + \mathbb{E}[r(x)^2]$ and $r(x)^2 \leq \|\bar{x}\|^2 \leq \tau R L_X$. For the third, $U_X \geq 0$ in Eq. 11 gives $A_X/\tau = U_X - N_X \geq -N_X = \frac{1}{\tau} - R\, L_X$. □

Lemma 2 (The loss controls the gap from below)**.** $\Delta_X \;\geq\; a - b\, L_X$*, with* a, b *as in Eq. 14.*

Proof. For $\hat{A}_X$,

$$\hat{A}_X = \mathbb{E}\big[e^{\omega(\langle x,x^+\rangle - 1)}\big] \;\geq\; e^{\omega(A_X - 1)} \;\geq\; e^{-\omega\tau R L_X} \;\geq\; 1 - \omega\tau R\, L_X,$$

where the first inequality is Jensen, then the third bound of Lemma 1, then $e^{-t} \geq 1 - t$.

For $\hat{S}_X = e^{-\omega}\mathbb{E}[e^{\omega\langle x,x'\rangle}]$, Lemma 10(i) with $\alpha = \omega$, $c = \mathbb{E}[\langle x, x'\rangle] = \|\bar{x}\|^2$

$$\hat{S}_X \leq e^{-\omega}\left(\cosh\omega + \|\bar{x}\|^2 \sinh\omega\right).$$

Applying the first bound of Lemma 1 gives

$$\hat{S}_X \;\leq\; e^{-\omega}\cosh\omega + e^{-\omega}\tau R L_X \sinh\omega.$$

Subtracting and using $1 - e^{-\omega}\cosh\omega = \frac{1}{2}(1 - e^{-2\omega}) = e^{-\omega}\sinh\omega$ gives $\Delta_X \geq a - b\, L_X$. □

Lemma 3 (Centered kernel and the augmentation gap). $\langle k_X^c, \mathsf{J}-1\rangle = \Delta_X$, *and* $\|\mathsf{J}-1\| = \chi$.

Proof. $\mathsf{J}-1$ is a function of (m,m') alone, so for any $f \in L^2$, writing $\bar{f}(m,m') = \mathbb{E}[f \mid m,m']$ and using $\mathbb{E}_{\mu\otimes\mu}[\bar{f}\,\mathsf{J}] = \mathbb{E}_\pi[\bar{f}]$,

$$\langle f, \mathsf{J}-1\rangle = \mathbb{E}_{\mu\otimes\mu}\big[\bar{f}\,(\mathsf{J}-1)\big] = \mathbb{E}_\pi[f] - \mathbb{E}[f]. \tag{16}$$

Apply Eq. 16 to $f = k_X^c$. By centering Eq. 6, $\mathbb{E}[k_X^c] = 0$. For the π-expectation,

$$\mathbb{E}_\pi[k_X^c] = \mathbb{E}_\pi[k_X] - \mathbb{E}_\pi\mathbb{E}[k_X \mid x] - \mathbb{E}_\pi\mathbb{E}[k_X \mid x^+] + \mathbb{E}_\pi\mathbb{E}[k_X].$$

Because the marginals of π are μ, $\mathbb{E}_\pi\mathbb{E}[k_X \mid x] = \mathbb{E}_\pi\mathbb{E}[k_X \mid x^+] = \mathbb{E}[k_X]$, hence

$$\mathbb{E}_\pi[k_X^c] = \hat{A}_X - \hat{S}_X - \hat{S}_X + \hat{S}_X = \Delta_X,$$

and Eq. 16 gives the claim.

The norm is Eq. 10: the marginal of L^2 on (m,m') is $\mu\otimes\mu$, so embedding a function of (m,m') into L^2 preserves its norm. □

Lemma 4 (Kernel scale). $\sigma_X^2 \;\le\; D\,L_X$, *with* D *as in Eq. 14.*

Proof. Let $Cg = g - \mathbb{E}[g \mid x] - \mathbb{E}[g \mid x'] + \mathbb{E}[g]$, so that $k_X^c = Ck_X$. Since $x \perp x'$ we may factor $C = (\mathrm{Id} - P_1)(\mathrm{Id} - P_2)$ with $P_1 g = \mathbb{E}[g \mid x']$ and $P_2 g = \mathbb{E}[g \mid x]$; each P_i is an orthogonal projection, so each $\mathrm{Id} - P_i$ is a contraction and $\|Cg\| \le \|g\|$. As $Cc = 0$ for every constant c, $\sigma_X = \|k_X^c\| = \|C(k_X - c)\| \le \|k_X - c\|$ for every constant c.

Take $c = e^{-\omega}$. By Eq. 5, $k_X - e^{-\omega} = e^{-\omega}\big(e^{\omega\langle x,x'\rangle} - 1\big)$, and Lemma 10(ii) with $\alpha = \omega$ gives $|k_X - e^{-\omega}| = e^{-\omega}|\big(e^{\omega\langle x,x'\rangle} - 1\big)| \le (1 - e^{-\omega})|\langle x, x'\rangle|$. Squaring and applying the second bound of Lemma 1 gives $\sigma_X^2 \le D\,L_X$. □

Theorem 1 (Implicit alignment). *Suppose* $\sigma_X\sigma_Y > 0$ *and* $\chi > 0$. *Define, with* a, b, D *as in Eq. 14,*

$$c(L) \;:=\; \frac{\big[\,a - bL\,\big]_+}{\chi\sqrt{D\,L}}, \qquad c_X = c(L_X), \quad c_Y = c(L_Y). \tag{17}$$

Then $c_X, c_Y \in [0,1]$ *and*

$$\boxed{\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\ge\; c_X c_Y - \sqrt{(1-c_X^2)(1-c_Y^2)} \;=:\; f(L_X, L_Y)}$$

Proof. Lemma 4 gives $L_X \ge \sigma_X^2/D > 0$ because $\sigma_X > 0$, so $c(L_X)$ is well defined; likewise $c(L_Y)$.

If $c_X = 0$ or $c_Y = 0$ the claimed right-hand side is ≤ 0, and the bound follows from $\mathrm{CKA} \ge 0$ (see Eq. 7). Assume therefore $a - bL_X > 0$ and $a - bL_Y > 0$, and work in L^2 with the three unit vectors

$$\hat{a} = \frac{k_X^c}{\sigma_X}, \qquad \hat{b} = \frac{k_Y^c}{\sigma_Y}, \qquad \hat{\jmath} = \frac{\mathsf{J}-1}{\chi},$$

and let $\theta_X = \angle(\hat{a}, \hat{\jmath})$, $\theta_Y = \angle(\hat{b}, \hat{\jmath})$. Lemma 3 evaluates the cosine, Lemma 2 bounds its numerator below by a positive quantity, and Lemma 4 bounds its denominator above:

$$\cos\theta_X = \frac{\langle k_X^c, \mathsf{J}-1\rangle}{\chi\,\sigma_X} = \frac{\Delta_X}{\chi\,\sigma_X} \;\ge\; \frac{a - bL_X}{\chi\sqrt{D\,L_X}} = c_X \;>\; 0, \tag{18}$$

and likewise $\cos\theta_Y \ge c_Y > 0$. In particular $c_X \le \cos\theta_X \le 1$, so $c_X \in (0,1]$ and $\theta_X \le \arccos c_X \le \pi/2$; the same for Y.

The angle $\angle(\cdot,\cdot) = \arccos\langle\cdot,\cdot\rangle$ is the geodesic distance on the unit sphere of L^2, hence satisfies the triangle inequality:

$$\angle(\hat{a}, \hat{b}) \;\le\; \angle(\hat{a}, \hat{\jmath}) + \angle(\hat{b}, \hat{\jmath}) = \theta_X + \theta_Y \;\le\; \arccos c_X + \arccos c_Y \;\le\; \pi.$$

Apply $\cos$, decreasing on $[0,\pi]$, and expand $\cos(\arccos c_X + \arccos c_Y) = c_X c_Y - \sqrt{(1-c_X^2)(1-c_Y^2)}$. Noting $\cos\angle(\hat{a},\hat{b}) = \mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y)$ gives

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\ge\; c_X c_Y - \sqrt{(1-c_X^2)(1-c_Y^2)}$$

□

A.4 IMPLICIT MISALIGNMENT

This section gives the complement direction and proves the implicit misalignment theorem (Thm. 2). That is, a large unimodal loss restricts the CKA to be small. The key observation is that only semantic variation can align across modalities. Lemma 5 formalizes this: given the shared semantics, the two embeddings are independent, so only the semantic part of each kernel contributes to the CKA. Hence, the CKA is upper-bounded by the semantic variance β_X, i.e., how much the embedding varies with the semantics, relative to its overall variation σ_X. It then remains to show that a large loss forces β_X to be small. Lemma 6 bounds the semantic variance by the augmentation gap Δ_X, up to how much the augmentations displace the observations, which is where the Lipschitz assumption (Assumption 1) is used. Lemma 7 shows that a large loss forces this gap to be small. Together, a large loss leaves much of the variation view-specific, and Thm. 2 chains these inequalities into an upper bound on the CKA.

A.4.1 CKA AND THE SEMANTIC VARIANCE

Notations. Let $\phi(x) = k(x, \cdot)$ be the canonical feature map of the Gaussian kernel (Eq. 5) into its RKHS $\mathcal{H}$ (Muandet et al., 2017), so that $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x')$ and $\|\phi(x)\|_{\mathcal{H}} = 1$ on the sphere. Put

$$\mu_m := \mathbb{E}[\phi(x) \mid m], \qquad \bar{\mu} := \mathbb{E}[\phi(x)], \qquad \delta_m := \mu_m - \bar{\mu}, \tag{19}$$

and define the *semantic kernel* and the *semantic variance* induced by X

$$h_X(m, m') := \mathbb{E}\big[k_X^c \mid m, m'\big], \qquad \beta_X := \mathbb{E}_{m \sim \mu}\big[h_X(m, m)\big]. \tag{20}$$

Lemma 5. *$h_X(m, m') = \langle \delta_m, \delta_{m'} \rangle_{\mathcal{H}}$ and $\beta_X = \mathbb{E}_m \|\delta_m\|_{\mathcal{H}}^2 \geq 0$. Moreover, writing $s_X := \|h_X\|$,*

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\leq\; \frac{s_X}{\sigma_X} \;\leq\; \frac{\beta_X}{\sigma_X}.$$

Proof. Expanding Eq. 6 with $\mathbb{E}[k_X \mid x] = \langle \phi(x), \bar{\mu} \rangle_{\mathcal{H}}$ gives the pointwise identity $k_X^c = \langle \phi(x) - \bar{\mu}, \phi(x') - \bar{\mu} \rangle_{\mathcal{H}}$. Given (m, m') the nuisances u, u' are independent, so taking conditional expectations inside the inner product yields $h_X(m, m') = \langle \delta_m, \delta_{m'} \rangle_{\mathcal{H}}$; in particular $h_X(m, m) = \|\delta_m\|_{\mathcal{H}}^2$ and $\beta_X = \mathbb{E}_m \|\delta_m\|_{\mathcal{H}}^2$.

Because (u, u') and (v, v') are independent of each other and of (m, m'), the pairs (x, x') and (y, y') are conditionally independent given (m, m'), so $\mathbb{E}[k_X^c k_Y^c \mid m, m'] = h_X h_Y$ and $\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_Y) = \langle h_X, h_Y \rangle$. Now h_X is the orthogonal projection of k_X^c onto the functions of (m, m'), hence $s_X \leq \sigma_X$ and likewise $s_Y \leq \sigma_Y$. Cauchy–Schwarz in L^2 gives $\mathrm{HSIC}(\mathcal{E}_X, \mathcal{E}_Y) \leq s_X s_Y \leq s_X \sigma_Y$; divide by $\sigma_X \sigma_Y$.

Finally, Cauchy–Schwarz in $\mathcal{H}$ and independence of m, m' give

$$s_X^2 = \mathbb{E}_{m,m'} \langle \delta_m, \delta_{m'} \rangle^2 \leq \mathbb{E}_m \|\delta_m\|^2 \, \mathbb{E}_{m'} \|\delta_{m'}\|^2 = \beta_X^2. \qquad \square$$

A.4.2 SEMANTIC DISPLACEMENT AND THE AUGMENTATION GAP

Assumption 1 (Lipschitz encoders)**.** $\|e_X(p) - e_X(q)\| \leq \ell_X \, d_{\mathcal{X}}(p, q)$ for all $p, q \in \mathcal{X}$, and likewise for e_Y with constant ℓ_Y.

Definition 2 (Semantic displacement)**.**

$$\varepsilon_X := \mathbb{E}_{\pi,\, u}\Big[d_{\mathcal{X}}\big(g_X(m, u), g_X(m^+, u)\big)^2 \Big], \tag{21}$$

the mean-square distance the augmentation moves an observation when the nuisance is held fixed, with $d_{\mathcal{X}}$ a metric on $\mathcal{X}$.

Lemma 6 (Semantic variance versus augmentation gap)**.** *Under Assumption 1,*

$$\beta_X \leq \Delta_X + \tfrac{1}{2} \, \omega \, \ell_X^2 \, \varepsilon_X.$$

Proof. By Lemma 3 and Eq. 16, $\Delta_X = \mathbb{E}_\pi[k_X^c] = \mathbb{E}_\pi[h_X] = \mathbb{E}_\pi \langle \delta_m, \delta_{m^+} \rangle$. Both marginals of π are μ, so $\mathbb{E}_\pi \|\delta_m\|^2 = \mathbb{E}_\pi \|\delta_{m^+}\|^2 = \beta_X$ and

$$\mathbb{E}_\pi \big\| \mu_m - \mu_{m^+} \big\|^2 = \mathbb{E}_\pi \big\| \delta_m - \delta_{m^+} \big\|^2 = 2\beta_X - 2\Delta_X.$$

For the bound, couple the two mean embeddings through the same nuisance: with $\psi = \phi \circ e_X$, $\mu_m - \mu_{m^+} = \mathbb{E}_u[\psi(g_X(m,u)) - \psi(g_X(m^+,u))]$, so by Jensen

$$\|\mu_m - \mu_{m^+}\| \le \mathbb{E}_u \|\psi(g_X(m,u)) - \psi(g_X(m^+,u))\|.$$

On the sphere, $1 - e^{-t} \le t$ gives

$$\|\phi(p) - \phi(q)\|_{\mathcal{H}}^2 = 2\big(1 - e^{-\omega\|p-q\|^2/2}\big) \le \omega\|p-q\|^2,$$

so ϕ is $\sqrt{\omega}$-Lipschitz and ψ is $\sqrt{\omega}\,\ell_X$-Lipschitz. Hence

$$\|\mu_m - \mu_{m^+}\| \le \sqrt{\omega}\,\ell_X\,\mathbb{E}_u\, d_{\mathcal{X}}(g_X(m,u), g_X(m^+,u)),$$

and a second application of Jensen gives $\mathbb{E}_\pi\|\mu_m - \mu_{m^+}\|^2 \le \omega\,\ell_X^2\,\mathbb{E}_{\pi,u}[d_{\mathcal{X}}^2] = \omega\,\ell_X^2\varepsilon_X$. □

Write

$$\Lambda(\tau) := \coth\tfrac{1}{\tau} - \tau + \tau\log\big(\tau\tanh\tfrac{1}{\tau}\big) - 1 \;\in\; (-1, 0). \tag{22}$$

Lemma 7 (The loss controls the gap from above)**.** *Suppose $\tau\,\omega \le 1$. Then,*

$$\Delta_X \;\le\; a\,\big(\Lambda(\tau) + \tau R(1 - L_X)\big).$$

Proof. The objective controls two quantities: the alignment A_X from above and the uniformity $T := \mathbb{E}\big[e^{\langle x,x'\rangle/\tau}\big] > 0$ from below. Chord bounds convert $\hat{A}_X$ and $\hat{S}_X$ into exactly those two, after which the constraint linking them makes T cancel.

First, Lemma 10(i) with $\alpha = \omega$ gives $e^{\omega}\hat{A}_X \le \cosh\omega + A_X\sinh\omega$. Second, Lemma 10(iii) with $b = \omega \le \alpha = 1/\tau$, gives,

$$e^{\omega\langle x,x'\rangle} \ge e^{-\omega} + \frac{\sinh\omega}{\sinh\frac{1}{\tau}}\big(e^{\frac{1}{\tau}\langle x,x'\rangle} - e^{-\frac{1}{\tau}}\big).$$

After averaging,

$$e^{\omega}\hat{S}_X = \mathbb{E}[e^{\omega\langle x,x'\rangle}] \ge e^{-\omega} + \frac{\sinh\omega}{\sinh\frac{1}{\tau}}\big(T - e^{-1/\tau}\big).$$

Subtracting $\hat{A}_X - \hat{S}_X$, and using $\cosh\omega - e^{-\omega} = \sinh\omega$,

$$\frac{e^{\omega}\Delta_X}{\sinh\omega} \;\le\; 1 + A_X - \frac{T - e^{-1/\tau}}{\sinh\frac{1}{\tau}}. \tag{23}$$

Now eliminate A_X. By Eq. 11 and Jensen ($U_X \le \log T$), $A_X = \tau(U_X - N_X) \le \tau(\log T - N_X)$. Bound the logarithm by its tangent at $c = \tau\sinh\frac{1}{\tau}$, namely $\log T \le T/c - 1 + \log c$, valid for every $T > 0$:

$$A_X \;\le\; \frac{T}{\sinh\frac{1}{\tau}} - \tau + \tau\log\big(\tau\sinh\tfrac{1}{\tau}\big) - \tau N_X.$$

Using $1 + e^{-1/\tau}/\sinh\frac{1}{\tau} = \coth\frac{1}{\tau}$,

$$\frac{e^{\omega}\Delta_X}{\sinh\omega} \;\le\; \coth\tfrac{1}{\tau} - \tau + \tau\log\big(\tau\sinh\tfrac{1}{\tau}\big) - \tau N_X.$$

Finally $N_X = R\,L_X - \frac{1}{\tau}$ and $\tau\log\cosh\frac{1}{\tau} = \tau R - 2$ by Eq. 12 and Eq. 13, so $\tau\log(\tau\sinh\frac{1}{\tau}) - \tau N_X = \tau\log(\tau\tanh\frac{1}{\tau}) - 1 + \tau R(1 - L_X)$, giving the claim. □

A.4.3 Connecting a high loss with low CKA

Theorem 2 (Implicit misalignment)**.** *Suppose $\sigma_X > 0$, $\tau\,\omega \le 1$, and Assumption 1 holds. Then*

$$\boxed{\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\le\; \frac{d + \ell_X^2\varepsilon_X}{2\sigma^2\sigma_X} + \frac{\tau R a}{\sigma_X}(1 - L_X) =:\; g(L_X)}$$

with $d = 2\sigma^2 a\,\big(\coth\frac{1}{\tau} - \tau + \tau\log\big(\tau\tanh\frac{1}{\tau}\big) - 1\big)$, and the same statement holds with Y in place of X.

Proof. Chain Lemma 5, Lemma 6 and Lemma 7:

$$\mathrm{CKA}(\mathcal{E}_X, \mathcal{E}_Y) \;\le\; \frac{\beta_X}{\sigma_X} \;\le\; \frac{\Delta_X + \frac{1}{2}\,\omega\,\ell_X^2\,\varepsilon_X}{\sigma_X} \;\le\; \frac{a\big(\Lambda(\tau) + \tau R\,(1 - L_X)\big) + \frac{1}{2}\,\omega\,\ell_X^2\,\varepsilon_X}{\sigma_X}.$$

Reorganization gives the bound. □

A.5 THE LOSS RANGE

Lemma 8 (Range of the normalized loss)**.** $L_X \in [0, 1]$. *Moreover:*

(i) $L_X = 0$ *if and only if* $\langle x, x^+\rangle = 1$ *and* $\langle x, x'\rangle = 0$ *almost surely;*

(ii) $L_X = 1$ *if and only if there is a unit vector* e *such that* $x \in \{e, -e\}$ *almost surely with* $\Pr(x = e) = \frac{1}{2}$*, and* $x^+ = -x$ *almost surely;*

(iii) If additionally Assumption 1 is satisfied and $\ell_X^2 \varepsilon_X < 4$*, then*

$$L_X \;\le\; 1 - \frac{1 - \frac{1}{4}\ell_X^2 \varepsilon_X}{\tau R} \;<\; 1,$$

with equality if and only if there is a unit vector e *such that* $x \in \{e, -e\}$ *almost surely with* $\Pr(x = e) = \frac{1}{2}$*, and the augmentation reverses* $q(m) = \mathbb{E}[\langle x, e\rangle \mid m]$ *(*$q(m^+) = -q(m)$*)* π*-almost surely.*

Proof. Lower bound. $\mathbb{E}[\langle x, x^+\rangle] \le 1$ and the uniformity term is nonnegative (Eq. 15), so $N_X \ge -1/\tau$, i.e. $L_X \ge 0$.

Case (i). Equality forces $\langle x, x^+\rangle = 1$ a.s. and, by Lemma 1, $\|\bar{x}\|^2 = 0$ and $\mathbb{E}[V] = 0$; then $r(x)^2 \le \|\bar{x}\|^2 = 0$ and $V(x) = 0$ a.s., so $\langle x, x'\rangle = 0$ a.s. The converse is a direct substitution.

Upper bound. Since both marginals of π are μ, the pair (x, x^+) is a coupling of ρ_X with itself. Writing $z = x - \bar{x}$ and $z^+ = x^+ - \bar{x}$, the cross terms vanish and $A_X = \|\bar{x}\|^2 + \mathbb{E}\langle z, z^+\rangle$; as $\|x\| = 1$ we have $\mathbb{E}\|z\|^2 = \mathbb{E}\|z^+\|^2 = 1 - \|\bar{x}\|^2$, so Cauchy–Schwarz gives

$$A_X \;\ge\; \|\bar{x}\|^2 - \sqrt{\mathbb{E}\|z\|^2}\sqrt{\mathbb{E}\|z^+\|^2} = \|\bar{x}\|^2 - (1 - \|\bar{x}\|^2) \;=\; 2\|\bar{x}\|^2 - 1. \tag{24}$$

By Jensen and Lemma 10(i) with $\alpha = 1/\tau$, $c = \langle x, x'\rangle$,

$$U_X = \mathbb{E}_x\Big[\log \mathbb{E}_{x'}\big[e^{\langle x, x'\rangle/\tau}\big]\Big] \le \log \mathbb{E}\big[e^{\langle x, x'\rangle/\tau}\big] \le \log\Big(\cosh\tfrac{1}{\tau} + \|\bar{x}\|^2 \sinh\tfrac{1}{\tau}\Big). \tag{25}$$

Denote $t = \|\bar{x}\|^2$. With Eq. 24 this gives $N_X \le f(t)$ for $f(t) = (1-2t)/\tau + \log(\cosh\frac{1}{\tau} + t\sinh\frac{1}{\tau})$, and $f'(t) \le -\frac{2}{\tau} + \tanh\frac{1}{\tau} < 0$ for $t \ge 0$, so $N_X \le f(0) = \frac{1}{\tau} + \log\cosh\frac{1}{\tau} = R - \frac{1}{\tau}$. This gives $L_X \le 1$.

Case (ii). $L_X = 1$, meaning $N_X = R - \frac{1}{\tau}$, forces equality throughout. That is, the chord bound in Eq. 25 is tight; Jensen in Eq. 25 is tight; Cauchy–Schwarz in Eq. 24 is tight; and $\|\bar{x}\|^2 = 0$.

Since $c \mapsto e^{c/\tau}$ is strictly convex, it meets its chord on $[-1, 1]$ only at the endpoints, so tightness of the chord bound forces $\langle x, x'\rangle \in \{-1, 1\}$ a.e., i.e. $x' \in \{x, -x\}$. Hence, there is a single e in the support with $\rho_X(\{e, -e\}) = 1$. Writing $p = \Pr(x = e)$ we get $\bar{x} = (2p - 1)e$, so $\|\bar{x}\|^2 = 0$ gives $p = \frac{1}{2}$. Then $\mathbb{E}_{x'}[e^{\langle x, x'\rangle/\tau}] = \cosh\frac{1}{\tau}$ for both values of x. With $\bar{x} = 0$ we have $z = x$, and equality in Eq. 24 forces $x^+ = -x$ a.s. Conversely these conditions give $A_X = -1$ and $U_X = \log\cosh\frac{1}{\tau}$, hence $N_X = \frac{1}{\tau} + \log\cosh\frac{1}{\tau} = R - \frac{1}{\tau}$ by direct substitution.

Case (iii). Let $c_m := \mathbb{E}[x \mid m]$ be the semantic centroid, $d_m := c_m - \bar{x}$, $\gamma := \mathbb{E}_m\|d_m\|^2$ and $\Gamma := \mathbb{E}_\pi\|d_m - d_{m^+}\|^2$. Since u and u^+ are drawn independently, $A_X = \mathbb{E}_\pi\langle c_m, c_{m^+}\rangle$, and since both marginals of π are μ, decomposing $\mathbb{E}_\pi\langle d_m, d_{m^+}\rangle$ gives,

$$A_X \;=\; \|\bar{x}\|^2 + \gamma - \tfrac{1}{2}\Gamma. \tag{26}$$

Now we bound Γ in two ways. Coupling the two centroids through the same nuisance, $c_m - c_{m^+} = \mathbb{E}_u[e_X(g_X(m, u)) - e_X(g_X(m^+, u))]$, so Jensen and Assumption 1 give

$$\Gamma = \mathbb{E}_\pi\|c_m - c_{m^+}\|^2 \le \mathbb{E}_\pi\mathbb{E}_u[\|e_X(g_X(m, u)) - e_X(g_X(m^+, u))\|]^2 \le \ell_X^2\varepsilon_X.$$

Also, $\|d_m - d_{m^+}\|^2 \le 2\|d_m\|^2 + 2\|d_{m^+}\|^2$ gives $\Gamma \le 4\gamma$. Feeding the smaller of the two into Eq. 26,

$$A_X - \|\bar{x}\|^2 \;\ge\; \gamma - \tfrac{1}{2}\min\{\ell_X^2\varepsilon_X, 4\gamma\} \;=\; \max\{\gamma - \tfrac{1}{2}\ell_X^2\varepsilon_X,\, -\gamma\} \;\ge\; -\tfrac{1}{4}\ell_X^2\varepsilon_X,$$

the last step because the two branches meet at $\gamma = \frac{1}{4}\ell_X^2\varepsilon_X$.

Replacing Eq. 24 by this and repeating the argument of the upper bound, $N_X \le \psi(\|\bar{x}\|^2)$ for $\psi(t) = (\frac{1}{4}\ell_X^2\varepsilon_X - t)/\tau + \log(\cosh\frac{1}{\tau} + t\sinh\frac{1}{\tau})$, and $\psi'(t) \le -\frac{1}{\tau} + \tanh\frac{1}{\tau} < 0$, so $N_X \le \psi(0) = R - \frac{1}{\tau} - \frac{1}{\tau}\big(1 - \frac{1}{4}\ell_X^2\varepsilon_X\big)$, which is the claim; it is < 1 because $\ell_X^2\varepsilon_X < 4$.

Equality forces $\|\bar{x}\|^2 = 0$, both steps of Eq. 25 tight, and $\gamma = \frac{1}{4}\ell_X^2\varepsilon_X$ with $\Gamma = 4\gamma$. As in case (ii) the first two give $\rho_X(\{e, -e\}) = 1$ with $\Pr(x = e) = \frac{1}{2}$, and $d_m = q(m)e$ for $q(m) = \mathbb{E}[\langle x, e\rangle \mid m]$; then $\Gamma = 4\gamma$ holds exactly when $q(m^+) = -q(m)$ π-a.s., and $\gamma = \mathbb{E}_m q(m)^2$. □

A.6 The shape of the bounds

A.6.1 The shape of the implicit alignment bound

This section analyses the shape of the implicit alignment bound from Thm. 1. Specifically, it proves that the bound $f(L_X, L_Y)$ is nonincreasing in both L_X and L_Y, and converges to the desired limits (Prop. 1).

Lemma 9 (Shape of the alignment coefficient)**.** *For $L \in (0, 1]$. Then:*

(i) c is nonincreasing on $(0, 1]$, and strictly decreasing on the set where $0 < c(L) < 1$.

(ii) $c(L) = 1$ for every $L \le L_1$, where

$$L_1 := \frac{2ab + \chi^2 D - \sqrt{\chi^2 D\,(4ab + \chi^2 D)}}{2b^2} \in (0, a/b),$$

and $c(L) < 1$ for $L > L_1$. In particular $\lim_{L\to 0^+} c(L) = 1$.

Proof. (i) Since $a, b, D, \chi > 0$, on $(0, \infty)$ the map $L \mapsto [a - bL]_+$ is nonnegative and nonincreasing and $L \mapsto \chi\sqrt{DL}$ is positive and increasing, so their ratio $g(L)$ is nonincreasing. Composing with the map $t \mapsto \min\{1, t\}$ preserves this. On $\{L < a/b\}$ we may drop the positive part and write $g(L) = (aL^{-1/2} - bL^{1/2})/(\chi\sqrt{D})$, whence $g'(L) = -\frac{1}{2}(aL^{-3/2} + bL^{-1/2})/(\chi\sqrt{D}) < 0$; since $c = g$ wherever $0 < c < 1$, the strict decrease follows.

(ii) For $L > 0$, $c(L) = 1$ iff $a - bL \ge \chi\sqrt{DL}$, which forces $L < a/b$ and is then equivalent, after squaring, to $q(L) \ge 0$ with $q(L) := b^2L^2 - (2ab + \chi^2 D)L + a^2$. The parabola q opens upwards, and

$$q(0) = a^2 > 0, \qquad q(a/b) = a^2 - \big(2ab + \chi^2 D\big)\tfrac{a}{b} + a^2 = -\chi^2 D\,\tfrac{a}{b} < 0,$$

so q has two positive roots $L_1 < a/b < L_2$ and $\{q \ge 0\} \cap (0, a/b) = (0, L_1]$. The quadratic formula gives the stated L_1, using $(2ab + \chi^2 D)^2 - 4a^2b^2 = \chi^2 D(4ab + \chi^2 D)$. Hence, $c \equiv 1$ on $(0, L_1]$, and since $L_1 > 0$, $c \equiv 1$ is attainable, and gives the limit.

□

Proposition 1 (Monotonicity of the bound)**.** Denote the right-hand side of Thm. 1 $f(L_X, L_Y)$. With L_1 as in Lemma 9:

(i) f is nonincreasing in L_X and in L_Y on $(0, 1]^2$, and is strictly decreasing in L_X at every point with $c(L_X) \in (0, 1)$ and $c(L_Y) > 0$.

(ii) $f \equiv 1$ on $(0, L_1]^2$.

Proof. Let $g(c(L_X), c(L_Y)) = f(L_X, L_Y)$. Write $\theta_X = \arccos c_X$, $\theta_Y = \arccos c_Y \in [0, \pi/2]$, so that

$$g(c_X, c_Y) = c_X c_Y - \sqrt{(1 - c_X^2)(1 - c_Y^2)} = \cos\theta_X \cos\theta_Y - \sin\theta_X \sin\theta_Y = \cos(\theta_X + \theta_Y).$$

Now $c_X \mapsto \theta_X$ is decreasing on $[0, 1]$, $\theta_X + \theta_Y \in [0, \pi]$, and cos is decreasing on $[0, \pi]$; a decreasing function of a decreasing function is nondecreasing, so g is nondecreasing in c_X, and by symmetry in c_Y. For the strict monotonicity of g on $c_X \in (0, 1)$ the chain rule gives

$$\frac{\partial g}{\partial c_X} = \frac{\sin(\theta_X + \theta_Y)}{\sin\theta_X} \ge 0,$$

and $\sin(\theta_X + \theta_Y) = 0$ with $\theta_X + \theta_Y \in [0, \pi]$ forces $\theta_X + \theta_Y \in \{0, \pi\}$, i.e. $c_X = c_Y = 1$ or $c_X = c_Y = 0$; at $c_X = 0$ the one-sided derivative is $\sin(\pi/2 + \theta_Y) = \cos\theta_Y = c_Y > 0$ unless $c_Y = 0$.

(i) From Lemma 9(i): f is a nondecreasing function of $c(L_X)$, which is nonincreasing in L_X. Strictness likewise combines the strict decrease of c on $\{0 < c < 1\}$ with the strict increase of g.

(ii) By Lemma 9(ii), $c \equiv 1$ on $(0, L_1]$ and $g(1, 1) = 1$. □

A.6.2 The shape of the implicit misalignment bound

This section analyses the shape of the implicit misalignment bound from Thm. 2. Specifically, it proves that the bound $g(L)$ is strictly decreasing L and converges to the desired limits (Prop. 2).

Proposition 2 (Monotonicity of the misalignment bound)**.** The bound g is strictly decreasing in L_X, and is minimized at the largest possible loss as in Lemma 8(iii), where it equals

$$g\left(1 - \frac{1 - \frac{1}{4}\ell_X^2\varepsilon_X}{\tau R}\right) = \frac{a\left(1 + \Lambda(\tau) - \frac{1}{4}\ell_X^2\varepsilon_X\right) + \frac{1}{2}\omega\ell_X^2\varepsilon_X}{\sigma_X},$$

if $\ell_X^2\varepsilon_X < 4$, and otherwise at $g(1) = \left(a\Lambda(\tau) + \frac{1}{2}\omega\ell_X^2\varepsilon_X\right)/\sigma_X$.

Proof. Monotonicity is immediate: g is affine in L_X with negative slope $-a\tau R/\sigma_X$. □

A.7 General inequalities

This section provides general mathematical inequalities used in the proofs above.

Lemma 10 (Chord bounds)**.** *Let $\alpha > 0$ and $c \in [-1, 1]$. Then*

(i) $e^{\alpha c} \le \cosh\alpha + c\sinh\alpha$*;*

(ii) $|e^{\alpha c} - 1| \le (e^{\alpha} - 1)|c|$*;*

(iii) for every b with $0 < b \le \alpha$, $e^{bc} \ge e^{-b} + \dfrac{\sinh b}{\sinh\alpha}\left(e^{\alpha c} - e^{-\alpha}\right)$.

Proof. (i) By convexity, $c \mapsto e^{\alpha c}$ lies below its chord on $[-1, 1]$, which is $\cosh\alpha + c\sinh\alpha$.

(ii) The difference quotient $c \mapsto (e^{\alpha c} - 1)/c = \alpha\int_0^1 e^{\alpha c s}\,ds$ (value α at $c = 0$) is positive and nondecreasing, hence at most its value $e^{\alpha} - 1$ at $c = 1$; multiply by $|c|$.

(iii) Put $t = e^{\alpha c} \in [e^{-\alpha}, e^{\alpha}]$. Then $e^{bc} = t^{b/\alpha}$ is concave in t because $b/\alpha \le 1$, so it lies above the chord joining $(e^{-\alpha}, e^{-b})$ and (e^{α}, e^{b}), whose slope is $\frac{e^b - e^{-b}}{e^{\alpha} - e^{-\alpha}} = \frac{\sinh b}{\sinh\alpha}$. The right side being affine in t, and the left concave, the inequality survives taking expectations. □

Lemma 11 (Variance floor for the log-MGF)**.** *Let δ have mean zero and support in an interval of length at most 2, and let $\lambda \ge 1$. Then* $\log\mathbb{E}[e^{\lambda\delta}] \ge \frac{\log 2}{4}\lambda\mathrm{Var}(\delta)$.

Proof. Write $V = \mathrm{Var}(\delta) = \mathbb{E}[\delta^2]$ and let $[\ell, u]$ contain the support with $0 \in [\ell, u]$ and $u - \ell \le 2$, so $\delta \ge -2$.

Taylor's formula with integral remainder gives $e^t = 1 + t + t^2\psi(t)$ with $\psi(t) = \int_0^1 (1 - s)e^{st}\,ds = (e^t - 1 - t)/t^2$ nondecreasing. Since $\lambda\delta \ge -2\lambda$,

$$\mathbb{E}\left[e^{\lambda\delta}\right] \;\ge\; 1 + \lambda^2\psi(-2\lambda)V \;\ge\; 1 + \tfrac{\lambda}{4}V, \tag{27}$$

because $\lambda^2\psi(-2\lambda) = \frac{e^{-2\lambda} - 1 + 2\lambda}{4} \ge \frac{2\lambda - 1}{4} \ge \frac{\lambda}{4}$ for $\lambda \ge 1$.

Case 1: $\lambda V/4 \le 1$. By concavity $\log(1 + s) \ge s\log 2$ on $[0, 1]$, so Eq. 27 gives the claim.

Case 2: $\lambda V/4 > 1$. Let $p = \Pr(\delta > 0)$, positive because $p = 0$ with $\mathbb{E}\delta = 0$ forces $V = 0$, and let $h = \mathbb{E}[\delta\mathbf{1}_{\delta>0}] = \mathbb{E}[-\delta\mathbf{1}_{\delta<0}]$. Bounding $\delta \le u$ on $\{\delta > 0\}$ and $-\delta \le -\ell$ on $\{\delta < 0\}$ gives

$V \le (u-\ell)h \le 2h$, so $\lambda h \ge \lambda V/2 > 2 \ge p$. Discarding the negative part and applying Jensen to the law of δ given $\{\delta > 0\}$, gives

$$\mathbb{E}[e^{\lambda\delta}] \;\ge\; p\,\mathbb{E}[e^{\lambda\delta} \mid \delta > 0] \;\ge\; p\,e^{\lambda h/p}.$$

The map $p \mapsto p\,e^{\lambda h/p}$ is nonincreasing on $(0, \lambda h]$, so at $p \le 1$ it is at least $e^{\lambda h}$, whence

$$\log \mathbb{E}[e^{\lambda\delta}] \;\ge\; \lambda h \;\ge\; \tfrac{\lambda V}{2} \;\ge\; \tfrac{\log 2}{4}\lambda V.$$

□

A.8 Constants Details

Thm. 1, as presented in the main text, incorporates several hyperparameter-dependent constants:

$$a_\sigma := e^{-\frac{1}{\sigma^2}} \sinh \frac{1}{\sigma^2},$$

which depends on the CKA-RBF scale σ. Recall $R_\tau = \frac{2}{\tau} + \log\cosh\frac{1}{\tau}$

$$b_{\sigma,\tau} := \big(\frac{1}{\sigma^2} + e^{-\frac{1}{\sigma^2}} \sinh \frac{1}{\sigma^2}\big)\tau R_\tau, \quad D_{\sigma,\tau} := \Big(1 + \frac{4}{\log 2}\Big)\big(1 - e^{-\frac{1}{\sigma^2}}\big)^2 \tau R_\tau,$$

which also rely on the InfoNCE temperature τ.

Similarly, Thm. 2 incorporates other hyperparameter-dependent constants:

$$h_{\sigma,\tau} = \tau R_\tau a_\sigma,$$

which depends on the InfoNCE temperature τ, and

$$d_{\sigma,\tau} = 2\sigma^2 a_\sigma \left(\coth\tfrac{1}{\tau} - \tau + \tau \log\left(\tau \tanh\tfrac{1}{\tau}\right) - 1\right),$$

which also incorporates the CKA-RBF scale σ.

B AudioCaps Results

In Sec. 4.2, the evaluation of unconstrained, pre-trained audio models focused primarily on the CLOTHO dataset. To further support our claim across audio datasets, we extend our analysis to AudioCaps.

Fig. 7a extends Fig. 6a from the main text to the AudioCaps dataset, while Fig. 7b extends Fig. 6b. Consistent with our primary findings, augmentation-based text models (Aug) consistently achieve higher cross-modal alignment on AudioCaps than next-token prediction (NTP) models of comparable or even substantially larger sizes. These results further support the role of unimodal augmentation objectives as a fundamental driver of implicit cross-modal alignment.

Fig. 7c extends Fig. 6c from the main text to the AudioCaps dataset. The strong correlation between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remains robust on AudioCaps. This finding further supports unimodal InfoNCE loss as a predictor of cross-modal alignment quality.

C Statistical Significance

In Sec. 4.2, the analysis of unconstrained, pre-trained models reports Spearman (ρ_s) and Pearson (r) correlations between CMA and unimodal model properties over the available encoders. To demonstrate that these correlations are not an artifact of the sample size, this section establishes their statistical significance.

For each panel, we test the null hypothesis that the alignment scores are exchangeable with respect to the property under consideration. We obtain a two-sided p-value from $200{,}000$ random permutations of the alignment scores against the property, where model size and training-corpus size are taken on a $\log_{10}$ scale and the InfoNCE loss as $-\mathcal{L}$, matching the axes of the corresponding figures. Since the three figures comprise 10 panels tested jointly, we control the family-wise error rate using the Holm

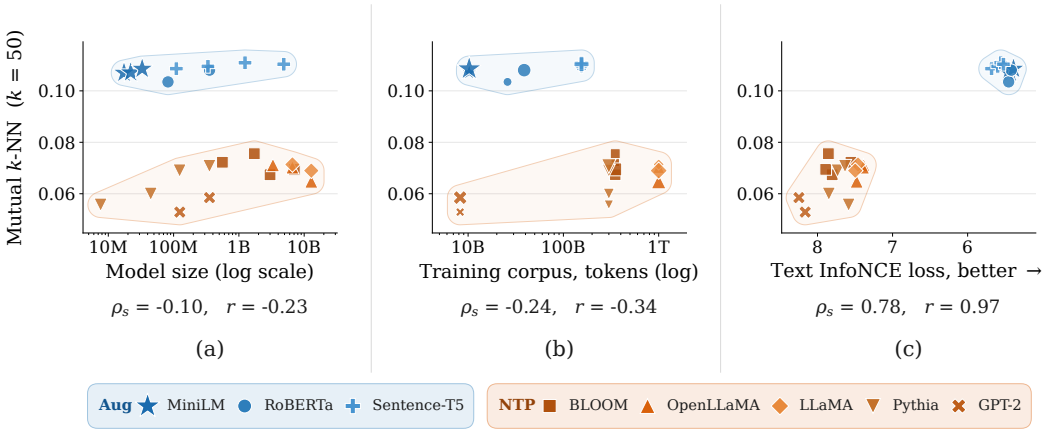


Figure 7: **Audio-text results on AudioCaps dataset.** CMA between text models and the BYOL-A audio encoder as a function of **(a) model size** (number of parameters), **(b) corpus size** (amount of tokens the model was trained on), and **(c) unimodal text InfoNCE loss**. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, while the unimodal loss is highly correlated, indicated by both Spearman (ρ_s) and Pearson (r) correlations.

Table 3: **Significance of the correlations in the three pre-trained figures 4, 5, 6.** Size and corpus size are correlated on $\log_{10}$ and the loss as $-\mathcal{L}$. p is two-sided from 200,000 permutations of the alignment scores against the property; ✓ marks $p < 0.05$ after Holm correction over the 10 panels, ✗ that it does not. Only the correlation with the unimodal InfoNCE loss is statistically significant.

		Spearman		Pearson	
Panel	Property	ρ_s	p	r	p
Fig. 4 — text encoders against DINOv2 ViT-B/14					
(a) MS-COCO	Model size ($\log_{10}$)	-0.030	0.889 ✗	-0.081	0.706 ✗
(b) MS-COCO	Training-corpus size ($\log_{10}$)	-0.233	0.282 ✗	-0.360	0.092 ✗
(b) WIT	Training-corpus size ($\log_{10}$)	-0.176	0.421 ✗	-0.173	0.431 ✗
Fig. 5 — text encoders against DINOv2 ViT-B/14 in (a); image towers against BLOOM-7b1 in (b)					
(a) MS-COCO	Text InfoNCE loss ($-\mathcal{L}$)	0.880	$<5.0\times10^{-6}$ ✓	0.925	$<5.0\times10^{-6}$ ✓
(a) WIT	Text InfoNCE loss ($-\mathcal{L}$)	0.820	1.5×10^{-5} ✓	0.857	$<5.0\times10^{-6}$ ✓
(b) MS-COCO	Vision InfoNCE loss ($-\mathcal{L}$)	0.855	2.6×10^{-4} ✓	0.939	$<5.0\times10^{-6}$ ✓
(b) WIT	Vision InfoNCE loss ($-\mathcal{L}$)	0.890	5.0×10^{-5} ✓	0.896	$<5.0\times10^{-6}$ ✓
Fig. 6 — text encoders against BYOL-A v2					
(a) Clotho	Model size ($\log_{10}$)	0.043	0.837 ✗	0.149	0.486 ✗
(b) Clotho	Training-corpus size ($\log_{10}$)	-0.196	0.368 ✗	-0.128	0.561 ✗
(c) Clotho	Text InfoNCE loss ($-\mathcal{L}$)	0.920	$<5.0\times10^{-6}$ ✓	0.860	$<5.0\times10^{-6}$ ✓

correction and consider a correlation significant when its corrected p-value is below 0.05. Values reported as $<5.0\times10^{-6}$ lie at the resolution limit of the permutation test.

Table 3 reports the results for the three pre-trained figures (4, 5, 6). Consistent with our primary findings, every correlation between the unimodal InfoNCE loss and CMA is significant, for both Spearman and Pearson, and across vision–text and audio–text alike, with all p-values below $2.6\text{x}10^{-4}$ before correction. This confirms that the unimodal Aug-SSL loss is a reliable predictor of cross-modal alignment, and that this relation is not driven by a small number of individual models.

In contrast, the correlations with model and corpus size are inconsistent. They are consistently insignificant across datasets and modalities. Notably, in no panel do we find evidence that larger models or larger training corpora yield higher CMA. Together with the loss results, this further

supports that unimodal augmentation objectives, rather than scale, act as a fundamental driver of implicit cross-modal alignment.

D Additional Text Augmentations

In Sec. 4, our evaluation of the text InfoNCE utilized LLM-augmentations generated by Qwen2.5. However, this holds universally across different text augmentations. This section extends the analysis in the main text to different LLM augmentations and back-translation augmentations. For the LLM augmentations, we used two additional LLMs (Phi-3.5-Mini (Haider et al., 2024) and Gemma 4 12b (Team et al., 2026)) to generate five rephrased versions of each text. For the back-translation, we used bilingual models to transfer the English text to another language (e.g., French) and back to English. Additional details are provided in App. L.

Fig. 8 and 9 extend Fig. 3a from the main text to the two additional LLMs and back-translation augmentations, respectively. Consistent with our primary findings, the results support Thm. 1: minimizing the unimodal Aug-SSL objectives leads to increasing CMA. In contrast, every pairing involving a loss-maximized encoder is consistent with Thm. 2, showing that a high unimodal loss in either modality prevents cross-modal alignment, with CMA approaching chance level.

Fig. 10 and 11 extend Fig. 5a from the main text to two additional LLMs and back-translation augmentations, respectively. The strong correlation between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remains robust across LLMs used to generate the augmentations.

E Different Vision Encoders

In Sec. 4.2, the evaluation of CMA to vision models focused on DINOv2, and to text models on BLOOM. This section shows that the same trends hold across different vision and text architectures and training paradigms. In particular, it extends the analysis to a broader suite of 14 vision models and 24 text models, spanning the model families discussed in Sec. 4.2.

Fig. 12 extends Fig. 4a to the 14 vision models, while Fig. 13 extends Fig. 1. Consistent with our primary findings, the results demonstrate that augmentation-based text models (Aug) consistently achieve higher cross-modal alignment across all tested vision encoders compared to next-token prediction (NTP) models of comparable or even substantially larger sizes.

Fig. 14 and 15 extend Fig. 4b to the 14 vision models, on COCO and WIT respectively. Consistent with our primary findings, the results demonstrate that augmentation-based text models (Aug) consistently achieve higher cross-modal alignment across all tested vision encoders compared to next-token prediction (NTP) models of comparable or even substantially smaller model and corpus sizes.

Fig. 16 and 17 extend Fig. 5a to the 14 vision models, on COCO and on WIT, respectively. The strong correlation between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remains robust regardless of the specific vision architecture or its scale.

Similarly, Fig. 18 and 19 extend Fig. 5b to the 24 text models, on COCO and on WIT, respectively. The strong correlation between lower unimodal vision Aug-SSL InfoNCE loss and higher cross-modal alignment remains robust regardless of the specific text architecture or its scale.

F Different Audio Encoders

In Sec. 4.2, the evaluation of audio models focused on BYOL-A-v2. This section extends our analysis to a broader suite of 15 audio models, spanning the six model families: wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), data2vec (Baevski et al., 2022), AudioMAE (Huang et al., 2022), BYOL-A (Niizumi et al., 2021).

Fig. 20 and 21 extend the model size analysis (Fig. 6a) to these additional audio models, on Clotho and AudioCaps, respectively. Consistent with our primary findings, the results demonstrate that augmentation-based text models (Aug) consistently achieve higher cross-modal alignment across

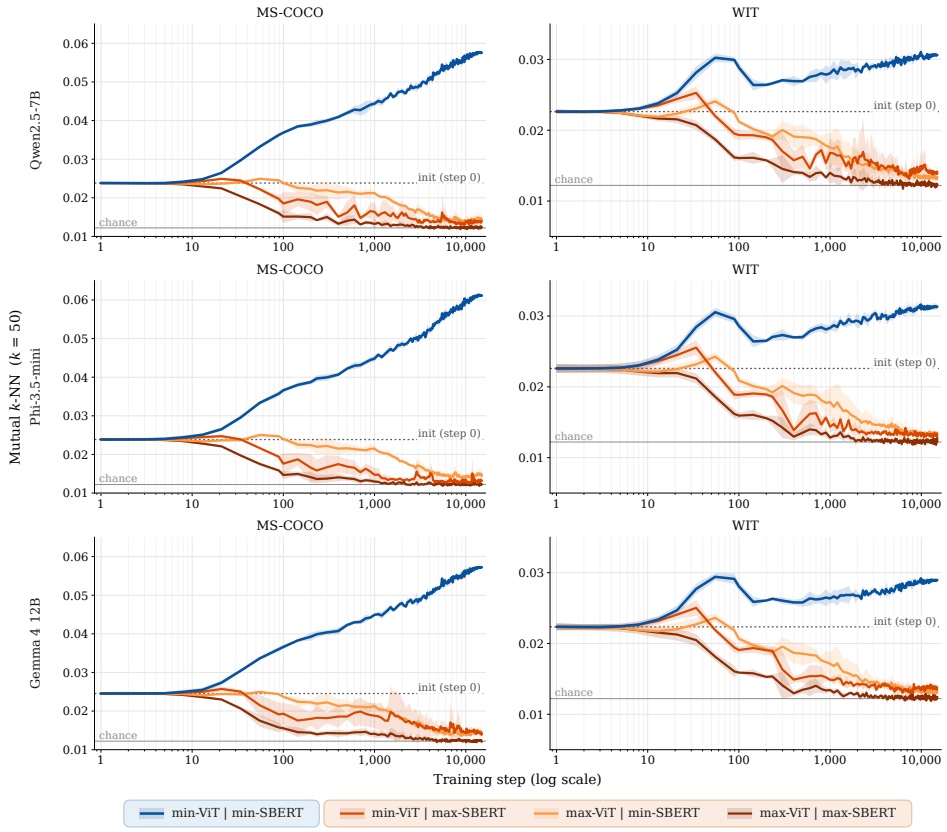


Figure 8: **Extended Empirical validation of Thm. 1 and 2 with different LLMs augmentations.** Vision (ViT) and text (SBERT) models trained to minimize/maximize the Aug-SSL loss (Eq. 3) using Phi-3.5-Mini and Gemma 4 12B text augmentations. Minimizing both unimodal losses improves CMA (Thm. 1), whereas maximizing either loss prevents alignment, with CMA converging to the chance level (Thm. 2).

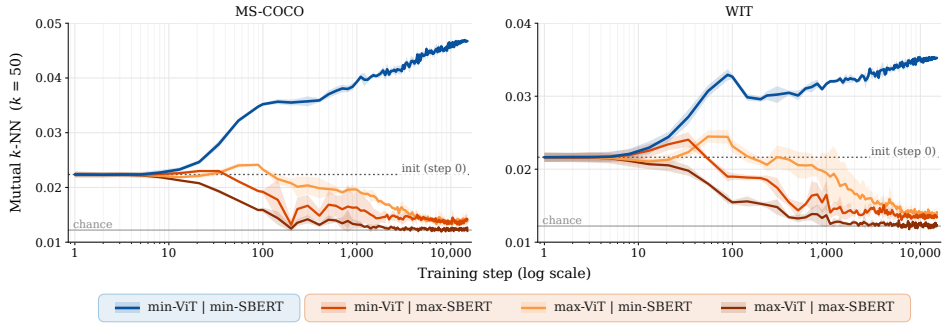


Figure 9: **Extended Empirical validation of Thm. 1 and 2 with back-translation augmentations.** Vision (ViT) and text (SBERT) models trained to minimize/maximize the Aug-SSL loss (Eq. 3) using back-translation text augmentations. Minimizing both unimodal losses improves CMA (Thm. 1), whereas maximizing either loss prevents alignment, with CMA converging to the chance level (Thm. 2).

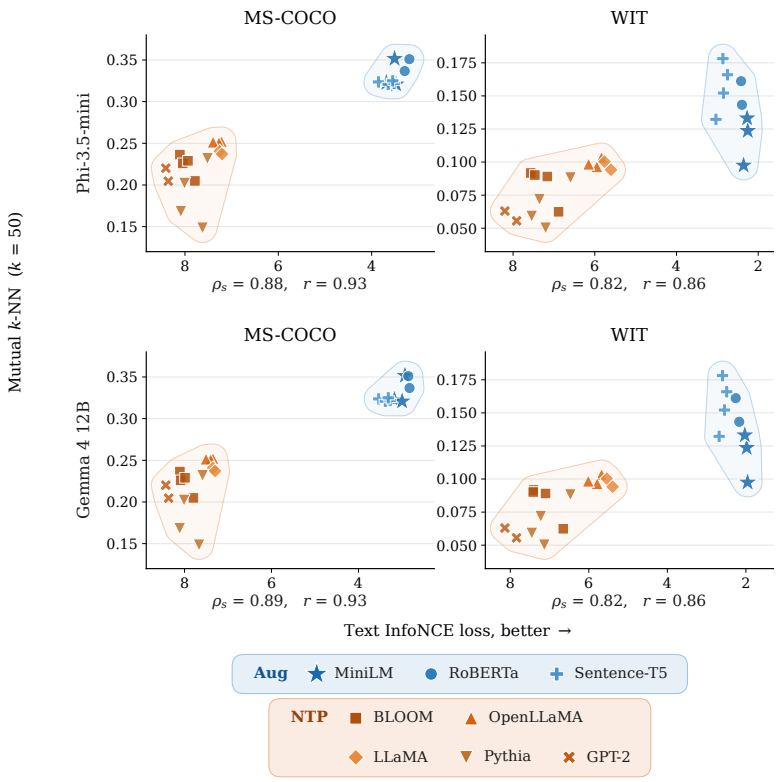


Figure 10: **CMA by InfoNCE across LLMs text augmentations on the COCO and WIT datasets.** CMA by unimodal text InfoNCE loss across different LLMs (Phi-3.5-mini, Gemma 4 12B) augmentations. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

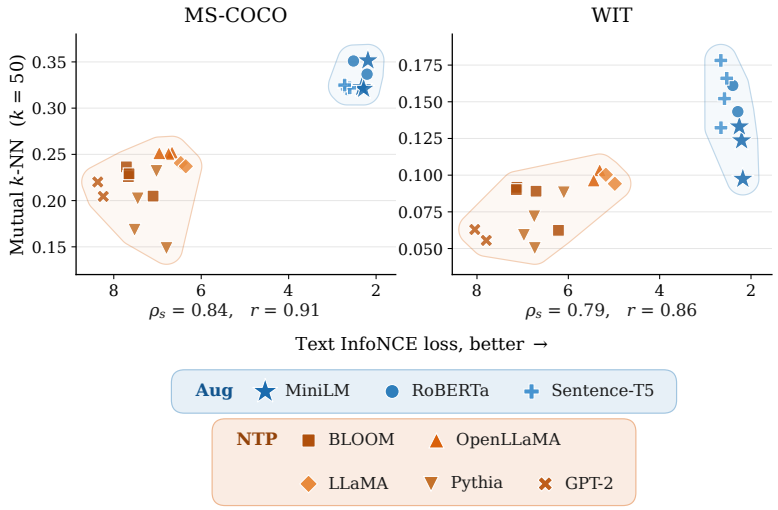


Figure 11: **CMA by InfoNCE across back-translation text augmentations on the COCO and WIT datasets.** CMA by unimodal text InfoNCE loss using back-translation augmentations. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

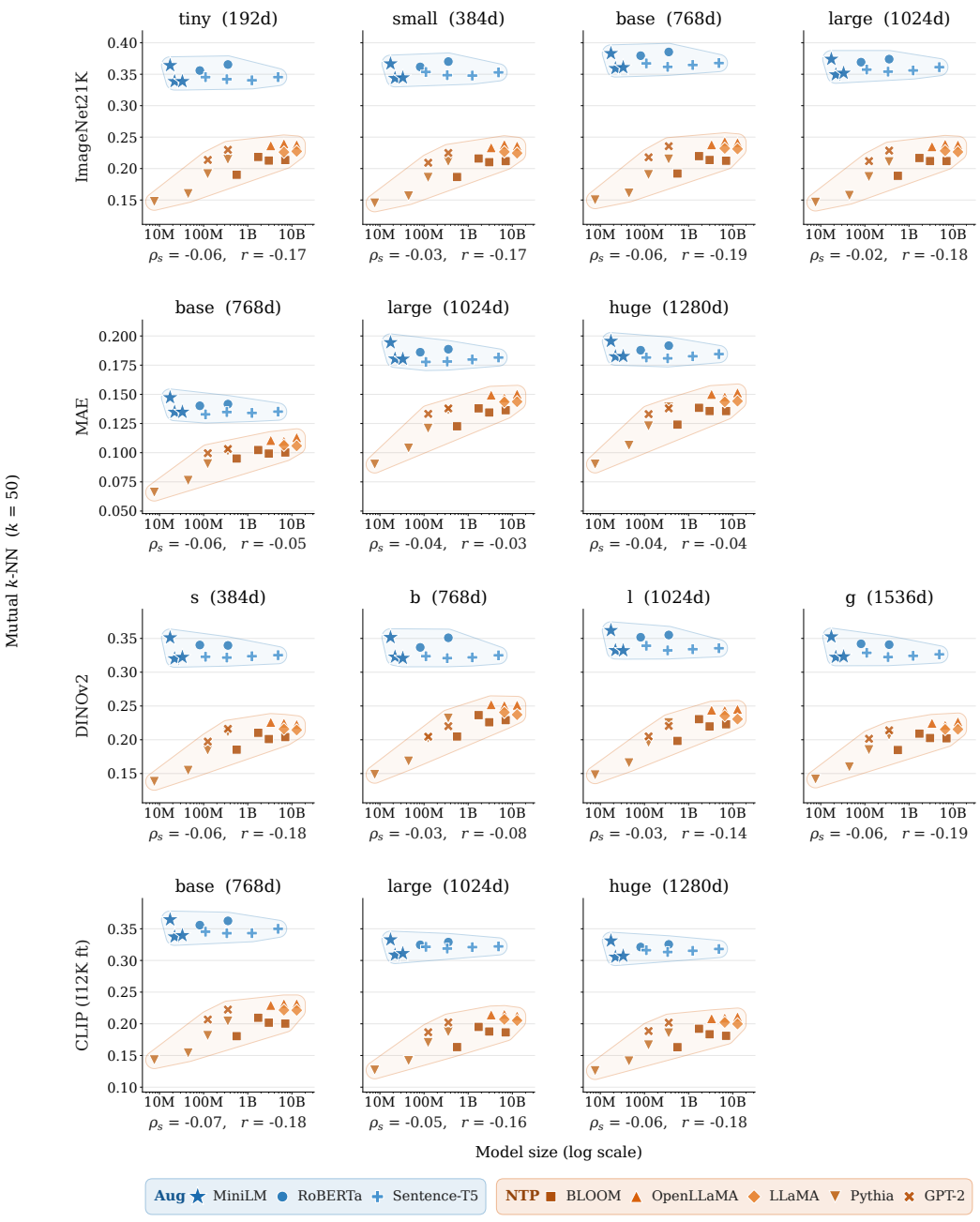


Figure 12: **CMA across vision architectures on the COCO dataset.** CMA by model size (number of parameters) across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

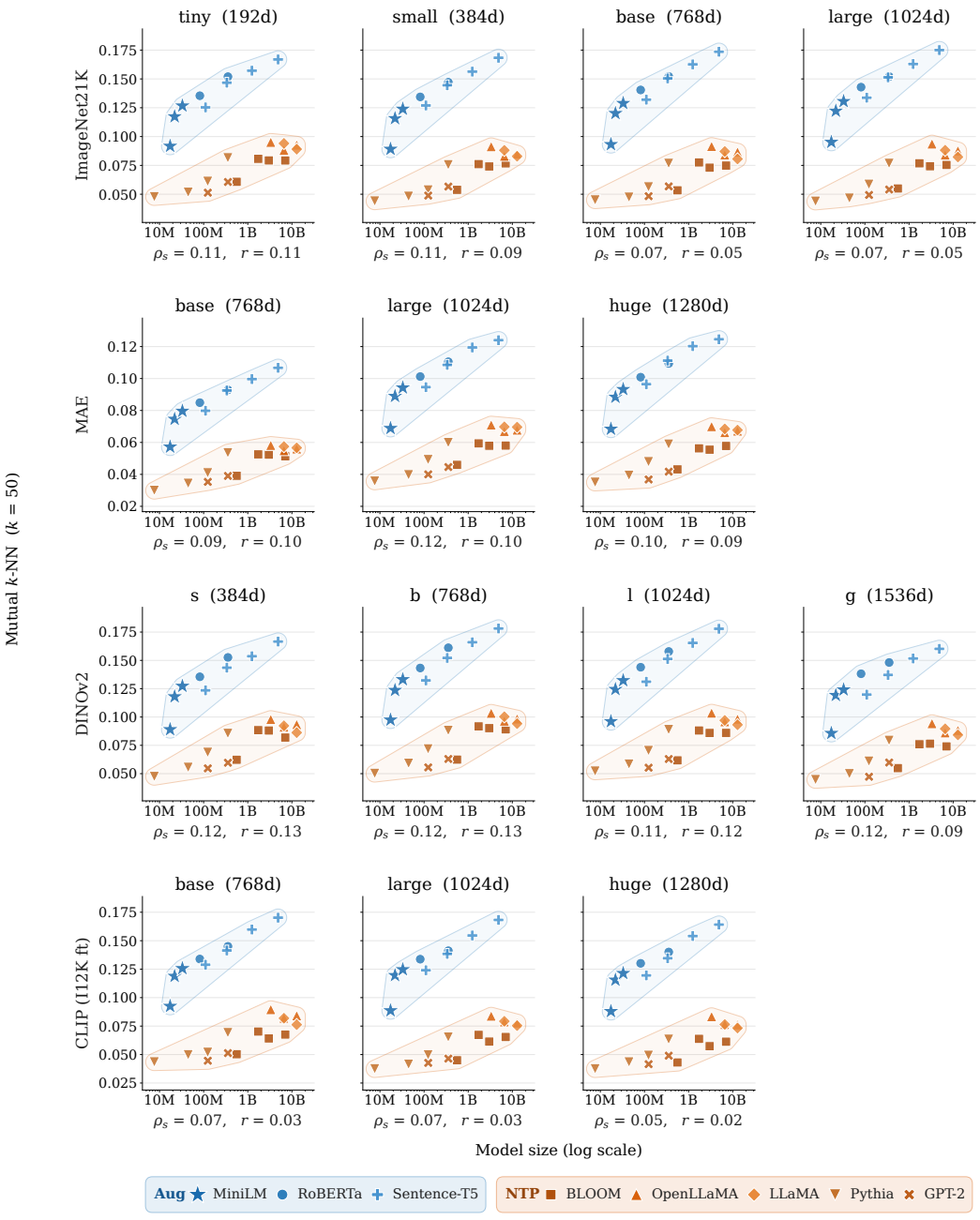


Figure 13: **CMA across vision architectures on the WIT dataset.** CMA by model size (number of parameters) across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

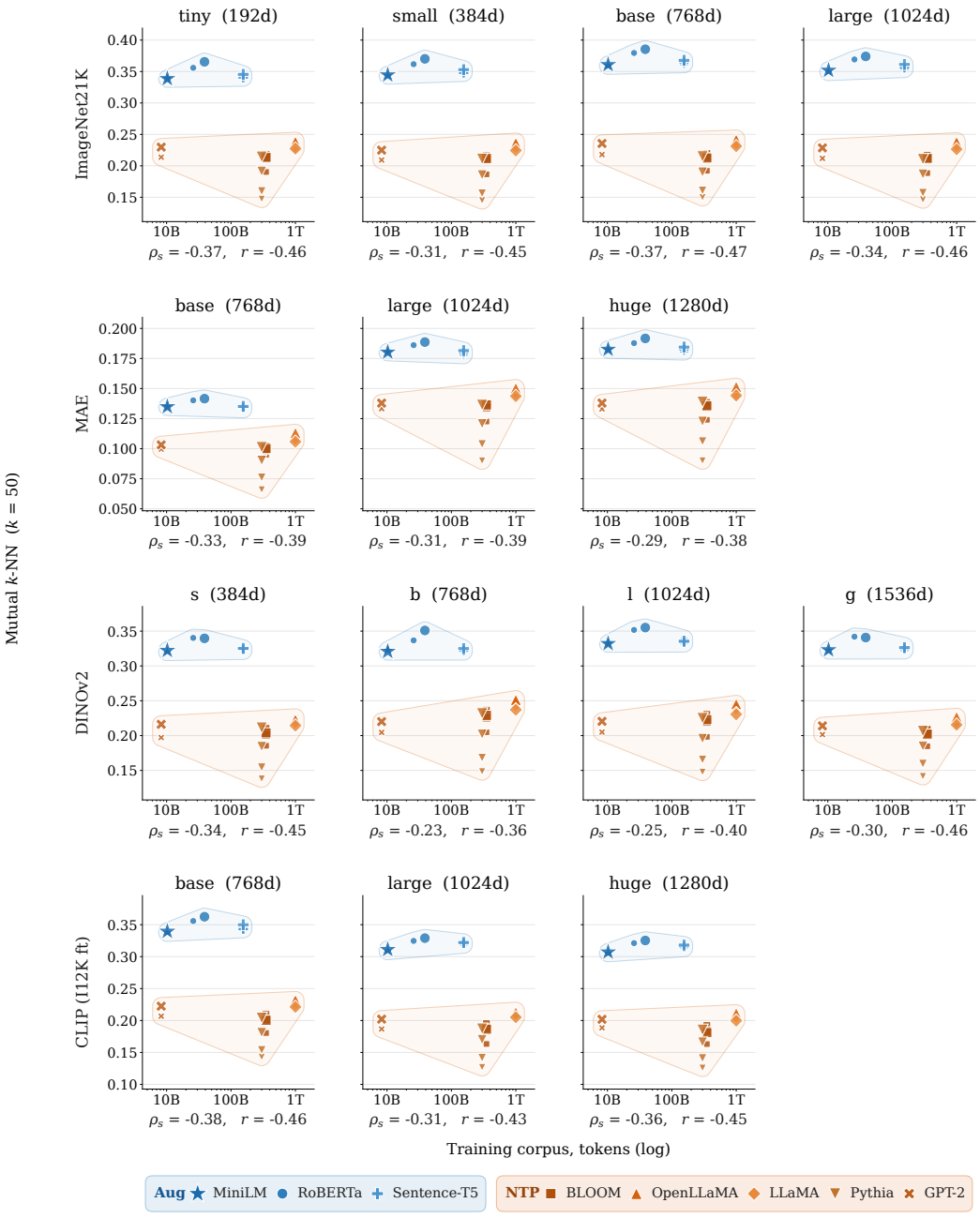


Figure 14: **CMA by corpus size across vision architectures on the COCO dataset.** CMA by corpus size (tokens trained on) across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

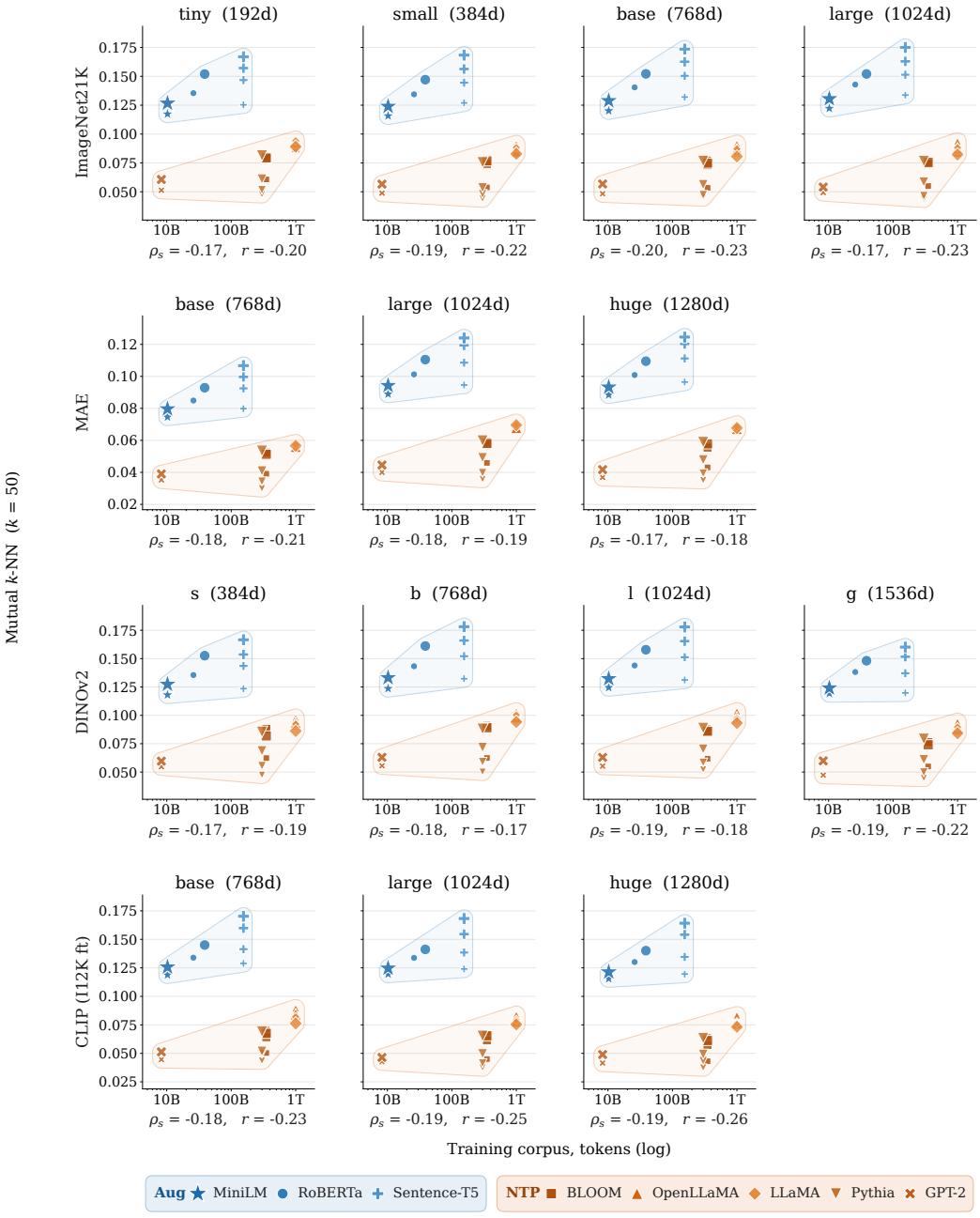


Figure 15: **CMA by corpus size across vision architectures on the WIT dataset.** CMA by corpus size (tokens trained on) across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

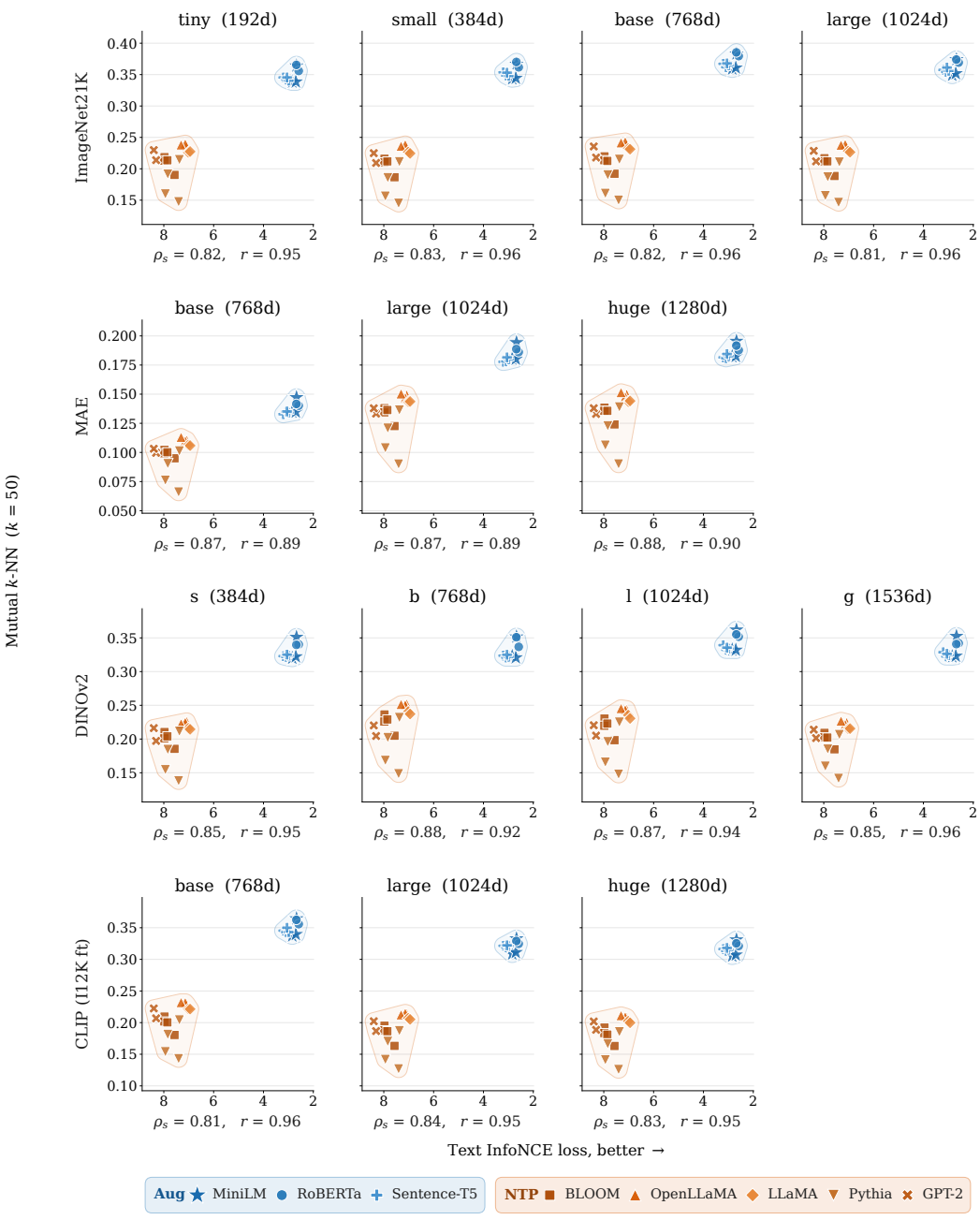


Figure 16: **CMA by InfoNCE across vision architectures on the COCO dataset.** CMA by unimodal text InfoNCE loss across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

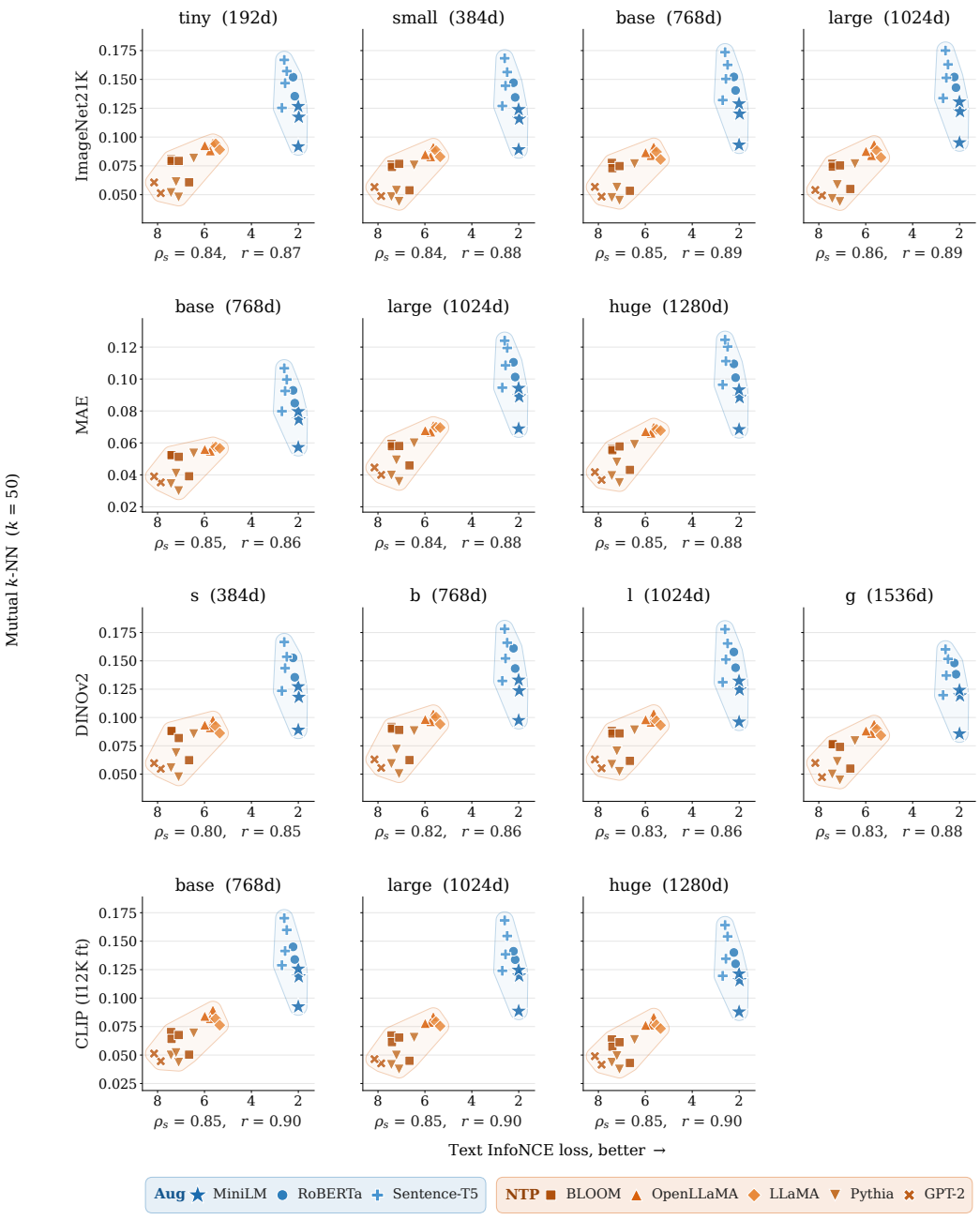


Figure 17: **CMA by InfoNCE across vision architectures on the WIT dataset.** CMA by unimodal text InfoNCE loss across different vision encoder families (ImageNet21K, MAE, DINOv2, and CLIP) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

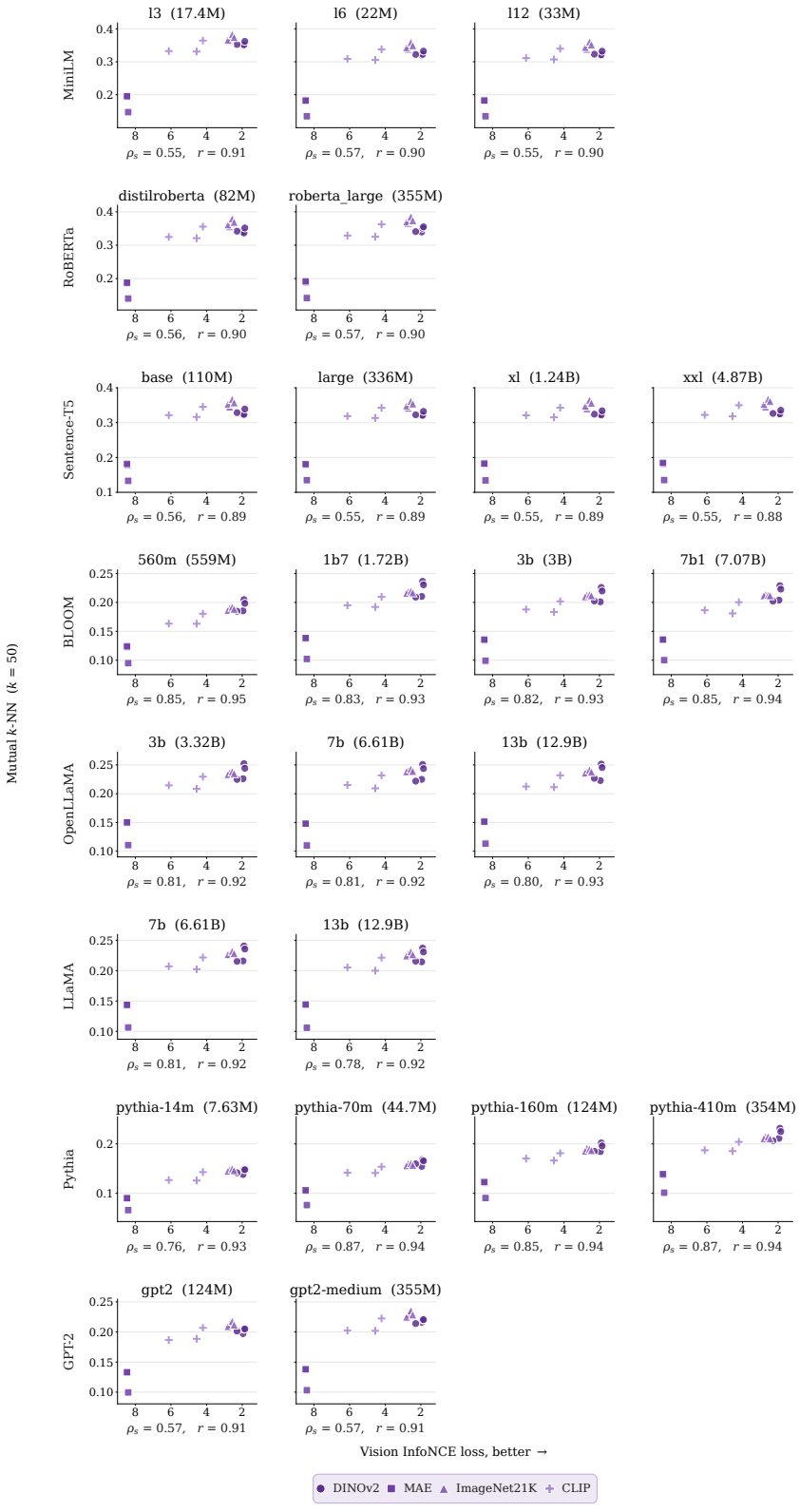


Figure 18: **CMA by Vision InfoNCE across text architectures on the COCO dataset.** CMA by unimodal vision InfoNCE loss across different text encoder families (MiniLM, RoBERTa, Sentence-T5, BLOOM, OpenLLaMA, LLaMA, Pythia, and GPT-2) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

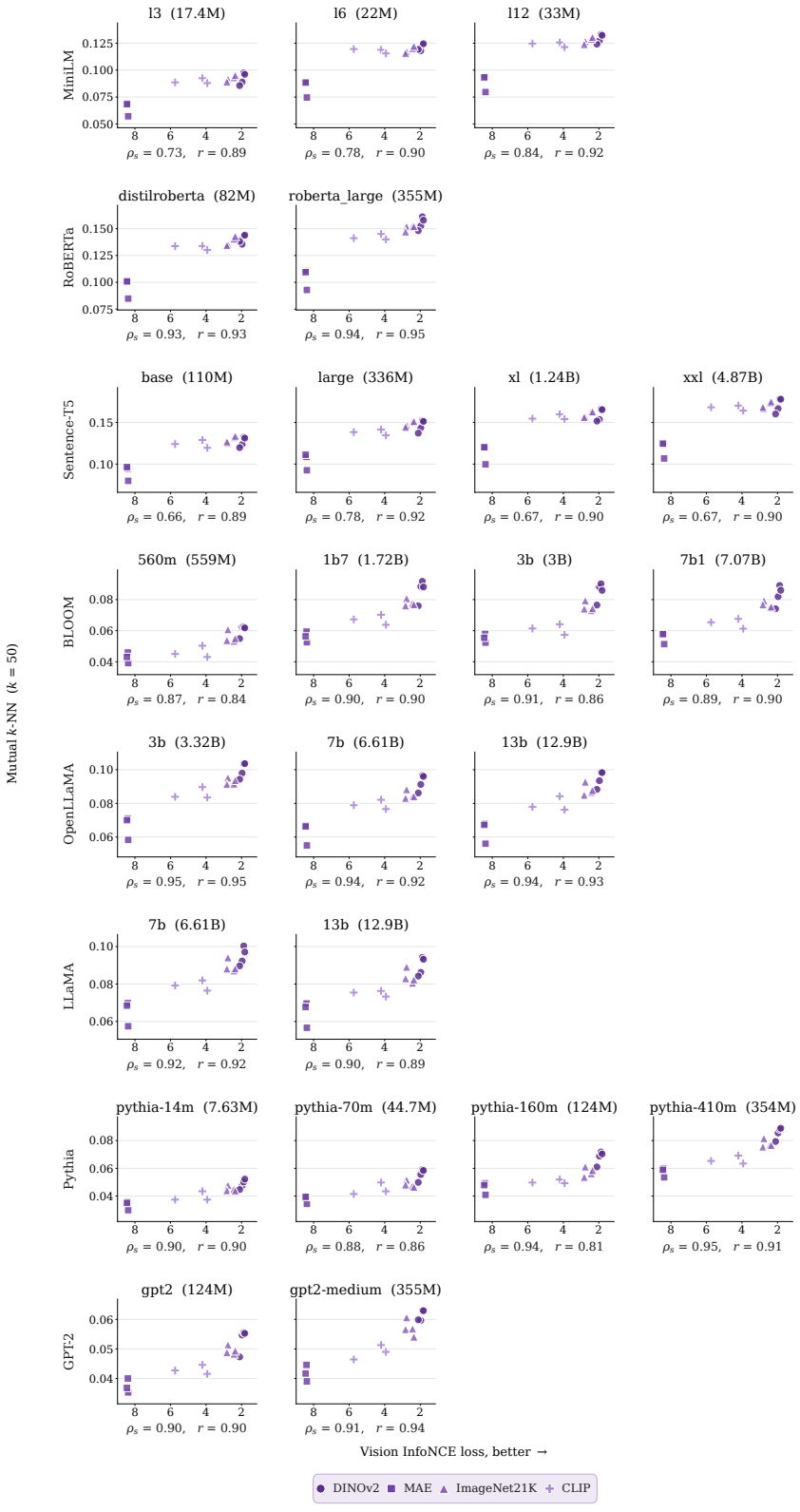


Figure 19: **CMA by Vision InfoNCE across text architectures on the WIT dataset.** CMA by unimodal vision InfoNCE loss across different text encoder families (MiniLM, RoBERTa, Sentence-T5, BLOOM, OpenLLaMA, LLaMA, Pythia, and GPT-2) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

all tested audio encoders compared to next-token prediction (NTP) models of comparable or even substantially larger sizes.

Fig. 22 and 23 extend Fig. 6b to these audio models, on Clotho and AudioCaps respectively. Consistent with our primary findings, the results demonstrate that augmentation-based text models (Aug) consistently achieve higher cross-modal alignment across all tested audio encoders compared to next-token prediction (NTP) models of comparable or even substantially smaller model and corpus sizes.

Fig. 24 and 25 extend our analysis of the unimodal loss (Fig. 6c) to these audio architectures, on Clotho and AudioCaps, respectively. The strong correlations between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remain remarkably robust across every evaluated audio encoder family and scale.

G Additional Cross-Modal Alignment Metrics

In Sec. 4.2, the evaluation focused on Mutual k-NN with k=50 for the CMA metric. To demonstrate that the connection between unimodal Aug-SSL objectives and CMA holds universally across different local CMA metrics on different k values, this section extends our analysis to a broader suite of Mutual k-NN and CKNNA on 4 different k values.

CKNNA. CKNNA is a popular local restriction of CKA: it evaluates the same kernel alignment, but only on the pairs that are mutual k-nearest neighbors in *both* embedding spaces. Given paired observations $\mathcal{S} = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{E}_X \times \mathcal{E}_Y$, let $K_{ij} = k_X\big(e_X(x_i), e_X(x_j)\big)$ and $L_{ij} = k_Y\big(e_Y(y_i), e_Y(y_j)\big)$ be the two kernel matrices (usually linear), and define the mutual-neighbor mask

$$M_{ij} \;=\; \mathbf{1}\big[x_j \in N_X^k(x_i)\big] \cdot \mathbf{1}\big[y_j \in N_Y^k(y_i)\big]\,, \qquad M_{ii} = 0. \tag{28}$$

Replacing each kernel by its masked counterpart in Eq. 7 gives

$$\mathrm{CKNNA}_k(\mathcal{E}_X, \mathcal{E}_Y) \;=\; \frac{\mathrm{HSIC}\big(M \odot K,\, M \odot L\big)}{\sqrt{\mathrm{HSIC}\big(M \odot K,\, M \odot K\big)\,\mathrm{HSIC}\big(M \odot L,\, M \odot L\big)}} \;\in\; [0, 1]. \tag{29}$$

Fig. 26 and 27 extend Fig. 4a to Mutual k-NN and CKNNA (respectively) with different k values on COCO and WIT. Consistent with our primary findings, the results demonstrate that augmentation-based text models (Aug) consistently achieve higher cross-modal alignment across the tested CMA metrics and across k values compared to next-token prediction (NTP) models of comparable or even substantially larger sizes.

Fig. 28 and 29 extend Fig. 5a to CKNNA with different k values, on COCO and on WIT. The strong correlation between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remains remarkably robust regardless of the specific local CMA metric or the k value.

H Larger Test Set Sizes

In Sec. 4.2, the evaluation used a test set of size 1024. This section extends our analysis to additional larger test set sizes.

Fig. 30 extends Fig. 4 to different test set sizes on COCO. Consistent with our primary findings, the results demonstrate that the text models keep about the same relation between them across different test set sizes and model capacities.

I Different Temperatures

In Sec. 4.2, the evaluation used InfoNCE loss with temperature $\tau = 0.1$. Fig. 31 extends Fig. 5a to different temperatures on COCO. The strong correlation between lower unimodal text Aug-SSL InfoNCE loss and higher cross-modal alignment remains robust regardless of the specific temperature of the InfoNCE loss.

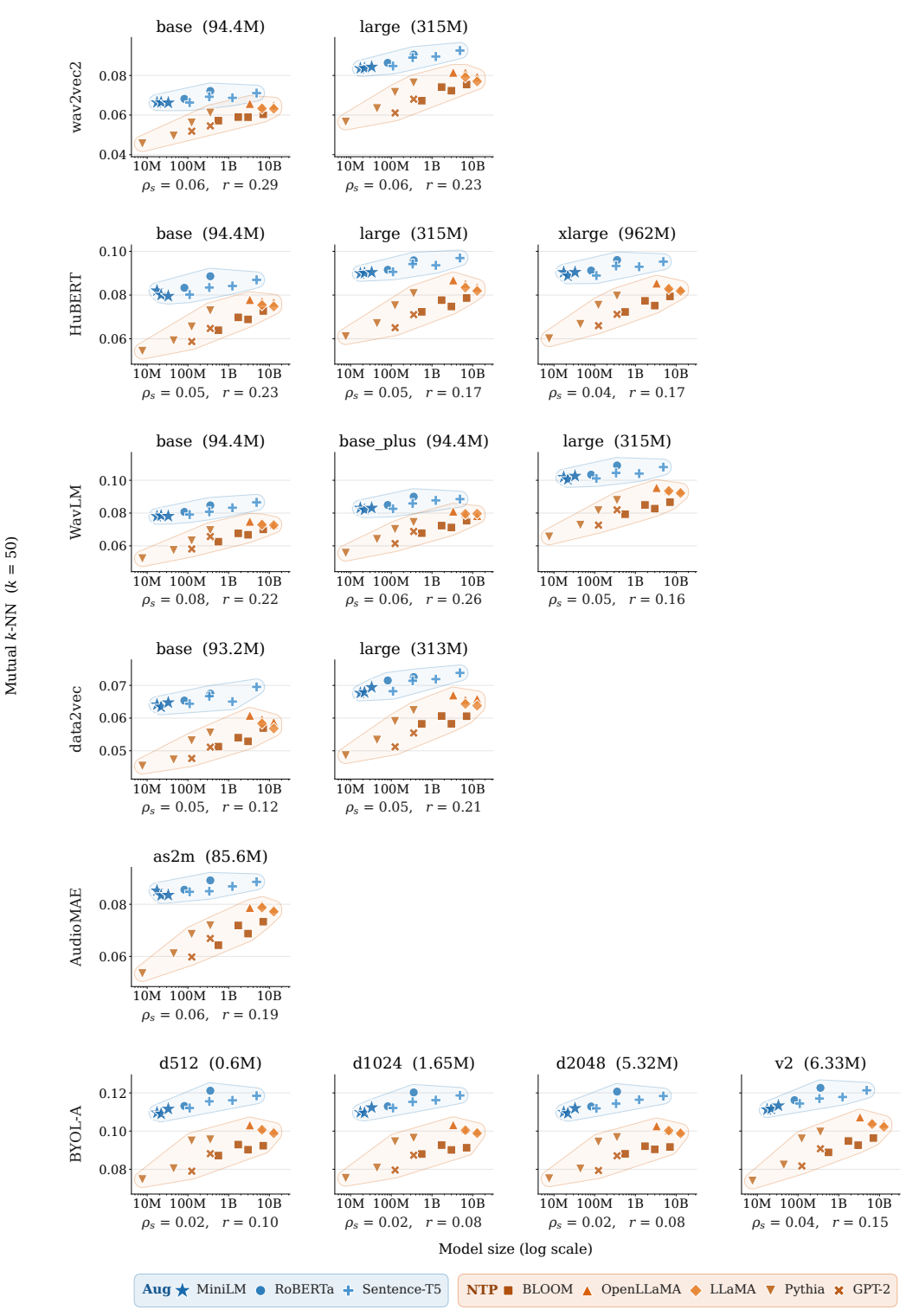


Figure 20: **CMA across audio architectures on the CLOTHO dataset.** CMA by model size (number of parameters) across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

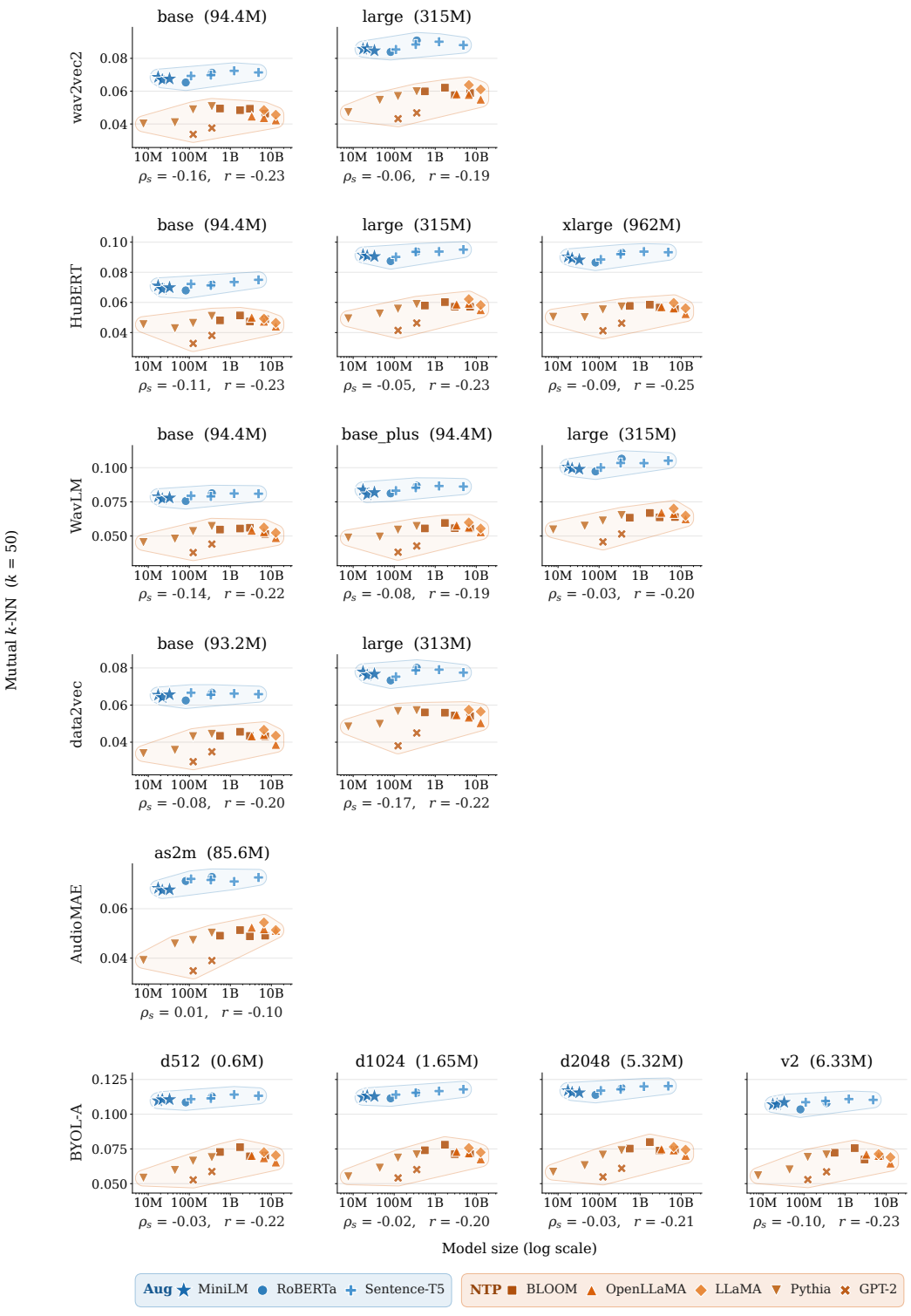


Figure 21: **CMA across audio architectures on the AudioCaps dataset.** CMA by model size (number of parameters) across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

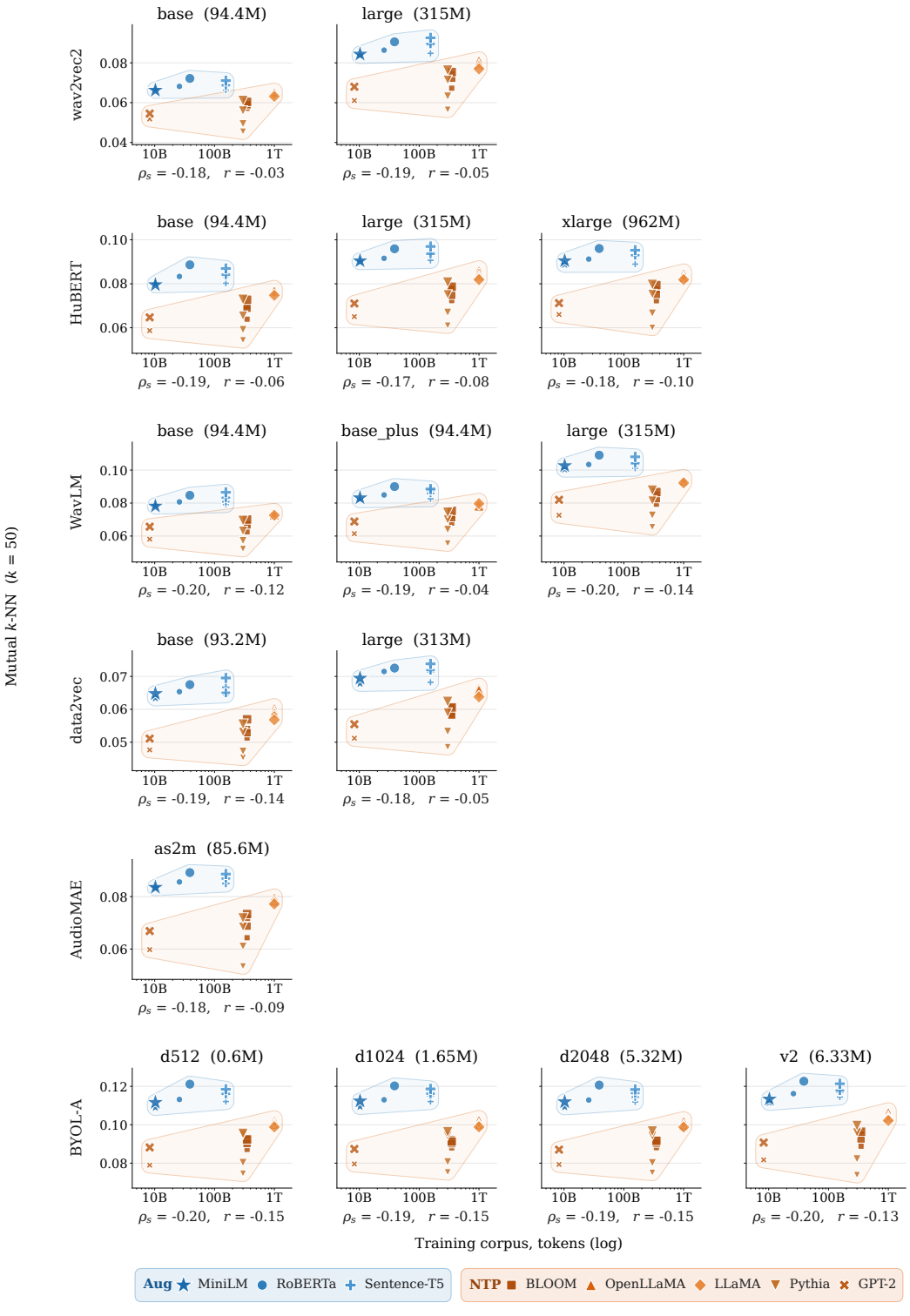


Figure 22: **CMA by corpus size across audio architectures on the CLOTHO dataset.** CMA by corpus size (tokens trained on) across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

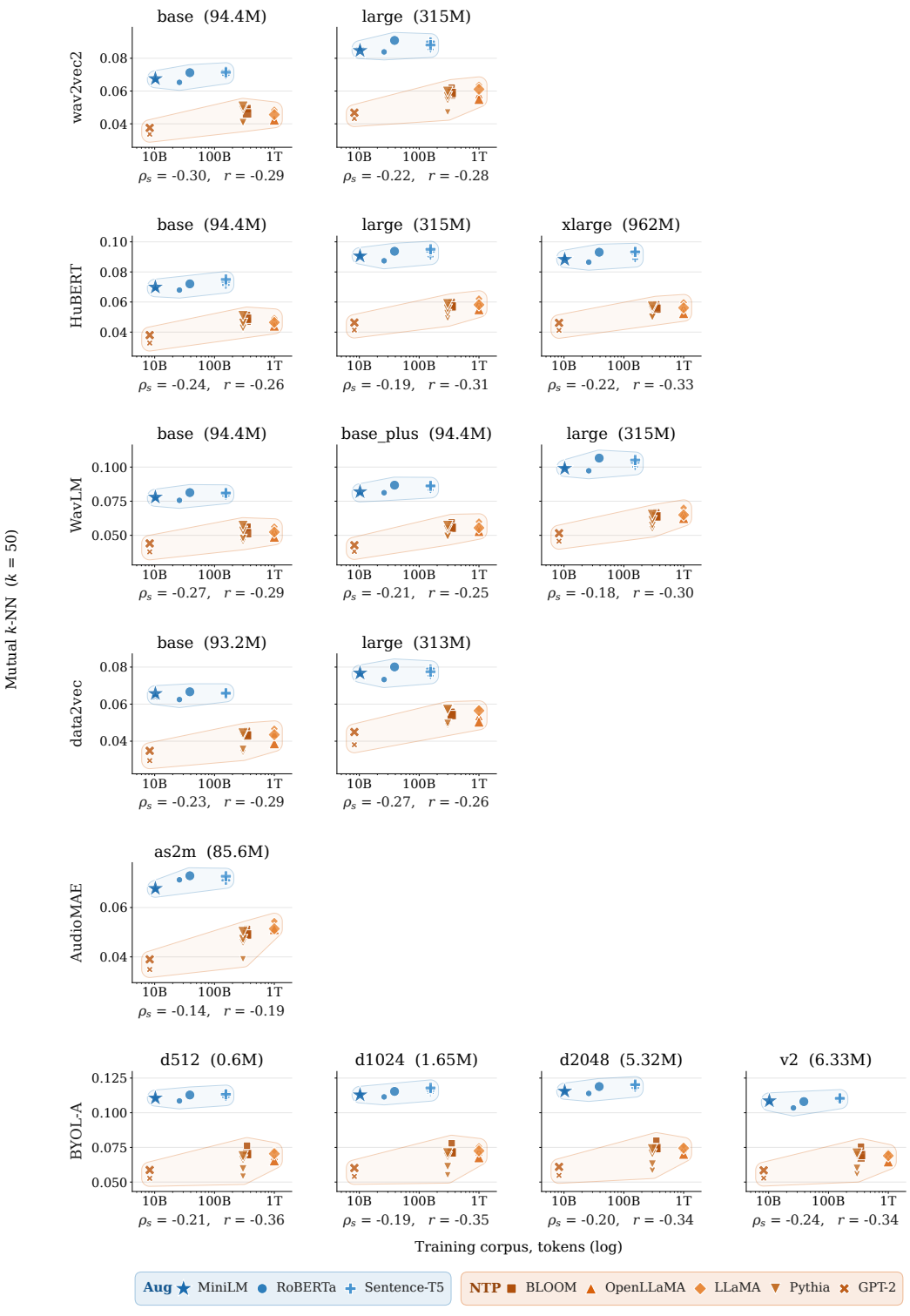


Figure 23: **CMA by corpus size across audio architectures on the AudioCaps dataset.** CMA by corpus size (tokens trained on) across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all models jointly, model/corpus sizes are uncorrelated with CMA, as indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

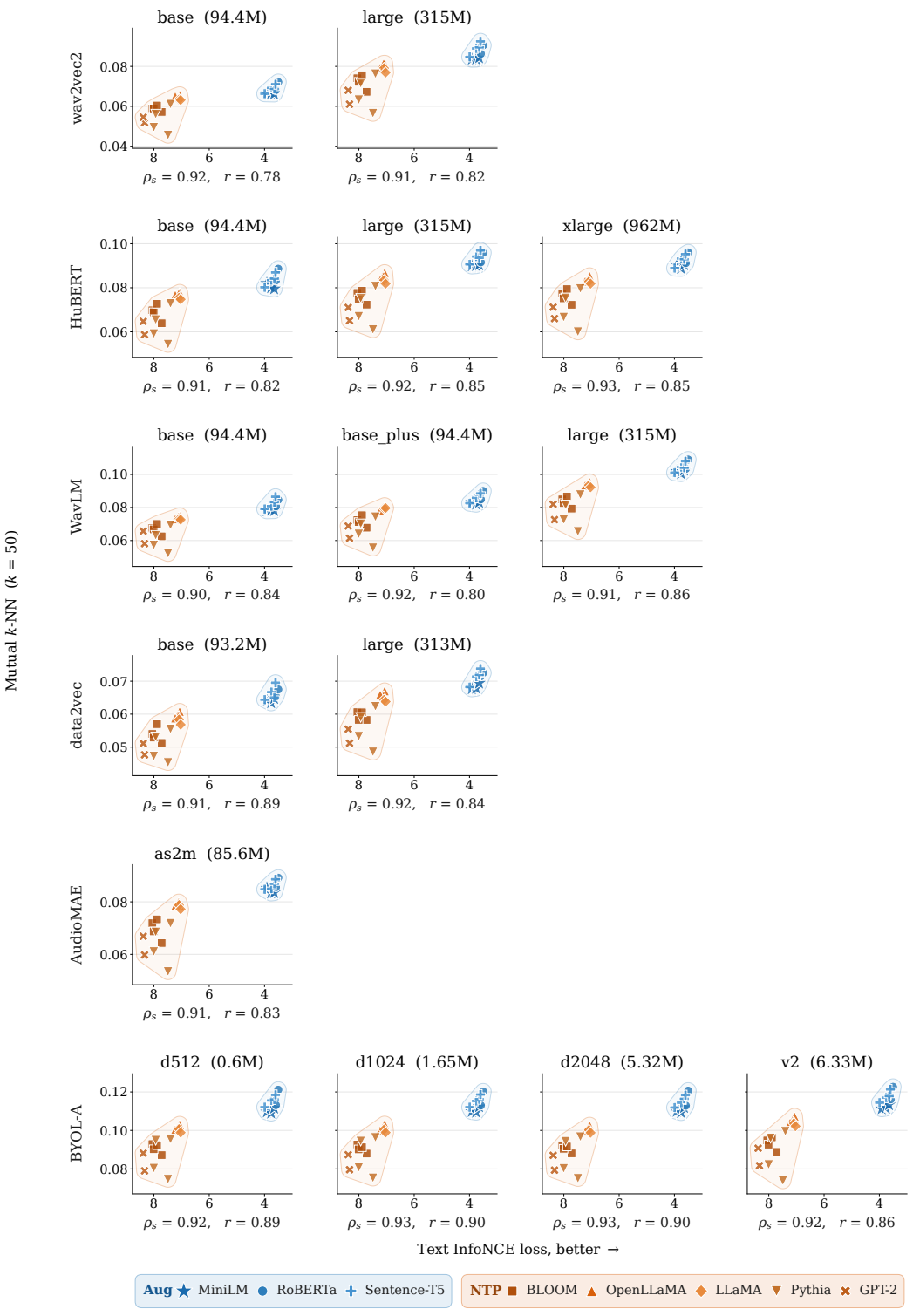


Figure 24: **CMA by InfoNCE across audio architectures on the CLOTHO dataset.** CMA by unimodal text InfoNCE loss across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

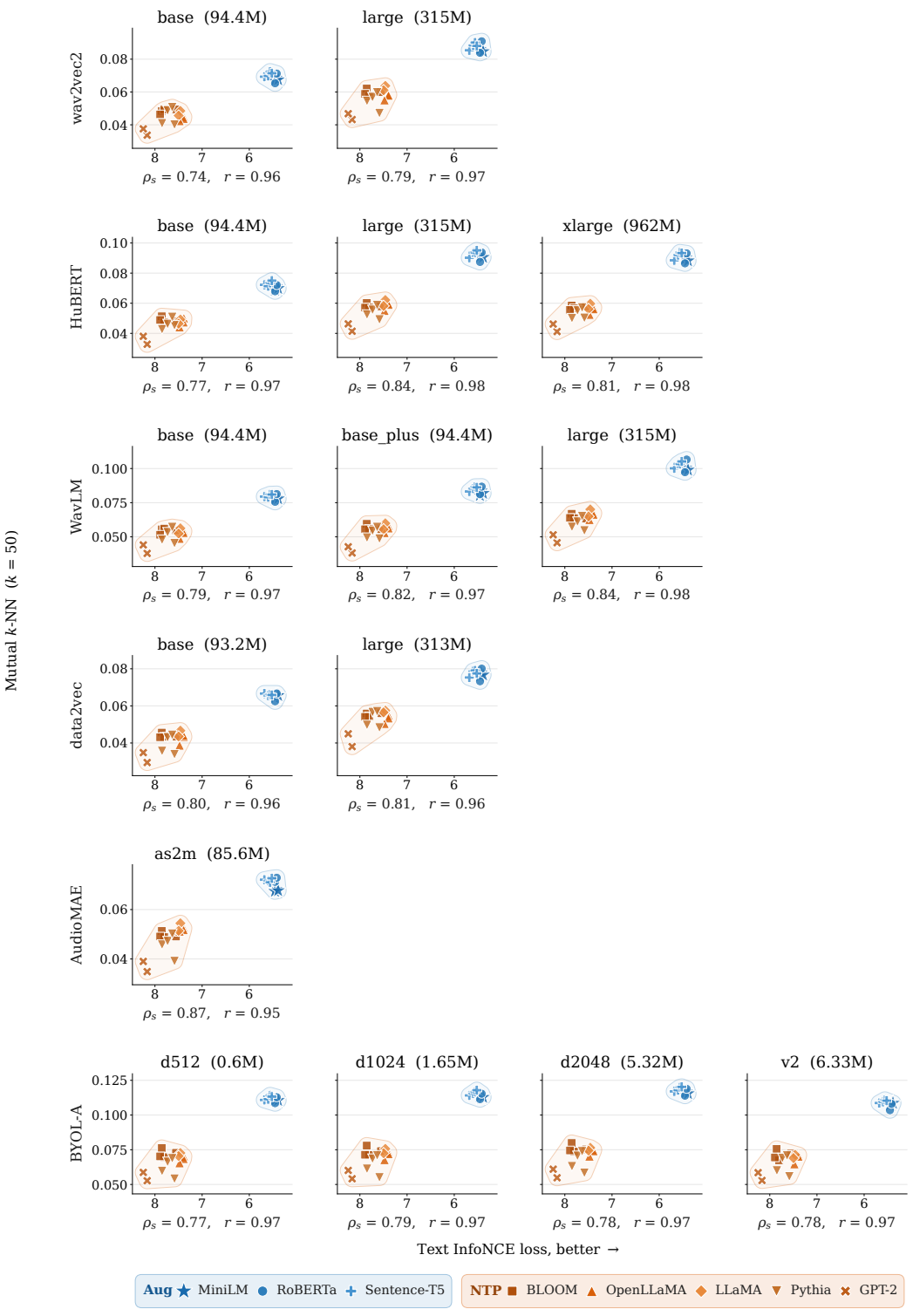


Figure 25: **CMA by InfoNCE across audio architectures on the AudioCaps dataset.** CMA by unimodal text InfoNCE loss across different audio encoder families (wav2vec2, HuBERT, WavLM, data2vec, AudioMAE, and BYOL-A) at various capacity scales. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

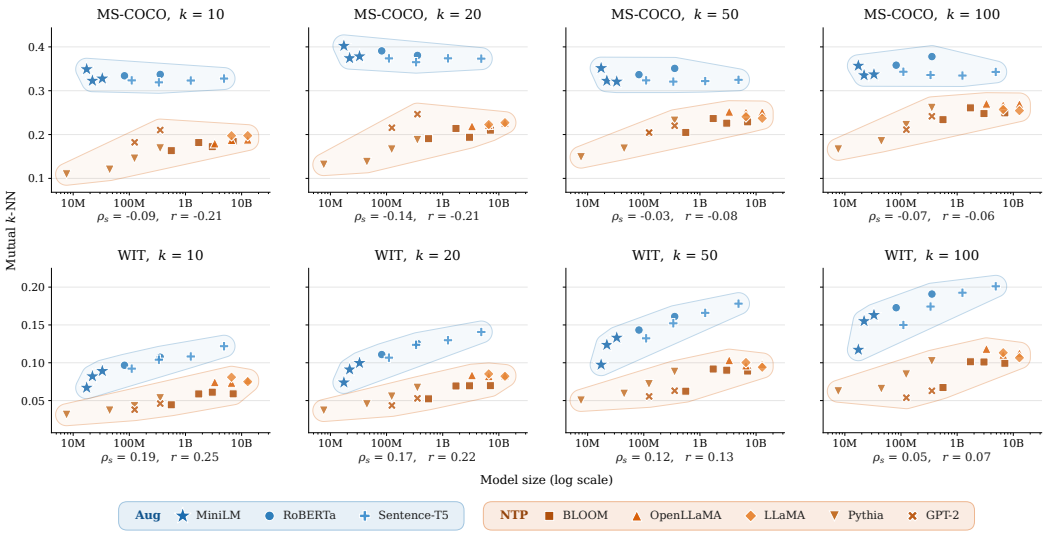


Figure 26: **Mutual k-NN across k values on the COCO and WIT datasets.** Mutual k-NN by model size (number of parameters) across different k values for the metric. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all k values jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

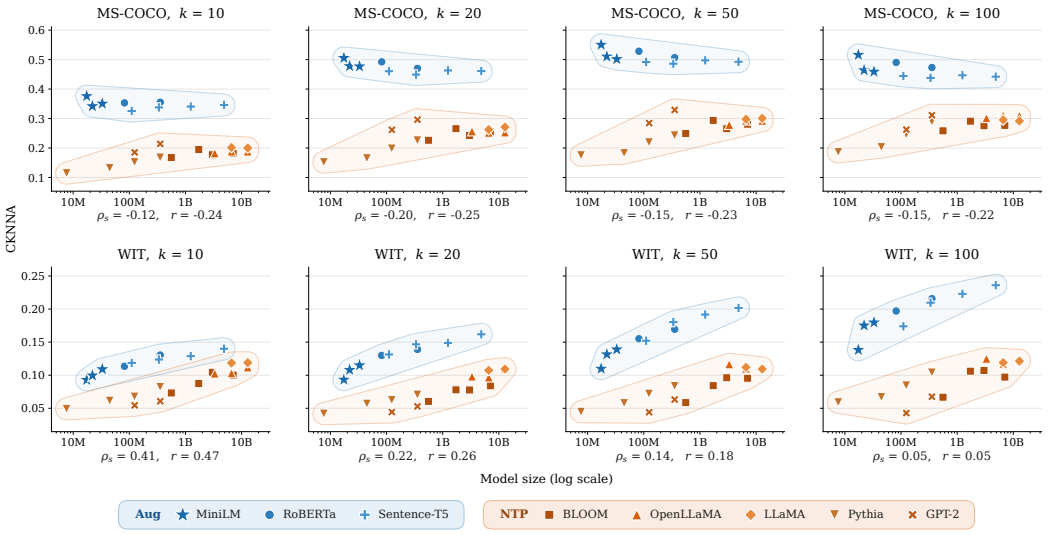


Figure 27: **CKNNA across k values on the COCO and WIT datasets.** CKNNA by model size (number of parameters) across different k values for the metric. Notably, Aug models consistently achieve higher CMA, even at substantially smaller model and corpus sizes. Also, considering all k values jointly, model/corpus sizes are uncorrelated with CMA, indicated by both weak Spearman (ρ_s) and Pearson (r) correlations.

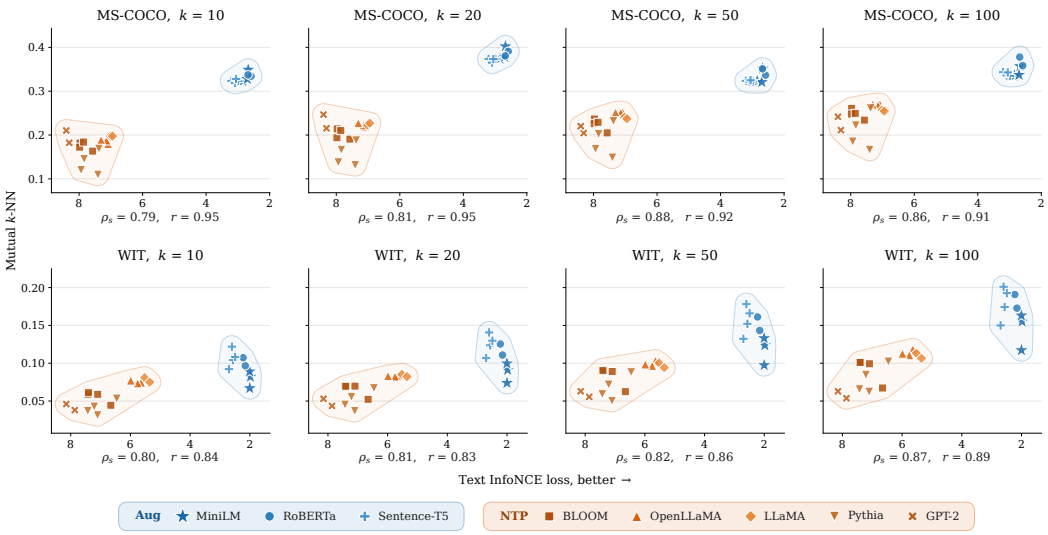


Figure 28: **Mutual k-NN by InfoNCE across k values on the COCO and WIT datasets.** Mutual k-NN by unimodal text InfoNCE loss across different k values for the metric. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

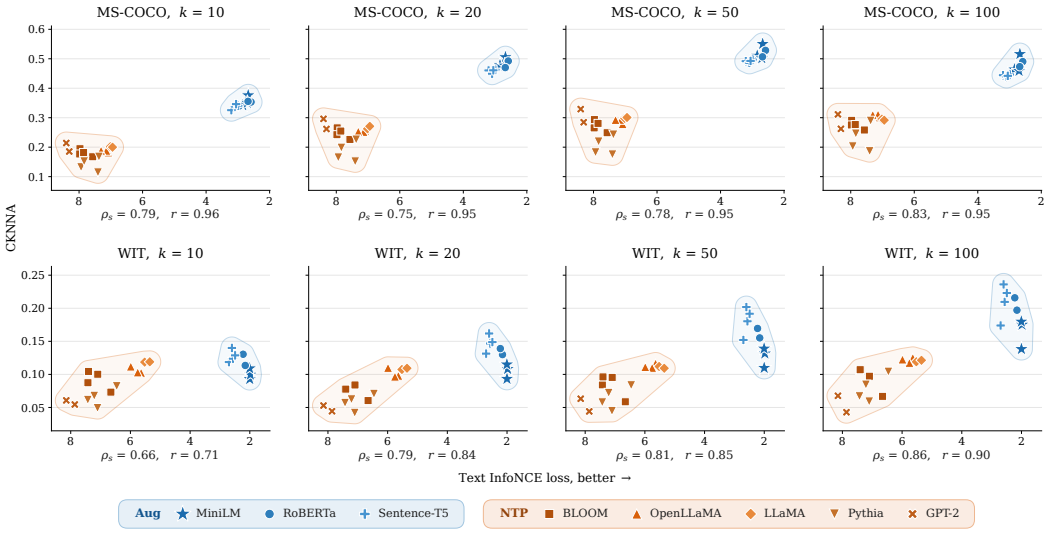


Figure 29: **CKNNA by InfoNCE across k values on the COCO and WIT datasets.** CKNNA by unimodal text InfoNCE loss across different k values for the metric. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

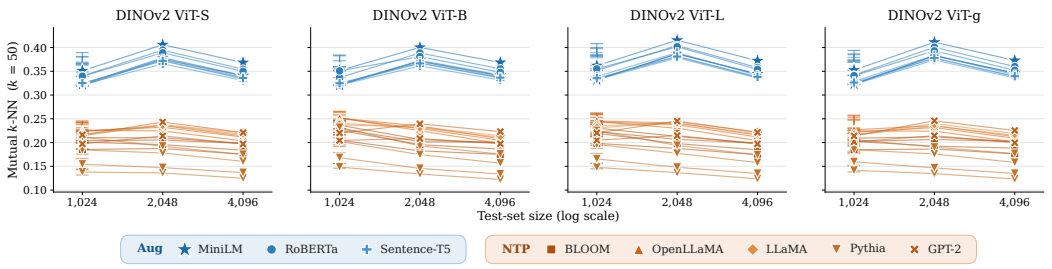


Figure 30: **CMA by test set size across capacity scales on the COCO dataset.** CMA by test size across different DINOv2 capacity scales. Notably, all text models keep about the same relation across different test set size.

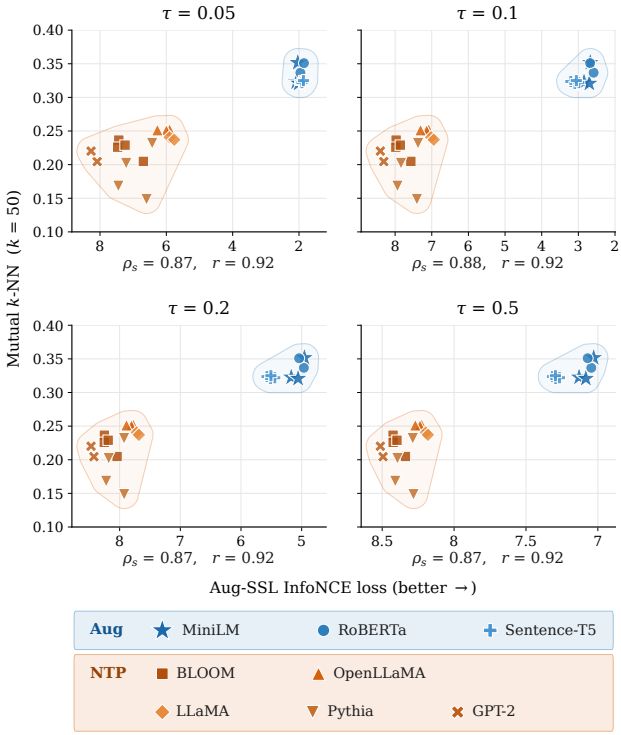


Figure 31: **CMA by InfoNCE across different temperature on the COCO dataset.** CMA by unimodal text InfoNCE loss across different temperature for the InfoNCE loss. Notably, the unimodal Aug-SSL loss is highly correlated with CMA, very consistent across different temperature, clearly separating low- and high-alignment models, and having large Spearman (ρ_s) and Pearson (r) correlations.

J CKA-RBF AND MkNN

Fig. 32 extends Fig. 2b from Sec. 3, evaluating the correlation between Mutual k-NN and CKA-RBF, to different k values. The strong correlation between Mutual k-NN and CKA-RBF remains robust regardless of the specific k value, while adapting the scale σ accordingly. This confirms that CKA-RBF can be viewed as a continuous relaxation of the Mutual k-NN.

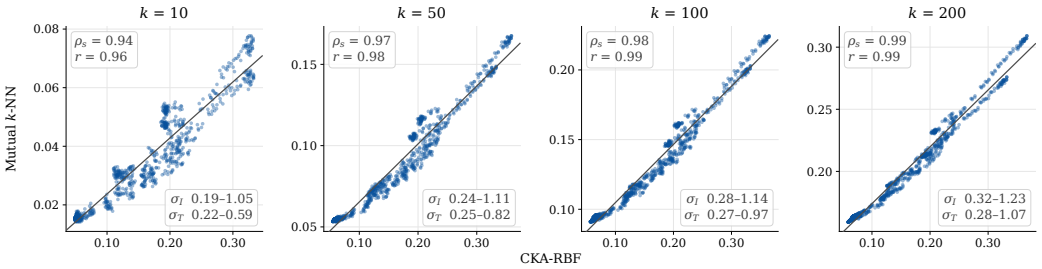


Figure 32: **Mutual k-NN by CKA-RBF across k values on the COCO dataset.** Mutual k-NN by CKA-RBF across different k values for Mutual k-NN. Notably, Mutual k-NN is highly correlated with CKA-RBF, very consistent across different k, and having large Spearman (ρ_s) and Pearson (r) correlations.

K Visualizing the Bounds

This section empirically visualizes the bounds from Thm. 1 and Thm. 2 across different parameter options.

Fig. 33 visualizes the bounds in $\tau = 0.3$ and $\sigma^2 = 1$. As depicted in Fig. 33a, the lower bound is effective for sufficiently low losses, and in Fig. 33b, the upper bound is effective for sufficiently high losses. The latter confirms that low unimodal augmentation-based losses guarantee alignment, whereas only one high loss prevents alignment.

Fig. 34 and 35 extends 33 to different values of τ and σ^2, on Thm. 1 and Thm. 2, respectively. This show robustness of the bounds across loss and CMA parameters.

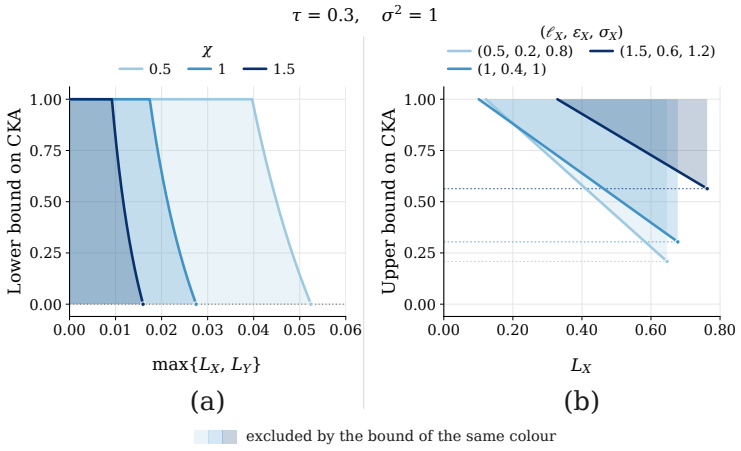


Figure 33: **Visualization of the (a) lower and (b) upper bounds from Thm. 1 and 2.**

L Implementation Details

This section specifies the experimental configurations used in Sec. 4.

L.1 Models

Pre-trained encoders. Tables 4, 5, and 6 summarize the 24 text, 14 vision, and 15 audio frozen encoders used in our experiments. L denotes the number of hidden layers, excluding the embedding output, and d denotes the corresponding representation width. The model

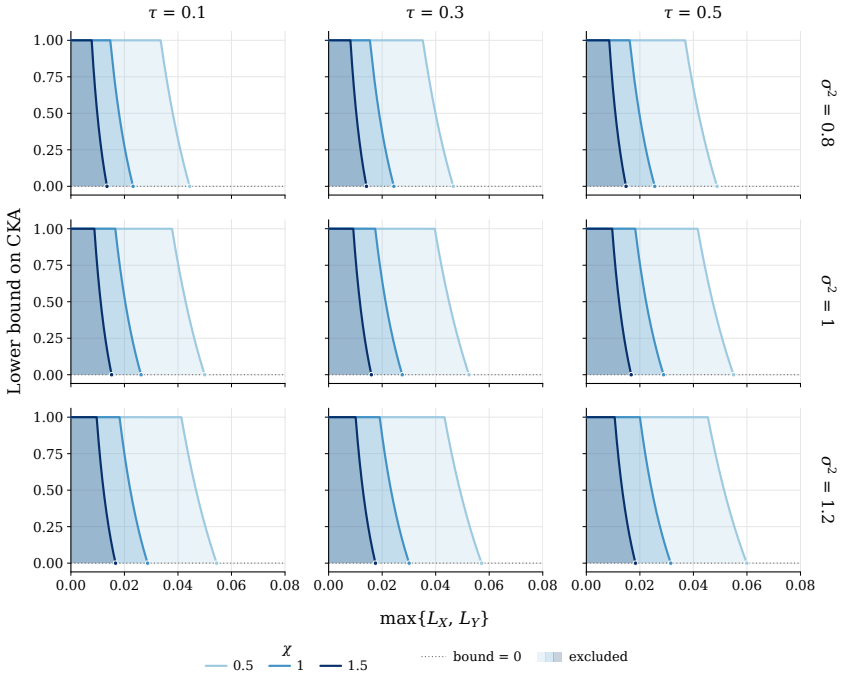


Figure 34: **Extended visualization of the lower bound from Thm. 1,** across different choices of τ and σ^2.

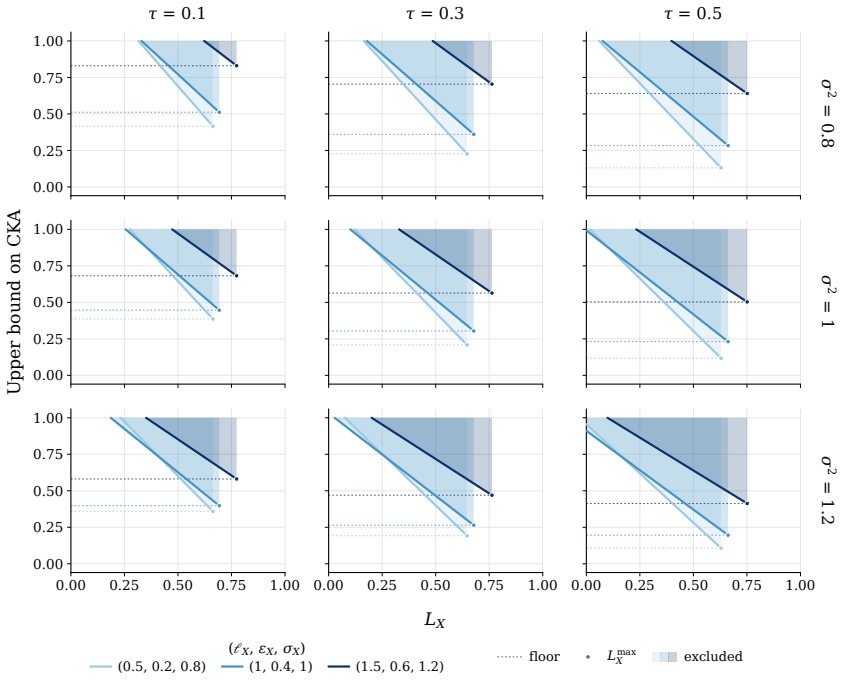


Figure 35: **Extended visualization of the upper bound from Thm. 2,** across different choices of τ and σ^2.

Table 4: **The 24 frozen text encoders.** Aug denotes augmentation-based sentence encoders and NTP denotes next-token-prediction language models. Corpus size is measured as the number of unique tokens across all training stages, as described below.

Checkpoint	Family	Type	Params	L / d	Corpus (tokens)
sentence-transformers/paraphrase-MiniLM-L3-v2	MiniLM	Aug	17.4 M	3 / 384	-
sentence-transformers/all-MiniLM-L6-v2	MiniLM	Aug	22.0 M	6 / 384	1.03×10^{10}
sentence-transformers/all-MiniLM-L12-v2	MiniLM	Aug	33.0 M	12 / 384	1.03×10^{10}
sentence-transformers/all-distilroberta-v1	RoBERTa	Aug	82.0 M	6 / 768	2.59×10^{10}
sentence-transformers/all-roberta-large-v1	RoBERTa	Aug	355.4 M	24 / 1024	3.87×10^{10}
sentence-transformers/sentence-t5-base	Sentence-T5	Aug	110.2 M	12 / 768	1.55×10^{11}
sentence-transformers/sentence-t5-large	Sentence-T5	Aug	335.7 M	24 / 1024	1.55×10^{11}
sentence-transformers/sentence-t5-xl	Sentence-T5	Aug	1.24 B	24 / 1024	1.55×10^{11}
sentence-transformers/sentence-t5-xxl	Sentence-T5	Aug	4.87 B	24 / 1024	1.55×10^{11}
bigscience/bloom-560m	BLOOM	NTP	559.2 M	24 / 1024	3.50×10^{11}
bigscience/bloom-1b7	BLOOM	NTP	1.72 B	24 / 2048	3.50×10^{11}
bigscience/bloom-3b	BLOOM	NTP	3.00 B	30 / 2560	3.50×10^{11}
bigscience/bloom-7b1	BLOOM	NTP	7.07 B	30 / 4096	3.50×10^{11}
openlm-research/open_llama_3b	OpenLLaMA	NTP	3.32 B	26 / 3200	1.00×10^{12}
openlm-research/open_llama_7b	OpenLLaMA	NTP	6.61 B	32 / 4096	1.00×10^{12}
openlm-research/open_llama_13b	OpenLLaMA	NTP	12.85 B	40 / 5120	1.00×10^{12}
LLaMA-1 7B	LLaMA	NTP	6.61 B	32 / 4096	1.00×10^{12}
LLaMA-1 13B	LLaMA	NTP	12.85 B	40 / 5120	1.00×10^{12}
EleutherAI/pythia-14m	Pythia	NTP	7.6 M	6 / 128	3.00×10^{11}
EleutherAI/pythia-70m	Pythia	NTP	44.7 M	6 / 512	3.00×10^{11}
EleutherAI/pythia-160m	Pythia	NTP	123.7 M	12 / 768	3.00×10^{11}
EleutherAI/pythia-410m	Pythia	NTP	353.8 M	24 / 1024	3.00×10^{11}
openai-community/gpt2	GPT-2	NTP	124.4 M	12 / 768	8.25×10^{9}
openai-community/gpt2-medium	GPT-2	NTP	354.8 M	24 / 1024	8.25×10^{9}

Table 5: **The 14 frozen vision encoders.** All models are loaded through `timm`. Corpus size denotes the number of training images. Fig. 4 and Fig. 5a use `vit_base_patch14_dinov2`; the full 14-model comparison is reported in App. E.

`timm` checkpoint	Family	Params	L / d	Corpus (images)
vit_tiny_patch16_224.augreg_in21k	ImageNet21K	9.74 M	12 / 192	ImageNet-21k, 1.42×10^{7}
vit_small_patch16_224.augreg_in21k	ImageNet21K	30.08 M	12 / 384	ImageNet-21k, 1.42×10^{7}
vit_base_patch16_224.augreg_in21k	ImageNet21K	102.60 M	12 / 768	ImageNet-21k, 1.42×10^{7}
vit_large_patch16_224.augreg_in21k	ImageNet21K	325.69 M	24 / 1024	ImageNet-21k, 1.42×10^{7}
vit_base_patch16_224.mae	MAE	85.80 M	12 / 768	ImageNet-1k, 1.28×10^{6}
vit_large_patch16_224.mae	MAE	303.30 M	24 / 1024	ImageNet-1k, 1.28×10^{6}
vit_huge_patch14_224.mae	MAE	630.76 M	32 / 1280	ImageNet-1k, 1.28×10^{6}
vit_small_patch14_dinov2.lvd142m	DINOv2	22.06 M	12 / 384	LVD-142M, 1.42×10^{8}
vit_base_patch14_dinov2.lvd142m	DINOv2	86.58 M	12 / 768	LVD-142M, 1.42×10^{8}
vit_large_patch14_dinov2.lvd142m	DINOv2	304.37 M	24 / 1024	LVD-142M, 1.42×10^{8}
vit_giant_patch14_dinov2.lvd142m	DINOv2	1.14 B	40 / 1536	LVD-142M, 1.42×10^{8}
vit_base_patch16_clip_224.laion2b	CLIP	86.19 M	12 / 768	LAION-2B-en, 2.32×10^{9}
vit_large_patch14_clip_224.laion2b	CLIP	303.97 M	24 / 1024	LAION-2B-en, 2.32×10^{9}
vit_huge_patch14_clip_224.laion2b	CLIP	632.08 M	32 / 1280	LAION-2B-en, 2.32×10^{9}

`paraphrase-MiniLM-L3-v2` has no published corpus-size estimate. We therefore exclude it from analyses involving corpus size.

Corpus sizes. We obtained the corpus-sizes in Tables 4 and 5 from published sources and classified them as exact when the source reports the corresponding count, or approximate when the value is derived from a training schedule or byte count. For NTP encoders, corpus sizes reported in bytes are converted using 4.85 bytes per BPE token. For MiniLM students, the reported 3.3B-word corpus is converted using 1.30 WordPiece tokens per word. For augmentation-based encoders, corpus size includes the backbone pre-training corpus and the distinct examples encountered during the contrastive training stage. We estimate the latter using the average: 2.078 texts per training example and 53.9 tokens per text. This quantifies the number of distinct tokens encountered. Sentence-T5 is the only family for which the number of contrastive training steps is unpublished; we therefore use GTR's 800k steps for its corresponding 2×10^{9}-pair training stage. Vision corpus size is measured as the number of unique training images.

Table 6: **The 15 frozen audio encoders.** d denotes the widest read-out block. BYOL-A models are convolutional and use unequal block widths: 2048/1024/512/512/512 (`d512`), 2048/1024/512/1024/1024 (`d1024`), 2048/1024/512/2048/2048 (`d2048`), and 2048/1024/2048/2048/3072 (v2). BYOL-A checkpoints are obtained from the authors' repository.

Checkpoint	Family	Objective	Params	L / d
`facebook/wav2vec2-base`	wav2vec2	contrastive	94.37 M	12 / 768
`facebook/wav2vec2-large-lv60`	wav2vec2	contrastive	315.44 M	24 / 1024
`facebook/hubert-base-ls960`	HuBERT	masked	94.37 M	12 / 768
`facebook/hubert-large-ll60k`	HuBERT	masked	315.44 M	24 / 1024
`facebook/hubert-xlarge-ll60k`	HuBERT	masked	962.50 M	48 / 1280
`microsoft/wavlm-base`	WavLM	masked	94.38 M	12 / 768
`microsoft/wavlm-base-plus`	WavLM	masked	94.38 M	12 / 768
`microsoft/wavlm-large`	WavLM	masked	315.45 M	24 / 1024
`facebook/data2vec-audio-base`	data2vec	masked	93.16 M	12 / 768
`facebook/data2vec-audio-large`	data2vec	masked	313.28 M	24 / 1024
`AudioNTT2020-BYOLA-64x96d512`	BYOL-A	augmentation	0.60 M	5 / 2048
`AudioNTT2020-BYOLA-64x96d1024`	BYOL-A	augmentation	1.65 M	5 / 2048
`AudioNTT2020-BYOLA-64x96d2048`	BYOL-A	augmentation	5.32 M	5 / 2048
`AudioNTT2022-BYOLA-64x96d2048`	BYOL-A	augmentation	6.33 M	5 / 3072
`gaunernst/vit_base_patch16_1024_128.audiomae_as2m`	AudioMAE	masked	85.65 M	12 / 768

Table 7: **Embedding extraction.** All hidden layers are retained. Audio is converted to mono and resampled to 16 kHz, then divided into non-overlapping 10 s windows. Each window is embedded independently and the resulting vectors are averaged using weights proportional to the number of valid samples in each window. Clotho clips contain an average of 2.647 windows, whereas AudioCaps clips contain exactly one. The waveform-transformer models receive normalized waveforms; BYOL-A receives 64-band log-mel spectrograms normalized using evaluation-set statistics (mean -5.4703, standard deviation 4.3999); AudioMAE receives a 1024×128 spectrogram represented using a 64×8 patch grid.

Modality	Read-out per layer	Input
Text (Aug)	mean over real tokens (SBERT pooling)	128 tokens
Text (NTP)	mean over real tokens	128 tokens
Vision	CLS at the second residual add	each model's own preprocessing
Audio	mean over time (SUPERB convention)	10 s windows at 16 kHz

Embedding extraction. Table 7 summarizes the extraction settings.

Controlled architectures and training. Table 8 specifies the controlled architectures and training schedules. Both training objectives operate on the 128-dimensional projection-head representations, and alignment is evaluated in the same space. The vision models used a common preprocessing pipeline consisting of resizing the short side to 256 pixels, centre cropping to 224×224, and ImageNet normalization. The curves in Fig. 3 and the corresponding appendix figures report the mean $\pm$ one sample standard deviation over three training seeds. Alignment is evaluated on a fixed held-out draw of 4096 of the COCO 2017 validation set pairs.

Loss maximization and chance level. The loss-maximizing condition differs from the corresponding minimizing condition only in the sign of the backward pass. That is, initialization, mini-batch stream, optimization schedule, and gradient clipping remain unchanged. For mutual k-NN overlap, the chance level is the expected overlap of two independent k-neighborhoods among n instances, which is $k/(n-1)$. For $k = 50$ and $n = 4096$, this gives $50/4095 = 0.0122$.

Multi-positive InfoNCE objective. Equation 3 contains one positive per anchor, whereas our controlled experiments use five views per item (Tian et al., 2023). Let a batch contain B items with V views per item, let $z_1, \ldots, z_{BV}$ denote the unit-normalized projection-head outputs, and let $P(i)$ denote the $V - 1$ other views of the same item as anchor i. All views belonging to other items are

Table 8: **The controlled encoders.** Both models use AdamW with linear warmup followed by cosine decay to zero, 15 000 training steps, batches of 256 items with five views per item (1280 views per step), $\tau = 0.1$ as in Eq. 3, bfloat16 autocasting on compute capability ≥ 8.0, and float16 with gradient scaling otherwise. Contrastive logits are always computed in float32. No EMA is used. Training uses seeds 5000, 5001, and 5002.

	Vision encoder (ViT)	Text encoder (SBERT)
Backbone	ViT-S/16, randomly initialised	`bert-base-uncased`, mean pooling
Architecture	12 layers, 6 heads, width 384, MLP 1536	12 layers, width 768
Input	224^2, patch 16	40 tokens
Projection head	$384 \to 384 \to 128$ with GELU	$768 \to 768 \to 128$ with GELU
Learning rate	5×10^{-4}	2×10^{-5}
Weight decay	0.05	0.01
Warmup steps	1000	300
Gradient clip	none	global norm 1.0

treated as negatives. The resulting objective is

$$\ell^{(V)} \;=\; -\frac{1}{BV}\sum_{i=1}^{BV}\frac{1}{V-1}\sum_{j\in P(i)}\log\frac{\exp\big(\langle z_i, z_j\rangle/\tau\big)}{\sum_{l\neq i}\exp\big(\langle z_i, z_l\rangle/\tau\big)}. \tag{30}$$

We use $V = 5$. When $V = 2$, the objective reduces to Eq. 3 with $K = 2B - 2$.

Evaluating the unimodal loss of frozen encoders. The loss shown on the x-axis of Fig. 5 is Eq. 30 evaluated on the frozen representations of the five views in the evaluation draw. The complete evaluation draw is processed jointly rather than in mini-batches, so each anchor is compared against all other rows. With $n = 1024$ and $V = 5$, this gives $K = nV - 1 = 5119$ negatives. We use $\tau = 0.1$. For each encoder, the reported loss is the minimum across layers, mirroring the maximum used for alignment: both read out the model at its best layer. The minimum keeps the loss a property of the encoder alone, identical across all panels. The corresponding chance loss is $\log(nV - 1) = 8.54$ nats.

Supervised baselines. MS-COCO instance annotations are reduced to the largest object per image among the 80 categories. The cross-entropy baseline uses a single $\mathrm{Linear}(384, 80)$ classification head on the backbone representation. SupCon uses the same 128-dimensional projection head as the augmentation-based SSL objective, with $\tau = 0.1$; all same-label views in a batch are treated as positives, while anchors without another sample of the same label are omitted. All three objectives are trained from random initialization on the same training pool for 150 epochs (23,250 steps), using batch size 256, learning rate 5×10^{-4}, weight decay 0.05, and 775 warmup steps. Thus, the three baselines differ only in their objectives

To control for label memorization, classes are partitioned according to COCO supercategories. The *animal*, *furniture*, *food*, and *kitchen* categories constitute the seen set of 33 classes (39,867 images); the remaining 46 classes are held out. The *person* category is omitted. Fig. 3b evaluates CMA on the 1,716 validation images belonging to held-out classes.

Alignment metrics. The bandwidth of the CKA-RBF, σ, is determined for each modality as the median distance to the k-th nearest neighbor across samples.

Layer aggregation and null calibration. We follow Gröger et al. (2026a). For frozen encoders, the reported alignment score is the maximum across all image–text layer pairs. Since this maximum is affected by the number of layers, we calibrate each encoder pair as presented in Gröger et al. (2026a).

L.2 Datasets

Evaluation draw. Unless otherwise stated, each dataset is evaluated using a fixed subsample of $n = 1024$ aligned pairs drawn without replacement using a global seed. The same draw is reused

across experimental stages. The larger evaluation sets in App. H are nested supersets, $1024 \subset 2048 \subset 4096$, so that the sample-size analysis varies only n.

MS-COCO. Evaluation uses `val2017`. Images with all five captions provide 5,000 image–caption pairs; 4,952 of these images contain an instance annotation and therefore have an associated label. The evaluation draw is sampled from these images. Captions are ordered by annotation ID, and the first caption is used. Controlled experiments use `train2017`, containing 118,287 images.

WIT. Evaluation uses the `wit_1024` benchmark split of Huh et al. (2024), containing 1,024 image–text pairs with one caption per image. For controlled training, we construct a COCO-sized corpus from `wikimedia/wit_base` by retaining the first English reference caption for each image. Benchmark images are removed from the training corpus by caption matching. The remaining data are split using a seeded shuffle into 118,287 training images and 5,000 validation images, matching the sizes of `train2017` and `val2017`.

Clotho. We use Clotho v2.1 (Zenodo `10.5281/zenodo.4783391`), specifically its evaluation split containing 1,045 clips and 5,225 captions. Clips have durations between 15 and 30 seconds (mean 22.5 seconds) and a native sampling rate of 44.1 kHz. We sample 1,024 clips for evaluation. Audio is converted to mono, resampled to 16 kHz, and stored. Each clip has five human captions in a fixed column order, which defines the view order, and we use the first.

AudioCaps. Among the 1,470 clips in the official validation and test splits, we use all 856 available valid clips, containing a caption and valid WAV files. Each is exactly 10.00 seconds long at 48 kHz and is downsampled by a factor of three to 16 kHz, yielding a single 10-second window per clip.

Image augmentations. For the controlled experiments and frozen-encoder image-loss evaluation, each image is transformed using the following pipeline: random resized crop to 224^2 with area scale $(0.5, 1.0)$; horizontal flip with probability 0.5; color jitter with brightness, contrast, and saturation 0.2 and hue 0.05, applied with probability one; and ImageNet normalization. Five independent draws produce the five views of each image. The label-free control in Fig. 3b uses the DINO-style multi-crop augmentation associated with its baseline. Each image produces two global crops with scale $(0.4, 1.0)$ and four local crops with scale $(0.05, 0.4)$, all resized to 224^2. Horizontal flipping is applied to all crops. Color jitter $(0.4, 0.4, 0.2, 0.1)$ is applied with probability 0.8, grayscale with probability 0.2, and Gaussian blur with $\sigma \in (0.1, 2.0)$ is applied to the first global crop with probability one, to the second global crop with probability 0.1, and to local crops with probability 0.5. Solarization is applied to the second global crop with probability 0.2.

Text augmentations. The main text views are generated using Qwen with a single user message and no system message. The generator receives only the source caption and has no access to the corresponding image, other captions, or labels. The prompt is:

```
You rewrite a caption.  Output only JSON.
Rewrite the caption below as EXACTLY <N> alternatives with
the same meaning.  Vary word choice, syntax and clause
order; the <N> must differ from each other and from the
original by more than a synonym swap.
- Keep the original's register and length (a plain caption,
8-15 words).  No commentary.
- No preamble, no numbering.
Return only a JSON array of exactly <N> strings.
Caption:  "<CAPTION>"
```

Here $<\texttt{N}> = 4$, so the source caption constitutes view 0 and the four generated captions constitute views 1–4. Decoding uses temperature 0.9, top-p 0.95, and a fixed generation seed. Outputs are required to contain exactly four English strings. Outputs failing this constraint are regenerated up to two times; after two failed attempts, the source caption is used as fallback. This filtering prevents off-language outputs from artificially increasing within-group variation. For WIT and Clotho, one prompt constraint is modified to account for the properties of their captions. For WIT, whose captions have a median length of seven words, the length instruction is replaced with: Match the origi-

nal's register and length, however short it is. Keep proper nouns, titles and technical terms verbatim. No commentary. For Clotho, the instruction is replaced with: Keep the original's register and length (a plain caption, 8-20 words). Keep the same events in the same order; do not add, drop or reorder one. No commentary. All remaining prompt components are unchanged. App. D repeats the experiment using `microsoft/Phi-3.5-mini-instruct` and `google/gemma-4-12B-it`.

Back-translation views. Back-translation provides an alternative text augmentation procedure that does not rely on prompting or stochastic decoding. View 0 is the source caption, while views 1–4 are obtained by translating the source through German, French, Spanish, and Russian, respectively, and translating each result back to English. The translations use eight `Helsinki-NLP/opus-mt` checkpoints, corresponding to the four pivot languages in both translation directions. Decoding uses beam search with four beams and no sampling. The resulting corpus is deterministic. Each view uses a different pivot language, producing four translation systems rather than four stochastic samples from a single system.

L.3 The Bounds

Evaluation of the bounds. The bound figures visualize the analytical bounds derived in Sec. 3. No encoder training or sampling is involved. The lower bound depends on both modality losses, but a single curve suffices because it is given by $\cos(\theta_X + \theta_Y)$ and decreases monotonically in each loss. Thus, the worst case under $\max \mathcal{L}_X, \mathcal{L}_Y \leq \mathcal{L}$ occurs on the diagonal. The upper bound depends only on the loss of one modality and is therefore plotted directly against $\mathcal{L}_X$. A dotted line marks $\mathrm{CKA} = 0$. The excluded corner of each bound is also indicated: the lower bound excludes the lower-left region, whereas the upper bound excludes the upper-right region. Each closed-form implementation is additionally verified against a 50-digit evaluation of the same expression.

Parameter grids and valid domains. Both bound figures use $\tau \in \{0.1, 0.3, 0.5\}$ and $\sigma^2 \in \{0.8, 1.0, 1.2\}$. The lower-bound panels additionally use $\chi \in \{0.5, 1, 1.5\}$. The upper-bound panels use the parameter triples

$$(\ell_X, \epsilon_X, \sigma_X) \in (0.5, 0.2, 0.8), (1, 0.4, 1.0), (1.5, 0.6, 1.2).$$

The side-by-side comparison uses the slice $\tau = 0.3$ and $\sigma^2 = 1$. Curves are plotted only over the region in which the corresponding bound is non-vacuous. The lower bound terminates at its zero crossing, below which it is weaker than the trivial constraint $\mathrm{CKA} \geq 0$. The upper bound is cut where it crosses one and terminates at its maximum possible value as in Lemma 8.

L.4 Hardware and Software

Hardware and Software. All experiments use Python on a single GPU without distributed training or gradient accumulation. GPU selection is performed at runtime. The reported experiments were executed on NVIDIA RTX PRO 6000 Blackwell (98 GB), B200 (183 GB), H200 (144 GB), A100 80 GB (PCIe and SXM4), L40S (46 GB), RTX 6000 Ada (49 GB), and L4 (23 GB) GPUs, with 8–32 CPU cores and 32–150 GB host memory per job.

Determinism and seeding. Evaluation uses fixed random seeds that are independent of the training seeds. These seeds determine evaluation subsampling, loss-evaluation blocks, kernel-statistic samples, alignment subsamples, and frozen augmentation draws. Consequently, different training runs share the same evaluation data and augmentation realizations and differ only through their training seeds. Caption generation uses a separate generation seed, while back-translation is deterministic and requires no random seed.