BEYOND OBJECTIVE EXPRESSIVITY: GEOMETRY PRESERVATION IN MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning is increasingly moving toward settings with three or more modalities instead of image-text pairs. Yet, extending models from pairwise to higher-order multimodal alignment can introduce optimization and representation challenges. We identify encoder Jacobian conditioning as a key factor in trimodal contrastive learning: poorly conditioned encoders exhibit collapsing or amplified singular-value spectra, leading to exploding Jacobian condition numbers and degraded multimodal alignment. We introduce geometry-preserving encoders (GPEs) by directly conditioning the Jacobian through regularization and demonstrating that simple modifications like LeakyReLU activations and residual paths recover these geometric benefits. Across a synthetic benchmark and four real-world datasets including missing modalities, improving Jacobian conditioning boosts retrieval and linear probe performance across multiple contrastive objectives, whereas expressive objectives yield little benefit in linear probes. More broadly, our results show that multimodal contrastive learning depends not only on objective expressivity, but also on the geometric and optimization properties of the underlying encoders.[1]

1 INTRODUCTION

Multimodal contrastive learning has emerged as a powerful method for aligning representations across modalities. While early work focused on bimodal vision-language models (VLMs) (Radford et al., 2021), many real-world applications involve more modalities. For example, in life sciences, modalities may cover demographics, clinical timeseries, laboratory measurements, free-text reports, and neural or behavioral recordings (Acosta et al., 2022; Schneider et al., 2023). Motivated by these settings, recent work has proposed higher-order contrastive objectives that move beyond pairwise alignment (Saporta et al., 2024; Cicchetti et al., 2025a;b; Dufumier et al., 2025). At the same time, multimodal learning relies on heterogeneous encoder families, ranging from ResNets (He et al., 2016) and Transformers (Vaswani et al., 2017) to Multi-Layer Perceptron (MLP)-based encoders (Rosenblatt, 1957), which are frequently used to process tabular modalities and map pretrained representations into shared embedding spaces. Different encoder families exhibit different optimization characteristics, and some encoders such as MLPs may be more sensitive to optimization pathologies, as highlighted in Figure 1. This raises a complementary question to objective design: how do the geometric properties of modality encoders affect multimodal contrastive learning? In contrast to the growing multimodal contrastive learning literature on objective design and representation geometry (Wang & Isola, 2020; Liang et al., 2022; Jing et al., 2022; Yi et al., 2025; Cai et al., 2026), relatively little attention has been paid to the geometry of the encoders that produce these representations (Golovanevsky et al.,

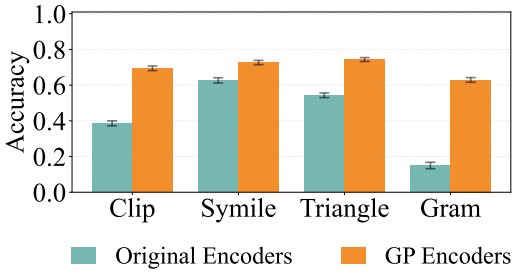


Figure 1: Well-tuned standard encoders vs. GPEs across multiple multimodal contrastive objectives on the UKB. GPEs improve retrieval performance for all evaluated contrastive objectives.

[1]The code repository is attached anonymously for review and will be uploaded to GitHub.

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

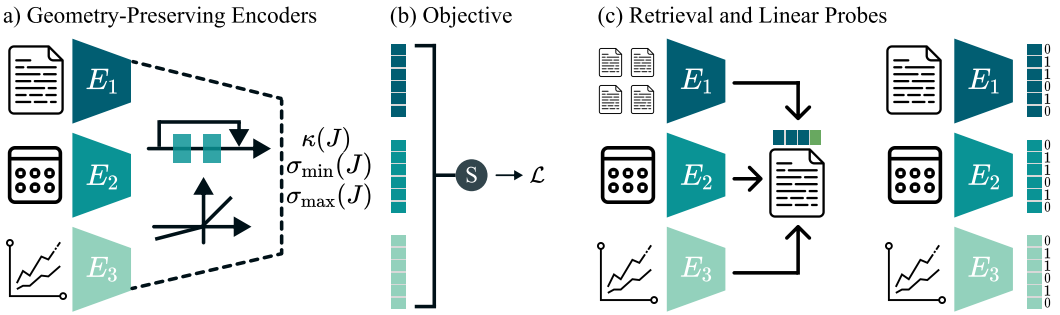


Figure 2: Multimodal contrastive learning with GPEs exemplified for MIMIC-IV (Johnson et al., 2023b;a). Radiology reports, demographics, and timeseries are processed by LoRA-finetuned RadBERT (Yan et al., 2022) and MLP encoders E_i. **(a)** GPEs maintain well-conditioned Jacobians through residual transport paths, and LeakyReLU activations, preventing the condition number κ defined by singular values $\sigma_{\min}$ and $\sigma_{\max}$ to explode. **(b)** GPEs can be combined with arbitrary multimodal contrastive objectives S, such as Clip. **(c)** The resulting models and representations are extensively tuned and evaluated on retrieval and downstream linear probing tasks. Our results demonstrate that preserving encoder geometry improves training dynamics, retrieval performance, and downstream linear probes across diverse multimodal contrastive objectives and datasets.

2025; Rheude et al., 2026b). Meanwhile, a large body of work has linked neural network trainability to Jacobian conditioning and dynamical isometry, showing that poorly conditioned Jacobians impair signal propagation, optimization, and representation learning (Saxe et al., 2014; Pennington et al., 2017; Burkholz & Dubatovka, 2019; Xiao et al., 2018; Tarnowski et al., 2019). We show that similar phenomena arise in multimodal contrastive learning: Alignment difficulties are characterized by exploding Jacobian condition numbers.

In this paper, we study encoder geometry in trimodal contrastive learning. We theoretically and empirically connect multimodal contrastive learning to ill-conditioned encoder Jacobians. This is particularly relevant when heterogeneous modalities are used. For example, tabular modalities might be encoded by MLPs, while radiology reports might be represented by a LoRA-finetuned BERT model followed by an MLP projection head (Figure 2). First, we test whether encoder Jacobian conditioning has a mechanistic effect on multimodal contrastive learning through Jacobian regularization. Then, we introduce geometry-preserving encoders (GPEs), a family of lightweight architectural interventions based on residual paths (He et al., 2016) and LeakyReLU activations (Maas et al., 2013) that provide similar geometric benefits at lower computational cost. Across a controlled synthetic benchmark (Rheude et al., 2026b), UK Biobank (UKB) cohorts with complete and missing modalities (Sudlow et al., 2015), and the MIMIC-IV benchmark (Johnson et al., 2023b;a), we show that preserving encoder geometry improves optimization stability, retrieval performance, and downstream linear probing for diverse contrastive objectives. Our results suggest that objective expressivity alone is insufficient: while increasingly expressive objectives primarily affect retrieval performance, preserving encoder geometry improves retrieval and downstream linear probes across objectives. Our contributions are summarized as follows:

- We identify encoder Jacobian conditioning as a key factor in multimodal contrastive learning and show that it influences the trainability of modality encoders.
- We introduce GPEs, *i.e.,* practical interventions based on residual paths and LeakyReLU activations that improve optimization stability.
- We demonstrate that ill-conditioned encoders appear across multiple contrastive objectives on both synthetic and real-world datasets including missing modalities.
- We show that GPEs consistently improve retrieval and downstream linear probes, whereas increasingly expressive multimodal objectives primarily improve retrieval.

2 Related Work

Multimodal Contrastive Learning Multimodal contrastive learning has become a central approach for learning representations across modalities with*, e.g.,* VLMs such as Clip (Radford et al., 2021) and Align (Jia et al., 2021). Subsequent work improved scalability and training formulations through approaches such as LiT (Zhai et al., 2022) and SigLIP (Zhai et al., 2023). More recent methods move beyond pairwise alignment by introducing higher-order or many-modal objectives, including Symile (Saporta et al., 2024), Triangle (Cicchetti et al., 2025a), CoMM (Dufumier et al., 2025), ConFu (Koutoupis et al., 2026), and Gram (Cicchetti et al., 2025b). Further, relatively little attention has been paid to the geometry and optimization properties of the encoders themselves with approaches like Gated Symile (Rheude et al., 2026b) and PiCME (Golovanevsky et al., 2025) highlighting the closest related work. In contrast, we study encoder Jacobian conditioning across multimodal contrastive objectives, evaluate both retrieval and linear probes, and show that expressive objectives can still underperform when the encoders producing the representations become poorly conditioned.

Geometry, Dynamical Isometry, and Signal Propagation Neural network optimization is commonly linked to the singular value spectrum of network Jacobians. Dynamical isometry studies show that networks train more reliably when Jacobian spectra remain well-conditioned throughout depth (Saxe et al., 2014; Pennington et al., 2017). Related work has investigated how initialization schemes can promote favorable Jacobian spectra and dynamical isometry (Pennington et al., 2017), how activation functions influence signal propagation and trainability (Burkholz & Dubatovka, 2019), how Jacobian regularization can improve local representation stability (Rifai et al., 2011), how network Jacobians relate to robustness and generalization (Sokolic et al., 2017), and how architectural design choices affect optimization dynamics (Hoffman et al., 2019). Further, Jacobian spectra have been connected to feature learning dynamics (Xiao et al., 2018), the effectiveness of Jacobian regularization (Xue et al., 2022), the role of data structure and activation functions in neural network learning dynamics (Sonthalia et al., 2026), and alignment stability in multimodal models (Garg et al., 2026). Separately, contrastive learning research has studied alignment and uniformity of learned representations (Wang & Isola, 2020), dimensional collapse in contrastive embeddings (Jing et al., 2022), modality-gap formation and multimodal representation structure (Liang et al., 2022; Yi et al., 2025), alignment under missing modalities (Poklukar et al., 2022), and geometric analyses of contrastive objectives (Cai et al., 2026). In contrast, we study the conditioning of modality-specific encoder Jacobians in multimodal contrastive learning and show its impact on both retrieval and downstream representation quality.

Residual Architectures and Optimization Stability Residual connections are central to modern deep architectures because they provide identity transport paths that stabilize optimization and gradient flow (He et al., 2016). Correspondingly, prior work has shown that residual networks can preserve favorable input-output Jacobian spectra and achieve dynamical isometry across a broad range of activation functions (Tarnowski et al., 2019). This principle underlies residual architectures such as Transformer architectures built around skip connections (Vaswani et al., 2017), and residual MLPs such as ResMLPs (Touvron et al., 2023). Related stabilization methods include learnable residual gating through ReZero (Bachlechner et al., 2021), initialization schemes such as Fixup (Huang et al., 2020), and normalization strategies such as DeepNorm (Wang et al., 2024). More recently, Ji et al. (2025) argue that skip connections are a primary driver of Transformer trainability, potentially more important than the expressivity of the attention blocks themselves. Complementing this line of work, we study how encoder conditioning affects multimodal contrastive learning. Through direct Jacobian regularization, we provide interventional evidence that improving encoder conditioning stabilizes optimization and demonstrate that residual transport provides a practical and computationally efficient mechanism for achieving similar geometric benefits.

3 Method

3.1 Multimodal Contrastive Learning

We consider trimodal observations (x_1, x_2, x_3) analogous to prior work on multimodal contrastive learning (Section 2) and modality-specific encoders

$$E_i : \mathcal{X}_i \to \mathbb{R}^d, \qquad i \in \{1, 2, 3\}. \tag{1}$$

The resulting modality representations are $z_i = E_i(x_i)$. A multimodal contrastive objective assigns a compatibility score $s(z_1, z_2, z_3)$ to aligned triples and contrasts it against mismatched triples sampled from the batch. Most existing work focuses on increasing the expressivity of the scoring function s. In contrast, we investigate a complementary question: how does the geometry of the modality encoders E_i influence whether a multimodal contrastive objective can be optimized successfully?

3.2 Encoder Geometry and Jacobian Conditioning

The geometry induced by encoder E_i is defined by its input-output Jacobian

$$J_i(x) = \frac{\partial E_i(x)}{\partial x}. \tag{2}$$

Let $\sigma_{\min}(J_i)$ and $\sigma_{\max}(J_i)$ denote the smallest and largest singular values of J_i, respectively. We quantify encoder conditioning through the Jacobian condition number

$$\kappa(J_i) = \frac{\sigma_{\max}(J_i)}{\sigma_{\min}(J_i)}. \tag{3}$$

For a small perturbation δ,

$$E_i(x + \delta) - E_i(x) \approx J_i(x)\delta. \tag{4}$$

The extremal singular values characterize the local amplification of perturbations:

$$\sigma_{\min}(J_i(x))\|\delta\|_2 \leq \|J_i(x)\delta\|_2 \leq \sigma_{\max}(J_i(x))\|\delta\|_2. \tag{5}$$

Small values of $\sigma_{\min}$ indicate that some input directions are strongly suppressed, whereas large values of $\sigma_{\max}$ indicate excessive amplification. Both effects degrade Jacobian conditioning and can impair information and gradient transport. Moreover, gradients from the contrastive objective with loss $\mathcal{L}$ propagate according to

$$\nabla_x \mathcal{L} = J_i(x)^\top \nabla_{z_i} \mathcal{L}, \tag{6}$$

showing that poorly conditioned Jacobians directly impair gradient transport from the multimodal objective back to the modality-specific inputs. For example, for an MLP with L layers, weights W_ℓ, ReLU activations, and diagonal activation Jacobians D_ℓ, the encoder Jacobian can be written as

$$J_i(x) = W_L D_{L-1} W_{L-1} \cdots D_1 W_1. \tag{7}$$

Since ReLU activation Jacobians contain only zeros and ones, entire directions can be removed during forward and backward propagation, leading to rank loss and collapsing $\sigma_{\min}$. Conversely, products of weight matrices may excessively amplify certain directions, resulting in large $\sigma_{\max}$.

Interestingly, multimodal contrastive learning requires encoders to preserve modality-specific information rather than immediately transform it. However, learning near-identity mappings is itself challenging for plain ReLU MLPs, whose Jacobians are products of weights and activation gates (Equation (7)) (He et al., 2016).

3.3 Jacobian Intervention

The analysis above suggests that encoder Jacobian conditioning may play a mechanistic role in multimodal contrastive learning. To test this hypothesis, we directly intervene on the encoder Jacobian through stochastic Jacobian regularization. Computing singular values of J_i exactly is computationally expensive for high-dimensional encoders. Instead, we employ an approximation based on Jacobian-vector products (JVPs). For a random direction v, we define the directional gain

$$r_i(x, v) = \frac{\|J_i(x)v\|_2^2}{\|v\|_2^2}. \tag{8}$$

Given K sampled directions, we estimate

$$r_{\min}(x) = \min_{k \leq K} r_i(x, v_k), \qquad r_{\max}(x) = \max_{k \leq K} r_i(x, v_k). \tag{9}$$

We then penalize deviations from a desired directional-gain band $[m, M]$:

$$\mathcal{L}_{\text{JVP}} = w_{\min} \underbrace{\Big[\max(0, m^2 - r_{\min}(x))\Big]^2}_{\text{prevent directional collapse}} + w_{\max} \underbrace{\Big[\max(0, r_{\max}(x) - M^2)\Big]^2}_{\text{prevent directional explosion}}. \tag{10}$$

Importantly, $r_{\min}$ and $r_{\max}$ are not estimates of the extremal singular values themselves. Rather, they summarize directional gains observed along randomly sampled directions and bias optimization toward better-conditioned Jacobians. Consequently, improvements obtained through Jacobian regularization provide evidence that encoder conditioning influences multimodal alignment performance rather than merely correlating with it. While effective, Jacobian regularization introduces additional computational overhead through repeated JVPs. We therefore seek lightweight architectural mechanisms that naturally preserve encoder geometry throughout training.

3.4 Geometry-Preserving Encoders

We introduce geometry-preserving encoders (GPEs)*, i.e.,* encoder modifications to maintain well-conditioned Jacobians throughout contrastive optimization. We find that two surprisingly simple modifications*, i.e.,* residual transport paths and LeakyReLU activations recover many of the benefits obtained through direct Jacobian intervention.

Residual Transport We introduce residual transport paths around modality-specific encoders. Given a nonlinear branch h_i, we define

$$E_i(x) = P_i x + h_i(x), \tag{11}$$

where P_i is a learned projection. The corresponding Jacobian is

$$J_{E_i}(x) = P_i + J_{h_i}(x). \tag{12}$$

By standard singular-value perturbation bounds (Section A),

$$\sigma_{\min}(J_{E_i}) \geq \sigma_{\min}(P_i) - \|J_{h_i}\|_2, \qquad \text{and} \qquad \sigma_{\max}(J_{E_i}) \leq \sigma_{\max}(P_i) + \|J_{h_i}\|_2. \tag{13}$$

Consequently, if $\|J_{h_i}\|_2 < \sigma_{\min}(P_i)$, then $\sigma_{\min}(J_{E_i}) > 0$. Thus, a well-conditioned projection P_i provides an additive transport path that preserves Jacobian conditioning by simultaneously supporting weak directions and limiting excessive amplification.

LeakyReLU Activations We replace ReLU with LeakyReLU activations (Maas et al., 2013)

$$\phi_\alpha(u) = \begin{cases} u, & u > 0, \\ \alpha u, & u \leq 0, \end{cases} \tag{14}$$

where the negative slope $\alpha > 0$. Unlike ReLU, whose derivative can become exactly zero,

$$\frac{\partial \phi_\alpha(u)}{\partial u} \in \{\alpha, 1\} \tag{15}$$

for $\alpha > 0$. Consequently, the activation Jacobian D_ℓ satisfies

$$\alpha \leq \sigma_{\min}(D_\ell) \leq \sigma_{\max}(D_\ell) \leq 1. \tag{16}$$

Unlike ReLU, LeakyReLU prevents activation derivatives from becoming exactly zero. Consequently, activation gates can no longer completely remove directions from the encoder Jacobian. While rank loss may still arise from weights or architectural bottlenecks, LeakyReLU removes an important source of singular-value collapse.

4 Experiments

4.1 Datasets

To assess the proposed GPEs in the multimodal contrastive setting, we consider a diverse set of datasets (Table 1). Synthetic-XNOR (Rheude et al., 2026b) provides a controlled environment for studying higher-order multimodal alignment and optimization dynamics. MIMIC-IV (Johnson et al., 2023b;a) and MIMIC-Symile (Saporta et al., 2024) serves as a clinically relevant benchmark combining heterogeneous modalities. Finally, the UKB (Sudlow et al., 2015) allows us to evaluate performance at population scale, while UKB-Union additionally introduces modality missingness.

Table 1: Overview of benchmark datasets for our evaluation. Retrieval modality is listed first.

Dataset	# Samples	Modalities	Linear Probe
Synthetic-XNOR (Rheude et al., 2026b)	$30,000$	A, B, C	n/a
MIMIC-Symile (Saporta et al., 2024)	$10,345$	Chest X-rays, ECGs, laboratory	n/a
MIMIC-IV (Johnson et al., 2023b;a)	$52,802$	Reports, Clinicals, Demographics	Mortality
UKB-Intersection (Sudlow et al., 2015)	$37,888$	Proteomics, Metabolomics, EHR	Mortality
UKB-Union (Sudlow et al., 2015)	$486,400$	Proteomics, Metabolomics, EHR	Mortality

Synthetic-XNOR We adopt Synthetic-XNOR (Rheude et al., 2026b), a controlled trimodal benchmark designed to expose failure modes of higher-order multimodal contrastive objectives. Rather than requiring all modalities to contribute equally, the task tests whether an objective can combine information selectively, including cases where one modality is weakly informative, misaligned, or should be effectively ignored for successful retrieval. This makes the benchmark useful for studying optimization dynamics and alignment failures in a controlled setting, isolated from real-world confounders. We use Synthetic-XNOR to analyze how encoder Jacobian conditioning and especially direct Jacobian regularization interacts with higher-order contrastive objectives.

MIMIC We construct a benchmark from MIMIC-IV (Johnson et al., 2023b;a; Goldberger et al., 2000) with radiology reports, clinical timeseries (Section B), and demographics recorded during the first 48 hours of admission. Radiology reports are encoded using a LoRA-finetuned RadBERT model (Hu et al., 2022; Yan et al., 2022) followed by an MLP. Timeseries and demographic variables are encoded using an MLP. Further, we use MIMIC-Symile (Saporta et al., 2024) which includes chest X-rays, electrocardiograms (ECGs), and laboratory values. Chest X-rays and ECGs are encoded with ResNets and laboratory values with an MLP (Saporta et al., 2024). We ablate them by removing the skip connections and comparing the resulting encoders against the standard variants (Figure 10).

UKB We use the UKB (Sudlow et al., 2015), comprising proteomics, metabolomics, and electronic health records (EHRs) (Rheude et al., 2025). For the EHR modality, we employ Qwen3-Embedding-8B embeddings (Zhang et al., 2025; Hegselmann et al., 2026), and all modality encoders are implemented as MLPs. We evaluate two complementary data-driven settings. First, UKB-Intersection contains only samples with complete observations across all modalities ($37{,}888$ samples). Second, UKB-Union contains $486{,}400$ samples and introduces substantial modality missingness. Following Saporta et al. (2024), missing modalities are represented through binary missingness masks appended to the corresponding inputs. The UKB experiments enable evaluation of GPEs at population scale, both with complete multimodal observations and under realistic missing-modality conditions.

4.2 Experimental Setup

All methods are implemented within a unified framework (Falcon & team, 2019) to ensure fair comparisons across objectives, encoders, and datasets. We follow best-practices for multimodal learning (Rheude et al., 2026a) including extensive hyperparameter tuning (Section E). We construct GPEs using both residual paths and LeakyReLU activations. Further, we adopt the use of a learnable logit scale, and a pair-sampling strategy (Rheude et al., 2026b) if applicable. Optimization is performed using ScheduleFree-AdamW (Defazio et al., 2024) with norm-based gradient clipping, helping distinguish Jacobian-induced optimization difficulties from instabilities caused by large gradients. LoRA adapters are applied to the query and value projections of the Transformer attention layers. Tabular results on real-world datasets are evaluated using 5-fold cross-validation with mutually exclusive sample identifiers across folds. For each fold, models are trained with three independent random seeds and reported metrics are averaged across folds and seeds (mean $\pm$ standard error (SE)). Jacobian statistics for the training dynamics are computed separately for each modality encoder and then averaged across modalities. Based on preliminary scaling experiments, we fix the batch size to 128 for Synthetic-XNOR, 280 for MIMIC-Symile, and 512 for MIMIC-IV and the UKB. We evaluate all methods in retrieval tasks and linear probes to assess whether improvements from objectives and GPEs translate to downstream performance, which is often more relevant than retrieval in life science applications.

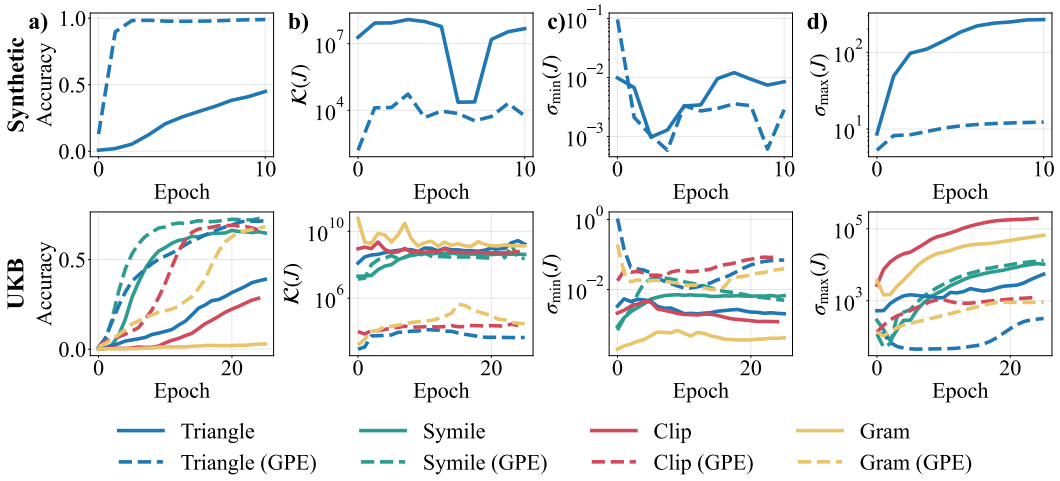


Figure 3: Training dynamics of well-tuned standard encoders and GPEs. The upper row focuses on Triangle and Synthetic-XNOR to highlight the effect of GPEs, whereas the lower row demonstrates their transferability to a real-world dataset, namely the UKB. Standard encoders exhibit severe geometric degeneration, characterized by **(a)** low accuracy, **(b)** exploding condition numbers $\kappa(J)$, **(c)** collapsing minimum singular values $\sigma_{\min}(J)$, and **(d)** rapidly increasing maximum singular values $\sigma_{\max}(J)$. In contrast, GPEs maintain performance and well-conditioned Jacobians throughout training. This demonstrates that objectives alone are insufficient for multimodal contrastive learning.

4.3 Jacobian Analyses

Cross-Objective Training Dynamics Having established Jacobian conditioning as a potential source of optimization instability, we next examine whether this reflects a broader encoder-level limitation. We therefore compare training dynamics across objectives and datasets. Despite their different scoring functions, standard encoders exhibit a general behavior (Figure 3): retrieval accuracy is lower, $\kappa(J)$ rapidly increase, $\sigma_{\min}(J)$ collapse, and $\sigma_{\max}(J)$ grow. In contrast, GPEs generally maintain better-conditioned Jacobians and achieve both faster convergence and higher retrieval accuracies. Interestingly, Symile on the UKB constitutes a notable exception, exhibiting comparatively stable optimization even without GPEs. This behavior is not observed for Symile on Synthetic-XNOR, MIMIC-IV, and MIMIC-Symile (Section B and Figures 7 to 9), suggesting that the interaction between objective design and encoder geometry may depend on datasets. In general, these results indicate that objective expressivity alone is insufficient for successful multimodal contrastive learning and suggest that encoder geometry is a shared bottleneck across objectives.

Jacobian Regularization The results above suggest a strong connection between degraded contrastive performance and ill-conditioned encoder Jacobians. To determine whether this relationship reflects an underlying mechanism rather than a simple correlation, we perform a direct intervention on encoder geometry using the Jacobian regularizer introduced in Section 3. Constraining, *e.g.,* the Triangle objective with local directional gains improves Jacobian conditioning throughout training, and leads to both faster convergence and higher retrieval performance exemplified on Synthetic-XNOR (Figure 4). These improvements mirror those obtained by GPEs. Because the intervention acts directly on encoder Jacobians rather than modifying the contrastive objective, these results provide evidence consistent with a mechanism in which encoder Jacobian conditioning is a key factor governing multimodal contrastive learning performance.

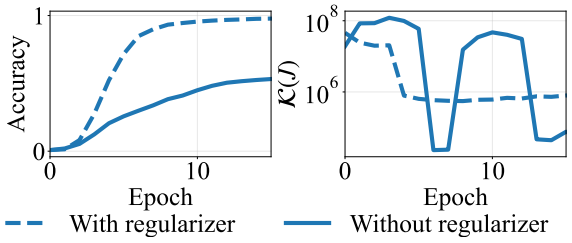


Figure 4: Well-tuned training dynamics with and without Jacobian regularization. Without regularization, $\kappa(J)$ fluctuates to extremes, resulting in performance stagnation. In contrast, Jacobian regularization stabilizes $\kappa(J)$ and leads to maximum performance.

Table 2: Comparison of well-tuned state-of-the-art (sota) multimodal contrastive objectives with and without GPEs across different datasets. Values represent top-1 accuracy of the retrieval task (mean ± SE). The results for the UKB are highlighted in Figure 1. Best non-overlapping value bold.

Method	Synthetic ↑	UKB ↑	UKB-Union ↑	MIMIC-Symile ↑	MIMIC-IV ↑
Clip	0.2434	0.3860 ± 0.014	0.3157 ± 0.007	0.3075 ± 0.008	0.6547 ± 0.009
+ GPE	**0.5814**	$\mathbf{0.6944 \pm 0.013}$	$\mathbf{0.5448 \pm 0.013}$	$\mathbf{0.3604 \pm 0.009}$	$\mathbf{0.6985 \pm 0.003}$
Triangle	0.6093	0.5432 ± 0.013	0.3474 ± 0.011	0.1785 ± 0.015	0.6226 ± 0.008
+ GPE	**0.9952**	$\mathbf{0.7434 \pm 0.011}$	$\mathbf{0.5647 \pm 0.009}$	$\mathbf{0.3045 \pm 0.013}$	$\mathbf{0.7089 \pm 0.003}$
Gram	0.4864	0.1498 ± 0.018	0.2222 ± 0.007	0.1370 ± 0.004	$\mathbf{0.6276 \pm 0.002}$
+ GPE	**0.9967**	$\mathbf{0.6293 \pm 0.013}$	$\mathbf{0.4375 \pm 0.023}$	$\mathbf{0.2325 \pm 0.007}$	0.6087 ± 0.006
Symile	0.3310	0.6266 ± 0.015	0.5165 ± 0.008	0.3943 ± 0.007	0.6874 ± 0.001
+ GPE	**0.5385**	$\mathbf{0.7267 \pm 0.012}$	$\mathbf{0.6309 \pm 0.007}$	$\mathbf{0.4529 \pm 0.011}$	$\mathbf{0.7125 \pm 0.001}$

LeakyReLU Analysis To better understand Jacobian conditioning in case of the LeakyReLU GPE, we vary the negative slope parameter α. As shown in Figure 5, larger values of α improve retrieval performance for the Triangle objective. This trend coincides with improved Jacobian conditioning and reduced singular value collapse or explosion, supporting the hypothesis that preserving local encoder geometry is critical for effective multimodal alignment. Notably, these gains are obtained without modifying the contrastive objective itself, further emphasizing the importance of encoder geometry as an independent factor governing multimodal learning success.

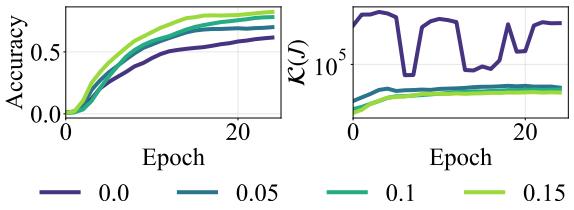


Figure 5: Well-tuned analysis of increasing LeakyReLU's negative slope α on the Synthetic-XNOR dataset. Higher α improves accuracy by preserving non-exploding condition numbers $\kappa(J)$.

4.4 Downstream Performance

Retrieval GPEs generally improve retrieval performance across multimodal contrastive objectives and datasets (Table 2). On Synthetic-XNOR, all objectives benefit from GPEs, especially higher-order objectives such as Triangle and Gram. Similar trends are observed on both UKB cohorts, where GPEs yield consistent improvements under complete observations as well as under missing-modality conditions. On MIMIC-IV, GPEs remain beneficial despite the presence of a LoRA-adapted BERT encoder, indicating that preserving geometry within the trainable projection encoders continues to improve multimodal alignment. However, on MIMIC-IV, Gram represents an outlier and does not benefit from GPEs. On MIMIC-Symile, GPEs consistently improve retrieval across all objectives, suggesting that geometry-preservation also benefits image-, signal-, and laboratory-based alignment. The results demonstrate a general benefit of GPEs across objectives and retrieval tasks.

Linear Probes We further evaluate representation quality through linear probing on mortality prediction tasks (Tables 3 and 4 and Figure 6). On the UKB, GPEs improve multimodal probe performance across CLIP, Triangle, and Gram, while maintaining competitive unimodal representations. Similar improvements are observed on MIMIC-IV, where geometry preservation generally yields higher multimodal AUROC despite more modest changes at the unimodal level. Notably, the gains observed for multimodal probes are often larger and more consistent than those observed for individual modalities, suggesting that geometry preservation primarily improves cross-modal alignment. These findings

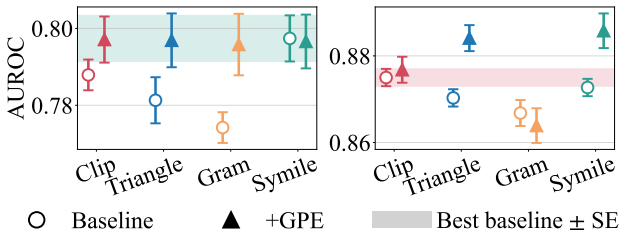


Figure 6: GPEs match or exceed non-GPE linear probe baselines. Results for UKB (left) and MIMIC-IV (right).

Table 3: Well-tuned modality probes on the UKB. Values represent the AUROC for 10 year mortality prediction (mean ± SE). Best non-overlapping value bold, best overlapping value underlined.

Method	Metabolomics ↑	EHR ↑	Proteomics ↑	Multimodal ↑
Clip	$\underline{0.7153} \pm 0.014$	0.7687 ± 0.005	0.7929 ± 0.003	0.7879 ± 0.004
+ GPE	0.7015 ± 0.013	$\underline{0.7743} \pm 0.008$	$\mathbf{0.8006} \pm 0.004$	$\underline{0.7971} \pm 0.006$
Triangle	0.6905 ± 0.014	0.7492 ± 0.005	0.7806 ± 0.003	0.7813 ± 0.006
+ GPE	$\underline{0.6998} \pm 0.014$	$\mathbf{0.7689} \pm 0.007$	$\mathbf{0.7949} \pm 0.005$	$\mathbf{0.7969} \pm 0.007$
Gram	0.6566 ± 0.015	0.7552 ± 0.004	0.7734 ± 0.004	0.7742 ± 0.004
+ GPE	$\mathbf{0.6981} \pm 0.013$	$\mathbf{0.7747} \pm 0.008$	$\mathbf{0.8039} \pm 0.002$	$\mathbf{0.7958} \pm 0.008$
Symile	0.7042 ± 0.014	0.7769 ± 0.007	0.7991 ± 0.003	$\underline{0.7974} \pm 0.006$
+ GPE	$\underline{0.7048} \pm 0.014$	$\underline{0.7781} \pm 0.006$	$\underline{0.8004} \pm 0.004$	0.7966 ± 0.007

Table 4: Well-tuned modality probes on MIMIC-IV. Values represent the AUROC for in-hospital mortality prediction (mean ± SE). Best non-overlapping value bold, best overlapping value underlined.

Method	Reports ↑	Timeseries ↑	Demographics ↑	Multimodal ↑
Clip	$\mathbf{0.7834 \pm 0.013}$	$\underline{0.8747 \pm 0.002}$	0.5992 ± 0.003	0.8750 ± 0.002
+ GPE	0.7613 ± 0.006	0.8744 ± 0.003	$\underline{0.6002 \pm 0.003}$	$\underline{0.8768 \pm 0.003}$
Triangle	0.7733 ± 0.009	0.8716 ± 0.002	0.5991 ± 0.004	0.8703 ± 0.002
+ GPE	$\underline{0.7827 \pm 0.009}$	$\mathbf{0.8768 \pm 0.002}$	$\underline{0.6005 \pm 0.003}$	$\mathbf{0.8841 \pm 0.003}$
Gram	$\mathbf{0.7735 \pm 0.013}$	$\underline{0.8685 \pm 0.003}$	0.6005 ± 0.003	$\underline{0.8668 \pm 0.003}$
+ GPE	0.7403 ± 0.005	0.8653 ± 0.003	$\underline{0.6011 \pm 0.003}$	0.8639 ± 0.004
Symile	$\underline{0.7827 \pm 0.009}$	0.8718 ± 0.002	0.5935 ± 0.004	0.8727 ± 0.002
+ GPE	0.7782 ± 0.007	$\mathbf{0.8792 \pm 0.003}$	$\mathbf{0.6011 \pm 0.003}$	$\mathbf{0.8858 \pm 0.004}$

complement the retrieval results and indicate that the improved optimization dynamics induced by GPEs translate into more useful multimodal representations for downstream prediction tasks. While objective choice can strongly affect alignment quality, representations learned by different objectives achieve comparatively similar downstream predictive performance after linear probing (Figure 6). Similar observations have been reported in generative modeling, where more expressive objectives do not necessarily translate into superior linear probe performance (Esmati et al., 2026). In contrast, GPEs generally improve linear probe performance across objectives and reduce the dependence of downstream representation quality on objective choice (Figure 6), suggesting that encoder geometry plays an important role alongside objective expressivity.

5 Conclusion, Limitations & Future Work

We showed that multimodal contrastive learning depends strongly on encoder Jacobian conditioning. Across objectives and datasets, training is characterized by collapsing minimum singular values, exploding maximum singular values, and rapidly increasing condition numbers. Through direct Jacobian regularization, we demonstrated that encoder conditioning stabilizes optimization. Building on this observation, we introduced GPEs based on residual paths and LeakyReLU activations. These architectural modifications improved retrieval and linear probing performance across multimodal contrastive objectives and datasets. We focus on trimodal contrastive learning, *i.e.,* more modalities remain important for future work. Although we evaluate both structured and unstructured modalities, our experiments are centered on healthcare datasets, motivated by the availability of diverse modalities in this domain. Finally, our analysis focuses on MLP-based projections, which remain common in multimodal learning. In contrast, other architectures already incorporate architectural mechanisms that facilitate information and gradient transport and appear more robust to the Jacobian degeneration. We therefore provide a preliminary analysis of the mechanisms (Section C), leaving a more detailed investigation to future work. More broadly, multimodal contrastive learning may benefit from a shift from increasingly expressive objectives toward a deeper understanding of encoder geometry, optimization dynamics, and information transport across modalities.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on methodology or experiments, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. The remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create or edit software code and to edit parts of the research paper to improve readability. We did not use generative AI tools for the other activities for which disclosure is recommended. We have reviewed all AI-assisted work. Generative AI was used during the implementation of methods and the creation or editing of software code to support the development and refinement of the required code. All AI-assisted code and implementations were manually reviewed, executed, and checked for correctness and consistency with the intended methodology before being incorporated into the final work. Generative AI was also used to assist with translation. AI-generated or AI-assisted translations were manually reviewed and revised where necessary to ensure that the intended meaning, technical terminology, and context were preserved. Finally, generative AI was used to improve the readability of parts of the manuscript, including aspects such as phrasing, grammar, and clarity. All such edits were reviewed by the authors to ensure that they did not alter the scientific meaning, claims, interpretation, or conclusions of the work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the optimization and geometry of multimodal contrastive learning systems. Our primary contribution is a better understanding of how encoder Jacobian conditioning influences multimodal alignment and how simple architectural modifications can improve training stability. As such, the proposed methods are general-purpose and are not tied to a specific application domain. The datasets considered in this work are either synthetic or derived from existing biomedical resources that have been extensively used for machine learning research. We do not introduce new sensitive data sources, nor do we propose methods intended for clinical decision-making without further validation. While improved multimodal representation learning may ultimately benefit downstream healthcare applications, any deployment in high-stakes settings would require rigorous evaluation for fairness, robustness, privacy, and clinical safety. Overall, we believe the primary impact of this work is methodological: improving the reliability, stability, and scientific understanding of multimodal contrastive learning systems.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of our results. We provide the source code used for our experiments, including a direct-start example based on a synthetic dataset to illustrate the complete workflow. The paper and supplementary material describe the model architectures and implementation details in detail, and we report the full hyperparameter optimization spaces used in our experiments. Together, these materials are intended to make the experimental setup, model configurations, and optimization procedure transparent and to enable independent reproduction and extension of our results.

REFERENCES

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, September 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-022-01981-2.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Gary Cottrell, and Julian J. McAuley. Rezero is all you need: fast convergence at large depth. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July*

2021, Proceedings of Machine Learning Research, pp. 1352–1361. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Rebekka Burkholz and Alina Dubatovka. Initialization of relus for dynamical isometry. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d9731321ef4e063ebbee79298fa36f56-Paper.pdf.

Yichao Cai, Zhen Zhang, Yuhang Liu, and Javen Qinfeng Shi. The geometric mechanics of contrastive representation learning: Alignment potentials, entropic dispersion, and cross-modal divergence. *CoRR*, abs/2601.19597, 2026. doi: 10.48550/ARXIV.2601.19597. URL https://doi.org/10.48550/arXiv.2601.19597. arXiv: 2601.19597.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=3Hjfzh5Eyk.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=ftGnpZrW7P.

Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/136b9a13861308c8948cd308ccd02658-Abstract-Conference.html.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pe3AxLq6Wf.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. arXiv:2606.05328 [cs.GR], June 2026. URL http://arxiv.org/abs/2606.05328.

William Falcon and The PyTorch Lightning team. Pytorch lightning, March 2019. URL https://github.com/Lightning-AI/lightning.

Arpit Garg, Hemanth Saratchandran, and Simon Lucey. Sineproject: Machine unlearning for stable vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31735–31745, June 2026.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep learning tuning playbook, 2023. URL http://github.com/google-research/tuning_playbook. Version 1.0.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, June 2000.

Michal Golovanevsky, Pranav Mahableshwarkar, Carsten Eickhoff, and Ritambhara Singh. Picme: Pipeline for contrastive modality evaluation and encoding in the mimic dataset. arXiv:2507.03165 [cs.LG], July 2025. URL http://arxiv.org/abs/2507.03165.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Stefan Hegselmann, Georg Von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *npj Digital Medicine*, 9(1):530, July 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02915-9.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. arXiv:1908.02729 [stat.ML], August 2019. URL http://arxiv.org/abs/1908.02729.

Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4475–4483. PMLR, July 2020. URL https://proceedings.mlr.press/v119/huang20f.html.

Yiping Ji, Hemanth Saratchandran, Peyman Moghadam, and Simon Lucey. Always skip attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23115–23123, October 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, January 2023a. doi: 10.13026/6mm1-ek67. URL https://doi.org/10.13026/6mm1-ek67. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, January 2023b. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x.

Stefanos Koutoupis, Michaela Areti Zervou, Konstantinos Kontras, Maarten De Vos, Panagiotis Tsakalides, and Grigorios Tsagkatakis. The more, the merrier: Contrastive fusion for higher-order multimodal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8825–8835, June 2026.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, Atlanta, Georgia, USA, 2013. JMLR.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp.

4785–4795, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d9fc0cdb67638d50f411432d0d41d0ba-Abstract.html.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. Geometric multimodal contrastive representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17782–17800. PMLR, 2022. URL https://proceedings.mlr.press/v162/poklukar22a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Cohort-based active modality acquisition. arXiv:2505.16791 [cs], December 2025. URL http://arxiv.org/abs/2505.16791.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Fusion or confusion? multimodal complexity is not all you need. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=RtOPHuqmpj.

Tillmann Rheude, Stefan Hegselmann, Roland Eils, and Benjamin Wild. Hidden in the multiplicative interaction: Uncovering fragility in multimodal contrastive learning. arXiv:2604.05834 [cs.LG], May 2026b. URL http://arxiv.org/abs/2604.05834.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 833–840. Omnipress, 2011. URL https://icml.cc/2011/papers/455_icmlpaper.pdf.

F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton: (Project Para)*. Report / Cornell Aeronautical Laboratory. Cornell Aeronautical Laboratory, 1957. URL https://books.google.de/books?id=P_XGPgAACAAJ.

Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8PWvdaRQAu.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6120.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06031-6.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Trans. Signal Process.*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039.

Rishi Sonthalia, Michael Murray, and Guido Montufar. Low rank gradients and where to find them. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=de3kwOXQ9e.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779.

Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrz˛ebski, Jacek Tabor, and Maciej Nowak. Dynamical isometry is achieved in residual networks in a universal way for any activation function. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2221–2230. PMLR, April 2019. URL https://proceedings.mlr.press/v89/tarnowski19a.html.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):5314–5321, 2023. doi: 10.1109/TPAMI.2022.3206148.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(10):6761–6774, October 2024. ISSN 0162-8828. doi: 10.1109/TPAMI.2024.3386927.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, July 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402. PMLR, July 2018. URL https://proceedings.mlr.press/v80/xiao18a.html.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=s436PHXRzMm.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, July 2022. ISSN 2638-6100. doi: 10.1148/ryai.210258.

Lingjie Yi, Raphael Douady, and Chao Chen. Decipher the modality gap in multimodal contrastive learning: From convergent representations to pairwise alignment, 2025. URL https://arxiv.org/abs/2510.03268.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18123–18133, June 2022.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176 [cs], June 2025. URL http://arxiv.org/abs/2506.05176.

A Residual Transport and Singular-Value Perturbation

The residual transport analysis in Section 3 relies on standard singular-value perturbation bounds. Let

$$J_E = P + H, \tag{17}$$

where P denotes the residual projection and H the Jacobian of the nonlinear branch. Weyl's inequality for singular values implies

$$|\sigma_k(P + H) - \sigma_k(P)| \leq \|H\|_2 \tag{18}$$

for all singular values σ_k. Applying this result to the smallest and largest singular values yields

$$\sigma_{\min}(J_E) \geq \sigma_{\min}(P) - \|H\|_2 \tag{19}$$

and

$$\sigma_{\max}(J_E) \leq \sigma_{\max}(P) + \|H\|_2. \tag{20}$$

Consequently, if

$$\|H\|_2 < \sigma_{\min}(P), \tag{21}$$

then

$$\sigma_{\min}(J_E) > 0, \tag{22}$$

showing that the residual projection preserves locally non-degenerate directions despite perturbations introduced by the nonlinear branch.

B Additional Details

In the following, we provide additional details for our experiments, setups, and results.

Training Dynamics The full scorer training dynamics on the Synthetic-XNOR dataset from Figure 3 is visualized in Figure 7. The training dynamics analyses for MIMIC-IV and MIMIC-Symile are visualized in Figures 8 and 9. All Jacobian statistics are based on the encoders, with raw inputs or, in case of MIMIC-IV, RadBERT embeddings.

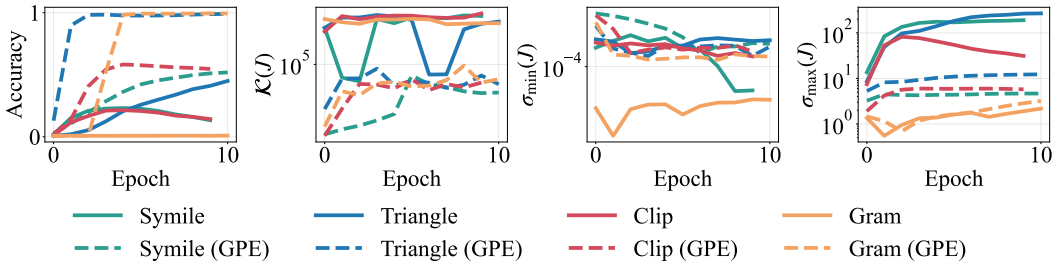


Figure 7: Training dynamics of standard encoders and GPEs with different scorers on the Synthetic-XNOR dataset. Full visualization from focused view in Figure 3.

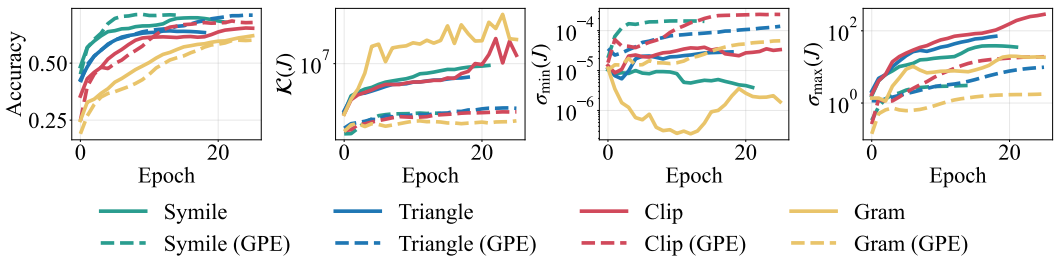


Figure 8: Training dynamics of standard encoders and GPEs with different scorers on the MIMIC-IV dataset.

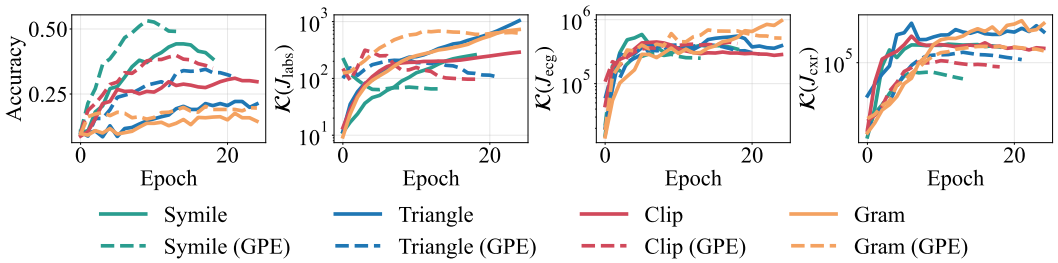


Figure 9: Training dynamics of standard encoders and GPEs with different scorers on the MIMIC-Symile dataset for each modality. For GPEs, labs are encoded by residaul MLPs, and ECGs and chest X-rays are encoded by ResNets. For non-GPE, we use traditional MLPs and removed the skip connections from the ResNets.

Modality Setup in MIMIC-IV For our MIMIC-IV trimodal contrastive learning setup, we use three modalities per ICU stay: radiology text, clinical chart timeseries, and demographics. The radiology modality is restricted to notes with charttime in the interval from the first 48 hours, with notes after discharge excluded and notes after death excluded. Among valid notes, we keep the earliest report per stay. The clinical timeseries modality is restricted to measurements in the same interval, with observations after discharge removed and observations after death removed. We use the following 32 charted variables (item IDs) exactly: 220045 (Heart Rate), 220050 (Arterial Blood Pressure systolic), 220051 (Arterial Blood Pressure diastolic), 220052 (Arterial Blood Pressure mean), 220074 (Central Venous Pressure), 220179 (Non Invasive Blood Pressure systolic), 220180 (Non Invasive Blood Pressure diastolic), 220181 (Non Invasive Blood Pressure mean), 220210 (Respiratory Rate), 220277 (O_2 saturation pulseoxymetry), 220739 (GCS Eye Opening), 223761 (Temperature Fahrenheit), 223900 (GCS Verbal Response), 223901 (GCS Motor Response), 224054 (Braden Sensory Perception), 224055 (Braden Moisture), 224056 (Braden Activity), 224057 (Braden Mobility), 224058 (Braden Nutrition), 224059 (Braden Friction/Shear), 224168 (Parameters Checked), 224641 (Alarms On), 225664 (Glucose finger stick), 228096 (Richmond-RAS Scale), 228299 (Goal Richmond-RAS Scale), 228305 (ST Segment Monitoring On), 228409 (Strength L Arm), 228410 (Strength L Leg), 228411 (Strength R Leg), 228412 (Strength R Arm), 229321 (Activity/Mobility, JH-HLM), and 229381 (Orientation). For each item ID and ICU stay, we compute six summary statistics similar to Saporta et al. (2024): count, mean, std, min, max, and last. Concatenating these statistics yields a fixed-length vector of $32 \times 6 = 192$ features. The demographic modality uses structured patient-level covariates, *e.g.,* age and sex.

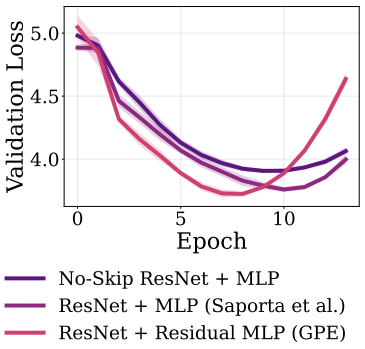


Figure 10: Loss comparison on the MIMIC-Symile dataset for the originally published encoders, non-GPE, and GPEs. GPEs show faster convergence and lower loss values.

MIMIC-Symile Comparisons Figure 10 compares validation loss trajectories across CV folds and independent seeds on MIMIC-Symile. Compared with the originally published encoders (MLP and ResNets) and the non-GPE (MLP and non-skip ResNets) variants, GPEs (residual MLP and ResNets) converge more rapidly and reach lower validation losses.

Linear Encoder Ablation Table 5 compares GPEs against linear modality encoders on the UKB. Linear encoders perform competitively for several objectives and even match Triangle with GPEs, indicating that a substantial part of the multimodal retrieval signal is accessible through stable low-complexity projections. However, the strong degradation

Table 5: Linear encoder ablation of well-tuned objectives on the UKB. Best non-overlapping top-1 accuracy bold and best overlapping underlined.

Ablation	**Top-1 Accuracy** ↑
Clip w/ GPE	$\mathbf{0.6944 \pm 0.013}$
Clip w/ Linear	0.5969 ± 0.015
Triangle w/ GPE	0.7434 ± 0.011
Triangle w/ Linear	$\underline{0.7438 \pm 0.012}$
Gram w/ GPE	$\underline{0.6293 \pm 0.013}$
Gram w/ Linear	0.6116 ± 0.015
Symile w/ GPE	$\mathbf{0.7267 \pm 0.012}$
Symile w/ Linear	0.3111 ± 0.010

for Symile and the lower performance for Clip and Gram show that linear encoders are not uniformly sufficient, suggesting that GPEs can retain the stability of simple projections while preserving objective-dependent nonlinear flexibility. Moreover, our MIMIC-Symile results show that GPEs also transfer to settings with intrinsically nonlinear modalities such as chest X-ray images and ECGs.

C Transformers Contain Geometry-Preserving Mechanisms

In the main paper, we observed that geometry-preserving interventions are particularly beneficial for modality encoders implemented as MLPs and optionally on top of LoRA-finetuned RadBERT models. In contrast, Transformer encoders exhibited stable optimization behavior even without explicitly introducing geometry-preserving modifications. To better understand this phenomenon, we investigate which architectural components are responsible for this robustness. Further, we refer to Ji et al. (2025) for related work in this direction. Since the original Synthetic-XNOR benchmark consists of fixed-size feature vectors, we extend the dataset to a two-dimensional sequential variant that can be processed by sequence models. We then evaluate a series of Transformer-inspired architectures that progressively remove components commonly believed to contribute to optimization stability (Table 6). *Self-Attention* corresponds to a standard Transformer encoder layer with learned query, key, and value projections. *Uniform Attention* removes the learned attention mechanism and instead replaces each token representation by the mean of all tokens, thereby preserving token mixing while eliminating learned attention weights. Finally, *Token-wise MLP* processes each token independently through identical feed-forward layers and therefore removes both learned attention and cross-token information exchange. Standard self-attention Transformers converge both with and without residual transport paths. Interestingly, replacing learned self-attention by fixed uniform token mixing does not prevent convergence as long as residual transport remains present. In contrast, removing residual transport from the uniform mixing architecture leads to optimization collapse. Likewise, a token-wise MLP without token mixing and without residual transport fails to learn the task. These findings suggest that the robustness of Transformer architectures cannot be explained solely by the expressivity of learned self-attention. Instead, the results indicate that dense information transport across tokens, together with residual transport paths, provides a strong inductive bias against the Jacobian degeneration observed in standard MLP encoders. This observation is consistent with recent work arguing that skip connections play a central role in Transformer optimization and may be more important than attention expressivity itself for stable training (Ji et al., 2025). While these experiments do not isolate every component of the Transformer architecture, they provide additional evidence for the central hypothesis of this work: successful multimodal contrastive learning depends critically on preserving trainable encoder geometry. Transformer architectures appear naturally robust to ill-conditioned Jacobians because they already contain architectural mechanisms that facilitate information and gradient transport throughout the network.

Table 6: Effect of token mixing and residual transport on convergence for Transformer encoders. Learned self-attention is not strictly necessary: fixed uniform token mixing also converges when residual transport is present. In contrast, removing both residual transport and token mixing causes optimization collapse. These results suggest that dense transport paths, rather than attention expressivity alone, are the primary stabilizing mechanism.

Architecture	Token Mixing	Residual	Convergence
Self-Attention	✓	✓	✓
Self-Attention	✓	✗	✓
Uniform Attention	✓	✓	✓
Uniform Attention	✓	✗	✗
Token-wise MLP	✗	✗	✗

D Compute Environment

We computed every experiment on a high-performance cluster (HPC) with the following environment:

1. 21 Dell PowerEdge R7525 compute nodes, each with 64 AMD Epyc cores (Rome), 512GB RAM and 1 NVIDIA A100 40G GPU
2. 2 Dell PowerEdge XE8545 compute nodes, each with 128 AMD Epyc cores (Milan), 512GB RAM, 4 NVIDIA A100 40G and 4 NVIDIA A100 80G GPUs (NVLink-connected)

E HYPERPARAMETER TUNING

Hyperparameters are selected using Bayesian optimization with validation loss as the optimization objective. Each method is tuned using 100 optimization trials. Following common practice (Godbole et al., 2023), batch size is fixed in advance and excluded from the search space. For Synthetic-XNOR, hyperparameters are retuned for each experimental setting, including different values of LeakyReLU's negative slope α. Due to the substantially larger computational requirements of UKB-Union, the number of Bayesian optimization trials is reduced to 50 for this dataset. For the Jacobian regularizer experiments, we use $K = 32$ random probe directions per sample (computed every training step) across all three modalities. We set $m = 0.5$, $M = 5.0$, $w_{\min} = 10.0$, and $w_{\max} = 1.0$. The intervention is added to the base contrastive objective as $\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{JVP}}$ with $\lambda = 1.0$. The complete search spaces for all methods, linear probes, and datasets are provided in Listings 1 to 7.

Listing 1 Hyperparameters related to Clip, Triangle, Gram and Symile.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

modelname.logit_scale_init:
    min: -3
    max: 0
    distribution: "uniform"
optimizer.lr:
    min: 0.00001
    max: 0.01
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 10, 50, 100, 200, 500, 1000, 1200]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
```

Listing 2 Hyperparameters related to GPEs of the objectives.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

encoders.residual_path:
    values: [True]
encoders.leaky_relu_negative_slope:
    values: [0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]
```

Listing 3 Hyperparameters related to Synthetic-XNOR.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

modelname.emb_dim:
    values: [256]  # initially tuned from 32-1024
modelname.embedding_norm:
    values: [True]
# Encoders fixed to MLPs
```

Listing 4 Hyperparameters related to MIMIC-IV.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

modelname.emb_dim:
    values: [1024]  # initially tuned from 256-8196
modelname.embedding_norm:
    values: [True]
# Encoders fixed to MLPs + LoRA-finetuned RadBERT
```

Listing 5 Hyperparameters related to MIMIC-Symile.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

modelname.emb_dim:
    values: [1024]  # initially tuned from 256-8196
modelname.embedding_norm:
    values: [True]
# Encoders fixed to MLPs + ResNets
```

Listing 6 Hyperparameters related to the UKB.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize

modelname.emb_dim:
    values: [6144]  # initially tuned from 256-8196
modelname.embedding_norm:
    values: [True]
encoders.nmr.mlp.hidden_dims:
    values: [[1024,2048,4096]]  # initially tuned with 128-4096
encoders.nmr.mlp.hidden_dropouts:
    values: [[0.2,0.2,0.2]]  # initially tuned with 0.0-0.6
encoders.ehr.mlp.hidden_dims:
    values: [[1024,2048,4096]]  # initially tuned with 128-4096
encoders.ehr.mlp.hidden_dropouts:
    values: [[0.6,0.6,0.6]]  # initially tuned with 0.0-0.6
encoders.olink.mlp.hidden_dims:
    values: [[1024,2048,4096]]  # initially tuned with 128-4096
encoders.olink.mlp.hidden_dropouts:
    values: [[0.4,0.4,0.4]]  # initially tuned with 0.0-0.6
# Encoders fixed to MLPs
```

Listing 7 Hyperparameter search space for linear probes on MIMIC-IV and the UKB.

```yaml
method: grid
selection_metric:
  values: [roc_auc]

probe_type:
  values: [logreg_saga]
C:
  values: [1e-3, 1e-2, 1e-1, 1.0]
class_weight:
  values: [balanced]
max_iter:
  value: 1000
tol:
  value: 1e-3
```