PRIOR-FITTED AND FROZEN SCORERS FOR MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning typically relies on manually specified scoring functions, such as cosine similarity, to quantify the similarity between representations from different modalities. Several such functions have been recently proposed to extend contrastive learning beyond two modalities. In principle, however, any function that assigns high compatibility to matched samples and low compatibility to mismatched ones could serve as a scorer. Motivated by this broader view, we instantiate the scorer with a Transformer and find that even a randomly initialized model, when kept frozen, can outperform strong handcrafted multimodal contrastive baselines, whereas optimizing the scorer jointly with the encoders is less effective. Building on this observation, we introduce the Prior-Fitted Scorer (PFS): a Transformer scorer pre-trained on synthetic multimodal data generated from structural causal model (SCM) priors before transferring it to real-world data. The SCM exposes the scorer to diverse dependency structures, including shared, pairwise, and private information, providing a structured prior over higher-order cross-modal similarity. Across four real-world datasets, frozen Transformer scorers are competitive to contrastive baselines, and prior-fitting further improves the random initialization. More broadly, our results suggest that multimodal contrastive learning need not rely on manual similarity functions: Transformers can provide strong inductive biases, while causal pretraining can further shape these biases.[1]

1 INTRODUCTION

Contrastive learning has become a central method for learning transferable multimodal representations. A famous example is CLIP (Radford et al., 2021), which aligns image and text embeddings by contrasting matched and mismatched pairs using cosine similarity as a scoring function. However, many real-world problems involve more than two modalities, motivating objectives that explicitly capture higher-order dependencies. Symile (Saporta et al., 2024) introduces the multilinear inner product (MIP) extending the standard dot product to capture higher-order multimodal interactions. Other approaches, such as Gram (Cicchetti et al., 2025b) and Triangle (Cicchetti et al., 2025a), replace pairwise cosine similarity with geometry-based measures that jointly score multiple modality embeddings. These methods differ in how multimodal similarity is defined, but they share a common design choice: the scoring function is manually defined. This raises a more fundamental question: if similarity can, in principle, be expressed by many different functions, why restrict the scorer to a particular hand-designed manual form rather than represent and learn it more generally?

Neural networks provide a natural way to instantiate such a general scoring function. Particularly, a Transformer (Vaswani et al., 2017) is a convenient realization of this idea, since modality embeddings can be represented as tokens and processed jointly by the same scorer across different modality counts. This avoids designing a new manual compatibility function for each setting and also allows missing modalities to be handled, which is particularly attractive in domains such as medicine (Acosta et al., 2022). Using a Transformer scorer introduces additional computation compared with closed-form similarities, but this cost is limited by the small number of modality tokens and by keeping the scorer frozen during downstream training. Even without any task-specific training, a randomly initialized and frozen Transformer already defines a nontrivial nonlinear function over its inputs. We therefore investigate whether such Transformer-based functions can serve as effective multimodal scoring rules.

[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.

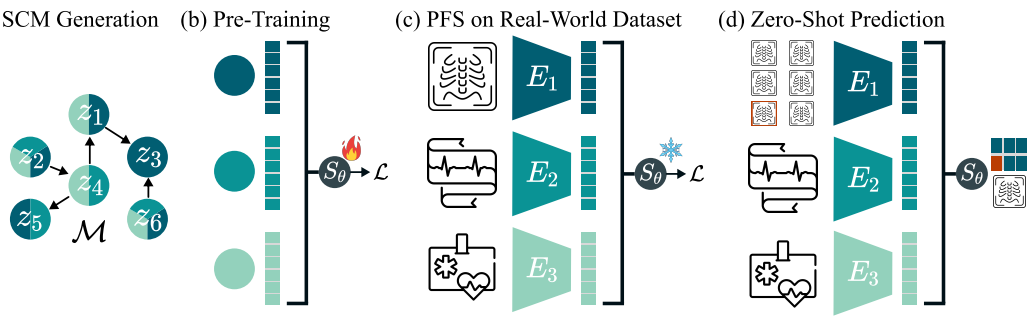


Figure 1: Overview of the Prior-Fitted Scorer (PFS), illustrated on the trimodal MIMIC-Symile setting (Saporta et al., 2024). **(a)** We sample latent causal structures and generate synthetic multimodal tuples from an SCM prior $\mathcal{M}$. **(b)** A Transformer scoring function s_θ is pretrained on these synthetic tuples to distinguish compatible from mismatched multimodal combinations with loss $\mathcal{L}$, thereby shaping its inductive bias. **(c)** The PFS is transferred to the real dataset and kept frozen while the modality encoders E_i are optimized through the contrastive objective. **(d)** After extensive tuning, the PFS is evaluated on zero-shot predicting the future chest X-ray. Overall, a PFS replaces a manually specified similarity function with a Transformer whose structure is learned from a synthetic causal prior.

Surprisingly, we find that a randomly initialized Transformer already provides a strong multimodal scorer when kept frozen during training. In this setting, all optimization pressure is placed on the modality encoders.

Random initialization, however, leaves the scoring function entirely unguided. We therefore ask whether its inductive bias can be shaped before exposure to real data. Inspired by Prior-Data Fitted Networks (PFNs) (Müller et al., 2022), which learn inference procedures by training Transformers across synthetic tasks sampled from a prior, we pretrain the scorer on synthetic multimodal embeddings generated from SCMs (Figure 1). Unlike meta-learning approaches that pretrain a model to perform in-context learning, we pretrain only the multimodal scorer. The modality-specific encoders are learned from scratch for each dataset because the underlying inputs can differ, *e.g.,* images, time series, text, or tabular measurements. Related work has demonstrated that synthetic task distributions can support effective pretraining across a wide range of problems, including tabular prediction (Hollmann et al., 2025), Bayesian optimization (Müller et al., 2023), clustering (Zhao et al., 2026), survival analysis (Böhm et al., 2026), uncertainty prediction (Shi et al., 2026), and causal inference (Robertson et al., 2025; Ma et al., 2026). Here, we use SCMs for a different purpose: to shape the inductive bias of a multimodal contrastive scorer before transferring it to real-world data. Concretely, our SCM prior generates latent causal variables and maps them into modalities with shared, pairwise, and modality-specific information. We use these synthetic multimodal tuples to pretrain a Transformer model as a contrastive scorer and subsequently transfer the resulting PFS to real-world datasets, exemplified in the trimodal setting.

Across four datasets, frozen randomly initialized Transformer scorers are competitive with well-tuned baselines trained by manual scoring functions, while SCM-based prior-fitting consistently improves upon random initialization. These results indicate that the scorer architecture itself contains an important inductive bias, and that this bias can be shaped through synthetic pretraining. Together, these findings broaden the design space of multimodal contrastive learning from choosing among manual similarity functions to designing, learning, and prior-fitting expressive scoring architectures.

Our main contributions are:

- We show that a frozen Transformer scorer can serve as a strong competitor to manual multimodal similarity functions despite having randomly initialized parameters.
- We introduce the Prior-Fitted Scorer (PFS), which pretrains the frozen Transformer scorer on synthetic SCM-generated multimodal data in order to shape its inductive bias toward higher-order multimodal similarity.
- We demonstrate that SCM-based prior-fitting further improves the already strong frozen Transformer scorer across real-world multimodal retrieval tasks.

2 Related Work

Multimodal Contrastive Learning Multimodal contrastive learning has largely relied on manually designed similarity functions. Vision-language models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) use pairwise similarity, with later approaches such as LiT (Zhai et al., 2022) and SigLIP (Zhai et al., 2023) refining the corresponding training objectives. Extensions to more than two modalities introduce increasingly specialized constructions, 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), whose scoring rules impose specific assumptions on how modalities interact and can depend on the modality configuration. Learned similarity functions have also been explored more directly, *e.g.,* by replacing fixed inner-product similarity with a learned parametric measure (Liu et al., 2019), and learning an MLP-based similarity function between representations within a self-supervised objective (Zhang et al., 2022). Learned critics provide an alternative case: Deep InfoMax (Hjelm et al., 2019) and CPC (Oord et al., 2018) parameterize the contrastive scoring function rather than fixing it analytically. Our finding that even frozen random Transformers can be effective is also related to work on random features and untrained neural networks, where fixed random nonlinear mappings can preserve useful structure (Rahimi & Recht, 2007; Giryes et al., 2016; Amid et al., 2022). We build on these ideas by using a Transformer as a flexible multimodal scorer that can operate across varying modality combinations without redesigning the scoring rule.

Causal Models for Representation Learning Structural causal models (SCMs) describe data-generating processes through a directed acyclic graph (DAG) and structural assignments that map parent variables and exogenous noise to their descendants (Peters et al., 2017). Causal structure has increasingly been connected to contrastive learning, *e.g.,* by interpreting augmentations as interventions and enforcing invariance across causal mechanisms (Mitrovic et al., 2021), or by constructing counterfactual positive pairs to improve robustness under distribution shifts (Roschewitz et al., 2025). Recent work has also studied latent causal models specifically for multimodal contrastive learning, providing theoretical links between multimodal alignment and the recovery of shared latent factors (Liu et al., 2026). In contrast to these approaches, we neither aim to recover causal structure from observed multimodal data nor use interventions to impose causal invariances. Instead, we use SCMs as structured generators of synthetic multimodal dependency patterns, providing a controllable source of shared, pairwise, private, linear, and nonlinear interactions on which a multimodal scoring function can be pretrained.

Prior-Data Fitted Networks Prior-Data Fitted Networks (PFNs) introduced the idea of training Transformers on synthetic datasets sampled from a predefined prior, enabling them to learn inference procedures that generalize across a distribution of tasks (Müller et al., 2022). This principle has been successfully scaled to, *e.g.,* tabular prediction with TabPFN (Hollmann et al., 2025), and to, *e.g.,* causal inference with Do-PFN (Robertson et al., 2025) and CausalFM (Ma et al., 2026). Beyond tabular inference, CauKer demonstrates that causally structured synthetic data can also provide effective pretraining distributions for time-series foundation models (Xie et al., 2026). We adopt the central idea of fitting a neural model to a synthetic prior, but distinguish from existing PFNs in both objective and use: rather than learning an in-context predictor or causal estimator, we prior-fit a Transformer that serves directly as the compatibility function of a multimodal contrastive objective. The resulting scorer is subsequently frozen and transferred to real-world multimodal data.

3 Method

Most contrastive objectives define multimodal similarity through a manually chosen analytical score, such as cosine similarity or a particular higher-order geometric construction. Yet there is, in principle, no reason why a specific function should be defined a priori: if one handcrafted interaction is useful, another may be as well. This motivates a more general view in which the scorer is itself treated as a nonlinear function over the modality embeddings. We denote a multimodal dataset $\mathcal{D}$ with N samples and M modalities by

$$\mathcal{D} = \left\{ \left(\mathbf{x}_m^{(n)} \right)_{m=1}^{M} \right\}_{n=1}^{N}, \tag{1}$$

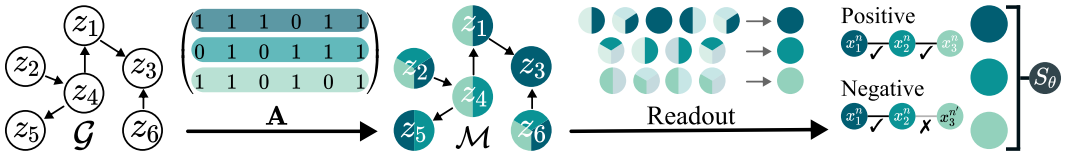


Figure 2: Overview of synthetic multimodal data generation and contrastive prior-fitting. A DAG $\mathcal{G}$ defines causal dependencies among latent variables, while the binary assignment matrix $\mathbf{A}$ selects the latent subsets contributing to each modality, inducing shared, subset-specific, and modality-private information. The selected latent variables are combined through linear and nonlinear readouts to produce modality vectors. During prior-fitting, matched tuples contain modalities from the same synthetic sample, whereas negatives are formed by replacing the target modality with a sampled mismatched candidate before scoring with s_θ.

where $\mathbf{x}_m^{(n)}$ denotes modality $m \in \{1, \dots, M\}$ for sample n. Each modality is mapped into a common embedding space by a modality-specific encoder E_m,

$$\mathbf{e}_m^{(n)} = E_m\left(\mathbf{x}_m^{(n)}\right) \in \mathbb{R}^{d_{\text{in}}}. \tag{2}$$

Transformer Scorer Given the resulting modality embeddings $\{\mathbf{e}_m^{(n)}\}_{m=1}^M$, we define their similarity through a Transformer-based scoring function s_θ. The modality embeddings are processed together with a learnable CLS token, and the resulting CLS representation $\mathbf{h}_{\text{CLS}}^{(n)}$ is mapped to a scalar score,

$$s_\theta\left(\mathbf{e}_1^{(n)}, \dots, \mathbf{e}_M^{(n)}\right) = \mathbf{w}^\top \mathbf{h}_{\text{CLS}}^{(n)} + b. \tag{3}$$

Here, θ denotes the parameters of the Transformer and its modality-specific input projections, while $\mathbf{w}$ and b parameterize the final linear readout. In contrast to prescribing a fixed manual similarity function, this formulation allows the scorer to represent a broad class of nonlinear and higher-order interactions between modality embeddings. This formulation replaces a manually specified similarity rule with a flexible function class, but leaves open which particular scoring function should be used. In the following, we address this question by shaping the Transformer scorer through prior-fitting on synthetic multimodal data generated from SCMs.

Structural Causal Model Prior To generate samples from a multimodal SCM-based prior (Figure 2), we sample a DAG $\mathcal{G}$ over K vector-valued latent variables $\{\mathbf{z}_1, \dots, \mathbf{z}_K\}$. We first sample a random topological ordering and then include each forward edge independently with probability p_{edge}. Consequently, a node may have zero, one, or multiple parents and children, while the number of root nodes is determined by the sampled graph. Each non-root latent variable is generated according to

$$\mathbf{z}_j = f_j(\{\mathbf{z}_k : k \in \text{Pa}(j)\}) + \sigma_z \boldsymbol{\epsilon}_j, \tag{4}$$

where $\text{Pa}(j)$ denotes the parents of node j, $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is standard Gaussian noise, and $\sigma_z \geq 0$ controls its magnitude. Root variables are sampled directly from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The structural mechanisms, *i.e.,* edges of the DAG, f_j are sampled from a family of linear and nonlinear functions, allowing the prior to represent heterogeneous causal dependencies. A binary assignment matrix $\mathbf{A} \in \{0, 1\}^{M \times K}$ determines which latent variables directly affect each modality, *i.e.,* each latent node is assigned to modalities by $\mathbf{A}$. Both root and non-root latent variables can be selected by $\mathbf{A}$. The assignment matrix can create latent factors that are shared across all modalities, shared by arbitrary subsets of modalities, or private to individual modalities. For each sample n, we independently draw the noise variables and evaluate the DAG to obtain $\{\mathbf{z}_1^{(n)}, \dots, \mathbf{z}_K^{(n)}\}$. The m-th row of $\mathbf{A}$ selects the complete set of latent variables used to construct modality m,

$$\mathcal{Z}_m^{(n)} = \left\{\mathbf{z}_j^{(n)} : A_{mj} = 1\right\}, \qquad m \in \{1, \dots, M\}. \tag{5}$$

The selected latent variables are then mapped into a modality vector $\mathbf{x}_m^{(n)}$. For notational simplicity, we suppress the sample index below. Given the latent variables, modality m is created using a mixture

Table 1: Trimodal benchmark datasets with retrieval modality as first listed modality.

Dataset	# Samples	Modalities
MIMIC-IV (Johnson et al., 2023b;a)	52,802	Reports, Clinicals, Demographics
MIMIC-Symile (Saporta et al., 2024)	10,345	Chest X-rays, ECGs, laboratory
UKB (Sudlow et al., 2015)	486,400	Proteomics, Metabolomics, EHR
INSPECT (Huang et al., 2023)	22,449	CTs, Reports, EHR

of linear and nonlinear readouts,

$$\widetilde{\mathbf{x}}_m = \underbrace{(1-\alpha)\sum_{j=1}^{K} A_{mj}\mathbf{z}_j\mathbf{P}_{mj}}_{\text{linear readout}} + \underbrace{\alpha\, g_m\big(\mathrm{vec}\big([A_{mj}\mathbf{z}_j]_{j=1}^{K}\big)\big)}_{\text{nonlinear MLP readout}} + \underbrace{\sigma_m\boldsymbol{\eta}_m}_{\text{modality-specific noise}} \quad . \tag{6}$$

The resulting modality vector is subsequently ℓ_2-normalized before being passed to the scorer. Here, $A_{mj} \in \{0,1\}$ indicates whether latent variable $\mathbf{z}_j$ contributes to modality m, $\mathbf{P}_{mj} \in \mathbb{R}^{d\times D}$ projects the d-dimensional latent variable into the D-dimensional modality space, and $g_m : \mathbb{R}^{Kd} \to \mathbb{R}^D$ is an Multi-Layer Perceptron (MLP) readout applied to the ordered concatenation of the masked latent variables. The coefficient $\alpha \in [0,1]$ controls the mixture of the linear and nonlinear readouts, while $\boldsymbol{\eta}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$ denotes standard modality-specific Gaussian noise and $\sigma_m \geq 0$ controls its magnitude. Across sampled SCMs, we vary the graph structure, latent dimensionality, structural mechanisms, modality assignments, and noise levels. This defines a prior over multimodal dependency structures rather than a single synthetic data-generating process.

Prior-Fitted Scorer We prior-fit the Transformer scorer across multiple datasets generated from the SCM prior. Sampling from the SCM yields a modality vector $\mathbf{x}_m^{(n)} \in \mathbb{R}^D$ for each modality, where $D = d_{\text{in}}$ is chosen to match the scorer input dimension. Consequently, during prior-fitting, no encoders are required. For each sampled SCM $\mathcal{M}$, we generate a matched M-modal dataset $\mathcal{D}_{\mathcal{M}}$. For a target modality m, a positive tuple contains modality vectors with the same sample index n, whereas negative tuples replace $\mathbf{x}_m^{(n)}$ with $\mathbf{x}_m^{(n')}$, $n' \neq n$, sampled from another observation generated by the same SCM. The Transformer outputs a scalar compatibility score for each candidate tuple. We use a pair-sampling approach (Rheude et al., 2026c) which enables contrastive learning beyond bimodal datasets with candidate-dependent scorers such as Transformers. For a retrieval direction, the remaining modality embeddings of an anchor are kept fixed, while the target embedding is varied over one matched candidate and sampled negatives. The scorer is evaluated independently for each resulting tuple, and a cross-entropy loss $\mathcal{L}$ encourages the matched candidate to receive the highest score. We apply this objective symmetrically over all retrieval directions. For the frozen PFS, the Transformer, including its input projections and scalar readout, is held fixed while the real-data modality encoders are trained to map their inputs into the scorer's input space.

4 Experiments

4.1 Datasets

We focus on clinical and biomedical datasets because they provide a heterogeneous multimodal setting with images, physiological signals, tabular measurements, text-derived representations, and substantial modality missingness: MIMIC-IV (Johnson et al., 2023a;b; Goldberger et al., 2000), MIMIC-Symile (Saporta et al., 2024), INSPECT (Huang et al., 2023), and the UK Biobank (UKB) (Sudlow et al., 2015) (Table 1). These datasets represent different cohort sizes and clinical settings (Table 1), combine raw inputs with precomputed embeddings, and include both complete and missing modality configurations. Together, they enable comparing established multimodal objectives while testing whether the proposed scorers transfer across heterogeneous datasets.

MIMIC-IV We use MIMIC-IV (Johnson et al., 2023a;b; Goldberger et al., 2000) with radiology reports, clinical time series, and demographic variables analogously to Rheude et al. (2026b). Radiology reports are represented using a pretrained RadBERT encoder (Yan et al., 2022), followed by a

trainable residual MLP (Touvron et al., 2023) projection head. Clinical time series and demographic features are encoded using modality-specific residual MLPs.

MIMIC-Symile We additionally evaluate on MIMIC-Symile (Saporta et al., 2024), which combines chest X-rays, electrocardiograms (ECGs), and laboratory measurements as a benchmark for higher-order multimodal contrastive learning. Following the original setup, chest X-rays and ECGs are encoded using ResNet encoders (He et al., 2016; Saporta et al., 2024), while laboratory measurements are represented using a residual MLP. Each sample pairs an ECG and laboratory measurements collected within 24 hours of admission with a chest X-ray acquired 24-72 hours after admission. Retrieving this future chest X-ray from the ECG and laboratory values using the learned score therefore represents a zero-shot, prospective cross-modal prediction task (Figure 1) (Saporta et al., 2024).

UKB We further consider the UKB (Sudlow et al., 2015), using proteomics, metabolomics, and electronic health records (EHRs) (Rheude et al., 2025). The EHR modality is represented using Qwen3-Embedding-8B embeddings (Zhang et al., 2025; Hegselmann et al., 2026), and all three modality encoders are implemented as residual MLPs. We implement the union setting with $486{,}400$ participants, which contains substantial modality missingness and therefore reflects a realistic large-scale multimodal scenario. Retrieval evaluation is filtered for samples with the available proteomics modality. For modality availabilities, metabolomics is almost complete with a missing rate of only 0.87%. EHR data is missing for 25.82%. The expensive proteomics modality is substantially sparser, with a missing rate of 89.25%.

INSPECT The INSPECT dataset (Huang et al., 2023) comprises $22,449$ samples for pulmonary embolism (PE) diagnosis, combining computed tomography pulmonary angiography (CTPA) volumes, EHRs, and radiology reports. We filter to complete multimodal samples with available binary PE labels. EHRs are represented using MOTOR embeddings (Steinberg et al., 2024), projected with a residual MLP, CTPA volumes use spatial feature maps from the final Med3DVLM encoder stage (Xin et al., 2025), aggregated by a lightweight 3D ResNet, and report impressions are encoded with RadBERT (Yan et al., 2022), adapted using LoRA and a residual MLP. For vision retrieval conditioned on EHRs and reports, negative candidates are drawn from different patients but restricted to the same PE label as the query, preventing shortcut retrieval based solely on PE status.

4.2 Experimental Setup

Training and Evaluation All methods are implemented in a common PyTorch Lightning framework (Falcon & team, 2019) to ensure consistent training, optimization, encoder choices, and evaluation across methods and datasets. We perform extensive hyperparameter tuning for all methods (Section B), design encoders to be geometry preserving, and use a learnable logit scale (Rheude et al., 2026a;b;c). Models are optimized with ScheduleFree-AdamW (Defazio et al., 2024) together with norm-based gradient clipping. For the Transformer scorer, we do not include positional encodings because the modality tokens have no spatial or temporal ordering. Modality identity is retained through modality-specific projection networks. Each modality embedding is first mapped to the Transformer dimension through a modality-specific projection, after which a learnable CLS token is concatenated and the resulting tokens are processed by a pre-layer-normalized Transformer encoder. All linear layers are initialized with Xavier initialization (Glorot & Bengio, 2010), while the CLS token is initialized from $\mathcal{N}(0, 0.02^2)$. Missing modalities are zero imputed and concatenated with binary indicators identifying which input features are missing (Saporta et al., 2024). For the Transformer scorer missing modality tokens are excluded from self-attention using a key-padding mask. Results on real-world datasets are computed using 5-fold cross-validation with non-overlapping sample identifiers across folds and by reporting mean $\pm$ standard error (SE). We use a common batch size of 128 for all datasets.

Prior-Fitting The SCM prior emphasizes linear structural mechanisms while introducing small nonlinear contributions through the modality-specific readouts (Section B). Latent and modality-specific noise are sampled to expose the scorer to varying noise levels and encourage robust transfer. Training and model selection follow two stages. First, the SCM prior-fitting hyperparameters are selected by minimizing the validation loss on held-out synthetic SCMs. General SCM parameters,

Table 2: Comparison of well-tuned models optimized with different multimodal contrastive objectives. Values represent top-1 accuracy of the retrieval task (mean ± SE). Best non-overlapping value bold.

Scorer	INSPECT ↑	UKB ↑	MIMIC-Symile ↑	MIMIC-IV ↑
Baselines				
Clip	0.5548 ± 0.005	0.4738 ± 0.008	0.3109 ± 0.006	0.6503 ± 0.001
Triangle	0.5183 ± 0.014	0.5259 ± 0.009	0.4477 ± 0.005	0.6476 ± 0.003
Gram	0.4586 ± 0.012	0.4623 ± 0.009	0.3240 ± 0.005	0.6454 ± 0.002
Symile	0.6979 ± 0.003	0.5554 ± 0.005	0.4707 ± 0.008	0.6562 ± 0.003
Proposed				
Transformer (frozen)	0.6979 ± 0.002	0.4462 ± 0.005	$\mathbf{0.5021 \pm 0.019}$	$\mathbf{0.6626 \pm 0.003}$
PFS (frozen)	$\mathbf{0.7173 \pm 0.002}$	$\mathbf{0.5745 \pm 0.004}$	$\mathbf{0.5164 \pm 0.007}$	$\mathbf{0.6743 \pm 0.002}$

e.g., the structural mechanisms, number of nodes, and latent dimensionality, are fixed based on preliminary experiments to avoid overly simple synthetic tasks and potential shortcut solutions. For example, reducing the number of nodes can lower the synthetic validation loss by simplifying the prior while yielding a scorer that transfers less effectively to real-world datasets. The resulting single PFS is then frozen and reused for all real-world datasets, without dataset-specific prior-fitting or checkpoint selection. Second, the real-data encoders, learnable logit scale, and downstream optimization hyperparameters are tuned separately for each dataset by minimizing its validation loss. Thus, selection of the PFS is independent of all real-world datasets. The complete search spaces and selected values are reported in Section B.

4.3 Downstream Performance

We compare well-tuned models optimized with different multimodal contrastive objectives with a randomly initialized frozen Transformer and an SCM-prior-fitted Transformer (PFS) across different datasets (Table 2).

Frozen Transformer Even without prior-fitting or downstream updates to the scorer, the frozen Transformer is competitive on three of the four datasets. It matches the best baseline, *i.e.,* Symile, on INSPECT and further exceeds all baselines on MIMIC-Symile and MIMIC-IV. This indicates that the modality encoders can adapt effectively to a fixed nonlinear scoring function. However, its substantially weaker UKB performance shows that the architecture and masking mechanism alone do not provide uniformly strong transfer.

PFS SCM-based prior fitting improves mean retrieval accuracy on all four datasets while also reducing variability across folds and seeds, as reflected by equal or lower SE. The PFS consequently achieves the highest accuracy across all benchmarks. Overall, prior-fitting is particularly valuable in settings where a frozen Transformer scorer transfers poorly, as demonstrated on the UKB. The PFS benefits from more SCMs during prior fitting, with performance improving as the number of sampled SCMs increases from 2,000 to 10,000 (Figure 3a). Beyond this range, *e.g.,* up to 100,000 SCMs, we do not observe further gains, suggesting that scaling the prior may require more architectural optimization. In addition to higher retrieval scores, the models using the PFS also demonstrate faster convergence compared to the baselines (Figure 3b).

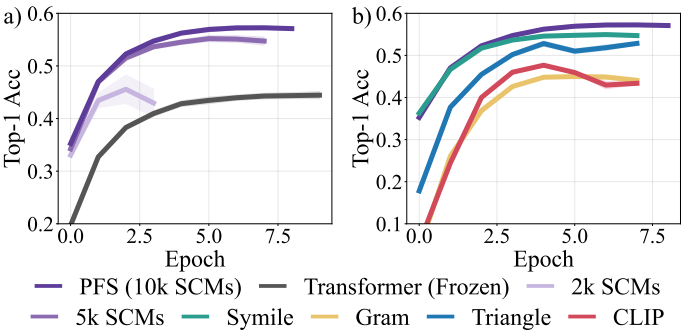


Figure 3: Validation performance of well-tuned baselines, the frozen Transformer and the PFS on the UKB. **a)** Scaling the prior-fitting with more sampled SCMs improves transfer to real-world data. **b)** Using the frozen PFS leads to faster convergence and higher retrieval accuracy than the baselines.

Table 3: MIMIC-Symile: Ablating the strong performance of frozen Transformers.

Ablation	Top-1 Acc ↑
Strongest Baseline (Symile)	0.4707 ± 0.008
Scorer Architecture	
MLP scorer (frozen)	0.2010 ± 0.018
Residual MLP scorer (frozen)	0.2351 ± 0.013
Transformer scorer	0.4203 ± 0.011
MLP scorer	0.4304 ± 0.014
PFS	0.4860 ± 0.011
Transformer scorer (frozen)	0.5021 ± 0.019
PFS (frozen)	0.5164 ± 0.007
Scorer Initializations	
Mimetic	0.2520 ± 0.016
Kaiming	0.3815 ± 0.009
Xavier	0.5021 ± 0.019
PFS (frozen)	0.5164 ± 0.007

Table 4: UKB: Ablating the unusually weak performance of frozen Transformers.

Ablation	Top-1 Acc ↑
Strongest Baseline (Symile)	0.5554 ± 0.005
Scorer Architecture	
MLP scorer (frozen)	0.2969 ± 0.074
Residual MLP scorer (frozen)	0.3840 ± 0.009
Transformer scorer (frozen)	0.4462 ± 0.005
PFS	0.4589 ± 0.029
Transformer scorer	0.4736 ± 0.025
PFS (frozen)	0.5745 ± 0.004
MLP scorer	0.6115 ± 0.013
Scorer Initializations	
Mimetic	0.3652 ± 0.006
Xavier	0.4462 ± 0.005
Kaiming	0.4998 ± 0.004
PFS (frozen)	0.5745 ± 0.004

4.4 Ablation Study

We demonstrate scorer and prior-fitting ablations. Hyperparameters of every ablation variant are retuned. (Section B). We ablate the performance of the frozen Transformer scorer on MIMIC-Symile and the UKB to get insights for the strong and unusually weak performance respectively (Tables 3 and 4). For the PFS, we ablate on the UKB (Table 5). Here, prior-fitting demonstrates its largest improvement, making the UKB better suited to distinguish between differnt SCMs.

Scorer Ablations We compare scorer architectures and initialization schemes on MIMIC-Symile and the UKB (Tables 3 and 4). Varying the scorer architecture ablates whether the benefit is specific to the Transformer or instead reflects more general properties of neural scorers. Likewise, because a frozen scorer is entirely determined by its weights, comparing initialization schemes is interesting since PFSs can be interpreted as a data-driven, prior-informed initialization of the weights. On MIMIC-Symile, residual and standard MLP variants perform substantially worse than the frozen Transformer, and making either the Transformer or MLP trainable does not improve retrieval. The frozen Transformer is highly sensitive to initialization, with Xavier (Glorot & Bengio, 2010) outperforming Kaiming (He et al., 2015) and mimetic initialization (Trockman & Kolter, 2023), while PFS yields the strongest result. On the UKB, the overall picture stays the same: we observe prior-fitting benefits and optimization difficulties when training the scorer as well. However, the trainable MLP clearly benefits from end-to-end optimization and reaches the highest accuracy. We do not observe this behavior on the other datasets, suggesting that the UKB data can be learned with a comparatively simple trainable scorer which is unusually effective. This is consistent with prior work indicating that UKB shows different multimodal behavior from MIMIC-Symile (Rheude et al., 2026b).

Prior-Fitting Ablations Although UKB is atypical with respect to scorer trainability, it is highly discriminative for studying prior-fitting: PFS improves top-1 accuracy over the frozen Transformer by 12.83 percentage points. This large gap provides substantially more resolution for distinguishing individual SCM prior components than datasets on which random and prior-fitted Transformer scorers perform similarly. With the private-information ablation, each latent variable is assigned directly to only one modality. Although this removes direct latent overlap, DAG edges can still induce indirect

Table 5: Ablation study for the SCM design and prior-fitting the PFS on the UKB dataset.

Ablation	Top-1 Acc ↑
Only private information SCMs	0.3170 ± 0.007
Transformer scorer (frozen, Xavier)	0.4462 ± 0.005
Permute BC	0.4643 ± 0.003
2k SCMs	0.4723 ± 0.015
Remove DAG edges	0.5431 ± 0.008
5k SCMs	0.5527 ± 0.003
Only linear SCMs	0.5733 ± 0.003
PFS (10k shared/pairwise SCMs)	0.5745 ± 0.004

Table 6: Wall time for the frozen Transformer, PFS, and the baselines averaged over 10 epochs.

Metric	MLP	Transformer	PFS	Clip	Triangle	Gram	Symile
Training step wall time (s) ↓	0.049	0.244	0.244	0.032	0.031	0.035	0.038

dependencies between modality-specific latent variables. This reduces accuracy to 0.3170, confirming that prior-fitting requires cross-modal dependencies. Randomly shuffling the sample alignment between modalities during prior fitting limits performance to 0.4643, only slightly above the frozen Transformer, highlighting the importance of preserving correct cross-modal correspondences in the synthetic data. Performance also scales with the diversity of the prior-fitting distribution: increasing the number of sampled SCMs from $2{,}000$ to $5{,}000$ and $10{,}000$ improves accuracy (Figure 3). Removing all DAG edges makes every latent variable a root sampled independently from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Nevertheless, shared and pairwise assignments in $\mathbf{A}$ still induce cross-modal dependence by assigning multiple modalities to the same latent variables. This variant still achieves 0.5431, indicating that overlapping latent assignments alone provide a substantial transferable prior, while structural dependencies through the edges of the DAG offer an additional benefit. Restricting the prior to purely linear dependencies by setting $\alpha = 0$ (Equation (6)) and using only linear structural mechanisms along the DAG edges yields 0.5733, effectively matching the full PFS. Thus, cross-modal information sharing, correct correspondence, and SCM diversity are the primary contributors, whereas nonlinear mechanisms provide no measurable benefit in this setting.

5 Limitations, Conclusion & Future Work

Limitations Transformer-based scoring incurs greater wall-clock training time than analytical scoring functions (Table 6), while the PFS additionally requires a separate SCM prior-fitting stage. The effect of making a neural scorer trainable is not yet fully understood. Training the MLP scorer improves over its frozen counterpart, whereas training the Transformer reduces performance, even with a separate scorer-specific learning rate and warm-up schedule. This suggests an unresolved interaction between scorer, initialization, and joint optimization with the modality encoders. Our experiments are restricted to trimodal clinical and biomedical settings and retrieval-based evaluation. Consequently, generalization to more modalities and other domains remains to be established. Moreover, the effectiveness of prior-fitting depends on how well the synthetic SCM distribution captures relevant properties of real multimodal data. Selecting and tuning this prior remains an open problem, since prior-fitting is validated on synthetic tasks rather than directly on the downstream real-world distribution. Finally, any medical application would require additional validation, including analyses of demographic and institutional biases, subgroup performance, robustness to distribution shifts, and potential failure modes, before clinical use could be considered.

Conclusion We investigated whether contrastive learning must be specified through a manually designed analytical function or can instead be provided by a Transformer-based scorer. Our results establish a randomly initialized frozen Transformer as a surprisingly strong baseline: it matches or exceeds well-tuned models with manual objectives. We further shape this initial inductive bias by prior-fitting the Transformer on multimodal data generated from an SCM prior. The resulting PFS improves over the frozen Transformer across all evaluated datasets.

Future Work Prior-fitting remains relatively underexplored for multimodal contrastive learning, leaving considerable scope for richer and more targeted priors. Future work could incorporate broader families of structural mechanisms, modality counts, information-sharing patterns, missingness processes, and domain-informed data-generating assumptions. It could also investigate how properties of the prior determine the interactions learned by the scorer and how these properties should be selected. A major bottleneck is the computational cost of prior-fitting. For every sampled SCM, each positive tuple must be contrasted against multiple negative candidates for every target modality, requiring a separate Transformer evaluation for many candidate combinations. The cost therefore grows with the number of SCMs, samples, modalities, and negatives, and becomes expensive when approximating the full contrastive candidate set. Nevertheless, our results indicate positive scaling effects, motivating work on more efficient candidate sampling.

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 methodological aspects of multimodal contrastive learning and does not involve the collection of new human-subject data. Our real-world experiments use established biomedical datasets and are conducted in accordance with their respective access, governance, and data-use requirements. We report results only at the aggregate level and do not attempt to identify individual participants. Although our experiments involve clinically relevant data, the proposed methods are intended for methodological research and are not designed or evaluated for direct clinical decision making. As with other representation-learning methods trained on observational biomedical data, learned representations may inherit biases present in the underlying datasets. Consequently, performance reported here should not be interpreted as evidence of clinical validity or fairness across patient populations.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our experiments throughout the main paper and supplementary material. The model architectures, Transformer scoring function, SCM prior, and prior-fitting procedure are described in Section 3. Dataset construction and preprocessing are documented in Section 4.1 and the corresponding appendix sections. We report optimization settings, hyperparameter search spaces, evaluation protocols, and compute resources in the appendix. All reported real-world experiments use the same cross-validation and evaluation framework across competing methods. An anonymized implementation is provided with the submission and will be made publicly available upon publication.

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.

Ehsan Amid, Rohan Anil, Wojciech Kotłowski, and Manfred K. Warmuth. Learning from randomly initialized neural network features. , 2022. URL https://arxiv.org/abs/2202.06438.

Samuel Böhm, Lennart Purucker, Frank Hutter, and Pascal Schlosser. Survpfn: Towards foundation models for survival predictions. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=kDEHp7ytr2.

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://arxiv.org/pdf/2409.07402.

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

Raja Giryes, Guillermo Sapiro, and Alex M. Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13): 3444–3457, 2016. doi: 10.1109/TSP.2016.2543202.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and D. Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 249–256. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/glorot10a.html.

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.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

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.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-08328-6.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis Langlotz, Serena Yeung, Nigam Shah, and Jason Alan Fries. Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=3sRR2u72oQ.

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.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2023a. URL https://physionet.org/content/mimiciv/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.

Weiyang Liu, Zhen Liu, James Rehg, and Le Song. Neural similarity learning. 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/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

Yuhang Liu, Zhen Zhang, Dong Gong, Erdun Gao, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Beyond dags: A latent partial causal model for multimodal learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bZqCBgm2N0.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/pdf/2506.10914.

Jovana Mitrovic, Brian McWilliams, Jacob C. Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/pdf/2010.07922.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://arxiv.org/pdf/2112.10510.

Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. Pfns4bo: In-context learning for bayesian optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25444–25470. PMLR, 2023. URL https://proceedings.mlr.press/v202/muller23a.html.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748. arXiv: 1807.03748.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017. URL https://library.oapen.org/bitstream/id/056a11be-ce3a-44b9-8987-a6c68fce8d9b/11283.pdf.

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.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

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, Roland Eils, and Benjamin Wild. Beyond objective expressivity: Geometry preservation in multimodal contrastive learning. arXiv:2607.17673 [cs.LG], July 2026b. URL http://arxiv.org/abs/2607.17673.

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 2026c. URL http://arxiv.org/abs/2604.05834.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/pdf/2506.06039v1.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, 105: 103668, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103668.

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.

Yan Shi, Cheng Liu, Aodi Yu, Zhenzhou Lu, Said Elias, Kai Cheng, Jiaqing Kou, Xin Chen, Yu Liu, Hong-Zhong Huang, and Michael Beer. A multi-fidelity tabular prior-data fitted network model for accurate prediction and uncertainty quantification. *Nature Communications*, 17(1):8308, 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-75163-w.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

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.

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.

Asher Trockman and J. Zico Kolter. Mimetic initialization of self-attention layers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34456–34468. PMLR, 2023. URL https://proceedings.mlr.press/v202/trockman23a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 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. 5998–6008, 2017.

Shifeng Xie, Vasilii Feofanov, Jianfeng Zhang, Themis Palpanas, and Ievgen Redko. Cauker: Classification time series foundation models can be pretrained on synthetic data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/pdf/2508.02879.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. arXiv: 2503.20047, 2025. URL https://doi.org/10.48550/arXiv.2503.20047.

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.

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.

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16559–16568. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01608. URL https://doi.org/10.1109/CVPR52688.2022.01608.

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.

Tianqi Zhao, Guanyang Wang, Yan Shuo Tan, and Qiong Zhang. Tabclustpfn: A prior-fitted network for tabular data clustering. (arXiv:2601.21656), May 2026. doi: 10.48550/arXiv.2601.21656. URL http://arxiv.org/abs/2601.21656. arXiv:2601.21656 [cs.LG].

A Compute Environment

All experiments were conducted on a shared high-performance cluster (HPC) cluster comprising the following compute resources:

1. 21 Dell PowerEdge R7525 nodes, each equipped with 64 AMD EPYC Rome CPU cores, 512 GB of RAM, and one NVIDIA A100 40 GB GPU.
2. 2 Dell PowerEdge XE8545 nodes, each equipped with 128 AMD EPYC Milan CPU cores, 512 GB of RAM, four NVIDIA A100 40 GB GPUs, and four NVIDIA A100 80 GB GPUs connected via NVLink.

B Hyperparameter Tuning

Hyperparameter selection is performed with Bayesian optimization using validation loss as the search objective. Unless stated otherwise, we allocate 50 optimization runs per method. In line with established tuning recommendations (Godbole et al., 2023), the batch size is chosen beforehand and is not included in the search space. We repeat the hyperparameter search independently for every main experiment and ablation to avoid carrying over configurations across settings. For UKB-Union, which is substantially more computationally demanding, we reduce the search budget to 30 trials and to 10% of the training dataset size. The bayesian search spaces for all methods are listed in the following (Listings 1 to 3). Encoders for the datasets are fixed as explained in Section 4.

Listing 1 Baseline hyperparameters related to Clip, Triangle, Gram, and Symile.

```
method: bayes
metric:
  name: val/min_loss
  goal: minimize
early_terminate:
  type: hyperband
  min_iter: 10

modelname.logit_scale_init:
    min: -3
    max: 0
    distribution: "uniform"
modelname.emb_dim:
    values: [128]
optimizer.lr:
    min: 0.00001
    max: 0.01
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
```

Listing 2 Hyperparameters related to the frozen Transformer and PFS.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize
early_terminate:
  type: hyperband
  min_iter: 10

modelname.logit_scale_init:
    min: -3
    max: 0
    distribution: "uniform"
modelname.emb_dim:
    values: [128]
modelname.transformer_params.d_model:
    values: [512]
modelname.transformer_params.nhead:
    values: [2]
modelname.transformer_params.num_layers:
    values: [5]
modelname.transformer_params.dropout:
    values: [0.0]
modelname.pretrained_load_logit_scale:
    values: [True, False]
modelname.freeze_scorer:
    values: [True]
optimizer.lr:
    min: 0.00001
    max: 0.01
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
```

Listing 3 Hyperparameters related to pretraining the PFS on SCMs.

```yaml
method: bayes
metric:
  name: val/min_loss
  goal: minimize
early_terminate:
  type: hyperband
  min_iter: 10

modelname.logit_scale_init:
    min: -3
    max: 0
    distribution: "uniform"
modelname.emb_dim:
    values: [128]
datamodule.samples_per_scm:
    values: [32]
datamodule.scms_per_batch:
    values: [64]
encoders.latent_noise_std:
    values: [0.0,0.1]
encoders.modality_readout:
    values: ["mlp"]
encoders.modality_readout_alpha_choices:
    values: ["[0.0,0.1,0.2]"]
encoders.n_latents_choices:  # number of nodes
    values: ["[16,32]"]
encoders.latent_dim_choices:
    values: ["[2,4,8]"]
encoders.modality_noise_std_choices:
    values: ["[0.0,0.1]"]
encoders.edge_probability_choices:
    values: ["[0.2,0.3,0.4]"]
encoders.modality_allocation_choices:
    values:
      - "[{shared_all: 0.5, pairwise: 0.5, private: 0.0}]"
encoders.mechanism_type_choices:
    values: ["[[linear],[linear,tanh]]"]
optimizer.lr:
    min: 0.00001
    max: 0.01
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
```