HIDDEN IN THE MULTIPLICATIVE INTERACTION: UNCOVERING FRAGILITY IN MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning has become a standard approach for unsupervised learning from paired data, as demonstrated by CLIP for image-text matching. However, many domains involve more than two modalities and require objectives that capture higher-order dependencies beyond pairwise alignment. Symile extends CLIP to this setting by replacing the dot product with the multilinear inner product (MIP) over modality embeddings. In this work, we show that there is a fragility which is *hidden in the multiplicative interaction*: a single weakly informative, misaligned, or missing modality can propagate through the objective and distort cross-modal retrieval scores. We propose Gated Symile, a contrastive gating mechanism that adapts modality contributions on an attention-based, per-candidate basis. The gate suppresses unreliable inputs by interpolating embeddings toward learnable neutral directions with an explicit `NULL` option when reliable cross-modal alignment is unlikely. Across a controlled synthetic benchmark that *uncovers this fragility* and three real-world trimodal datasets, Gated Symile achieves higher top-1 retrieval accuracy than well-tuned state-of-the-art (sota) baselines. More broadly, our results highlight gating as a step toward robust multimodal contrastive learning beyond two modalities in the presence of noise, misalignment, or missing inputs.[1]

1 INTRODUCTION

Contrastive learning has become a standard tool for bimodal learning exemplified with image-text pairs (Radford et al., 2021). However, many real-world problems require reasoning over more than two modalities. For instance, medical prediction tasks often combine more modalities (Acosta et al., 2022), including imaging (Tak et al., 2026), laboratory measurements (Saporta et al., 2024), proteomics (Rheude et al., 2025), metabolomics (Buergel et al., 2022), electrocardiograms (ECGs), and electronic health records (EHRs) (Steinfeldt et al., 2025). Hence, recent work extends bimodal contrastive learning with objectives that model higher-order interactions across modalities (Saporta et al., 2024; Cicchetti et al., 2025a;b; Dufumier et al., 2025). The objective for multimodal contrastive learning beyond bimodality could be*, e.g.,* a triangle (Cicchetti et al., 2025a) or a parallelotope (Cicchetti et al., 2025b) between modality embeddings to model all interactions instead of only pairwise interactions. These can be generalized with Symile (Saporta et al., 2024) which extends CLIP's dot product to multimodality with the MIP as the contrastive objective.

However, real-world datasets often contain modalities that may be complementary, conflicting, weak, or missing. For example, a patient with early-stage pneumonia may show signs in chest X-rays and lab tests, while the clinical notes do not yet document respiratory symptoms, leading to conflicting cross-modal signals for the same patient. We find that existing multimodal contrastive learning methods treat all modalities symmetrically and cannot explicitly model reliability differences. This results in a distortion of the cross-modal objective for sota baselines. Therefore, a central challenge in multimodal contrastive learning is to exploit higher-order interactions when modalities are informative, while remaining robust to weak, noisy, or missing inputs.

In this paper, we highlight an under-explored limitation of multimodal contrastive learning: weak or misaligned modalities can propagate through the objective and distort the learned signal. While

[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.

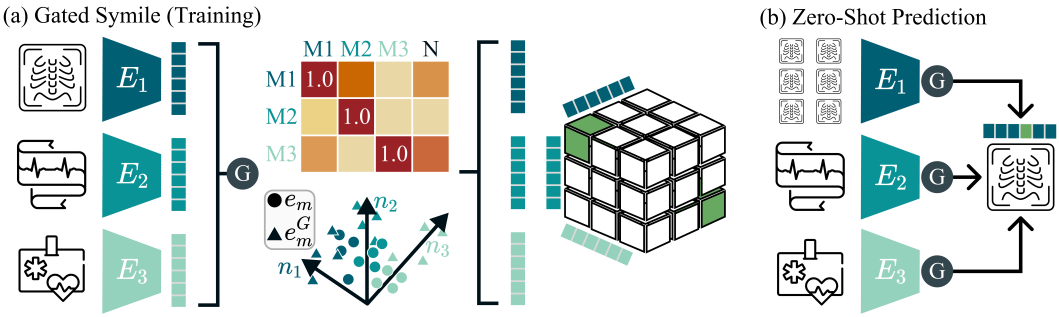

Figure 1: Illustrative overview of Gated Symile exemplified with the trimodal Symile-MIMIC (Saporta et al., 2024) dataset. **(a)** During training, modality-specific encoders E_m create embeddings e_m and our proposed gate G produces target-conditioned weights over the available modalities and a `NULL` option N (heatmap). The gate forms gated embeddings by weighting and interpolating each modality embedding with a modality-specific neutral direction n_m (coordinate system), enabling the model to suppress unreliable modalities while preserving useful signal. The gated embeddings are then combined via Symile's MIP (cube with positives on the diagonal). **(b)** At inference, the gating is applied to compute candidate scores for zero-shot prediction.

one might expect modality encoders to circumvent this, we show that this is not guaranteed. For CLIP's multimodal generalization, *i.e.,* Symile, this effect is *hidden in the multiplicative interaction*. Because the MIP combines modalities multiplicatively, errors from a single modality are amplified by the others. As a result, aggregate performance can remain relatively strong even though individual samples are affected by unreliable modalities, masking architectural fragility and leading to silent failures when modalities are misaligned or uninformative. Motivated by this observation, we introduce a gating mechanism on top of Symile to adaptively modulate modality contributions. The resulting method, Gated Symile (Figure 1), computes candidate-conditioned gate weights to attenuate unreliable modalities. The gate interpolates misaligned embeddings toward neutral directions, with an explicit `NULL` option when the target embedding indicates that reliable cross-modal alignment is unlikely. We find that Gated Symile strongly improves the robustness for misaligned modalities while preserving Symile's original higher-order interactions.

Taken together, our approach improves multimodal contrastive learning with modality-specific reliability and higher-order interactions. Our key contributions are:

- We identify and theoretically derive that **Symile's MIP implicitly treats all modalities symmetrically**, which can hide fragility, *e.g.,* under modality misalignment.
- We propose Gated Symile, **an attention-based per-candidate gating mechanism** that downweights unreliable modalities by interpolating embeddings toward learnable neutral directions and incorporating a `NULL` option for uninformative evidence.
- We propose **a new synthetic benchmark** with controlled modality misalignment, enabling systematic analysis of robustness to unreliable inputs.
- We demonstrate on synthetic and three real-world datasets that **Gated Symile consistently improves top-1 retrieval accuracy** over well-tuned sota baselines.
- Beyond retrieval performance, we **analyze gate weights, embedding geometries, scaling behavior, model efficiency, and ablate gate features** to better understand the impact of modality misalignment and the performance gains enabled by gating.

2 Related Work

Contrastive Learning Contrastive learning has become a dominant paradigm for representation learning. Its optimizing objective pulls together embeddings of related views while pushing unrelated ones apart. In the unimodal setting, this is typically done with instance discrimination under data augmentations, *e.g.,* with the InfoNCE objective (Oord et al., 2019), SimCLR (Chen et al., 2020), and MoCo (He et al., 2020). Bimodal contrastive learning extends this to cross-modal representations,

in which pairs are positives across modalities exemplified by image-text pairs in CLIP (Radford et al., 2021) and follow-ups such as SigLIP (Zhai et al., 2023). Moving beyond two modalities, multimodal methods can benefit from objectives enforcing agreement across all modalities*, e.g.,* by matching relational structure across modalities in addition to instance-level pairing. This is exemplified with methods like AudioClip (Guzhov et al., 2022), ImageBind (Girdhar et al., 2023), Gram (Cicchetti et al., 2025b), Triangle (Cicchetti et al., 2025a), CoMM (Dufumier et al., 2025), and Symile (Saporta et al., 2024). However, prior methods do not explicitly model robustness to varying modality interactions. In such cases, naive alignment objectives may be suboptimal when one modality is weakly informative or noisy.

Gating Mechanism Gating modulates information flow by selecting, reweighting, or routing representations. Early and widely used examples include gates in recurrent networks such as LSTMs (Hochreiter & Schmidhuber, 1997) and GRUs (Cho et al., 2014), which regulate how much past state is retained and how new evidence is incorporated. Beyond sequence models, gating is often used for conditional feature modulation. For example, FiLM (Perez et al., 2018) applies conditioning-dependent scaling and shifting to intermediate activations. In Transformers (Vaswani et al., 2017), attention weights similarly implement soft selection by controlling how strongly tokens contribute to representations. Channel- and spatial-wise gating has also been used, most prominently in Squeeze-and-Excitation blocks (Hu et al., 2018), which learn per-channel importance weights to emphasize informative feature maps. Further, routing-based gates enable conditional computation by selecting subsets of expert modules, as in mixture-of-experts models (MoEs) (Jordan & Jacobs, 1994; Jacobs et al., 1991) and product-of-experts models (PoEs) (Hinton, 2002). Most closely related to our setting is contrastive gating such as CDG (Meng et al., 2022), CR-MoE (Jiang et al., 2024), and MCMR (Lu et al., 2026) to suppress less informative inputs (Wang et al., 2019; Zohra et al., 2026; Wan et al., 2025; Gorti et al., 2022). We apply gating (i) within a contrastive objective, (ii) beyond bimodal settings, and (iii) explicitly to address modality misalignment. This combination has not been studied in prior work.

Selective Prediction and Explicit Abstention Early work formalized selective classification*, i.e.,* prediction with a reject option, as a principled risk-coverage trade-off (Chow, 1970; El-Yaniv & Wiener, 2010), and later instantiated it for deep models such as SelectiveNet (Geifman & El-Yaniv, 2019). Further related work include open set recognition*, i.e.,* prediction with unknowns at test time (Scheirer et al., 2013), exemplified by architectures such as OpenMax (Bendale & Boult, 2016). This intersects with Out-of-Distribution (OOD) detection, where abstention is often implemented through an explicit `NULL` pathway or confidence-based rejection (Hendrycks & Gimpel, 2017; Liang et al., 2018). In contrast, in our work, selection and rejection is represented more locally*, e.g.,* as a probability mass and without supervision. This connects to the broader idea of learned neutral placeholders with special tokens such as `CLS`, `MASK`, and `REG` in BERT (Devlin et al., 2019) and Vision Transformers (ViTs) (Dosovitskiy et al., 2021; Darcet et al., 2024), as well as learned prototypes such as discrete latents in VQ-VAEs (van den Oord et al., 2017) and class prototypes in prototypical networks (Snell et al., 2017).

3 METHOD

3.1 SENSITIVITY OF THE MIP TO MODALITY MISALIGNMENT

Symile maximizes a multi-sample contrastive lower bound on total correlation (TC), a measure of higher-order dependence among modalities (Watanabe, 1960; Saporta et al., 2024). For M modalities,

$$\mathrm{TC}(x^{(1)}, \ldots, x^{(M)}) = D_{\mathrm{KL}}\left(p(x^{(1)}, \ldots, x^{(M)}) \,\middle\|\, \prod_{m=1}^{M} p(x^{(m)})\right), \tag{1}$$

which is zero under mutual independence. Symile uses an InfoNCE-style objective to distinguish a positive tuple from negatives formed by sampling from the product of marginals, yielding a tractable lower bound on TC with a learned critic g. To instantiate g, Symile replaces CLIP's dot product with the multilinear inner product (MIP). Let $e_m = E_m(x^{(m)}) \in \mathbb{R}^D$ denote the embedding of modality

m with encoders E_m and the shared embedding dimension D, then the MIP is

$$\langle e_1, \ldots, e_M \rangle = \sum_{j=1}^{D} \prod_{m=1}^{M} e_{m,j}, \tag{2}$$

and Symile scores tuples via $g(x^{(1)}, \ldots, x^{(M)}) = \langle e_1, \ldots, e_M \rangle / \tau_{\text{MIP}}$ with temperature τ_{MIP} (Saporta et al., 2024). For example, for a retrieval task with target modality t, misalignment in a single non-target modality can strongly distort scores because the MIP multiplies contributions across modalities:

$$g(x^{(1)}, \ldots, x^{(M)}) = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} \left(e_{t,j} \prod_{\substack{m=1 \\ m \neq t}}^{M} e_{m,j} \right). \tag{3}$$

If a non-target modality $c \neq t$ is perturbed, *i.e.,* its embedding changes as $\hat{e}_c = e_c + \delta$, then

$$g_{\text{corr}} - g_{\text{clean}} = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} e_{t,j} \left(\prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_{m,j} \right) \delta_j. \tag{4}$$

Thus the score error is linear in δ but scaled by the product of the remaining modalities. This perturbation does not assume any specific source and can reflect, *e.g.,* misalignment to the ideal cross-modal tuple. Rewriting this as an inner product and applying Cauchy-Schwarz yields

$$|g_{\text{corr}} - g_{\text{clean}}| \leq \frac{1}{\tau_{\text{MIP}}} \|\delta\|_2 \left\| e_t \odot \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_m \right\|_2, \tag{5}$$

highlighting how multiplicative interactions can cause perturbations from a single unreliable modality to scale with the remaining embeddings during both training and inference (Details in Section A).

3.2 Gate Mechanism

We introduce a gate that modulates the contribution of each modality in Symile's MIP. For a retrieval direction, the gate outputs gated embeddings $e_1^G, \ldots, e_M^G$ by using gate weights $\{w_{t \to m}\}_{m=1}^M$ that control how strongly each modality should influence the MIP score. Intuitively, the gate aims to suppress non-target modalities whose current sample provides unreliable evidence for the retrieval target, *e.g.,* because the modality is misaligned, weakly informative, or missing. The gate is illustrated in Figure 2 and formalized in Algorithm 1. Embeddings e_m, projected queries/keys, neutral prototypes n_m, and gated embeddings e_m^G are ℓ_2-normalized.

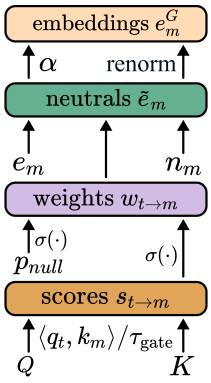

Figure 2: Sigmoid gate with NULL option and neutral directions.

Attention-Based, Candidate-Dependent Gating The proposed gate is candidate-dependent, *i.e.,* the weights are computed conditioned on the target embedding e_t and the candidate's non-target embeddings $\{e_m\}_{m \neq t}$. Concretely, we form a query vector from the target modality, $q_t = Q_t(e_t)$, and key vectors for each non-target modality, $k_m = K_m(e_m)$ for $m \neq t$. Due to ℓ_2-normalization, the relevance score is a scaled cosine similarity where $\tau_{\text{gate}} > 0$ controls the sharpness of the gating decisions[2],

$$s_{t \to m} \;=\; \langle q_t, k_m \rangle / \tau_{\text{gate}}, \tag{6}$$

and is mapped to a gate weight with an activation function σ, *e.g.,* a sigmoid or softmax function,

$$w_{t \to m} \;=\; \sigma(s_{t \to m}) \in (0, 1). \tag{7}$$

We set $w_{t \to t} = 1$ so that the target modality is never suppressed. To disentangle the effect of candidate dependence from the act of reweighting itself, we also consider an ablation with a lightweight baseline that replaces attention scores with a learned static matrix of gating logits (per target-modality pair).

[2]We use the same temperature τ_{gate} for the modality relevance scores and the NULL gating logit, so that τ_{gate} jointly controls the sharpness of both gating weights and the NULL decision.

Neutral Directions We introduce a per-modality neutral prototype $n_m \in \mathbb{R}^D$ to make downweighting explicit in representation space. For each modality, we interpolate between the current embedding and its neutral direction:

$$\tilde{e}_m \;=\; w_{t\to m}\, e_m \;+\; (1 - w_{t\to m})\, n_m. \tag{8}$$

Thus, a small $w_{t\to m}$ pushes a modality m toward a learned neutral embedding, making its contribution to the MIP closer to a non-informative baseline rather than injecting noise.

Gate Strength and Renormalization We include a strength parameter $\alpha \in [0, 1]$ that blends between the identity and the fully gated embedding analogous to a residual connection (He et al., 2016):

$$e_m^G \;=\; (1 - \alpha)\, e_m + \alpha\, \tilde{e}_m. \tag{9}$$

We ℓ_2-normalize e_m^G after gating to keep magnitudes comparable across settings and prevent the gate from trivially changing the score via norm scaling. Hence, the gate primarily affects the direction of each modality embedding.

Null Option Since our main gate uses independent sigmoid activations (and softmax as an ablation), multiple non-target modalities can be downweighted simultaneously. We additionally include a `NULL` option in the gating mechanism (not an additional embedding in the MIP) to allow the model to down-weight cross-modal evidence when the target embedding indicates that reliable cross-modal alignment is unlikely. Specifically, we compute a `NULL` logit with a projection head h_t and a bias u_t as

$$z_t \;=\; \big(h_t(e_t) + u_t\big)/\tau_{\text{gate}}, \tag{10}$$

and set $p_{\text{null}} = \sigma(z_t)$ for the sigmoid case, which multiplicatively shrinks all non-target weights by $(1 - p_{\text{null}})$. Under the softmax gate, `NULL` is implemented as an additional logit category appended to the softmax; under the sigmoid gate, it is implemented as an independent probability shared across non-target modalities. In both cases, `NULL` affects the MIP only indirectly by suppressing non-target contributions, thereby pushing the corresponding gated embeddings toward their neutral directions.

Algorithm 1 Attention-based gate with sigmoid, `NULL` option, and neutral directions.

Require: $e_1, \ldots, e_M \in \mathbb{R}^D$, target index t
Require: $Q_t : \mathbb{R}^D \to \mathbb{R}^{d_k}$, $K_m : \mathbb{R}^D \to \mathbb{R}^{d_k}$
Require: $h_t : \mathbb{R}^D \to \mathbb{R}$ and $u_t \in \mathbb{R}$
Require: $n_1, \ldots, n_M \in \mathbb{R}^D$
Require: $\tau_{\text{gate}} > 0$, $\alpha \in [0, 1]$
Require: $\|e_m\|_2 = \|n_m\|_2 = 1, \quad \forall m \in [M]$

$q_t \leftarrow \text{norm}(Q_t(e_t))$
for $m \in \{1, \ldots, M\} \setminus \{t\}$ **do**
  $k_m \leftarrow \text{norm}(K_m(e_m))$
  $s_{t\to m} \leftarrow \langle q_t, k_m \rangle / \tau_{\text{gate}}$
  $w_{t\to m} \leftarrow \sigma(s_{t\to m})$ ▷ sigmoid weight
end for
$z_t \leftarrow (h_t(e_t) + u_t)/\tau_{\text{gate}}$ ▷ `NULL` logit
$p_{\text{null}} \leftarrow \sigma(z_t)$
for $m \neq t$ **do**
  $w_{t\to m} \leftarrow (1 - p_{\text{null}})\, w_{t\to m}$
end for
$w_{t\to t} \leftarrow 1$
for $m = 1$ to M **do**
  $\tilde{e}_m \leftarrow w_{t\to m}\, e_m + (1 - w_{t\to m}) n_m$ ▷ neutral
  $e_m^G \leftarrow (1 - \alpha)\, e_m + \alpha\, \tilde{e}_m$ ▷ gate strength
  $e_m^G \leftarrow \text{norm}(e_m^G)$ ▷ renorm
end for
return gated embeddings $e_1^G, \ldots, e_M^G$

Robustness of the Gated MIP to Modality Misalignment Analogous to the ungated case, if a non-target modality is perturbed, the pre-normalized gated embedding gets edited by the gate:

$$\hat{e}_c^G = e_c^G + \beta_{t\to c}\delta \qquad \text{with} \qquad \beta_{t\to c} = 1 - \alpha + \alpha(1 - p_{\text{null}}) w_{t\to c}. \tag{11}$$

Therefore, the gated score satisfies

$$|g_{\text{corr}}^G - g_{\text{clean}}^G| \le \frac{\beta_{t\to c}}{\tau_{\text{MIP}}} \|\delta\|_2 \left\| e_t^G \odot \prod_{\substack{m=1 \\ m\neq t,c}}^{M} e_m^G \right\|_2. \tag{12}$$

Thus, when the gate suppresses an unreliable modality, the embedding entering the MIP is contracted by $\beta_{t\to c}$, reducing the multiplicative score distortion in Equation (4) (derivation in Section B).

4 Experiments

4.1 Datasets

We focus on trimodal retrieval settings ($M = 3$) analogous to prior work on multimodal contrastive learning (Saporta et al., 2024; Cicchetti et al., 2025a;b; Dufumier et al., 2025). We select datasets to

Table 1: Overview of benchmark datasets for our evaluation. For the Synthetic-XNOR dataset, $u, v \in \{0,1\}^K$ with $K = 16$, and $uv = \texttt{XNOR}(u, v)$ applied element-wise.

Dataset	# Samples	Modalities	Retrieval
Synthetic-XNOR	$30,000$	$A = [u, v, uv]$, $B = [u, 1, u]$, $C = [1, v, v]$	A
Symile-MIMIC	$10,345$	Chest X-ray, Laboratory, ECG	Chest X-ray
UKB	$37,888$	Proteomics, Metabolomics, EHR	Proteomics
UKB-Union	$486,400$	Proteomics, Metabolomics, EHR	Proteomics

probe complementary aspects (Table 1). Synthetic-XNOR isolates modality alignment in a controlled setting. Symile-MIMIC (Saporta et al., 2024) provides a well-established, non-saturated benchmark. The UK Biobank (UKB) (Sudlow et al., 2015) enables evaluation at larger scale, while UKB-Union further stresses the method under large-scale and missing-modality conditions.

Synthetic-XNOR We introduce a synthetic benchmark to study retrieval under controlled modality misalignment with a known ground-truth interaction. The aim is to evaluate retrieval when one of the two non-target modalities is partly misleading. Although one non-target modality remains informative, Symile's MIP entangles both non-target signals, so a misaligned modality can dominate the interaction and prevent learning from the clean evidence. We sample binary vectors $u, v \in \{0,1\}^K$ ($K = 16$) with i.i.d. $\text{Bernoulli}(0.5)$ bits and define the interaction $uv := \texttt{XNOR}(u, v)$. The target modality encodes $A = [u, v, uv]$, while the non-target modalities encode complementary signals $B = [u, 1, u]$ and $C = [1, v, v]$, so that for clean samples $B \odot C = [u, v, uv]$ matches the signal coordinates of A. Each bit is embedded as $\{-s, +s\}$ on signal coordinates ($s = 1$) and remaining dimensions contain Gaussian distractors ($\sigma = 3$), producing a low signal-to-noise setting. With probability p, we replace one modality in $\{B, C\}$ with the modality from another sample to simulate in-distribution misalignment. The swapping preserves marginal statistics but breaks cross-modal alignment, preventing trivial noise detection which is an easy task for the encoders to learn.

Symile-MIMIC The Symile-MIMIC dataset (Saporta et al., 2024) comprises $10,345$ samples collected from patients in an intensive care unit. The dataset contains three modalities: laboratory tests, chest X-ray images, and ECGs. The retrieval task is set up for the future chest X-rays (Figure 1b), so the evaluation can be interpreted as zero-shot prediction and retrieval of an expensive target modality from cheaper complementary evidence. For the whole setup including the actual implementation of the retrieval task and the encoders, we follow Saporta et al. (2024) (Multi-Layer Perceptron (MLP) for laboratory tests, ResNets (He et al., 2016) for the vision and ECGs modalities).

UK Biobank (UKB) The UKB (Sudlow et al., 2015) is a large prospective biomedical cohort. We focus on proteomics, metabolomics and EHRs for the trimodal modalities. We choose proteomics to be retrieved analogous to Rheude et al. (2025). We use both the intersection of modalities*, i.e.,* no missing modalities and $37,888$ samples, and the union of modalities*, i.e.,* missing modalities and $486,400$ samples. Missing modalities are implemented analogous to Saporta et al. (2024) by appending a binary mask indicating missingness to the input modality. We use normalized, raw modality inputs except for the EHR modality. Here, we use QWEN (Zhang et al., 2025) embeddings (Hegselmann et al., 2025). Modalities are encoded with MLPs. Details in Section D.

4.2 Experimental Setup

We follow best practices for multimodal learning (Rheude et al., 2026) (Section F). For non-synthetic datasets, we use 5-fold cross-validation and report mean $\pm$ standard error (SE) over three random seeds per fold. For Synthetic-XNOR, we use fixed splits.

Optimization We use ScheduleFree-AdamW (Defazio et al., 2024) and apply gradient clipping to stabilize training. We use a learned logit scale $s = \exp(\gamma)$ to control the softmax temperature, and for Symile-style objectives we additionally apply a fixed (d, M)-dependent normalization to the MIP before the learned scaling to stabilize training across embedding dimensions and numbers of modalities. We optimize gate parameters jointly with the encoders but use a separate learning rate multiplier for the gate module parameters. Details are in Section D.

Table 2: Comparison with well-tuned sota baselines on synthetic ($p = 1.0$) and real-world datasets. Values represent top-1 accuracy of the retrieval task (mean ± SE). More baselines in Section D.

Method	**Synthetic-XNOR ↑**	**Symile-Mimic ↑**	**UKB ↑**	**UKB-Union ↑**
CLIP	0.2434	0.4103 ± 0.016	0.4089 ± 0.015	0.0516 ± 0.007
TRIANGLE	0.6093	0.0948 ± 0.003	0.5651 ± 0.012	0.3597 ± 0.020
GRAM	0.4864	0.2516 ± 0.015	0.1848 ± 0.007	0.2008 ± 0.014
Symile	0.3310	0.4556 ± 0.006	0.6570 ± 0.012	0.5278 ± 0.009
Gated Symile	**0.8733**	$\mathbf{0.4670 \pm 0.005}$	$\mathbf{0.6819 \pm 0.010}$	$\mathbf{0.6000 \pm 0.007}$

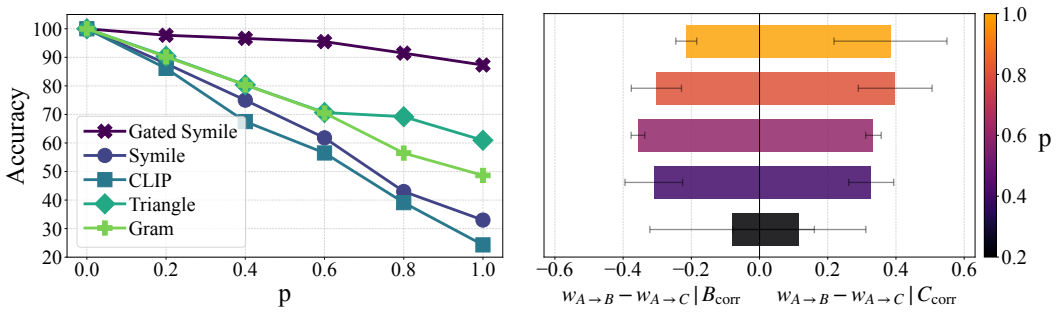

Figure 3: Analyses of well-tuned models on the Synthetic-XNOR dataset with probability p of one non-target modality being misaligned. **(Left)** Decreasing retrieval accuracy under increasing misalignment. Gated Symile preserves the accuracy and prevents a collapse of the MIP, demonstrating that gating improves robustness to misaligned modalities. **(Right)** Gate selects the reliable modality under misalignment. With two separate bar charts (toward left and right), we report the mean gate weight difference $w_{A\to B} - w_{A\to C}$. When B is misaligned (left), the difference becomes negative, *i.e.,* the gate assigns a smaller weight to B than to C. When C is misaligned (right), the difference becomes positive, *i.e.,* the opposite behavior. Results are averaged over six random seeds.

Sampling Symile supports two negative-sampling regimes: n (shuffled negatives) and n^2 (all pairings) (Saporta et al., 2024). In both cases, negatives are defined over combinations of the non-target modalities: in n-sampling, these are mismatched batch tuples, whereas in n^2-sampling all pairwise combinations are considered. However, we introduce candidate-dependent scoring and gating to Symile. To make this tractable, we use a *pair* formulation inspired by vision-language pretraining (Li et al., 2021; Bao et al., 2022; Li et al., 2023; 2022; Yu et al., 2022). Here, the target modality alone is varied while the remaining modalities are held fixed. For each query, this yields a candidate set consisting of the true positive and K uniformly sampled negatives from the target modality ($K = 128$). Based on preliminary scaling experiments, we fix the batch size to 128 (Synthetic-XNOR), 280 (Symile-MIMIC, analogous to Saporta et al. (2024)), and 512 (UKB).

4.3 Performance on Synthetic & Real-World Datasets

For comparing final performances, we report top-1 retrieval accuracy on the Synthetic-XNOR (Figure 3) and the three real-world trimodal datasets (Table 2). Across all datasets, Gated Symile yields the best performance compared to Symile and CLIP. The largest gain occurs on Synthetic-XNOR (from 0.3310 to 0.8733), consistent with the benchmark design in which exactly one non-target modality is intermittently misleading and the gate can suppress the unreliable factor before the multiplicative interaction. For the UKB, Gated Symile improves over Symile from 0.6570 ± 0.012 to 0.6819 ± 0.010, indicating that modulating modality contributions remains beneficial in a heterogeneous real-world cohort. On Symile-MIMIC, the improvement is smaller but consistent (from 0.4556 ± 0.006 to 0.4670 ± 0.005), and gating does not degrade performance. On UKB-Union, Gated Symile improves top-1 accuracy from 0.5278 ± 0.009 to 0.6000 ± 0.007. This setting highlights the benefit of adaptive gating when modalities are missing. Overall, our results suggest that Gated Symile improves retrieval accuracy. The strongest benefits appear in settings in which selective suppression is advantageous. More details and missing modality baselines in Section D.

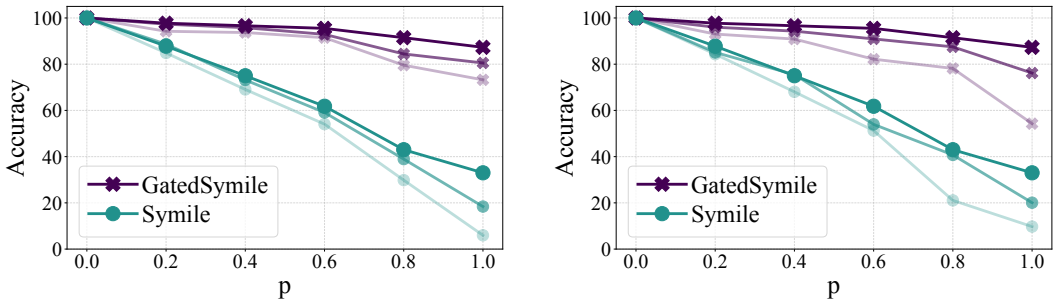

Figure 4: Scaling analyses of well-tuned models on the Synthetic-XNOR dataset under both increasing misalignment probability p and batch sizes (128, 256, 512 illustrated with decreasing brightness). **(Left)** Joint scaling of B and negatives per anchor K. Symile degrades substantially under misalignment, whereas Gated Symile remains markedly more robust, indicating increased fragility at larger contrastive scales. **(Right)** Scaling of B while keeping negatives per anchor K constant. This mirrors the joint-scaling regime, suggesting that the dominant scaling pathology is driven by the enlarged candidate pool: as B grows, fragility becomes more prevalent. Gating mitigates this effect by suppressing unreliable factors, thereby preserving retrieval performance as scale increases.

4.4 Alignment, Weight and Embedding Analyses

Synthetic Modality Misalignment As the misalignment probability p on the Synthetic-XNOR dataset increases, baseline sota methods degrade, whereas Gated Symile remains near-ceiling accuracy (Figure 3). This indicates that suppressing unreliable modalities prevents the MIP from collapsing under misleading inputs. Further, we report the mean weight difference $w_{A\to B} - w_{A\to C}$ conditioned on which non-target modality is misaligned. When B is misaligned, the difference is negative (the gate assigns smaller weight to B than to C), and when C is misaligned the difference is positive. This shows that the gate consistently shifts emphasis toward the reliable modality. Larger batch sizes typically improve contrastive learning performance (Chen et al., 2022; Cheng et al., 2024). However, in the presence of modality misalignment we observe the opposite trend (Figure 4): as the batch size B and candidate pool grow, performance degrades with increasing p. This appears both when scaling B and the number of negatives per anchor K, and when increasing B alone while keeping K fixed. This suggests that the main driver is the enlarged candidate pool. Gated Symile remains more stable across scaling regimes, indicating that suppressing unreliable modalities mitigates the fragility.

Table 3: Representation-level gate behavior for aligned and deliberately misaligned metabolomics inputs on the UKB.

Metric	Aligned	Misaligned
$\cos(e_{\mathrm{m}}^G, e_{\mathrm{m}})$	0.6732 ± 0.005	-0.0383 ± 0.0048
$\cos(e_{\mathrm{m}}^G, n_{\mathrm{m}})$	0.5982 ± 0.011	0.9869 ± 0.001

Real-world Modality Misalignment To test whether gating transfers to real-world data, we introduce controlled participant-level misalignment into the UKB by permuting metabolomics and EHR independently for 15% of the training samples (Tables 3 and 4). During validation, we permute metabolomics for 50% of the samples while retaining the corresponding proteomics and EHR. For paired diagnostics, we compare the metabolomics gate weight for each aligned sample, w_i^{clean}, with the weight obtained after replacing its metabolomics input with that of another participant, w_i^{corrupt}. Misalignment reduces the gate weight by 0.4382 ± 0.004 on average, with $w_i^{\text{corrupt}} < w_i^{\text{clean}}$ in $99.6\% \pm 0.1\%$ of the comparisons (Table 4). Using the fifth percentile of clean validation weights as a suppression threshold, the gate suppresses $98.79\% \pm 0.1\%$ of misaligned samples (true positive rate (TPR)) while flagging $4.94\% \pm 0.3\%$ of aligned samples (false positive rate (FPR)). At the representation level, gating preserves substantially more of the supplied metabolomics embedding when modalities are aligned than when they are misaligned, with cosine similarity decreasing from 0.6732 to -0.0383.

Table 4: Real-world and sample-level gate-weight diagnostics on the UKB.

Metric	Value
Mean weight reduction	0.4382 ± 0.004
Weight-reduction rate	0.9960 ± 0.001
Suppression rate (TPR)	0.9879 ± 0.001
Clean flag rate (FPR)	0.0494 ± 0.003

Conversely, alignment with the learned neutral direction increases from 0.5982 to 0.9869 under misalignment (Table 3). These results show that, on real-world multimodal data with controlled misalignment, the gate selectively suppresses the mismatched modality by moving its representation toward the learned neutral direction.

4.5 Ablation Study

We mainly report a retuned and cross-validated ablation of the proposed gate on the UKB (Table 5, Details and other ablations in Section E). Overall, the attention-based sigmoid gate with the `NULL` option and trainable neutral directions performs best, indicating that candidate-dependent suppression and an explicit neutral fallback are both important.

Table 5: Well-tuned ablation of the gate on the UKB (mean $\pm$ SE, details in Section E).

Ablation	**Top-1 Accuracy** $\uparrow$
Gated Symile	0.6819 ± 0.010
w/ neutral ones	0.6708 ± 0.014
w/o `NULL` option	0.6644 ± 0.013
w/ neutral frozen	0.6629 ± 0.013
w/ softmax (*w/o* sigmoid)	0.6622 ± 0.012
w/o renorm	0.6578 ± 0.013
w/o gate (Symile, pair)	0.6570 ± 0.012
w/o attention (*w/* matrix)	0.6446 ± 0.014
w/o gate (Symile, n)	0.6419 ± 0.024
w/o neutral & renorm	0.6314 ± 0.014

Neutral Directions and Renormalization Using a fixed neutral direction (all-ones or frozen random), remains competitive but trails the model. The all-ones variant is the next-best option, which is consistent with the intuition that the MIP contribution of a modality becomes weak in this case. Dropping renormalization reduces accuracy, and removing both neutral interpolation and renormalization yields the worst gated variant. This aligns with the view that unconstrained magnitudes can exacerbate multiplicative effects rather than suppress them.

Attention, Matrix, Sigmoid and Softmax Gating We find that sigmoid gating outperforms the softmax alternative. Sigmoid allows multiple modalities to be weighted simultaneously, whereas softmax enforces competition. Finally, the matrix-based (candidate-independent) gate can be harmful, implying that static global weights are insufficient to capture sample-dependent misalignment.

Sampling Choice Interestingly, the *pair* sampling slightly outperforms the *n* sampling. *Pair* sampling draws K negatives per anchor from the global candidate pool, which can be larger and more diverse than the in-batch shuffle used by *n* sampling. We therefore interpret the *pair* vs *n* comparison as both an efficiency and negative-diversity trade-off rather than a pure architectural ablation.

5 Conclusion, Limitations & Future Work

We studied robustness in multimodal contrastive learning beyond the bimodal setting and identified a failure mode of Symile-style objectives based on multiplicative interactions: misalignment in a single non-target modality can propagate through product terms and distort training. To address this, we proposed Gated Symile, a candidate-conditioned gating mechanism that adaptively downweights unreliable modalities. The introduced gating mechanism interpolates embeddings toward learnable neutral directions and allows a `NULL` option when the target embedding indicates that reliable cross-modal alignment is unlikely. Across a synthetic benchmark and three real-world trimodal retrieval datasets, Gated Symile improves top-1 retrieval over ungated Symile and well-tuned sota baselines. Our analyses revealed that the gate provides useful aggregate signals about modality reliability under misalignment. Our study has several limitations. While runtimes are only slightly increased (Section D, Table 8), empirical evaluation is limited to trimodal retrieval, predominantly biomedical datasets. Scalability, downstream tasks, and model-capacity effects require further study. Finally, global modality suppression is a coarse mechanism and could hide important conflicting evidence in clinical settings, motivating more finegrained cross-modal reasoning. Future work includes transferring the robustness induced by gating back into the encoders to study downstream tasks beyond retrieval and a deeper mechanistic interpretability analysis of the gate.

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

We do not see clear societal impact concerns raised by the paper itself. The work is methodological and evaluates robustness in multimodal contrastive learning, with the stated goal of making learning under imperfect modalities more reliable rather than enabling a new high-risk application.

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.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32897–32912. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d46662aa53e78a62afd980a29e0c37ed-Paper-Conference.pdf.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1563–1572. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.173. URL https://doi.org/10.1109/CVPR.2016.173.

Thore Buergel, Jakob Steinfeldt, Greg Ruyoga, Maik Pietzner, Daniele Bizzarri, Dina Vojinovic, Julius Upmeier Zu Belzen, Lukas Loock, Paul Kittner, Lara Christmann, Noah Hollmann, Henrik Strangalies, Jana M. Braunger, Benjamin Wild, Scott T. Chiesa, Joachim Spranger, Fabian Klostermann,

Erik B. Van Den Akker, Stella Trompet, Simon P. Mooijaart, Naveed Sattar, J. Wouter Jukema, Birgit Lavrijssen, Maryam Kavousi, Mohsen Ghanbari, Mohammad A. Ikram, Eline Slagboom, Mika Kivimaki, Claudia Langenberg, John Deanfield, Roland Eils, and Ulf Landmesser. Metabolomic profiles predict individual multidisease outcomes. *Nature Medicine*, 28(11):2309–2320, November 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-022-01980-3.

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33860–33875. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/db174d373133dcc6bf83bc98e4b681f8-Paper-Conference.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Zesen Cheng, Hang Zhang, Kehan Li, Sicong Leng, Zhiqiang Hu, Fei Wu, Deli Zhao, Xin Li, and Lidong Bing. Breaking the memory barrier: Near infinite batch size scaling for contrastive loss. arXiv:2410.17243 [cs], October 2024. URL http://arxiv.org/abs/2410.17243.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi (eds.), *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pp. 103–111. Association for Computational Linguistics, 2014. doi: 10.3115/V1/W14-4012. URL https://aclanthology.org/W14-4012/.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1224. URL https://doi.org/10.18653/v1/D19-1224.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL http://proceedings.mlr.press/v97/geifman19a.html.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, June 2023.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR. URL https://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. Version 1.0.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 4996–5005. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00495. URL https://doi.org/10.1109/CVPR52688.2022.00495.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980, 2022. doi: 10.1109/ICASSP43922.2022.9747631.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1026–1034, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. 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, Jun 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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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. arXiv:2502.17403 [cs], October 2025. URL http://arxiv.org/abs/2502.17403.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Ziyu Jiang, Guoqing Zheng, Yu Cheng, Ahmed Hassan Awadallah, and Zhangyang Wang. Cr-moe: Consistent routed mixture-of-experts for scaling contrastive learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=qKIvn9xL1R.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 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*, Proceedings of Machine Learning Research, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Xuan Lu, Kangle Li, Haohang Huang, Rui Meng, Wenjun Zeng, and Xiaoyu Shen. Beyond global similarity: Towards fine-grained, multi-condition multimodal retrieval. arXiv:2603.01082 [cs], March 2026. URL http://arxiv.org/abs/2603.01082.

Jian Meng, Li Yang, Jinwoo Shin, Deliang Fan, and Jae-Sun Seo. Contrastive dual gating: Learning sparse features with contrastive learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12247–12255, 2022. doi: 10.1109/CVPR52688.2022.01194.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748 [cs], January 2019. URL http://arxiv.org/abs/1807.03748.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

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*, volume 139 of *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*, 2026. URL https://openreview.net/forum?id=RtOPHuqmpj.

Adriel Saporta, Aahlad Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/pdf/2411.01053.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. 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/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

Jakob Steinfeldt, Benjamin Wild, Thore Buergel, Maik Pietzner, Julius Upmeier Zu Belzen, Andre Vauvelle, Stefan Hegselmann, Spiros Denaxas, Harry Hemingway, Claudia Langenberg, Ulf Landmesser, John Deanfield, and Roland Eils. Medical history predicts phenome-wide disease onset and enables the rapid response to emerging health threats. *Nature Communications*, 16(1): 585, January 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-55879-x.

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.

Divyanshu Tak, Biniam A. Garomsa, Anna Zapaishchykova, Tafadzwa L. Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, Maryam Mahootiha, Mitchell Parker, Luke R. G. Pike, Ceilidh Smith, Ariana M. Familiar, Kevin X. Liu, Sanjay Prabhu, Omar Arnaout, Pratiti Bandopadhayay, Ali Nabavizadeh, Sabine Mueller, Hugo Jwl Aerts, Raymond Y. Huang, Tina Y. Poussaint, and Benjamin H. Kann. A generalizable foundation model for analysis of human brain mri. *Nature Neuroscience*, February 2026. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-026-02202-6. URL https://www.nature.com/articles/s41593-026-02202-6.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. 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/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

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.

David Wan, Han Wang, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Clamr: Contextualized late-interaction for multimodal content retrieval. arXiv:2506.06144 [cs], June 2025. URL http://arxiv.org/abs/2506.06144.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5763–5772. IEEE, 2019. doi: 10.1109/ICCV.2019.00586. URL https://doi.org/10.1109/ICCV.2019.00586.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960. doi: 10.1147/rd.41.0066.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=Ee277P3AYC.

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.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. β-clip: Text-conditioned contrastive learning for multi-granular vision-language alignment. arXiv:2512.12678 [cs.CV], 2026. URL https://arxiv.org/abs/2512.12678.

A Relation to the Cauchy-Schwarz Bound

To quantify the sensitivity of the MIP critic to corruption in a single modality, we compare its score on a clean tuple and on a corrupted tuple and study the score deviation $\Delta g := g_{\text{corr}} - g_{\text{clean}}$. This difference isolates the effect of the corruption and admits a simple closed form because the MIP is multilinear in its arguments. Starting from

$$g_{\text{corr}} - g_{\text{clean}} = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} e_{t,j} \left(\prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_{m,j} \right) \delta_j, \tag{13}$$

define the vector

$$a \;:=\; e_t \odot \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_m \in \mathbb{R}^D, \qquad \text{i.e.,} \quad a_j = e_{t,j} \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_{m,j}. \tag{14}$$

Then Equation (13) can be written as an inner product,

$$g_{\text{corr}} - g_{\text{clean}} = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} a_j \, \delta_j \tag{15}$$

$$= \frac{1}{\tau_{\text{MIP}}} \langle a, \delta \rangle. \tag{16}$$

Taking absolute values yields

$$\left| g_{\text{corr}} - g_{\text{clean}} \right| = \frac{1}{\tau_{\text{MIP}}} \left| \langle a, \delta \rangle \right|, \tag{17}$$

where we use $\tau_{\text{MIP}} > 0$. By the Cauchy–Schwarz inequality, $|\langle a, \delta \rangle| \leq \|a\|_2 \, \|\delta\|_2$. Since $\tau_{\text{MIP}} > 0$, multiplying both sides by $1/\tau_{\text{MIP}}$ preserves the inequality direction, and thus

$$\left| g_{\text{corr}} - g_{\text{clean}} \right| \leq \frac{1}{\tau_{\text{MIP}}} \|a\|_2 \, \|\delta\|_2 = \frac{1}{\tau_{\text{MIP}}} \|\delta\|_2 \left\| e_t \odot \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_m \right\|_2. \tag{18}$$

Here $\prod e_m$ denotes elementwise multiplication across modalities, i.e.,

$$\Big(\prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_m \Big)_j = \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_{m,j}. \tag{19}$$

Applying Cauchy-Schwarz yields a worst-case upper bound on corruption-induced score distortion, separating the perturbation magnitude $\|\delta\|_2$ from a multiplicative amplification term $\left\| e_t \odot \prod_{\substack{m=1 \\ m \neq t,c}}^{M} e_m \right\|_2$.

B Gated MIP Sensitivity Derivation

We derive the perturbation bound for the gated MIP analogously to Section 3. For clarity, we consider the pre-normalized gated embedding. The normalization step is used in the implementation to keep embedding magnitudes comparable.

Recall from Equations (8) and (9) that

$$\tilde{e}_m = w_{t \to m} e_m + (1 - w_{t \to m}) n_m, \tag{20}$$

and

$$e_m^G = (1 - \alpha) e_m + \alpha \tilde{e}_m. \tag{21}$$

Substituting $\tilde{e}_m$ gives

$$e_m^G = (1 - \alpha) e_m + \alpha \big(w_{t \to m} e_m + (1 - w_{t \to m}) n_m \big) \tag{22}$$

$$= \big(1 - \alpha + \alpha w_{t \to m} \big) e_m + \alpha (1 - w_{t \to m}) n_m. \tag{23}$$

We define the effective coefficient on the original embedding as

$$\beta_{t\to m} = 1 - \alpha + \alpha w_{t\to m}. \tag{24}$$

Then the coefficient on the neutral direction is

$$1 - \beta_{t\to m} = 1 - \big(1 - \alpha + \alpha w_{t\to m}\big) \tag{25}$$
$$= \alpha - \alpha w_{t\to m} \tag{26}$$
$$= \alpha(1 - w_{t\to m}). \tag{27}$$

Hence, the gated embedding can be written equivalently as

$$e_m^G = \beta_{t\to m} e_m + (1 - \beta_{t\to m}) n_m. \tag{28}$$

If the sigmoid-based gate with the `NULL` option is active, the effective non-target gate weight becomes

$$\bar{w}_{t\to m} = (1 - p_{\text{null}}) w_{t\to m}. \tag{29}$$

In this case, the interpolation is

$$\tilde{e}_m = \bar{w}_{t\to m} e_m + (1 - \bar{w}_{t\to m}) n_m, \tag{30}$$

and therefore

$$e_m^G = (1 - \alpha) e_m + \alpha\big(\bar{w}_{t\to m} e_m + (1 - \bar{w}_{t\to m}) n_m\big) \tag{31}$$
$$= \big(1 - \alpha + \alpha \bar{w}_{t\to m}\big) e_m + \alpha(1 - \bar{w}_{t\to m}) n_m. \tag{32}$$

Thus, with the `NULL` option, we define

$$\beta_{t\to m} = 1 - \alpha + \alpha(1 - p_{\text{null}}) w_{t\to m}. \tag{33}$$

Again, the neutral coefficient is the complement:

$$1 - \beta_{t\to m} = 1 - \big(1 - \alpha + \alpha(1 - p_{\text{null}}) w_{t\to m}\big) \tag{34}$$
$$= \alpha\big(1 - (1 - p_{\text{null}}) w_{t\to m}\big). \tag{35}$$

Therefore, both with and without the `NULL` option, the pre-normalized gated embedding can be written as

$$e_m^G = \beta_{t\to m} e_m + (1 - \beta_{t\to m}) n_m, \tag{36}$$

where $\beta_{t\to m}$ denotes the effective coefficient on the original embedding.

Now suppose that a non-target modality $c \neq t$ is perturbed as $\hat{e}_c = e_c + \delta$ (Equation (4)), while the gate weights, $\beta_{t\to c}$, and the neutral direction n_c are held fixed. Then

$$\hat{e}_c^G = \beta_{t\to c} \hat{e}_c + (1 - \beta_{t\to c}) n_c \tag{37}$$
$$= \beta_{t\to c}(e_c + \delta) + (1 - \beta_{t\to c}) n_c \tag{38}$$
$$= \beta_{t\to c} e_c + \beta_{t\to c}\delta + (1 - \beta_{t\to c}) n_c \tag{39}$$
$$= e_c^G + \beta_{t\to c}\delta. \tag{40}$$

The gated Symile score is

$$g^G(x^{(1)}, \ldots, x^{(M)}) = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} \prod_{m=1}^{M} e_{m,j}^G. \tag{41}$$

Therefore, the difference between the corrupted and clean gated scores is

$$g_{\text{corr}}^G - g_{\text{clean}}^G = \frac{1}{\tau_{\text{MIP}}} \sum_{j=1}^{D} e_{t,j}^G \left(\prod_{\substack{m=1 \\ m\neq t,c}}^{M} e_{m,j}^G \right) (\hat{e}_{c,j}^G - e_{c,j}^G) \tag{42}$$

$$= \frac{\beta_{t\to c}}{\tau_{\text{MIP}}} \sum_{j=1}^{D} e_{t,j}^G \left(\prod_{\substack{m=1 \\ m\neq t,c}}^{M} e_{m,j}^G \right) \delta_j. \tag{43}$$

Writing this as an inner product yields

$$g^G_{\text{corr}} - g^G_{\text{clean}} = \frac{\beta_{t\to c}}{\tau_{\text{MIP}}} \left\langle \delta,\, e^G_t \odot \prod_{\substack{m=1 \\ m\neq t,c}}^{M} e^G_m \right\rangle. \tag{44}$$

Applying Cauchy–Schwarz gives

$$|g^G_{\text{corr}} - g^G_{\text{clean}}| \leq \frac{\beta_{t\to c}}{\tau_{\text{MIP}}} \|\delta\|_2 \left\| e^G_t \odot \prod_{\substack{m=1 \\ m\neq t,c}}^{M} e^G_m \right\|_2. \tag{45}$$

For the basic gate, $\alpha \in [0,1]$ and $w_{t\to c} \in [0,1]$ imply

$$\beta_{t\to c} = 1 - \alpha + \alpha w_{t\to c} \in [1-\alpha, 1]. \tag{46}$$

With the sigmoid `NULL` option, the effective weight is $\bar{w}_{t\to c} = (1 - p_{\text{null}}) w_{t\to c}$, hence

$$\beta_{t\to c} = 1 - \alpha + \alpha(1 - p_{\text{null}}) w_{t\to c} \in [1-\alpha, 1]. \tag{47}$$

Increasing p_{null} reduces $\bar{w}_{t\to c}$ and therefore reduces $\beta_{t\to c}$. Thus, downweighting an unreliable modality contracts the effective perturbation entering the multiplicative score.

C Compute Environment

Our experiments are conducted on a High-Performance Cluster (HPC) with the following environment:

- 21 Dell PowerEdge R7525 compute nodes, each with 64 AMD Epyc cores (Rome), 512GB RAM and 1 NVIDIA A100 40G GPU
- 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)

D Additional Details

In the following, we provide additional details for our proposed method, implementations, and comparisons.

MIP Normalization Following standard practice in contrastive learning, we use a learned logit scale to control the sharpness of the softmax over candidates (Radford et al., 2021). Concretely, we parameterize the scale as $s = \exp(\gamma) > 0$ and form logits $L = s \cdot S$ from a raw score matrix S. For Symile-style objectives, S is given by the MIP critic (Saporta et al., 2024), whose variance increases with embedding dimension d and number of modalities M due to multiplicative interactions. To stabilize early training and make temperature initialization comparable across (d, M), we additionally apply a fixed (d, M)-dependent normalization to the raw MIPs (a variance-style scaling analogous to variance-preserving initialization schemes (Glorot & Bengio, 2010; He et al., 2015)). After this normalization, we multiply by the learned scale s and apply cross-entropy.

Weight-Based vs. Representation-Based Diagnostics We additionally report aggregate gate diagnostics computed over all sampled retrieval tuples, including both matching and non-matching candidates, without filtering for positive pairs (Table 6). Consequently, these quantities summarize coarse, dataset-level gating trends rather than sample-level responses to modality misalignment, and should not be interpreted as misalignment-detection statistics. We compare two complementary gate diagnostics: mean gate weights $w_{t\to m}$ and cosine-based measures of representation change, $\cos(e^G_m, e_m)$ and $\cos(e^G_m, n_m)$ (Table 6). The mean weights alone can be hard to interpret because they average over heterogeneous, sample-dependent decisions and are influenced by remaining gate features. Cosine similarities, in contrast, directly quantify whether the gate edits a modality embedding. On Symile-MIMIC, where improvements are smallest, $\cos(e^G_m, e_m) \approx 1$ and $\cos(e^G_m, n_m)$ remains relatively low, indicating that the gate leaves embeddings largely unchanged. On the UKB,

Table 6: Analysis of mean gate statistics for the non-target modalities. Rows correspond to datasets and columns to gate diagnostics computed for the non-target modalities m (2nd column) relative to the target modality t. We highlight $\cos(e_m^G, e_m)$: The relative changes across datasets mirror downstream performance*, i.e.,* larger gains are associated with stronger deviations from the original embedding, reflected by smaller cosine similarity. This underscores the gate's effectiveness.

Dataset	Modalities	$w_{t\to m}$	$\cos(e_m^G, e_m)$	$\cos(e_m^G, n_m)$
Synthetic-XNOR	B / C	0.3428 / 0.1599	0.4385 / 0.1656	0.7171 / 0.9581
Symile-MIMIC	Lab. / ECG	0.3656 / 0.4568	0.9366 / 0.9465	0.3596 / 0.4837
UKB	Metabol. / EHR	0.5679 / 0.3867	0.8921 / 0.8203	0.6541 / 0.7451
UKB-Union	Metabol. / EHR	0.5808 / 0.4884	0.7819 / 0.6489	0.4854 / 0.6408

the gate shows moderate editing and a higher neutral-direction cosine which matches its intermediate performance gain. For the UKB-Union, $\cos(e_m^G, e_m)$ decreases further compared to UKB, indicating stronger embedding edits, but $\cos(e_m^G, n_m)$ is not larger, suggesting a richer transformation than simple interpolation toward the neutral direction. Finally, on Synthetic-XNOR, where gating yields the largest gains, $\cos(e_m^G, e_m)$ is substantially reduced and $\cos(e_m^G, n_m)$ is high for at least one non-target modality, consistent with pushing unreliable inputs toward the neutral direction.

UKB Modality Availability Modality availability in the UKB cohort is heterogeneous. Among $486,400$ participants, metabolomics is nearly complete, with $482,182$ available samples and a missing rate of only 0.87%. EHR data are available for $360,824$ participants, corresponding to a missing rate of 25.82%. Proteomics is substantially sparser, with measurements for only 52,305 participants and a missing rate of 89.25%. This imbalance motivates evaluating multimodal methods beyond the complete-case intersection, where robustness to unavailable modalities is essential for retaining the broader cohort. For proteomics retrieval, we restrict validation and testing to participants for whom proteomics are available, ensuring that every query has a valid matching proteomics candidate.

Paired Tests To test our hypothesis that gating improves Symile, we computed a one-sided Wilcoxon signed-rank tests on paired cross-validation split means, after averaging the three seed runs within each split (Table 7). The evidence consistently favored the gated variants: they achieved higher mean Top-1 accuracy on all three datasets and outperformed Symile on 14 of the 15 paired splits. The gains were statistically significant on UKB and UKB-Union. On MIMIC, GatedSymile also improved performance and outperformed Symile on four of five splits percentage points, providing positive but not conventionally significant evidence in favor of gating.

Table 7: Paired comparison of Gated Symile and Symile across five runs. We report mean performance, the paired mean difference, the number of runs won by Gated Symile, and the p-value from a one-sided Wilcoxon signed-rank test.

Dataset	Gated Symile	Symile	Mean diff.	Wins	One-sided p
MIMIC	0.4670 ± 0.0052	0.4556 ± 0.0064	$+0.0113$	$4/5$	0.06250
UKB	0.6819 ± 0.0104	0.6570 ± 0.0115	$+0.0249$	$5/5$	0.03125
UKB-Union	0.6000 ± 0.0073	0.5278 ± 0.0093	$+0.0723$	$5/5$	0.03125

Runtime Analyses Gated Symile retains a computational profile close to that of Symile during retrieval (Table 8). Overall, Gated Symile provides its robustness and retrieval improvements with modest inference-time and memory overhead.

Missing-Modality Baselines We compare Gated Symile with two standard approaches for handling missing modalities on UKB-Union: mean imputation and modality dropout ($p = 0.3$) (Table 9). Gated Symile achieves 0.6000 ± 0.0073, outperforming all evaluated baselines. These results show that Gated Symile remains effective relative to established missing-modality strategies under highly heterogeneous modality availability.

Table 8: Computational cost of Symile and Gated Symile. We report the mean and standard deviation across repeated measurements.

Metric	Symile	Gated Symile
Retrieval runtime (s) ↓	11.047 ± 0.206	12.335 ± 0.718
Retrieval peak allocated memory (GB) ↓	6.47 ± 0.00	6.52 ± 0.00
Candidates per second ↑	683.8 ± 12.7	614.7 ± 34.0
Training step wall time (s) ↓	0.6358 ± 0.0354	1.0488 ± 0.0024

Table 9: Retrieval performance on UKB-Union under heterogeneous modality availability. Baselines use either mean imputation or modality dropout to handle missing modalities.

Missing-modality strategy	Method	UKB-Union ↑
Mean imputation	CLIP	0.3471 ± 0.003
	GRAM	0.2122 ± 0.007
	Triangle	0.0002 ± 0.000
	Symile	0.5243 ± 0.011
Modality dropout	CLIP	0.0326 ± 0.006
	GRAM	0.1780 ± 0.003
	Triangle	0.2380 ± 0.020
	Symile	0.5780 ± 0.007
Proposed	Gated Symile	$\mathbf{0.6000 \pm 0.007}$

E Details About the Ablation

Ablation Details We primarily conduct ablations of the proposed gating mechanism on UKB without missing modalities. This choice is motivated by two factors: first, performance gains on Symile-MIMIC are marginal, making detailed ablations less informative. Second, the scale of UKB-Union with missing modalities renders exhaustive ablation and retuning computationally prohibitive. While synthetic datasets do not fully reflect real-world conditions, we additionally report a re-tuned ablation on Synthetic-`XNOR` in Table 10 to provide a controlled analysis of the method including different seeds. Further, we report a matched-capacity ablation in which all variants use the same capacity hyperparameters (and only optimizer parameters are re-tuned) (Table 11).

Table 10: Well-tuned ablation of the gate on the Synthetic-`XNOR` dataset.

Ablation	Top-1 Accuracy ↑
Gated Symile	0.8958 ± 0.0033
w/o `NULL` option	0.8843 ± 0.0030
w/ neutral ones	0.8579 ± 0.0083
w/ softmax (*w/o* sigmoid)	0.8236 ± 0.0248
w/ neutral frozen	0.7745 ± 0.0499
w/o renorm	0.7191 ± 0.1422
w/o gate (Symile, pair)	0.6466 ± 0.0108
w/o neutral & renorm	0.3884 ± 0.0653
w/o gate (Symile, n)	0.3171 ± 0.0040
w/o attention (*w/* matrix)	0.2453 ± 0.0040

Ablation Hyperparameter Tuning Ablation studies can be misleading if components are removed while keeping the original hyperparameters fixed: changing the model, *e.g.,* removing a gate, `NULL`, renormalization, or attention, can substantially shift the optimal learning rate, regularization, temperature, and even effective capacity, so performance differences may reflect suboptimal tuning rather than the true contribution of the ablated component (Rheude et al., 2026; Dodge et al., 2019). To avoid conflating architectural changes with mismatched hyperparameters, we re-run dataset-specific

Table 11: Matched-capacity ablation of the gate on the UKB.

Ablation	**Top-1 Accuracy ↑**
Gated Symile	0.6819 ± 0.010
w/ neutral ones	0.6576 ± 0.011
w/ neutral frozen	0.6642 ± 0.018
w/ softmax (*w/o* sigmoid)	0.6633 ± 0.013
w/o gate (Symile, pair)	0.6570 ± 0.012
w/o `NULL` option	0.6493 ± 0.006
w/o attention (*w/* matrix)	0.6446 ± 0.014
w/o renorm	0.6437 ± 0.015
w/o gate (Symile, n)	0.6419 ± 0.024
w/o neutral & renorm	0.6413 ± 0.012

hyperparameter tuning for every ablation. The ablation experiments are swept with 50 runs. Parameter counts are listed in Table 12.

Table 12: Parameter counts of (Gated) Symile and the ablated variants.

Ablation	**Parameters ↓**
Gated Symile	44.1M
w/ neutral ones	44.1M
w/o `NULL` option	44.1M
w/ neutral frozen	44.1M
w/ softmax (*w/o* sigmoid)	44.1M
w/o renorm	44.1M
w/o gate (Symile, pair)	0.0
w/o attention (*w/* matrix)	18.4K
w/o gate (Symile, n)	0.0
w/o neutral & renorm	44.1M

F HYPERPARAMETER TUNING

We maximize the validation retrieval accuracy by using Bayesian optimization without incorporating the batch size (Godbole et al., 2023). The methods are swept with 100 runs. For experiments on the Synthetic-`XNOR` dataset, hyperparameters are retuned, *e.g.,* for different values of p. For the UKB-Union results, sweep runs are reduced to 50 due to longer runtimes. Listings 1 to 5 show the search spaces with respect to methods and datasets.

Listing 1 Hyperparameters related to Symile-MIMIC.

```
method: bayes
metric:
  name: val/max_acc_top1
  goal: maximize

modelname.emb_dim:
    values: [1024]  # initially tuned from 256-8196
modelname.embedding_norm:
    values: [True]
# Encoders fixed to ResNets + MLP analogous to Saporta et al.
```

Listing 2 Hyperparameters related to Synthetic-XNOR.

```yaml
method: bayes
metric:
  name: val/max_acc_top1
  goal: maximize

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

Listing 3 Hyperparameters related to the UKB.

```yaml
method: bayes
metric:
  name: val/max_acc_top1
  goal: maximize

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
```

Listing 4 Hyperparameters related to Clip, Triangle, Gram and Symile.

```yaml
method: bayes
metric:
  name: val/max_acc_top1
  goal: maximize

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 5 Hyperparameters related to Gated Symile.

```yaml
method: bayes
metric:
  name: val/max_acc_top1
  goal: maximize

modelname.logit_scale_init:
    min: -3
    max: 0
    distribution: "uniform"
modelname.gate_strength_init:
    min: -1
    max: 6
    distribution: "uniform"
modelname.neutral_type:
    values: ["random_trainable"]
modelname.gate_mode:
    values: ["attention"]
modelname.use_gate:
    values: [True]
modelname.use_null:
    values: [True]
modelname.renormalize:
    values: [True]
modelname.gate_type:
    values: ["sigmoid"]
modelname.gate_temp:
    min: 0.2
    max: 1.2
    distribution: "uniform"
optimizer.lr_gate_mul:
    min: 1.0
    max: 20.0
    distribution: "log_uniform_values"
modelname.gate_d_k:
    values: [1024, 3072, 6144]
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]
```